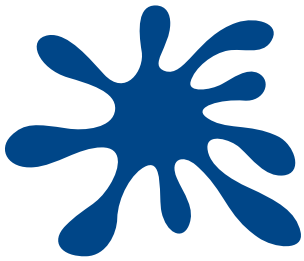
SLPAT 2016

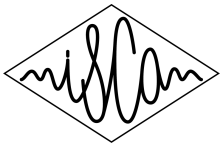
7th Workshop
on
Speech and Language Processing
for Assistive Technologies
(SLPAT)



Workshop Proceedings

13 September, 2016
San Francisco, USA





ISSN 2411-9962

Introduction

We are pleased to bring you the Proceedings of the 7th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT), held in San Francisco, USA, on 13 September, 2016. We received 24 paper submissions, of which 17 were accepted; 8 papers were chosen for oral presentation and 9 for presentation as posters or demos. All 17 papers are included in this volume.

This workshop was intended to bring researchers from all areas of speech and language technology with a common interest in making everyday life more accessible for people with physical, cognitive, sensory, emotional or developmental disabilities. This workshop builds on six previous such workshops (co-located with conferences such as ACL, NAACL, EMNLP and Interspeech); it provides an opportunity for individuals from research communities, and the individuals with whom they are working, to share research findings, and to discuss present and future challenges and the potential for collaboration and progress.

While Augmentative and Alternative Communication (AAC) is a particularly apt application area for speech and natural language processing technologies, we purposefully made the scope of the workshop broad enough to include assistive technologies (AT) as a whole, even those falling outside of AAC. Thus we have aimed at broad inclusivity, which is also manifest in the diversity of our Program Committee. We are also very delighted to have Prof. Helen Meng from The Chinese University of Hong Kong, as invited speaker.

The success of SLPAT 2016 was due to the authors who submitted such interesting and diverse work and which generated so intense discussions. Finally, we must thank all the people who made this event possible especially members of the Program Committee for completing their reviews promptly, and for providing useful feedback for deciding on the program and preparing the final versions of the papers.

Heidi Christensen, François Portet, Thomas Quatieri, Frank Rudzicz, and Keith Vertanen

Co-organizers of SLPAT 2016

List of People

Organizers:

Heidi Christensen, University of Sheffield, UK.
François Portet, University of Grenoble Alpes, France.
Thomas Quatieri, Massachusetts Institute of Technology, USA.
Frank Rudzicz, University of Toronto, Canada.
Keith Vertanen, Michigan Technological University, USA.

Program committee:

Jan Alexandersson, DFKI GmbH, Germany.
Jean-Yves Antoine, Université François Rabelais de Tours, France.
John Arnott, University of Dundee, UK.
Stefan Bott, IMS - University of Stuttgart, Germany.
Annelies Braffort, LIMSI-CNRS, France.
Heriberto Cuayahuitl, University of Lincoln, UK.
Stuart Cunningham, University of Sheffield, UK.
Michael Elhadad, Ben Gurion University, Israel.
Corinne Fredouille, CERI/LIA - University of Avignon, France.
Kallirroi Georgila, University of Southern California, USA.
Stefan Goetze, Fraunhofer Institute for Digital Media Technology, Germany.
Björn Granström, KTH, Sweden.
Mark Hasegawa-Johnson, University of Illinois at Urbana-Champaign, USA.
Per-Olof Hedvall, Lund University, Sweden.
Matt Huenerfauth, Rochester Institute of Technology, USA.
Sofie Johansson, Institutionen för svenska språket, Sweden.
Benjamin Lecouteux, University of Grenoble Alpes, France.
William Li, Massachusetts Institute of Technology, USA.
Peter Ljunglöf, University of Gothenburg, Chalmers University of Technology, Sweden.
Eduardo Lleida, University of Zaragoza, Spain.
Ornella Mich, Fondazione Bruno Kessler, Italy.
Climent Nadeu, Universitat Politècnica de Catalunya, Spain.
Vigouroux Nadine, IRIT, France.
Torbjørn Nordgård, Lingit AS, Norway.
Ehud Reiter, University of Aberdeen, UK.
Brian Roark, Google, USA.
Bitte Rydeman, Lund university, Sweden.
Rubén San-Segundo, Universidad Politécnica de Madrid, Spain.
Michel Vacher, CNRS, France.
Ravichander Vipperla, Nuance Communications, USA.
Maria Klara Wolters, University of Edinburgh, UK.
Hernisa Kacorri, Carnegie Mellon University, USA.
Phil Green, University of Sheffield, UK.
Mauro Nicolao, University of Sheffield, UK.
Bhusan Chettri, University of Sheffield, UK.

Table of Contents

Tag Thunder: Web Page Skimming in Non Visual Environment Using Concurrent Speech

Jean-Marc Lecarpentier, Elena Manishina, Fabrice Maurel
Stéphane Ferrari, Emmanuel Giguet, Gael Dias, Maxence Busson

Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, France
{firtsname.lastname}@unicaen.fr,

Abstract

Skimming and scanning are two strategies generally used for *speed reading*. Skimming allows a reader to get a first glance of a document; scanning is the process of searching for a specific piece of information in a document. While both techniques are available in visual reading mode, it is rather difficult to use them in non visual environments. In this paper, we introduce the concept of *tag thunder*, which provides speed reading non-visual techniques similar to skimming and scanning. A tag thunder is the oral transposition of the tag cloud concept. Tag cloud key terms are presented using typographic effects which reflect their relevance and number of occurrences. Within a tag thunder, the relevance of a given key term is translated into specific speech effects and its position on the page is reflected in the position of the corresponding sound on a 2D stereo space. All key terms of a tag thunder are output according to a concurrent speech strategy, which exploits the *cocktail party effect*.

In this paper, we present our implementation of the tag thunder concept. The results of the evaluation campaign show that tag thunders present a viable non-visual alternative to visual speed reading strategies.

Index Terms: non-visual web navigation, human-computer interaction, text-to-speech synthesis, key term extraction

1. Introduction

Most users share a similar mental process when accessing informative content of web pages. They get a first glance of the page content (skimming), followed by a quick search for specific information (scanning). Then the reader spots different areas of interest and seeks for specific information in identified areas using a zoom-in zoom-out strategy.

Although several factors may influence whether skimming and scanning are successful, such document properties as layout, logical structure and typographic effects play an important role in the perception process. However, this information is usually not available to users in non-visual environments [1]. Figure 1 illustrates how a web page is perceived in visual and non-visual environments using a screen reader.

Figure 1: Perception of the same web page in visual and non-visual environments.

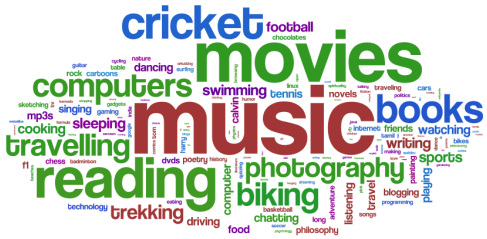


Figure 2: Example of a tag cloud.

To solve this problem, a number of non-visual replacement strategies [2, 3] have been proposed by screen readers such as faster speech rate depending on the textual block size, shortcuts which allow to jump from heading to heading, reading the beginning and the end of a paragraph, etc. However, these solutions are far from providing the reading capabilities of the visual mode [4].

This paper focuses on developing a strategy for a fast access to web page content in non-visual situations, that takes into account page layout and typographic clues. In particular, we transpose the visual concept of *tag cloud* to its audio version, called *tag thunder*.

In order to apply this concept to skimming and scanning, let us first consider a web page as a set of blocks. Figure 3 illustrates the result of a page segmentation. Each

SLPAT 2016, 7th Workshop on Speech and Language Processing for Assistive Technologies, pages 1–8,
San Francisco, USA, 13 September, 2016.

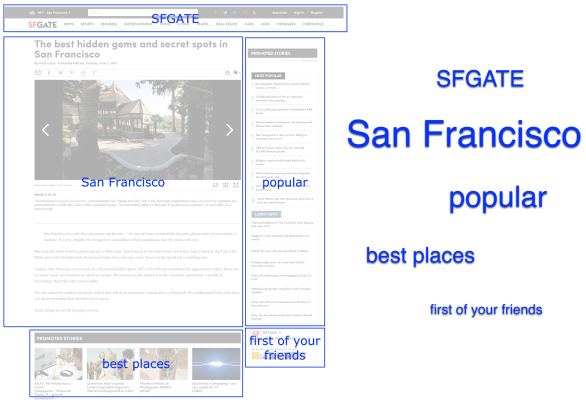


Figure 3: Segmentation of a page into zones and extracting key terms

zone is represented by key terms extracted from this zone which are combined into a tag cloud along with spatial and typographic effects that reflect the importance and relevance of each specific term, as shown in Figure 2[1]. Similarly, a tag thunder adds spatial and audio effects to key terms.

Tag thunders use concurrent speech strategy in order to represent the dense visual stimulus embodied by tag clouds. This strategy is based on the *Cocktail Party Effect*: users may identify key terms pronounced simultaneously or focus their attention on key terms that interest them among all the others.

This paper is structured as follows. In Section 2 we provide an outline of the related work in the area of non-visual content access strategies. In Section 3 we introduce our implementation of the tag thunder concept, specifically the three main steps in tag thunder creation: web page segmentation, key word extraction and vocal synthesis. Sections 4 presents the evaluation campaign we organized in order to assess the performance of our tag thunder implementation, as well as the potential of the tag thunder concept in general. We conclude this paper with a discussion and some directions for future work.

2. Related work

This Section presents some strategies developed in the field of assistive technologies, which facilitate the access to web content in non-visual environments.

Existing solutions for non-visual web page browsing often use Text-To-Speech (TTS) and Braille mode. Text-To-Speech has been used to convey document structure to users in non-visual situations via the content vocalization [5, 6]. To increase TTS efficiency, [7] proposed an oral transposition model based on layout reformulation strategies. These strategies combine a model of the written document [8], used to develop discursive forms from structured texts, and a prosodic model [9] used to reduce this new set of sentences in a more speech-adapted way. This approach brings a significant improvement in memorizing and understanding TTS output for strongly structured documents. But according to [10], the cognitive load is still hard to handle in comparison to visual reading.

Some early studies proposed to use summarization techniques to provide visually impaired people (VIP) with web pages skimming strategies [11]. However, a linearization step destroys the page layout which is at the core of the perception/action loop.

In the Accessibility through Simplification & Summarization project [12] (AcceSS), the content perceived as less important is removed from pages, thus modifying the page layout. A navigation page is then built, named *guide dog page*, which serves as a summary. Experiments show positive results when this method is combined with a JAWS screen reader. One of the limitations of this method is the incapacity of the pattern matching algorithm to correctly identify page sections. Furthermore, no simplification is proposed at textual level, providing no solutions to quickly browse large textual content.

SeEbrowser (Semantically Enhanced Browser) is a VIP-adapted audio web browser [13]. Manual semantic annotations are used to build ontologies modeling hierarchical relationships between elements within web pages. As the web page is loaded, the user may ask for Browser Shortcuts (BSs), go through them and interact using keyboard and audio feedback. Experiments show that this alleviates the information overload. However, the scanning strategy is still very long since users tend to listen to all BSs before choosing a relevant one.

Hearsay [14] is a non-visual multi-modal web browser which has been developed at Stony Brook University (New York, USA) since 2004. It supports different input modes: voice, keyboard and tactile interfaces. Possible output modalities are audio, screen and Braille. The browser provides many features: a segmentation module which analyzes web page structure and layout, a system of annotations which enables the addition of alternative text for pictures and other content blocks, algorithms detecting the changes between visited web pages, a context analyzer which detects the main content and identifies relevant information using hyperlinks. Experiments show a significant gain of time in finding the main content of a web page. In addition the system avoids repeating static content such as menus. Globally, most of these features made valuable contributions to improving user experience. Nevertheless, despite the positive results, two aspects still require improvement in comparison to visual reading. First, the page structure overview is not complete because it focuses on main content; the el-

[1]Image by Anand S, https://flic.kr/p/5BFE3V, CC-BY-2.0

ements are presented sequentially, making their browsing long. Thus, this method does not provide real skimming and scanning reading modes.

More and more work is now done using tactile strategies [15, 16, 17]. [18] incorporate patterns into web pages, thus enabling some elements and their relationships to be felt by running fingers over them. Such transformed documents are then given to VIPs using special paper with heat-sensitive ink. Putting the paper on a touch screen makes it possible to interact with it and obtain the oral transposition of a chosen web page section. Limitations come from the need to use a special paper with a heater. A similar concept is based on vibrotactile perception [19]. A special device captures contrast variations on the screen as fingers *browse* the content on a tablet. These variations are transformed into vibrations felt in a glove device worn on the other hand.

In recent years, some work has been carried out using Text-To-Speech tools within concurrent speech paradigm, exploiting the fact that human ears may concentrate on a specific audio source among many others [20]. The *Cocktail Party Effect* is a perfect example: even when many people are speaking simultaneously, we may concentrate our attention on one specific voice [21]. Variations in spatial location [22], as well as speech parameters (synchrony [23], frequencies [24]) may influence the perception of different voices. Using concurrent speech proved to accelerate blind people's scanning for relevant information [25, 26].

To resume this section, two main approaches (content summarization and concurrent speech synthesis) represent two interesting scanning strategies. However, they are not sufficient in providing real skimming abilities. The tag thunder concept combines both strategies: using segmentation and extraction techniques to give a summary of the page content and using concurrent speech synthesis to provide a quick overview.

3. Architecture

This Section presents our implementation of the tag thunder concept. It comprises three modules: web page segmentation, key term extraction and key term vocalization using concurrent speech synthesis.

3.1. Page segmentation

There exist numerous approaches to webpage segmentation [27, 28, 29, 30, 31]. We opted for the K-means++ algorithm [32, 33] [2]. The choice for unsupervised clustering algorithm was dictated, among other things, by the lack of unified web page layout, and robustness of K-Means algorithm in similar tasks [34]. It groups visible HTML elements into a desired number of zones based on their Euclidian distance. To optimize convergence and efficiency, each HTML element is enhanced with its computed styles based on underlying CSS and Javascript code [35]. Elements that are not part of the visual layout are ignored.

For the purpose of our experiment, the enhanced HTML is clustered into 5 zones. This choice of the number of zones was made with the objective to avoid a working memory overload, in accordance with the Miller's Law [36].

[2] http://scikit-learn.org/stable/modules/clustering.html#k-means

3.2. Key terms extraction and weighting

Each zone is represented by its key terms in the tag thunder. In our current implementation, key terms are n-grams of different lengths with a maximum order of 6.

For each n-gram, we compute $tf - idf$ [37] (term frequency – inverse document frequency). Tf is the frequency of a given term in a zone. The idf is computed using a corpus C containing 953 551 articles of the "Le Monde" newspaper dating from 1987 to 2006. Similar to [38], our solution couples $tf - idf$ metric with additional parameters. We use Formula (1) to compute the final score for each key term.

$$Score = tf(term, zone) \cdot idf(term, C) \cdot \sum_{i=1}^{n} \sigma(c_i) \quad (1)$$

where $tf(term, zone)$ is the frequency of the term within its zone, $idf(term, C)$ is the number of documents in our corpus C containing the term. $\sigma(c_i)$ is the weight for a characteristic c_i such as font weight, size, variant, style, etc. σ values were assigned empirically and reflect the visual perception of a given element. $\sigma(c_i)$ values range from 0.5 to 5.

For the purpose of our experiment, each zone is represented by one key term only.

3.3. Concurrent speech vocalization

This module generates the audio signal from a given key term and its zone properties. Based on [21, 23, 24], specific voice, volume, prosody, pitch, speech rate and synchronization characteristics are combined to build an audio track for a given key term.

Our synthesis module uses the Kali TTS [39] tool, developed at the University of Caen Normandie by the CRISCO laboratory. Kali supports speech rate acceleration without loss in intelligibility and sound quality, which is a very important feature in non-visual web browsing.

To vocalize the key terms, we use several cocktail party effect metaphors. Thus, we consider each zone as a discussion group in a cocktail party. Each metaphor provides rules which assign repetition frequency (Figure 4), volume (Figure 5) and a spot in a 2D audio space (Figure 6) to each key term. Vocalization of all the key terms

with their specific parameters produces the final tag thunder.

3.3.1. Repetition frequency

Metaphor 1: the larger the group talking about a topic, the more often related terms emerge.
Rule 1: vocalized key terms are played in a loop. Zone size influences repetition frequency within the loop.

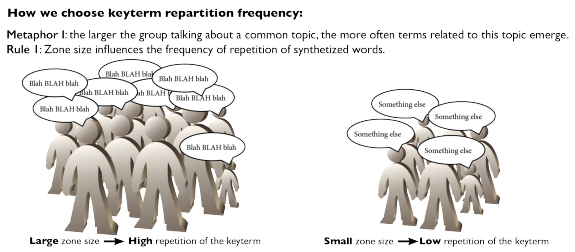


Figure 4: Repetition frequency metaphor

For each key term, the silence between two repetitions in the loop is proportional to the relative size of its zone. The larger the zone, the shorter the silence. In our experiment, silence duration has been empirically set between 0.5 second and 5 seconds.

3.3.2. Volume

Metaphor 2.a (distinctiveness): the more a voice in a group stands out, the easier it is to detect its source.
Metaphor 2.b (relevance): the more the words are repeated in a group, the more relevant they are.
Rule 2: volume is determined by zone contrast and key terms frequency in the zone.

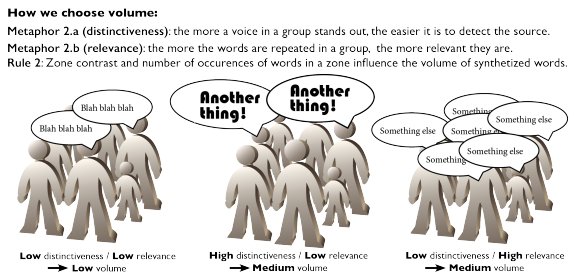


Figure 5: Volume metaphor

For each zone, contrast is computed based on the difference between the background color and the text. Volume is set within a $[min, max]$ interval, using the average of normalized contrast value and key term frequency. In our experiment, TTS constraints and perceptive tests led to setting the values for min and max to 4 and 8 points respectively, with each point representing 2 amplitude tones.

3.3.3. Spatialization

Metaphor 3: sound spatialization helps to physically place and distinguish several discussion groups.
Rule 3: zone coordinates influence the type of output voice and 2D spatialization of vocalized key terms.

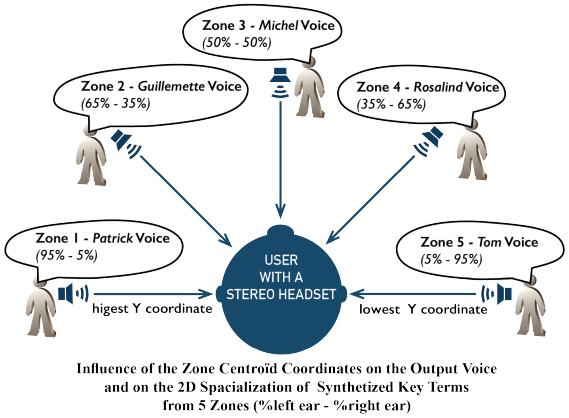


Figure 6: Sound spatialization metaphor

Voices are equally distributed in the 2D stereo space depending on the zone's centroid coordinates. In our experiment, sounds originate from 5 sources (i.e. 5 corresponding zones), as illustrated in Figure 6 .

4. Evaluation

We conducted an experiment in order to test the viability of the tag thunder concept and the quality of our implementation. In this Section, we present the experimental setting and the results.

4.1. Experimental setting

Our goal is to evaluate the system's capacity to provide fast skimming reading strategies. Here we present the results of the first experiment with sighted participants. The goal of this experiment is two-fold: to evaluate the relevance of the extracted key terms and to test the efficiency of tag thunder concept as a skimming strategy.

The experiment unfolds as follows. A participant sees a tag cloud followed by a web page, 15 seconds each. The page may or may not be the corresponding web page. The participant is asked whether the tag cloud corresponds to the displayed page.Possible answers are: definitely yes, probably yes, probably no, definitely no. Another participant is presented with the same data, but in the form of a tag thunder instead of the tag cloud and is asked to answer the same question. The experiment modalities were as follows:

(a) Tag Cloud and Tag Thunder output

(b) Webpage with a question form

Figure 7: User evaluation: web-based interface

- 18 sighted participants, each with 16 different stimuli (8 tag clouds - 8 tag thunders);
- 24 web pages from various web sites were used to generate a tag cloud and a tag thunder for each page;
- 24 other web pages were selected to create stimuli where the page and tags do not match;

Each couple (web page, tag set) was shown to 3 different participants; each participant evaluated an equal number of correct (matching) and incorrect couples.

Participants took the test autonomously, with a supervisor close by. The evaluation interface is shown in Figure 7.

4.2. Results

We present the evaluation results for Tag Clouds (TC) and Tag Thunders (TT) separately, as well as the combined overall results. We also separate the analysis of the correct (matching) and incorrect pages.

Figure 8 shows the dispersion of the total of 288 answers. It seems more difficult for participants to definitely validate a correct page than to definitely reject an incorrect page.

4.2.1. Agreement

We split the analysis of agreement statistics into three interpretations: *4-var* with four different answers; *3-var* where 'probably yes' and 'probably no' are combined into 'not sure'; and *2-var* where the answers 'definitely yes' and 'probably yes' are combined into 'yes' and 'probably no' and 'definitely no' into 'no'.

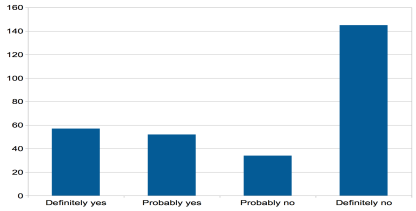


Figure 8: Dispersion of the 288 answers

	Web page	4-var	3-var	2-var
TC	Correct	12.8	20.8	70.8
	Incorrect	75.0	75.0	91.7
TT	Correct	20.8	33.3	75.0
	Incorrect	66.6	66.6	95.8
all		43.8	48.96	83.3

Table 1: Percentage of stimuli with the same answer

Table 1 presents the agreement statistics. The *2-var* interpretation shows a very high agreement rate when the incorrect page was shown, for both TT and TC. The *3-var* interpretation shows differences only for the correct pages, thus indicating that hesitations concerned correct pages only. This might mean that key terms were not always well suited to represent their zones in case of correct web pages, which created hesitation between 'probably yes' and 'yes' TTs tend to have a better agreement than TCs. Our hypothesis is that, in our experiment, textual key terms in a TC were displayed with fewer typographic effects whereas key terms in TTs had a full set of audio (or 'typophonic') effects described above. In general, the modality of the stimuli (TT vs TC) does not seem to influence the agreement rate between users.

4.2.2. Precision and Recall

Precision and recall are computed on the *2-var* interpretation. Table 2 presents the results. There is a significant difference in the perception of TCs or TTs between cases where the page was the correct or incorrect one. For the correct pages, the precision is very high, which means that the participants manage to associate a given page to a TT/TC. On the contrary, the recall is somewhat lower: as discussed before, users find it difficult to validate a correct page. This suggests that a number of correct pages were labeled as incorrect, which in turn might suggest the insufficiency of the TT/TC representation in these cases. This is especially apparent for tag thunders: 31% of correct pages were labeled as incorrect. Again, these results suggest that the extraction module needs further improvement.

Overall, participants found the exercise difficult but made few mistakes. In general, the results of TTs are

Format	TagCloud		TagThunder	
Web page	**Correct**	**Incor.**	**Correct**	**Incor.**
Precision	**0.96**	0.81	**0.98**	0.76
Recall	0.78	0.97	0.69	0.98
F-score	0.86	0.89	0.81	0.86
Accuracy	0.875		0.84	

Table 2: 2-var results: precision, recall and F-score

comparable in the overall accuracy with the results of the TCs. We can conclude that the tag thunder concept is valid and that certain limitations originate from the internal implementation of each module. We discuss these limitations in the following Section.

5. Discussion and future work

The first objective of this work was to implement the concept of tag thunders. Evaluation results demonstrate the viability of this concept. However, each module requires separate thorough evaluation.

5.1. Page segmentation

According to evaluation results, most errors come from pages where the number of distinct informational sections is larger than the default number of expected zones (5 in this experiment). In this case, the obtained zones contain multiple sections of content handling distinct subjects. Selected key terms therefore do not fully represent that zone as a whole, rather one of the zone sections.

In the future work, we consider two potential improvements of our segmentation module. The first one mixes DOM based and image based approaches to page segmentation. The second one uses the Gestalt theory [40] to simulate the similarity, proximity and complexity principles.

5.2. Key term extraction

One of the main issues with key terms extraction was the maximum size of the n-gram order, which we fixed to 6. As a result we do not always obtain coherent phrases: abrupt endings, missing beginnings, etc. At the same time augmenting n-gram order would lead to longer key terms which might affect the user's ability to comprehend and retain the information contained in these n-grams.

As already mentioned, another issue was the complex multi-section structure of certain zones which does not allow to extract one key term that would represent the zone. One possible solution is to extract several short key terms, one per zone section, and join them into one compound key term. Some zones, like menus and footers, usually contain list items, making it difficult to extract one key term per zone. A solution is, again, to produce a key term which would either contain several elements (several menu items) or a meta key term, for example 'navigation menu', which would summarize the content.

Finally, several issues are related to the corpus used to compute idf. In this implementation, it was composed of news articles, produced between 1986 and 2006. One way to extend the coverage of the corpus is to acquire new vocabulary dynamically.

5.3. Vocalization

The evaluation results indicate that our audio representations in a form of tag thunders were comparable to their visual counterparts in clarity and intelligibility (accuracy values of 0.875 vs. 0.84). However, some users indicated a somewhat artificial sound of the generated tag thunders. More experiments with different sound settings and spatialization modes are in process. Binaural recording techniques may be used to render spatial variations in tag thunders with simple stereo headsets. Since the Kali TTS is not compatible with markup languages such as VoiceXML and SSML, our solution needs to integrate a compatible TTS so that we can use industry standards.

More experiments using different prosodic strategies will need to be made in order to determine which combination of sound effects give a user the best representation of the typography and page layout.

6. Conclusion

In this article, we proposed a strategy to facilitate skimming of web pages in non-visual environments. Our solution, which we call tag thunder, involves several processing steps: segmentation of a web page into zones, extraction of key terms from each zone and finally, vocalization of the key terms in a tag thunder. Evaluation results show that participants were able to measure the correspondence between a tag thunder and a web page.

The next step is to find the best compromise between the number of zones and key terms and the perceptive capacity of users. We intend to evaluate our concept with VIPs and use their feedback to direct our future work.

Our final objective is to integrate human computer interaction into our system, specifically for in-page navigation: once a zone is selected, we want to be able to 'navigate' to and explore that zone. In that case, headsets with sensors may enable interactions with movements of the head. Combining our approach with vibro-tactile devices would lead to multi-modal systems which facilitate access to web content in non-visual situations.

7. Acknowledgments

This research work was funded by the 'Region Normandie' with the CPER NUMNIE project.

8. Website

Tag thunder generator: `https://tagthunder.greyc.fr/demo/`
Experiment (French version): `https://tagthunder.greyc.fr/demotest`

9. References

[1] G. Dias and B. Conde, "Accessing the web on handheld devices for visually impaired people," in *Advances in Intelligent Web Mastering*, ser. Advances in Soft Computing, K. Wegrzyn-Wolska and P. Szczepaniak, Eds., 2007, vol. 43, pp. 80–86.

[2] Y. Borodin, J. P. Bigham, G. Dausch, and I. Ramakrishnan, "More than meets the eye: A survey of screen-reader browsing strategies," in *International Cross Disciplinary Conference on Web Accessibility (W4A)*, 2010, pp. 1–10.

[3] F. Ahmed, Y. Borodin, A. Soviak, M. Islam, I. Ramakrishnan, and T. Hedgpeth, "Accessible skimming: Faster screen reading of web pages," in *25th Annual ACM Symposium on User Interface Software and Technology (UIST)*, 2012, pp. 367–378.

[4] J. P. Bigham, A. C. Cavender, J. T. Brudvik, J. O. Wobbrock, and R. E. Lander, "Webinsitu: A comparative analysis of blind and sighted browsing behavior," in *9th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS)*, 2007, pp. 51–58.

[5] S. Goose and C. Möller, "A 3d audio only interactive web browser: using spatialization to convey hypermedia document structure," in *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*. ACM, 1999, pp. 363–371.

[6] L. Sorin, J. Lemarié, N. Aussenac-Gilles, M. Mojahid, and B. Oriola, "Communicating text structure to blind people with text-to-speech," in *Computers Helping People with Special Needs*. Springer, 2014, pp. 61–68.

[7] F. Maurel, N. Vigouroux, M. Raynal, and B. Oriola, "Contribution of the transmodality concept to improve web accessibility," *Assistive Technology Research Series*, vol. 12, pp. 186–193, 2003.

[8] J. Virbel, "The contribution of linguistic knowledge to the interpretation of text structures," in *Structured documents*. Cambridge University Press, 1989, pp. 161–180.

[9] M. Rossi, *L'intonation: le système du français: description et modélisation*. Editions Ophrys, 1999.

[10] F. Maurel, M. Mojahid, N. Vigouroux, and J. Virbel, "Documents numériques et transmodalité," *Document numérique*, vol. 9, no. 1, pp. 25–42, 2006.

[11] F. Ahmed, Y. Borodin, Y. Puzis, and I. Ramakrishnan, "Why read if you can skim: towards enabling faster screen reading," in *International Cross-Disciplinary Conference on Web Accessibility - W4A2012, Article No. 39*, 2012.

[12] B. Parmanto, R. Ferrydiansyah, A. Saptono, L. Song, I. W. Sugiantara, and S. Hackett, "Access: accessibility through simplification & summarization," in *Proceedings of the 2005 international cross-disciplinary workshop on web accessibility (W4A)*. ACM, 2005, pp. 18–25.

[13] S. Michail and K. Christos, "Adaptive browsing shortcuts: Personalising the user interface of a specialised voice web browser for blind people," in *Data Engineering Workshop, 2007 IEEE 23rd International Conference on*. IEEE, 2007, pp. 818–825.

[14] Y. Borodin, F. Ahmed, M. A. Islam, Y. Puzis, V. Melnyk, S. Feng, I. Ramakrishnan, and G. Dausch, "Hearsay: a new generation context-driven multi-modal assistive web browser," in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 1233–1236.

[15] M. Ziat, O. Gapenne, J. Stewart, and C. Lenay, "Haptic recognition of shapes at different scales: A comparison of two methods of interaction," *Interacting with Computers*, vol. 19, no. 1, pp. 121–132, 2007.

[16] N. A. Giudice, H. P. Palani, E. Brenner, and K. M. Kramer, "Learning non-visual graphical information using a touch-based vibro-audio interface," in *Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility*. ACM, 2012, pp. 103–110.

[17] A. A. Ahmed, M. A. Yasin, and S. F. Babiker, "Tactile web navigator device for blind and visually impaired people," in *Applied Electrical Engineering and Computing Technologies (AEECT), 2011 IEEE Jordan Conference on*. IEEE, 2011, pp. 1–5.

[18] Y. B. Issa, M. Mojahid, B. Oriola, and N. Vigouroux, "Analysis and evaluation of the accessibility to visual information in web pages," in *Computers Helping People with Special Needs*. Springer, 2010, pp. 437–443.

[19] W. Safi, F. Maurel, J.-M. Routoure, P. Beust, and G. Dias, "Blind browsing on hand-held devices: Touching the web... to understand it better," in *Data Visualization Workshop (DataWiz 2014) associated to 25th ACM Conference on Hypertext and Social Media (HYPERTEXT 2014)*, 2014.

[20] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears." *Journal of the acoustical society of America, 25(5)*, pp. 975–979, 1953.

[21] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994.

[22] D. S. Brungart and B. D. Simpson, "Optimizing the spatial configuration of a seven-talker speech display," *ACM Transactions on Applied Perception (TAP)*, vol. 2, no. 4, pp. 430–436, 2005.

[23] M. Turgeon, A. S. Bregman, and B. Roberts, "Rhythmic masking release: effects of asynchrony, temporal overlap, harmonic relations, and source separation on cross-spectral grouping." *Journal of Experimental Psychology: Human Perception and Performance, 31(5)*, p. 939, 2005.

[24] C. J. Darwin, D. S. Brungart, and B. D. Simpson, "Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers." *The Journal of the Acoustical Society of America, 114*, p. 2913, 2003.

[25] J. Guerreiro and D. Gonçalves, "Text-to-speeches: evaluating the perception of concurrent speech by blind people," in *Proceedings of the 16th international ACM SIGACCESS conference on Computers & accessibility*. ACM, 2014, pp. 169–176.

[26] ——, "Faster text-to-speeches: Enhancing blind people's information scanning with faster concurrent speech," in *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*. ACM, 2015, pp. 3–11.

[27] A. Sanoja and S. Gançarski, "Block-o-matic: A web page segmentation framework," in *Multimedia Computing and Systems (ICMCS), 2014 International Conference on*. IEEE, 2014, pp. 595–600.

[28] D. Cai, S. Yu, J.-R. Wen, and W.-Y. Ma, "Vips: A vision-based page segmentation algorithm," Microsoft technical report, MSR-TR-2003-79, Tech. Rep., 2003.

[29] J. Cao, B. Mao, and J. Luo, "A segmentation method for web page analysis using shrinking and dividing," *International Journal of Parallel, Emergent and Distributed Systems*, vol. 25, no. 2, pp. 93–104, 2010.

[30] N. F. S. R. G. Adda, "Pré-segmentation de pages web et sélection de documents pertinents en questions-réponses," *TALN-RÉCITAL 2013*, p. 479, 2013.

[31] X. Liu, H. Lin, and Y. Tian, "Segmenting webpage with gomory-hu tree based clustering," *Journal of Software*, vol. 6, no. 12, pp. 2421–2425, 2011.

[32] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.

[33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *The Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[34] S. Tatiraju and A. Mehta, "Image segmentation using k-means clustering, em and normalized cuts."

[35] E. Giguet and N. Lucas, "The book structure extraction competition with the resurgence software at caen university," in *International Workshop of the Initiative for the Evaluation of XML Retrieval*. Springer, 2009, pp. 170–178.

[36] G. A. Miller, "The magical number seven, plus or minus two: Some limits on our capacity for processing information." *Psychological review*, vol. 63, no. 2, p. 81, 1956.

[37] K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[38] I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin, and C. G. Nevill-Manning, "KEA: Practical Automatic Keyphrase Extraction," *NRC/ERB-1057*, 1999.

[39] M. Morel and A. Lacheret-Dujour, "Kali, synthèse vocale à partir du texte : de la conception à la mise en oeuvre," *Traitement Automatique des Langues 42*, pp. 193–221, 2001.

[40] W. Köhler, *Gestalt Psychology.[British Ed.* G. Bells and sons, 1930.

Navigating the Spoken Wikipedia

Marcel Rohde, *Timo Baumann*

Universität Hamburg, Department of Informatics, Natural Language Systems Group, Germany
{2rohde,baumann}@informatik.uni-hamburg.de

Abstract

The Spoken Wikipedia project unites volunteer readers of encyclopedic entries. Their recordings make encyclopedic knowledge accessible to persons who are unable to read (out of alexia, visual impairment, or because their sight is currently occupied, e. g. while driving). However, on Wikipedia, recordings are available as raw audio files that can only be consumed linearly, without the possibility for targeted navigation or search. We present a reading application which uses an alignment between the recording, text and article structure and which allows to navigate spoken articles, through a graphical or voice-based user interface (or a combination thereof). We present the results of a usability study in which we compare the two interaction modalities. We find that both types of interaction enable users to navigate articles and to find specific information much more quickly compared to a sequential presentation of the full article. In particular when the VUI is not restricted by speech recognition and understanding issues, this interface is on par with the graphical interface and thus a real option for browsing the Wikipedia without the need for vision or reading.

Index Terms: accessibility, eyes-free interaction, voice user interface, Wikipedia, hyperlistening

1. Introduction

Accessibility on the web is primarily established through valid and semantically meaningful markup that can be rendered by web agents regardless of the presentation format. An auditory rendition of the web is available to persons who cannot read with screen readers which provide spoken access and rely on text-to-speech and speech synthesis. One of the problems of general text-to-speech is the broad variety of text that it has to deal with, whereas domain-restricted technology can perform better.

For Wikipedia, one of the 10 most heavily accessed websites on the web[1], there is a specific webservice (the Pediaphon[2] [1]) which offers to read out encyclopedic articles, without requiring any screen-reading software. However, while both the quality of speech synthesis itself (i.e., the process of producing artificial speech sound) and of text-to-speech technology (the process of inferring how some text should be spoken, e.g. wrt. abbreviations, phrasing, intonation, etc.) have advanced considerably in the past years [2], the quality of artificial speech still lacks compared to natural speech, even for read-out text [3]. Text-to-speech mostly performs sentence-by-sentence and hence is unable to adequately cover discourse and information structure (with some notable exceptions, e.g. [4]). Humans in contrast, do very well at presenting the information structure and this is crucial for understanding with little effort [5].

The Spoken Wikipedia[3] is a project in which volunteers read out articles from Wikipedia to provide high-quality aural access to Wikipedia for people who cannot read. Roughly a thousand articles for each of English, German and Dutch are available, each totalling around 300 hours of speech (with smaller amounts in another 25 languages). This data has recently been made accessible by Köhn et al. [6][4] who automatically aligned the audio recordings to their respective article texts using speech recognition technology. Using these alignments, we are able to relate what parts of the article are spoken at any moment in the recordings. While the resource can be useful for fostering speech technology research (e.g. training acoustic models for open-source speech recognition), we want to make the material more accessible for its original purpose, to bring natural speech to those who prefer speech over text but do not necessarily want to linearly listen to full recordings.

2. The Written and Spoken Wikipedia

Wikipedia is accepted as the standard source for encyclopedic knowledge on the web and comes in the form of a strongly interlinked *hypertext*. Hypertext adds to traditional text the means for reading along a self-chosen reading path (i.e., non-linearily), called *hyperreading* [7]. Wikipedia provides indices, extensive structural information, and – most importantly – associative links to enable hyperreading. A common strategy in hyperreading Wikipedia is *leaping* between sections of articles and between articles based on links or structure [7]. The recent advent of *find as you type* in most browsers has made *text*

[1] http://www.alexa.com/topsites
[2] www.pediaphon.org
[3] http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Spoken_Wikipedia
[4] https://nats-www.informatik.uni-hamburg.de/SWC/

SLPAT 2016, 7th Workshop on Speech and Language Processing for Assistive Technologies, pages 9–13,
San Francisco, USA, 13 September, 2016.

Table 1: Comparative statistics of spoken and written versions of the German and English Wikipedia.

		German	English
Written	# articles	1,950,022	5,174,458
	— distinguished	6,283	29,189
	average text size	5.3 kB	6.2 kB
Spoken	# articles	916	1,344
	— distinguished	314	213
	average text size	25.8 kB	26.0 kB
Spoken	articles	0.047 %	0.026 %
Coverage	— distinguished	5.0 %	0.73 %
	est. speech time	0.22 %	0.11 %

search a frequently used strategy to find information in web pages, including for users with disabilities [8].

The Spoken Wikipedia has previously only been available as a linear audio recording, omitting all the positive aspects of hypermedia and making navigation or search impossible. Our software sets out to change this.

As also mentioned by Zhang [7], a disadvantage of hyperreading is the possibility of getting lost due to the flexibility of what to read next. Getting lost may be of particular concern when *hyperlistening*, as speech is such an inherently linear medium. Our experiments below will hence focus on whether participants are able to leap through speech without getting lost (too much), by assessing whether they are successful in navigating to key information in the article.

Wikipedia contains millions of articles on all sorts of topics in the major languages, inviting the question of whether the Spoken Wikipedia's meager thousand articles per language (at least for English, German and Dutch) are of any practical relevance when browsing for spoken information, or whether a screen reader is needed in all practical use-cases anyway.

To address this concern, we compare the composition of the written and spoken collections for German and English in Table 1. As can be seen in the table, both language versions consist of several million articles each, with a small proportion of *distinguished* articles.[5] We estimate the average length of written articles on a random sample of 1,000 articles for both languages (using their size in bytes as a proxy for text length). We find that articles selected for being spoken are (a) much longer than average articles (4-5 times as long), and (b) more often come from one of the distinguished article categories. In the German Wikipedia, some 5 % of distinguished articles have been read. Nevertheless, only a tiny proportion of the full Wikipedia is available as a naturally read version (0.11–0.22 %) and we estimate that a fully read Wikipedia would have an audio duration of several decades – indicating the infeasibility of full coverage.

While high-quality synthetic voices can be rated as more natural than amateur speech [9], naturalness ratings have been shown to degrade when listening to synthesized speech for an extended period [10], making the advantage of natural speech particularly relevant for long and complex articles from the distinguished categories.

Distinguished articles also tend to be more stable with fewer relevant changes, and hence their recordings remain up-to-date for longer. Thus, while we have equipped our software with the ability to synthesize articles on-demand, our experiments reported below focus on natural speech and we focus on relatively long articles of around one hour of speech.

3. Implementation

We first explain how we postprocess the SWC to re-align text and HTML markup. We then describe the graphical and voice user interfaces of our application.[6]

3.1. Data model

The Spoken Wikipedia Corpus [6] contains per-article alignments of plain text to audio. Unfortunately, those alignments do not take into account the article structure (in terms of the HTML DOM). In addition, the text has partially been altered to ease alignment and does not fully match the text (and other elements) contained in the HTML version. We overcome this issue by using fuzzy matching to produce a document that contains all of:

- the structural hierarchy of the article,
- the timing of all time-aligned words in the article,
- the sentence segmentation from the corpus, and
- the hyperlinks contained in the article.

This enables the application to

- leap (by sentence, paragraph, or section),
- identify links close to the current timing in the article audio (and follow these links), and
- identify timings for all words (for searching).

Both the time-alignment and matching occasionally go astray or are missing some data. Our method is nevertheless robust to such errors and provides timings whenever possible. We synthesize the table of contents based on the observed article structure, as this is not spoken by the readers; other material that is not spoken by readers (e.g. tables, lists, bibliographies) remains left out.

3.2. GUI

The *graphical user interface* consists of multiple parts that can each be hidden for experimentation. It is implemented in JavaFX and depicted in Figure 1. It offers multiple ways of accessing and leaping the structure of the article, as well as access to close-by links.

[5] English articles can be distinguished as either 'good' or 'featured', where the corresponding German categories are 'lesenswert' (worth reading) and 'exzellent'.

[6] Available at `http://github.com/hainoon/wikipediareader`.

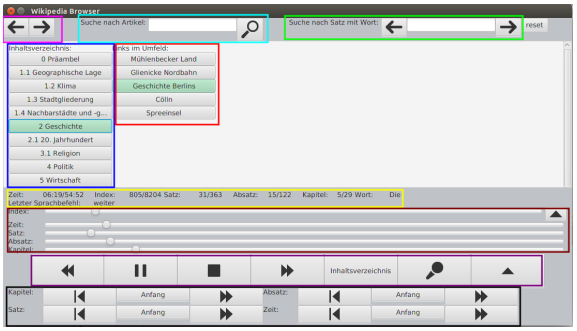


Figure 1: The full application GUI, including forward/backward jumps between articles (magenta), article search (cyan) and within-article search (green), the responsive table of contents (blue), a responsive list of currently relevant links (red), some status information (yellow), sliders indicating the relative position in the article (brown), buttons for standard audio navigation (forward/backward/pause), for listening to the table of contents, and for voice-based interaction (purple), and finally buttons to navigate the article structure: by chapter, paragraph, sentence, or jumping ahead/back by 10 seconds per click (black). In the experiments, only parts of the interface are available to users.

3.3. VUI

The *voice user interface* for navigating spoken articles consists of speech activation, recognition and rule-based language understanding with the aim of offering similar functionality as the graphical interface.

The user presses and holds down the only button in the interface to activate speech recognition. When the button is released, we decode the recording using Google's freely available Speech API [11][7].

Language understanding makes use all returned (n-best) hypotheses using a hierarchy of patterns. For robustness, patterns only need to match parts of what was spoken, allowing the user the freedom to add material such as "show me" or "now, go to". The hierarchy of rules is important as multiple rules may match a given input. N-best results are useful to deal with Google's variability in returning numbers (and other material). Users may say (variations of) the following:

- "[show me the] [table of] contents",
- "next/previous chapter/section/paragraph/sentence",
- "[go back to the] beginning of the chapter/section/paragraph/sentence" (or simply "repeat"),
- "[go to] chapter/section/subsection N",
- "*section name*" to go to the named section,
- "article *name*" to follow a link or search an article.

Our language understanding (as well as other parts of the software) currently work for English and German and would be easy to port to other languages.

[7] https://cloud.google.com/speech/

Figure 2: Setup of the user study: the experiment participant (right side) and the experimenter/wizard (left side) are separated by a dividing wall.

4. User Study

We conducted a user study to gain insight into the preferred modality for interaction, to see whether targeted navigation works as expected, and to learn about the overall usability of our software. For our experiment we disabled the search and link-following options in order to force users to stay within the article and to focus on structural navigation within the article.

Participants were given a choice of two articles so as to increase interest in the article in question. Participants where first allowed 2 minutes of 'free browsing' in the article. Afterwards, they were asked to use targeted navigation to answer three factual questions about the article in question. The facts were positioned anywhere in the article and sometimes required some combination (such as aggregation of denominations for the full proportion of religious affiliation). We compare three conditions:

GUI Users interacted using the graphical user interface as described in Subsection 3.2 above.

VUI Users interacted by speaking voice commands to the system described in Subsection 3.3. They were given a schema for possible commands.

Wizard-control As in the the VUI setting, users interacted by speaking, but were instructed to use commands as they saw fit for the task (lead to believe that this was a 'better' system). In this condition, the experimenter followed the Wizard of Oz paradigm and navigated the article according to how the speech interface *should* act absent of recognition (and ensuing understanding) errors.

12 participants (normally sighted, not regular TTS or screen reader users) took part in the study. Each participant used the system in all three conditions and we balanced for ordering effects. The first 6 participants were allowed no more than 2 minutes for each question, the remaining 6 participants were allowed a total of 15 minutes for the questions with gentle reminders to move on after 5 minutes per question. As participants were given a free choice of 2 articles for each condition, we could not balance the usage of every article.

In all conditions, users wore a headset to listen to the recording. The headset's microphone was used only in

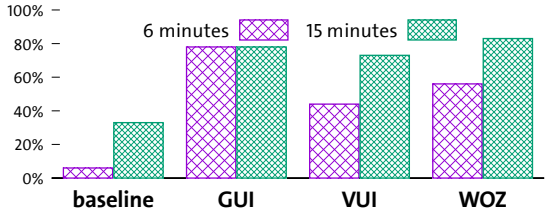


Figure 3: Proportion of questions answered after 6/ 15 minutes for the experimental conditions and a non-interactive baseline.

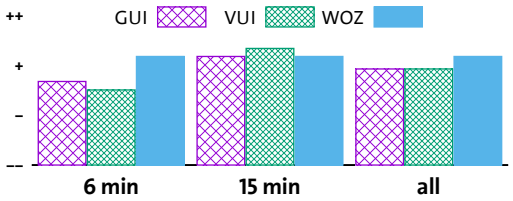


Figure 4: Average user ratings of overall interaction quality for the interaction conditions.

the VUI condition, whereas the wizard directly heard the speaker and performed commands using the GUI from a separate computer. See Figure 2 for a picture of the setup.

We asked the participants to fill out a questionnaire after the initial 'free browsing' and after targeted navigation for each interaction condition.

5. Results

We analyze our user study with respect to the participant answers to the given questions, their ratings in the questionnaire and the logged interaction behaviour. Given the low number of participants and the free choice of the read article, we do not expect results to be significant; they are, however, clearly indicative of general tendencies.

5.1. User Success

Figure 3 shows the proportion of (fully or partially) correct answers under the three experimental conditions for the first group (2 minutes per question, 6 in total) and second group (15 minutes). We add a baseline condition in which the user would not be able to navigate (and hence only be able to give answers that have occurred after a maximum of 6 and 15 minutes, respectively). As can be seen, targeted navigation greatly improves over linear listening. We find that voice-based navigation profits from longer interactions, then reaching results on par with the GUI. We want to add that a few questions were never answered correctly because the information was very hard to find given just structural navigation.

5.2. User Feedback

Figure 4 shows the overall interaction quality as reported in the questionnaires. All versions are rated as 'usable' with a slight tendency towards spoken interaction (possibly because there is no modality change between output and input as commented by one user). Users tend to rate better when they had more time to interact, indicating that only 2 minutes per question result in stress, whereas 5 minutes are sufficient. Stress could be lower in the WOZ condition in which interaction was more successful.

Users often commented that they would have liked to search by keywords, a functionality that we had excluded from the experiment. We believe that voice-based interaction will further improve when search is included.

5.3. User Behaviour

All participants interacted heavily (hyperlistened) in all conditions rather than listen linearly. In particular, they (a) navigate to sections, (b) skip ahead one section, paragraph or sentence, (c) go back one sentence when they notice that they found the desired information, or (d) pause playback. The GUI condition also shows interesting use of skipping words (presumably to save time), and in voice-based interactions users often call the table of contents (before then calling for a section). Unfortunately, we did not record statistics of whether participants prefer to call sections by name or number.

Users often pressed the push-to-talk button too late (and/or released it too early) which hindered recognition. This could easily be solved by voice activity detection. Likewise, while speech recognition worked well for some, VUI performance was greatly restricted by errors. This as well could be solved by better technology.

6. Summary and Conclusions

We have described a system for aural access to Wikipedia articles: spoken articles can be navigated via their structure, or searched by keywords and links can be followed to voice-browse the full Wikipedia (with articles synthesized if not available in a naturally spoken version). Our software enables *hyperlistening*, i.e. making use of the crucial hypertextuality of modern encyclopaedia usage without the need for reading.

We find that users are able to navigate to information in articles much quicker than if they had to listen linearly, and their usage patterns as well as comments indicate that they easily stay on top of things even without feedback about the current position in the article.

Both the graphical as well as the voice-based mode of interaction work well, at least when speech recognition error is low and enough time is available. This indicates that hyperlistening fits well with voice-based navigation and can hence be useful for persons without vision available for browsing.

Finally, while our interfaces enable browsing naturally read articles, the full Wikipedia experience includes user participation such as adding links and contents [7], or commenting on the 'talk' pages. Thus, ours are just initial steps towards a full eyes-free and speech-only access to Wikipedia.

Acknowledgments We wish to thank all Wikipedia authors and speakers for creating and maintaining the written and spoken data, as well as the participants in our experiments. We also thank Michael Blesel for selecting test articles and devising the questions used in the experiments. Finally, we wish to thank the reviewers for their helpful comments and pointers.

7. References

[1] A. Bischoff, "The Pediaphon-speech interface to the free Wikipedia encyclopedia for mobile phones, PDA's and MP3-players," in *18th International Workshop on Database and Expert Systems Applications (DEXA 2007)*. IEEE, 2007, pp. 575–579.

[2] J. Hirschberg and C. D. Manning, "Advances in natural language processing," *Science*, vol. 349, no. 6245, pp. 261–266, 2015.

[3] P. Taylor, *Text-to-Speech Synthesis*. Cambridge Univ Press, 2009.

[4] F. Kuegler, B. Smolibocki, and M. Stede, "Evaluation of information structure in speech synthesis: The case of product recommender systems," in *Speech Communication; 10. ITG Symposium; Proceedings of*, Sept 2012, pp. 1–4.

[5] J. Hirschberg and J. Pierrehumbert, "The intonational structuring of discourse," in *Proceedings of the 24th annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 1986, pp. 136–144.

[6] A. Köhn, F. Stegen, and T. Baumann, "Mining the spoken wikipedia for speech data and beyond," in *Proceedings of LREC 2016*, 2016.

[7] Y. Zhang, "Wiki means more: hyperreading in Wikipedia," in *Proceedings of the seventeenth conference on Hypertext and hypermedia*. ACM, 2006, pp. 23–26.

[8] L. Spalteholz, K. F. Li, and N. Livingston, "Efficient navigation on the world wide web for the physically disabled." in *WEBIST (2)*, 2007, pp. 321–327.

[9] K. Georgila, A. Black, K. Sagae, and D. R. Traum, "Practical evaluation of human and synthesized speech for virtual human dialogue systems." in *LREC*, 2012, pp. 3519–3526.

[10] E. Pincus, K. Georgila, and D. Traum, "Which synthetic voice should i choose for an evocative task?" in *16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, vol. 105, 2015.

[11] J. Schalkwyk, D. Beeferman, F. Beaufays, B. Byrne, C. Chelba, M. Cohen, M. Kamvar, and B. Strope, "Your word is my command: Google search by voice: A case study," in *Advances in Speech Recognition*. Springer, 2010, pp. 61–90.

Selecting Exemplar Recordings of American Sign Language Non-Manual Expressions for Animation Synthesis Based on Manual Sign Timing

Hernisa Kacorri

Carnegie Mellon University
Human Computer Interaction Institute
5000 Forbes Avenue
Pittsburgh, PA 15213 USA
hkacorri@andrew.cmu.edu

Matt Huenerfauth

Rochester Institute of Technology (RIT)
Golisano College of Computing
and Information Sciences
152 Lomb Memorial Dr, Rochester, NY 14623 USA
matt.huenerfauth@rit.edu

Abstract

Animations of sign language can increase the accessibility of information for people who are deaf or hard of hearing (DHH), but prior work has demonstrated that accurate non-manual expressions (NMEs), consisting of face and head movements, are necessary to produce linguistically accurate animations that are easy to understand. When synthesizing animation, given a sequence of signs performed on the hands (and their timing), we must select an NME performance. Given a corpus of facial motion-capture recordings of ASL sentences with annotation of the timing of signs in the recording, we investigate methods (based on word count and on delexicalized sign timing) for selecting the best NME recoding to use as a basis for synthesizing a novel animation. By comparing recordings selected using these methods to a gold-standard recording, we identify the top-performing exemplar selection method for several NME categories.

Index Terms: American Sign Language, non-manual expressions, exemplar selection, animation synthesis

1. Introduction

Being able to access information sources online has become necessary for employment, engaging in commerce, accessing government services, and in various other contexts in modern society. However, the majority of information content on the web is in the form of written-language text. There are many individuals who have difficulty reading text information sources online, including those with low literacy.

What may be less obvious is that even websites without any audio content present accessibility challenges for people who are deaf or hard of hearing (DHH). Due to a variety of factors, e.g., early language exposure or educational background, many DHH users have lower levels of written language literacy. In the U.S. context, standardized educational testing of secondary school graduates (i.e., students age 18+) has indicated that the majority of DHH graduates have English reading levels at the fourth grade or below [1], which would correspond to age 10 U.S. students. Although some DHH individuals may have difficulty reading written English, many have strong fluency in American Sign Language (ASL).

While presenting videos of ASL on websites is a simple solution, it can be difficult to update and maintain information content in the form of video. Therefore, technology to automate the creation of ASL content (in the form of animation) can make it easier and more cost-effective for companies and organizations to provide ASL content on their websites, as discussed in [2].

This paper focuses on methods for generating non-manual expressions (NMEs), i.e. face and head movements, for ASL animation. One method for producing linguistically accurate and natural NMEs is to select a pre-existing recording of a human ASL signer as a basis for the animation, as discussed in [3]. A challenge is selecting which recording in a corpus is the most suitable to serve as the basis for the face and head movements of the animated character, given that a sentence with specific lexical items (and their timings) must be synthesized. In this paper, we define four methods of considering the manual sign similarity between pairs of recordings, and we conduct an evaluation of how effective each technique is for identifying an exemplar human recording that could serve as a basis for synthesizing NMEs for ASL animations.

1.1. Background on American Sign Language and NMEs

As background, this section briefly summarizes ASL linguistics, with a focus on the use of non-manual expressions (NMEs) in the language. Researchers estimate that there are over a half-million people in the U.S. who use ASL as a primary means of communication [4]. As discussed above, many users of ASL are not fluent in written English; the two languages are linguistically distinct, with differences in word order, linguistic structure, and vocabulary. Generally speaking, movements of the hands and arms are used to indicate lexical items (ASL "manual signs"), but a complete production of ASL consists of much more than this, including head movement, facial expressions, eye-gaze, and torso movements, all of which can convey linguistic information. These additional channels of performance are commonly referred to as NMEs.

NMEs can convey a wide variety of information, including emotional connotation, variations in lexical meaning, or prosodic information. In this work, we focus on Syntactic NMEs, which are used to convey syntactic information about sentence structure. These Syntactic NMEs generally consist of movements of the upper face and movements of the head, and they are performed in parallel with phrases containing manual signs. Syntactic NMEs conveying essential grammatical information about individual words or about entire phrases or clauses [5].

In this paper, we examine five common Syntactic NMEs:

- **Negative:** The signer shakes his head left and right to indicate negated meaning (generally with some eyebrow

SLPAT 2016, 7th Workshop on Speech and Language Processing for Assistive Technologies, pages 14–19,
San Francisco, USA, 13 September, 2016.

furrowing). For instance, the addition of a Negative NME during the verb phrase "EAT APPLE" in the ASL sentence "TEACHER EAT APPLE" negates the meaning of the clause so that it means "The teacher is not eating the apple." There is a manual sign "NOT" which can optionally be inserted before the verb phrase: While the manual sign is optional, the NME is required.

- **Topic:** The signer raises his eyebrows and tilts his head backward during a clause-initial phrase that should be interpreted as a topic. For instance, a Topic NME would occur during "APPLE" in the sentence "APPLE TEACHER EAT," which translates to English as "As for the apple, the teacher is eating it."
- **Rhetorical:** The signer raises his eyebrows and tilts his head backward and to the side to indicate a rhetorical question. ASL Rhetoricals are immediately answered by signer. For instance, "TEACHER BUY WHAT APPLE" with Rhetorical NME during "WHAT" translates to English as "What is the teacher buying? An apple."
- **Yes-No Question:** The signer raises his eyebrows while tilting the head forward to indicate that the sentence is a yes-or-no question. For instance, the introduction of a Yes-No Question NME during the ASL declarative sentence "TEACHER EAT APPLE" (English translation: "The teacher is eating an apple.") creates a polar question: "Is the teacher eating an apple?"
- **WH Question:** The signer furrows his eyebrows and tilts his head forward during a sentence to indicate an interrogative question, typically with a "WH" word such as what, who, where, when, how, which, etc. For example, this NME would occur during the ASL sentence "TEACHER EAT WHAT," which translates to English as "What is the teacher eating?"

1.2. Prior Work on NME Animation Synthesis

As discussed in Section 1, posting videos of human signers is not a viable method for providing ASL content on websites. If information must be frequently updated, then re-recording a video of a human signer would be prohibitively expensive; furthermore, a video-based approach would not enable real-time generation of content from a user query. For this reason, "synthesis" software is needed that can convert from a script of an ASL sentence into a full animation of a virtual human performing ASL. This script of the sentence could be generated by a knowledgeable human author or by machine translation software (as the state-of-the-art of machine translation tools for ASL improve in the future). Given the sequence of words in the sentence, the synthesis software must plan the movements of the virtual human character so that the resulting animation is linguistically accurate, understandable, and acceptable by DHH users.

Many researchers have investigated the design of sign language synthesis systems, including research that has specifically focused on the generation of non-manual expressions [6, 7, 8, 9, 10]. Traditionally, researchers select a single recording of how a non-manual expression is performed, and they trigger this movement in parallel to the movements of the virtual human's hands.

In prior work, we have investigated data-driven methods for synthesizing the NMEs of the virtual human. Specifically, our prior work has made use of a small corpus of recordings of a female native signer performing ASL sentences with NMEs. This corpus is relatively small in size, and it has been divided into sub-corpora of sentence recordings for different categories of NMEs (Negation, Rhetorical, Topic, WH Question, Yes-No Question). See Table 1. (We note that sign language corpora are generally small in size, given the resource-intensive nature of obtaining these recordings and the annotation of manual-sign and NME information for individual frames of video.) This corpus was recorded and annotated at Boston University, as described in [3, 11]. The annotations include the timing and identity of manual signs and NMEs, and the videos have been processed by computer vision software [12] to create streams of MPEG4 Facial Action Parameters, which are numerical representations of the movements of various key points on the face [13].

Table 1: *NME corpus characteristics, including the duration of each recoding, in video frames and number of words.*

NME Category (Number of Recordings)	**Video Frames** min - max (mean)	**Num. of Signs** min - max (mean)
Negation (55)	10 - 76 (38.1)	2 - 7 (3.56)
Rhetorical (13)	11 - 46 (28.3)	1 - 4 (3.0)
Topic (96)	5 - 54 (15.5)	1 - 4 (1.43)
WH Question (14)	15 - 69 (31.2)	1 - 5 (2.2)
Yes-No Ques. (21)	9 - 78 (34.6)	2 - 6 (3.6)

Given this resource, our prior work has examined two possible methods for generating animations:

- We have used multidimensional dynamic time warping (DTW) on the MPEG4 FAP values to calculate pairwise similarity between all of the recordings in each sub-corpus, and we calculated the centroid recording in each set, with the minimum pairwise distance to all other members. Assuming that this recording was "most typical" of that category of NME, we used that recording as the basis for synthesizing animations of novel sentences [3].
- We subsequently investigated the use of a generative model of time-series data (Continuous Profile Models) to calculate an underlying "latent-trace" of a set of multiple recordings [11]. We used this latent-trace technique to intelligently "average" across multiple examples of each NME.

A common processing step that is necessary before using either of these two approaches listed above is that we must identify a set of recordings that will serve as the basis for producing a new animation. In prior work, we took the simplistic route of using all of the recordings in our corpus that included the specific category of NME (e.g. Topic) as the "basis set" for calculating our centroid or our latent-trace NME. However, some of those recordings may not have served as good examples of how our virtual human should move, perhaps due to differences between the sentence structure of the corpus recordings and the structure of the sentence we need to synthesize. The premise of this paper is that the selection of a basis set could be determined in a more sophisticated and discerning manner than simply using every recording of that NME category.

1.3. Input to NME Animation Synthesis

To better define the specific task that is the focus of this paper, we list the information and resources that are available during the generation of an ASL NME performance for an animation:

- We assume that the sequence of lexical items has already been determined for the sentence that must be generated. In addition to the identity of each word, we know the timing of when the lexical items begin and end (based partially on the timing information for each sign in the lexicon of our animation system).
- We assume that we already know which spans of lexical items in the sentence need to have an NME performed in parallel. For instance, given an ASL sentence "OLD BOOK I LIKE," we have already selected that a Topic NME should occur during the words "OLD BOOK." In fact, we presented initial research on how to perform this step of the process at SLPAT 2015 [14].
- Finally, we have our corpus of ASL sentence recordings, consisting of videos, the MPEG4 FAP values, and linguistic annotation of manual signs and NMEs (including the video frame numbers when each begins and ends).

2. Basis-Set Selection Techniques

Our task is to determine which of the recordings in our corpus should be included in the basis set for synthesizing an NME performance. Ideally, we would like to select recordings that are similar to the sentence we seek to synthesize. Given the few inputs to our task (listed in the previous section), there are limitations on the types of information that we may consider when defining strategies for selecting items for the basis set: namely, the identity and timing of the manual signs or NMEs. The intuition behind the basis-set selection strategies investigated in this paper is that we may prefer to select sentences with a similar number of words, a similar duration, or similarities in the patterns of the timing of the manual signs. Our selection metric should have the following properties:

1. Two phrases with a similar number of words or with a similar overall time duration should be scored as being similar.
2. Two phrases in which the beginning and ending timings of the words they contain align closely should be scored as similar.
3. Given the small size of our corpus, considering lexically specific information is impractical. Thus, we will consider the timing of manual signs in a "delexicalized" manner; that is, we will replace the sign labels such as "OLD" or "BOOK" in our corpus with a single token, e.g., "SIGN." This, we will not consider the labels of the specific words/glosses – only their timing.
4. A natural unit of time granularity for our analysis is the time duration of a single frame of video, since this is the basis for the linguistic annotation of word and NME timing for the recordings.

2.1. Comparing Temporal Language Signals

Prior to inventing a new metric for scoring the word-timing similarity of recordings of ASL sentences, we first examined the computational linguistic and automatic speech recognition (ASR) literature to examine the methods used to compare language signals with temporal information, specifically those techniques that have been used to evaluate the output of ASR systems against gold-standard annotations of the speech transcript. While there are a variety of metrics used to compare string output, e.g. [15], most techniques are focused on penalizing incorrect string transcription of the speech audio, and thus, scoring techniques rarely incorporate temporal alignment into the score. In our case, we are considering delexicalized word timing similarity.

Researchers focused on ASR temporal alignment accuracy have proposed a variety of metrics, e.g. average word boundary shift [16], and researchers studying speaker-segmentation in recordings of meetings have proposed metrics such as Diarisation Error Rate [17]. However, in both cases, these metrics assume that there will be some word label or speaker-ID correspondences across the two time-annotated transcriptions. Since, for our task, we are focused on delexicalized timing similarity, these previously invented metrics are not well-suited.

As discussed in the next section, some of our proposed metrics make use of Inside-Outside-Beginning (IOB) labelling. For this reason, we also considered comparison metrics in the named entity detection and information extraction literature. While the output of many systems consists of IOB labelling of the tokens in a string, the traditional evaluation metrics in this field are based on per-token precision, recall, or F-score [18]. Such metrics are ill-suited to evaluating fine-grained IOB similarity at the video-frame level, as in our situation. Some authors propose metrics to support evaluation of partial-matches (in which a system's named-entity tagging partially overlaps with the true gold-standard labeling) [19]. However, even these metrics do not consider the temporal dimension at a fine-enough granularity for our task.

2.2. Techniques Examined in This Paper

Since we did not find a suitable pre-existing metric for comparison of the delexicalized timing similarity of the manual component of two ASL sentences, we invented four sets of basis-set selection approaches (and a simplistic baseline), which we will investigate and compare in this paper:

- **Baseline Method.** This simplistic method for defining the basis set was used in our prior work [3]: We filter the corpus, leaving only those recordings containing the specific category of NME that we seek to generate (e.g., Topic). For this baseline approach (and in all of the other approaches listed below), we select and extract the portion of each recording that coincides with the span of time when the NME is occurring in that sentence (based on the linguistic annotation). Thus, if a Topic NME occurs during the first two words of some recording, then we extract the portion of the recording corresponding to this period of time for inclusion in the basis set.
- **Word Count.** This technique is based upon the intuition that an NME that occurs during a portion of a sentence with a large number of words may differ from an NME that occurs during a portion of a sentence containing few words. For instance, facial expressions with periodic movements, such as the head shaking that occurs during Negation, may consist of a larger number of individual movements when it occurs during a longer verb phrase. In this technique, we first filter for only those recordings that contain the category of NME we need to generate (e.g., Negation), as in the baseline approach above. Next, we count how many words co-occur with the NME in each recording, and we select items for the basis set that have a similar number of words within the timespan of the NME. Thus, if we must generate an ASL sentence with a Negation during a verb phrase consisting of five words, then we would prefer to select recordings

from our corpus that contain Negation performances during an identical (or similar) number of words.

- **Frame Count.** This technique is similar to the above, except we use the time duration of the NME (measured according to the number of video frames) as the similarity metric. This, if we needed to generate an ASL animation with a Topic facial expression that must last for 25 frames, then we would prefer to select Topic NME recordings of a similar duration from our corpus for inclusion in the basis set. (Our video recordings have a frame-rate of 30 frames per second.)
- **Levenshtein IOB.** In this technique, we pre-process each of the sentence recordings to generate a string consisting of the letters "I," "O," or "B," representing Inside, Outside, or Beginning, in the following manner: For each frame of video, we add one character to the string, based on whether this frame of video is the Beginning of a manual sign (the single video frame where this word begins), Inside (during) a manual sign, or Outside of a manual sign (i.e. during a period of time in-between signs or before/after all signs in the recording). Thus, an ASL sentence recording of duration 20 frames containing the words "BOOK" (frame 3 to 8) and "LOST" (frame 10 to 15) appears as: `OOBIIIIIOBIIIIIOOOOO`. We select all of the recordings in the corpus that contain the same category of NME (e.g., Topic) as the one we need to generate, and we focus on the IOB substring that corresponds to the time duration of each NME. To calculate similarity between pairs of substrings, we calculate the Levenshtein distance (with equal penalty for insertion, deletion, and substitution, with normalization based on the length of the shorter substring). The intuition behind this technique is that it may capture the temporal structure of a recording in a delexicalized manner such that we would prefer to include recordings in the basis set that consist of NME recordings with a similar number of words with similar word durations and timing.
- **Bigram IOB.** This technique uses a similar IOB string representation as above. After extracting the substrings that correspond to all examples of the category of NME we must generate (e.g. Rhetorical), then we count all character bigrams in each IOB substring. These counts are stored in a vector corresponding to each string; to calculate the similarity between a pair of recordings, we use the cosine similarity between their vectors. The intuition behind this approach is that it may capture some information about both word count (based on the number of n-grams containing the "B" character), and it would also indicate overall time duration (with longer recordings having higher counts in the vector cells).

3. Evaluation of Selection Techniques

Given the inputs described in section 1.3, a good basis-set selection technique would identify a subset of ASL recordings in our corpus that contain similar face and head movements to what a human would perform for the ASL sentence that we seek to synthesize.

3.1. Scoring Metric Used in This Evaluation

In prior work presented at SLPAT 2015, we demonstrated that multidimensional dynamic time warping (DTW) operating in the space of MPEG4 Facial Action Parameters can assign similarity scores to pairs of ASL NME recordings that correlate with the judgements of native ASL signers [20], and we defined a refined version of this scoring algorithm in [3]. This scoring algorithm provides a numerical score of the similarity in the face and head movements (specifically the eyebrows and head displacement/orientation) between any pair of ASL recordings. In the evaluation presented below, we use this multidimensional DTW scoring algorithm to evaluate how well each of the selection techniques is able to chose a basis set with recordings that are similar to gold-standard human performances.

3.2. Evaluation Methodology

We compared the efficacy of each of the five techniques listed in section 2.2, for each of the categories or NME in our corpus (Negation, Rhetorical, Topic, WH Question, Yes-No Question), using a leave-one-out evaluation paradigm, described below. To explain the process more clearly, we will discuss, by way of example, how the process occurs for the WH Question recordings.

1. We extracted a set of all the recordings in our corpus for this category of NME (e.g., there were 14 recordings of WH Question in our corpus). We iteratively held-out each of the recordings in this set (i.e., we repeated this process for all 14 items in the set of WH Question recordings), and we consider the held-out recording to be a gold-standard of how a human should move his face and head when performing the NME for the given sequence (and timing) of manual signs in this sentence. The remaining 13 recordings are used as the superset from which the basis set must be drawn, for this held-out recording.
2. For each of the basis-set selection techniques, we identify a subset of the recordings that are predicted to yield NME movements that are similar to the gold-standard held-out recording. We use each of the five selection techniques to identify a (potentially) different basis set.
   (a) For the baseline method, this is trivial: In the case of WH Question, we would simply use all 13 of our non-held-out recordings in the superset in order to form our basis set.
   (b) For the remaining four selection techniques, the similarity scoring methods defined in section 2.2 enable us to assign a score to each of the 13 WH Question recordings. For each of the selection techniques, we select the top 5 most similar recordings to form a basis set. Thus, each of the four selection techniques will be used to produce its own basis set (with cardinality 5), and each basis set may have different membership, as determined by that selection technique.
3. To evaluate the quality of the basis set chosen by each selection technique, we must compare how well the face and head movements of each of the recordings in the set matches the face and head movements of the held-out recording (considered as a gold standard). Using the DTW metric from [3] mentioned above, we calculate the distance between each of the recordings in the basis set and the held-out gold-standard recording. To produce a single score for each basis set, the individual distance-to-gold-standard scores for the members of the set are averaged to produce a single score.

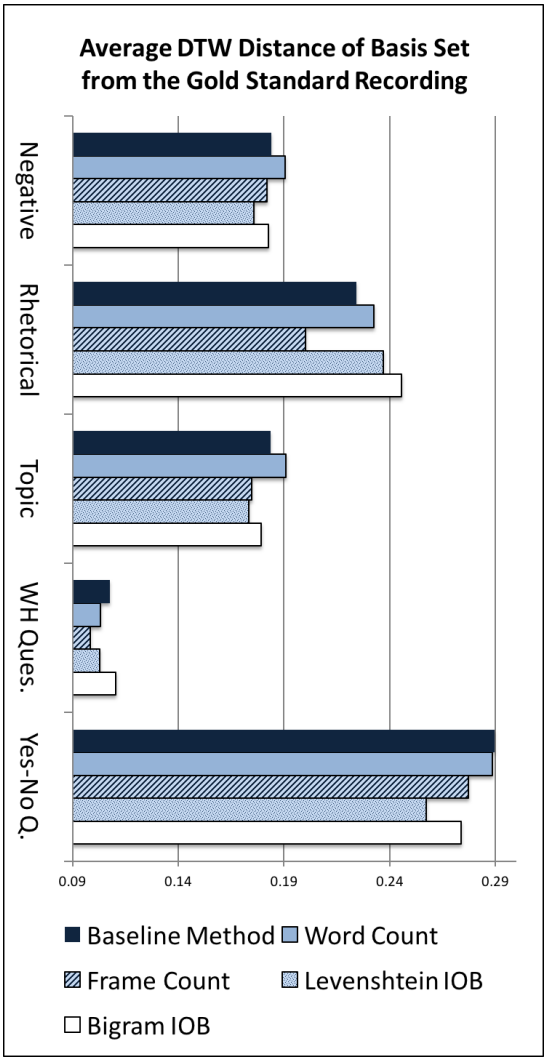


Figure 1: *Average DTW distance between basis set members and gold-standard sentences, for each NME category, for each selection technique. Note that smaller bars are better.*

4. Discussion of Results

As shown in Figure 1, at the end of the evaluation process, for each NME category, for each of the five selection techniques, we have a single score that represents how well that selection technique was able to identify a basis set of recordings from our corpus that were similar to human performance of that NME for the held-out gold-standard sentences.

For three of the NME categories (Negative, Topic, and Yes-No Question), the best performing selection technique was Levenshtein IOB. For the Rhetorical and WH Question categories, the best performing selection technique was Frame Count. (For WH Question, the performance of all of the selection algorithms was quite close, with Levenshtein IOB in second place.)

Our corpus contains relatively few recordings of Rhetorical (13) and WH Question (14), and due to the nature of how these NMEs are used in ASL, many of these recording examples occur during phrases consisting of a single word (e.g., often a single WH-word). We speculate that the difference in efficacy of the selection techniques for these two categories may relate to the relatively low cardinality of examples in our dataset and the relatively short duration of these NMEs.

No selection algorithm obtained the best (lowest) distance scores across all five categories of NME, and in principle, it is reasonable that a different selection technique could be best suited to each of the NME categories. This could be due to the way in which the lexical timing of manual signs may influence how that particular NME is performed by ASL signers.

5. Conclusions

This paper has investigated techniques for selecting a subset of recordings from a corpus that can be used as a basis for synthesizing the Syntactic NMEs for a sentence to be generated, based only on information about the delexicalized manual sign timing of the sentence. By identifying a set of similar recordings for inclusion in this basis set, various approaches can be used to select a single recording [11] or to identify a latent-trace of the set [11], in order to plan the face and head movements of a virtual human in the ASL animation. Ultimately, the goal of this work is to improve the state of the art of sign language animation synthesis technologies, especially since prior studies have demonstrated that the understandability of such animations is affected by the quality of the synthesized NMEs. Such technology has potential to make it easier for organizations to provide sign language content on websites in a manner that is more efficient and easier to maintain, which may increase the prevalence of such content online.

In future work, we plan to evaluate the efficacy of these basis-set selection techniques within the context of a full animation synthesis pipeline. By performing final animation production step, we can generate stimuli for display in user-based evaluation studies, in which native ASL signers could view animations generated using these selection algorithms as an intermediate pipeline stage. In this way, we can determine the degree to which the differences in efficacy identified in this study may influence DHH users' perception of the linguistic accuracy and understandability of the resulting animations.

In this study, we found that the Levenshtein IOB metric was most effective at selecting basis set recordings for three of the five NME categories in this study, and the number of recordings in our corpus for the remaining two categories (Rhetorical and WH Question) was relatively small. In future work, we are interested in acquiring additional ASL recordings of these NMEs from multiple signers so that we may repeat this analysis on a larger testing set.

6. Acknowledgements

This material is based upon work supported by the National Science Foundation under award number 1065009 and 1506786. This material is also based upon work supported by the Science Fellowship and Dissertation Fellowship programs of The Graduate Center, CUNY. We are grateful for support and resources provided by Ali Raza Syed at The Graduate Center, CUNY, and by Carol Neidle at Boston University.

7. References

[1] C. B. Traxler, "The stanford achievement test: National norming and performance standards for deaf and hard-of-hearing students," *Journal of deaf studies and deaf education*, vol. 5, no. 4, pp. 337–348, 2000.

[2] M. Huenerfauth, "Generating american sign language animation: overcoming misconceptions and technical challenges," *Universal Access in the Information Society*, vol. 6, no. 4, pp. 419–434, 2008.

[3] H. Kacorri, A. Syed Raza, M. Huenerfauth, and C. Neidle, "Centroid-based exemplar selection of asl non-manual expressions using multidimensional dynamic time warping and mpeg4 features," in *Proceedings of the 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining, The 10th International Conference on Language Resources and Evaluation (LREC 2016), Portoroz, Slovenia*, 2016.

[4] R. E. Mitchell, T. A. Young, B. Bachleda, and M. A. Karchmer, "How many people use asl in the united states? why estimates need updating," *Sign Language Studies*, vol. 6, no. 3, pp. 306–335, 2006.

[5] C. Neidle, D. Kegl, D. MacLaughlin, B. Bahan, and R. Lee, "The syntax of asl: functional categories and hierarchical structure," 2000.

[6] C. Schmidt, O. Koller, H. Ney, T. Hoyoux, and J. Piater, "Enhancing gloss-based corpora with facial features using active appearance models," in *International Symposium on Sign Language Translation and Avatar Technology*, vol. 2, 2013.

[7] S. Gibet, N. Courty, K. Duarte, and T. L. Naour, "The signcom system for data-driven animation of interactive virtual signers: Methodology and evaluation," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 1, no. 1, p. 6, 2011.

[8] N. Adamo-Villani and R. B. Wilbur, "Asl-pro: American sign language animation with prosodic elements," in *International Conference on Universal Access in Human-Computer Interaction*. Springer, 2015, pp. 307–318.

[9] M. Huenerfauth and P. Lu, "Modeling and synthesizing spatially inflected verbs for american sign language animations," in *Proceedings of the 12th international ACM SIGACCESS conference on Computers and accessibility*. ACM, 2010, pp. 99–106.

[10] M. Huenerfauth, L. Zhao, E. Gu, and J. Allbeck, "Evaluation of american sign language generation by native asl signers," *ACM Transactions on Accessible Computing (TACCESS)*, vol. 1, no. 1, p. 3, 2008.

[11] H. Kacorri and M. Huenerfauth, "Continuous profile models in asl syntactic facial expression synthesis," in *ACL 2016: the 54rd Annual Meeting of the Association for Computational Linguistics*. Curran Proceedings, 2016.

[12] T. Visage, "Face tracking," https://visagetechnologies.com/products-and-services/visagesdk/facetrack, 2016, accessed: 2016-03-10.

[13] I. S. Pandzic and R. Forchheimer, *MPEG-4 facial animation: The standard, implementation and applications*. Wiley, 2003.

[14] S. Ebling and M. Huenerfauth, "Bridging the gap between sign language machine translation and sign language animation using sequence classification," in *Proceedings of the 6th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, 2015.

[15] S. Dobrišek and F. Mihelic, "Criteria for the evaluation of automated speech-recognition scoring algorithms," *Electrotechnical Review*, vol. 75, no. 4, pp. 229–234, 2008.

[16] L. Chen, Y. Liu, M. P. Harper, E. Maia, and S. McRoy, "Evaluating factors impacting the accuracy of forced alignments in a multimodal corpus." in *LREC*, 2004.

[17] S. Tranter, K. Yu, G. Everinann, and P. C. Woodland, "Generating and evaluating segmentations for automatic speech recognition of conversational telephone speech," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 1. IEEE, 2004, pp. I–753.

[18] M. Marrero, S. Sánchez-Cuadrado, J. M. Lara, and G. Andreadakis, "Evaluation of named entity extraction systems," *Advances in Computational Linguistics, Research in Computing Science*, vol. 41, pp. 47–58, 2009.

[19] S. Atdağ and V. Labatut, "A comparison of named entity recognition tools applied to biographical texts," in *Systems and Computer Science (ICSCS), 2013 2nd International Conference on*. IEEE, 2013, pp. 228–233.

[20] H. Kacorri and M. Huenerfauth, "Evaluating a dynamic time warping based scoring algorithm for facial expressions in asl animations," in *6th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, 2015, p. 29.

Effect of Speech Recognition Errors on Text Understandability for People who are Deaf or Hard of Hearing

Sushant Kafle[1]*, Matt Huenerfauth*[1]

[1]Rochester Institute of Technology (RIT)
Golisano College of Computing and Information Sciences
20 Lomb Memorial Drive, Rochester, NY 14623
sxk5664@rit.edu, matt.huenerfauth@rit.edu

Abstract

Recent advancements in the accuracy of Automated Speech Recognition (ASR) technologies have made them a potential candidate for the task of captioning. However, the presence of errors in the output may present challenges in their use in a fully automatic system. In this research, we are looking more closely into the impact of different inaccurate transcriptions from the ASR system on the understandability of captions for Deaf or Hard-of-Hearing (DHH) individuals. Through a user study with 30 DHH users, we studied the effect of the presence of an error in a text on its understandability for DHH users. We also investigated different prediction models to capture this relation accurately. Among other models, our random forest based model provided the best mean accuracy of 62.04% on the task. Further, we plan to improve this model with more data and use it to advance our investigation on ASR technologies to improve ASR based captioning for DHH users.

Index Terms: Accessibility for People who are Deaf or Hard-of-Hearing; Captioning System; Speech Recognition; Human Computer Interaction; Computer Linguistics

1. Introduction

Captions provide a way to represent aural information in visual text for people who are Deaf or Hard-of-Hearing (DHH). Today there are more than 360 million people worldwide with hearing loss [1] and they use services such as captioning to get access to information existing in the form of speech such as information from mainstream classes, meetings, and live events. Several methods have been explored in providing such a service; a popular alternative includes the use of captionist to transcribe audio information to text using a keyboard, with the captions displayed on a screen for those in attendance. Captioning services produce a digital textual output which can be processed and represented in various forms easily, or it can be stored as a transcript, making it useful in various scenarios such as classrooms and meetings, where it could be reviewed later.

Over the past few decades, automated speech recognition (ASR) technologies have seen major progress in their accuracy and speed. With its increasing maturity, ASR technologies are now being used commercially for many consumer applications. Due to their cheap and scalable ability (compared to other captioning alternatives) to generate real-time text from live audio or recordings, ASR systems have a potential for the task of captioning. Researchers have begun to investigate the suitability of ASR to automate or semi-automate the process of captioning with the use of ASR systems [2, 3, 4, 5] in various application settings.

Despite the growing use of ASR systems, accurate, large-vocabulary, continuous speech recognition is still considered an unsolved problem; the performance of ASR system is not on par with humans [6], who currently provide most caption text for DHH users. Due to unpredictable ambiguity in human speech and ever existing noise, ASR systems often make errors, and it is likely that this technology will continue to be imperfect in the near future as well. Researchers have also argued that ASR generated errors on captions are more comprehension-demanding than human produced errors [7, 8]. While all users of ASR technology must cope with errors in the output, there is potential that this issue has greater significance when focusing on applications for DHH users. Past research has indicated that the majority of deaf high school graduates in the U.S. have an English literacy level at the fourth grade or below [9], and approximately 20% leave school with a reading level at or below second-grade [10]. This presents a huge challenge for caption acceptance by DHH individuals given the error-probable output from ASR.

For a successful use of an ASR system in captioning, errors that affect comprehension of a caption for DHH users might need to be appropriately reduced or at least sufficiently modulated. It may be the case that some classes of errors from an ASR system are especially problematic for DHH users (perhaps based on their unique English literacy profile), and other classes of errors are less problematic. Understanding this trade-off could make way for designing an adaptive ASR system optimized for the task of captioning, specifically for DHH users.

In this paper, we present a method to study the effect of different ASR-generated errors on the understandability of a text for DHH users. For our task, we formulate a user study with DHH users who are given imperfect English texts (containing ASR errors) and asked to answer some questions based on the information from the text. With the data collected from the user study, we model the relationship between ASR errors and the impact it has on the understandability of a text for DHH users. We also discuss the possible application of this model in designing a custom loss function that could be utilized during the decision making process of the ASR to produce better outputs for captioning for DHH users.

2. Background: N-best list Rescoring Technique

In an ASR system, the function of the decoder is to find the most likely word sequence given the sequence of audio features. Although decoders are designed primarily to find a single solution, in practice, it is relatively simple to generate not just the most

SLPAT 2016, 7th Workshop on Speech and Language Processing for Assistive Technologies, pages 20–25,
San Francisco, USA, 13 September, 2016.

	Transcription	WER loss	Avg. Understandability loss
Reference	The meeting today has been cancelled and is scheduled for next Thursday.	NA	NA
ASR Hypothesis 1	The *meet in* today has been cancelled *an* is scheduled for next Thursday.	25%	8.425%
ASR Hypothesis 2	The meeting today has been *capital* and is *skidoo* for next Thursday.	16.67%	46.425%

Table 1: Example shows how Understandability loss penalizes texts containing different errors as compared to WER loss. Higher loss value indicates worse output for the metric.

likely hypothesis but the n-best set of hypotheses. Therefore, in most ASR systems, along with the most likely word sequence, a list of n-best hypotheses can also be obtained as output. Other compact forms of representation of this n-best hypotheses list are also commonly used such as a word lattice representation [11] or a confusion network [12].

These representations have been popular especially because they provide a reduced search-space (out of all possible word sequence) that can be further decoded, with more flexibility, to improve the ASR output. This post processing technique of "rescoring" or "reranking" candidate hypotheses also allows for general-purpose hypothesis to be tuned in a domain-specific or user specific way without having to design the whole ASR engine to do so [13]. Furthermore, the n-best hypotheses generated as an output from the ASR system can be processed with complete independence from the ASR system; thus, it can be treated as a separate stage in an ASR pipeline.

Researchers [14, 15, 16, 17, 18, 19] have utilized various rescoring techniques to select the best hypothesis from an ASR n-best hypotheses. In [19], Stockle et al. presented an N-best list rescoring algorithm to improve upon the shortcomings of the ASR decoding process to produce more accurate output. A standard Hidden Markov Model (HMM) based ASR system uses Maximum A Posteriori (MAP) technique as a decoding criterion. The problem with the application of the MAP approach to speech recognition is that it is sub-optimal with respect to minimizing the number of word errors in the system output. Instead, it has been shown to minimize sentence error rate which is only loosely linked to the recognition Word Error Rate (WER) [19]. Subsequently, Stolcke et al. [19] proposed a rescoring algorithm that explicitly minimizes expected word error for recognition hypotheses. In [20] researchers provided a Decision Theoretic perspective to the work from [19] as a Bayes decision rule under word error loss, as shown in Equation (1).

$$\delta(X) = \underset{W \in W}{\operatorname{argmin}} \sum_{W' \in W} WER(W, \delta(X)) P(W'|X). \quad (1)$$

Goel et al. [20] proposed a modified loss function (as shown in Equation (2)) to be minimized during the modified decoding process by adding additional degree of freedom which can be "tuned" appropriately during training. Additionally, [20] also make simplifying assumptions to compute $P(W|X)$ with joint distribution $P(X, W)$ which are accessible from the n-best lists.

$$l(W, \delta(X)) = [WER(W, \delta(X))]^x. \quad (2)$$

This framework suggested by [20] provides a flexible way to incorporate a custom loss function in the decoding process of ASR, and this approach is how we intend to adapt ASR in our work. This Minimum Bayes Risk (MBR) based decoding has been shown to provide statistically significant improvements in recognition task compared to MAP based decoding as it explicitly incorporates task performance criterion to the decoding process of ASR. Successes of hypotheses scoring systems like ROVER [14] (and its variants) has been credited to MBR based decoding to directly improve WER. Several research groups have investigated this method of decoding in recent years [21, 22, 23].

3. Design & Implementation

The approaches discussed above utilize an n-best list rescoring technique to improve the WER of an ASR system. We propose to compare the efficacy of these rescoring approaches for optimizing ASR for real-time captioning, a task for which there may be better metrics than WER. We propose to learn a custom loss function (based on the analysis of data from experiments with DHH users) to optimize the comprehensibility of ASR output for DHH users. Unlike WER, our loss function may provide a better measure of text understandability for this group of users. Table (1) shows a comparative example of our loss function (based on the data and modeling presented later in Section 3.2 of the paper) against the traditional WER loss. In the example, we can see how our Understandability model prefers Hypothesis 1 over Hypothesis 2 as compared to WER metric which does the opposite.

This paper, in general, is about creating this loss function using a prediction model which captures the relation between different types of error and their impact on the understandability of sentence for DHH users. As a final step (in the future), we will be looking to see if this loss function can be incorporated into the decision-making process of an ASR system, following the framework provided by [20], such that the ASR can produce output that is optimized to be more comprehensive for our user group.

3.1. User Study

We performed a user study with a goal of understanding how ASR errors affect DHH users' performance on a comprehension task, given that a text contains some ASR generated errors. In this study, users were presented with imperfect English text passages (containing artificially inserted errors, based on real ASR errors for that passage) and were asked to answer questions that required understanding the information content of those passages. Based on the answers, we collected Comprehension Scores for the respective questions, which we subsequently used to model the relationship between errors in the text and its comprehensibility.

3.1.1. Error Categories used in Designing Stimuli

To guide our creation of stimuli for the user study, we established a hierarchical classification of various sub-types of ASR errors (based on a time-alignment between the ASR output and the gold-standard). Broadly, ASR errors can be categorized into three types: substitution, deletion and insertion errors. Further, we divided substitution errors into four types: one to one substitution, one to many substitution, many to one substitution and many to many substitution. One to one substitution refers to the errors when one word is substituted by the other. One to many substitution errors are the error due to substitution of one word by many (for e.g., *undistinguished* substituted by *on distinguished*). Similarly, many to one errors are the errors when many words are substituted by a single word. Many to many errors corresponds to a multi-word span of text in the reference transcript with inaccurate recognition such that none of the word boundaries within the span align with those within the corresponding span of ASR output. We further subcategorized one to one substitution errors into three types namely, morphologically similar substitution, phonetically similar substitution and remaining other types of substitution errors. The morphologically similar errors are the errors where the actual word is substituted by another word with an inflectional or derivational morphological relationship to the first (for e.g., *developed* substituted by *develop*). The phonetically similar errors are the errors due to the substitution of a word by another word with similar phoneme representation; for example, the words *table (T EY B AH L)* and *stable (S T EY B AH L)* have a very close ($\geq$ 60% match) phoneme structure so they are considered as a phone neighbor of each other.

These categories of different error types were meant to be a coarse categorization of the errors and was used as a basis for ensuring that the stimuli presented in our user study contained a good mixture of different error types.

3.1.2. Study Resources

For the user study, we created a dataset of 20 passages (average length 117 words), with each passage containing three sentences marked as our Region Of Interest (ROI). For example, the text below shows a sample text passage used in the study with three bold sentences representing the three ROIs in the text.

People who study film music often complain about the lack of recognition their field receives. The study of film music is an interdisciplinary field, falling in between cinema studies and musicology. *This is one of the reasons why it receives so little attention. For example, when film music scholars, who often do not have music-degree credentials on par with the pure musicologists, write about film soundtracks, their articles are often ignored by the musicologists.* ***Conversely, when the work of film music scholars touches on the visual aspects of film, the cinema studies people often treat it as the work of amateurs.*** *So with the members of the two fields most closely related to it ignoring it, it is easy to understand why members of the film music field feel a degree of frustration.*

The questions for passages was designed in such a way that each question was based on information from only one of the ROI sentences in that passage. In total, each passage had three text-explicit questions. As described in [9], text-explicit questions measure exact recall from the text without requiring any inferential use of information from the reader's memory.

The text below shows an example of a question asked during the user study. The question is based on the reading text shown above as an example. This question, in particular, is based on information from the first ROI sentence of the reading text.

A. According to the passage, what do film music students often complain about:

☐ *that their field doesn't receive the recognition they deserve.*
☐ *people who study film music are not recognized.*
☐ *film music study is not up-to the par.*
☐ *extra attention that their field receives.*

For each ROI sentence, an average of 8 different variations were generated where each variation was produced by inserting at most one category of ASR error into the ROI sentence. To produce each variation of the ROI, we began with a perfect text and inserted one of those errors. The text below shows an example of an ROI sentence without any errors:

Conversely, when the work of film music scholars touches on the visual aspects of film, the cinema studies people often treat it as the work of amateurs.

We produced different variations of this ROI text by adding ASR generated errors into the sentence. ASR generated errors were collected by creating an audio recording of a male English speaker performing each ROI sentence (multiple times) and running it against the ASR system. Since our goal was to obtain output containing a variety of errors, we used the CMU Sphinx system with its distributed trained models [24]. Some variations of the ROI text are shown below:

- *Conversely, when the work of film music scholars touches on the visual aspects of film, the cinema studies people often* ***cricket*** *as the work of amateurs.*
- *Conversely, when the* ***working*** *of film music scholars touches on the visual aspects of film, the cinema studies people often treat it as the work of amateurs.*
- *Conversely, when the work of film music scholars touches on the* ***region*** *aspects of film, the cinema studies people often treat it as the work of amateurs.*
- *Conversely, when the work of film music scholars touches on the visual aspects of film, the cinema studies people often treat it* ***has worked*** *amateurs.*

This procedure ensured that the artificially created variations of the ROI sentence agreed with the actual imperfect output produced by an ASR system.

3.1.3. Participants

Participants for the study were recruited from among associate degree students at the National Technical Institute for the Deaf (NTID) at Rochester Institute of Technology (RIT). We collected data from 30 DHH participants (age distribution with μ=22.63 and σ=2.63), 12 men and 18 women, where 26 participants self-identified as Deaf and 4 of participants as Hard-of-Hearing.

3.1.4. Procedure

Each participant was given 10 different comprehension passages to read, each containing three multiple choice questions that needed to be answered in a time period of 70 minutes. A pilot test with a DHH member of our research team helped us to determine an appropriate number of question items for the 70-minute experiment. The comprehension passages given to the participants were generated by replacing each ROI sentence by its erroneous counterpart (one of the variations). The number of errors of each category that were displayed to each participant was balanced among all participants in the study to ensure that individual human differences in task performance did not disproportionately affect the scores for any one category of error. Further, each ROI appeared several times throughout the entire study in a form without any errors inserted so that we could obtain baseline measurements for the difficulty of the particular comprehension question, to enable subsequent normalization of the collected scores. Scores of answers from each question were binary with correct answer receiving the Comprehension Score of 1 and incorrect answer receiving the score of 0.

3.2. Model Fitting

The data collected from the user study enabled us to determine whether there is a relation between the presence of an error with specific linguistic characteristics (see Table (2)) in a sentence and its impact on the comprehension of the sentence (whether or not participants answered the question referring to the sentence). However, the relation between the presence of an error and its impact on sentence is not straightforward. A wide variety of complex semantic factors can lead some ASR errors to be more confusing than others for end-users who are reading the text. For our automatic captioning application, we are interested in focusing on a subset of those aspects of a text that could be automatically computed, using modern computational linguistic software.

Table 2: List of features extracted from the error regions in the hypothesized text for analysis.

	Feature	Description	Type
1.	*WordLength*	Average length of the word in the region.	Numeric
2.	*SaliencyIndex*	Average TF-IDF score of the word in the region representing the importance of the word.	Numeric
3.	*POSTag*	Priority order based Part of Speech tag assigned to the region. The order is described in Section(3.2.2).	Categorical
4.	SyllableLength	Average number of syllables of the word in the region.	Numeric
5.	SentimentOrientation	Indicates whether the region alters the original sentiment (broadly, positive or negative) of the reference word(s) or not.	Categorical
6.	ContentOrFunction	Whether the region contains content word or not.	Categorical

3.2.1. Feature Identification

After consulting prior research on reading skills of deaf users [10, 25], we identified a list of 6 features of each error that we would examine as part of our analysis. The features are summarized in Table 2. Some features (for e.g. row 5 in Table (2)) are computationally more expensive than others. Since this model will eventually be used to produce a loss function to optimize a real-time ASR system, using these computationally expensive features may not be efficient. But, we considered these features in our preliminary analysis to understand their significance in the model.

3.2.2. Feature Extraction

Along with the Comprehension Scores for the text in the passages used in the study, we also extracted some linguistic features, summarized in Table (2). These features were obtained from the imperfect ROI texts in the passage which the users referred to when answering the questions provided during the study.

Each variation of ROI text contained at most one type of error which was created by replacing the actual (reference) word(s) from the error-free ROI text with a different (hypothesized) word(s). Thus, the first step of the feature extraction process involved alignment of error-free ROI text with its erroneous variation to identify the reference word(s) and the hypothesized word(s) pair. As the ROI texts were not time-aligned and there were few errors in each ROI text, we could utilize Levenshtein distance based word alignment technique to align the texts. We utilized CELEX2 [26] as our lexical database for syllable information for calculating the SyllableLength feature. A frequency-based Part-of-Speech (POS) tagger, Unigram Tagger [27], was utilized for POS tagging of words. The tagger was modified to output one of 11 different POS tags (in priority order: noun, verb, pronoun, adverb, adjective, preposition, conjunction, interjection, determiner, number and others) to an input word. The ContentOrFunction feature was calculated with the help of POS tag(s) of the word (a word is labeled as a Content word if it is a Noun, Verb, Adverb, or Adjective). The SaliencyIndex feature represented the general importance of the word and was estimated by calculating Term Frequency-Inverse Document Frequency (TF-IDF) score of a word(s). Scikit-learn's [28] TfidfVectorizer was used as our TF-IDF Scorer, and it was trained with a portion of dataset (N=18 books) from Project Gutenberg [29] corpus and Web Text corpus from NLTK [27]. TextBlob [30] library for python was used to compute the SentimentOrientation feature.

For each type of error, the features were extracted from the reference word (the actual word), except for the insertion error type (an insertion error doesn't have a reference word as it is produced due to an insertion of an extra word) whose features were extracted from hypothesized word(s).

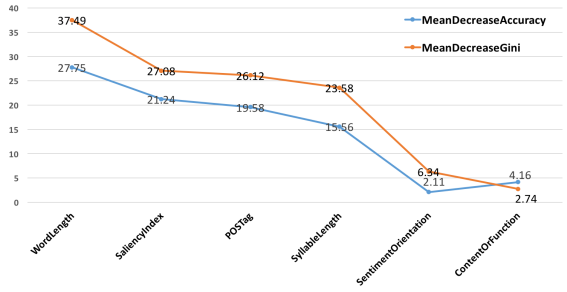


Figure 1: A plot showing the importance of each feature variable in-terms of their contribution to model accuracy and impurity.

3.2.3. Feature Selection

We utilized random forest to rank our 6 features and selected 3 features based on the measure of average accuracy decrease and average impurity decrease in the model without each of these features. As shown in Figure (1), features *WordLength*, *SaliencyIndex* and *POSTag* were among the best contributors to the Gini impurity and the accuracy of the model.

	Evaluation Metrics						
Models	**AUC**	**Cutoff**	**Accuracy**	**F-measure**	**Precision**	**Recall**	**Bal. Accuracy**
Logit (M_l)	0.496	0.539	0.618	0.754	0.618	0.968	0.498
Random Forest (M_{rf})	**0.572**	0.444	**0.620**	0.744	0.631	0.844	**0.533**
SVM (M_s)	0.496	0.605	0.617	0.738	0.625	0.919	0.497

Table 3: Summary of the evaluation of each prediction model on our test dataset. Value on each metric represents the average performance of the model in 5 different train and test partitions of our dataset.

3.2.4. Model Evaluation & Selection

We investigated three models for prediction and evaluated the performance of each model for our task. Table (3) summaries the result our evaluation. For the purpose, we selected 80% of our total observation (N= 862, excluding the baseline measurements) to train the model and used 20% of our remaining observation of test the model. For each model, five-fold cross validation with this 80/20 split was used to build each model, and the performance scores reported in Table (3) are based on the average of the models for each fold. We observed the performance of Random Forest model (M_{rf}) to be slightly better than other models with accuracy of ($\mu = 62.04\%$, $\sigma = 4.41$).

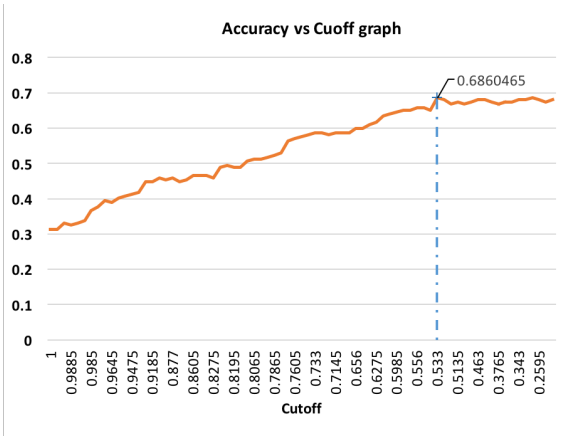


Figure 2: Example of Accuracy vs Cutoff graph for Random Forest Model on a test dataset. The marker represented by the *red-cross* represents the point of maximum accuracy at a cutoff value of 0.31.

During the testing process, the cutoff probability for each classification model, which was used to label output probability to our binary class, was chosen as the mode of the accuracy vs cutoff graph; the graph represented the accuracy of the model considering different cutoff values. Figure (2) shows the accuracy vs cutoff curve of Random Forest model on a test dataset.

4. Discussion

While its performance is above chance, the Random Forest model Accuracy results presented above are modest, but we view these results as preliminary. This study was based on a small amount of data (30 participants on 20 passages), and the set of features explored was relatively small. We view this effort as an initial proof-of-concept of our ability to identify useful features in a loss-function for predicting the comprehensibility of a text for DHH users.

Obviously, if we are to make use of this loss function in real-time captioning system, we would not know which words are errors. Our intention is to use the confidence value of the ASR system as a proxy for this information, and to use our loss-function to guide the hypothesis selection. Specifically, the prediction model (M_{rf}) we built from the user study results will be used in designing our loss function, as shown in Equation (3).

$$\ell(W, \delta(Y)) = -\Big(\sum_{f_i = f(W, \delta(Y))} M_{rf}(f_i) \Big) \tag{3}$$

where $\delta(Y)$ represents our decision rule that maps audio input (Y) to word sequence output ($\hat{W}$). We need a function $f(R, H)$ that returns set of features (listed is Table (2)) for each error type in the hypothesis text (H) when compared to the reference text (R).

This loss function looks to penalize the harsh errors that have significant 'predicted' impact on output comprehension (obtained from M_{rf}) for DHH users.

5. Conclusion & Future Work

The work described in the paper has been concerned with the development of a prediction model that represents the impact of ASR errors present in the text on its comprehension, specifically for DHH users. Beyond our intended application for ASR, we note that research on understanding the relationship between text characteristics and comprehensibility for DHH users may have other applications, such as automatic text readability detection software for these users. Further, we plan to extend our user study and improve our prediction model with more data. As we move on, we will look to investigate the Decision Theoretic framework for n-best list rescoring proposed by [20] to incorporate our custom loss function in to the ASR decoding process.

In addition, we will look to contrast its performance with other discriminative training techniques to optimize ASR components with our loss function. We also intend to do experimental analyses of the effectiveness of the final tool for DHH users.

6. Acknowledgements

This material was based on work supported by the National Technical Institute for the Deaf (NTID). We are grateful to Kellie Menzies, who assisted with data collection for this study, and to our collaborators Larwan Berke, Christopher Caulfield, Micheal Stinson, Lisa Elliot, Donna Easton, and James Mallory.

7. References

[1] W. H. Organization. (2015, March) Deafness and hearing loss. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs300/en/

[2] M. Wald, "Crowdsourcing correction of speech recognition captioning errors," in *Proceedings of the International Cross-Disciplinary Conference on Web Accessibility*, ser. W4A '11. New York, NY, USA: ACM, 2011, pp. 22:1–22:2. [Online]. Available: http://doi.acm.org/10.1145/1969289.1969318

[3] I. R. Forman, B. Fletcher, J. Hartley, B. Rippon, and A. Wilson, "Blue herd: Automated captioning for videoconferences," in *In Proc. ASSETS '12*. New York, NY, USA: ACM, 2012, pp. 227–228. [Online]. Available: http://doi.acm.org/10.1145/2384916.2384966

[4] M. Federico and M. Furini, "Enhancing learning accessibility through fully automatic captioning," in *Proceedings of the International Cross-Disciplinary Conference on Web Accessibility*, ser. W4A '12. New York, NY, USA: ACM, 2012, pp. 40:1–40:4. [Online]. Available: http://doi.acm.org/10.1145/2207016.2207053

[5] H. Takagi, T. Itoh, and K. Shinkawa, "Evaluation of real-time captioning by machine recognition with human support," in *In Proc. W4A'15*. New York, NY, USA: ACM, 2015, pp. 5:1–5:4. [Online]. Available: http://doi.acm.org/10.1145/2745555.2746648

[6] L. Deng and X. Li, "Machine learning paradigms for speech recognition: An overview," *IEEE Trans Audio Speech*, vol. 21, no. 5, pp. 1060–1089, 2013.

[7] R. S. Kushalnagar, W. S. Lasecki, and J. P. Bigham, "Accessibility evaluation of classroom captions," *ACM Trans. Access. Comput.*, vol. 5, no. 3, pp. 7:1–7:24, Jan. 2014. [Online]. Available: http://doi.acm.org/10.1145/2543578

[8] K. Bain, S. H. Basson, and M. Wald, "Speech recognition in university classrooms: Liberated learning project," in *In Proc. Assets '02*. New York, NY, USA: ACM, 2002, pp. 192–196. [Online]. Available: http://doi.acm.org/10.1145/638249.638284

[9] D. W. Jackson, P. V. Paul, and J. C. Smith, "Prior knowledge and reading comprehension ability of deaf adolescents," *Journal of Deaf Studies and Deaf Education*, pp. 172–184, 1997.

[10] J. L. Luckner and C. M. Handley, "A summary of the reading comprehension research undertaken with students who are deaf or hard of hearing," *American Annals of the Deaf*, vol. 153, no. 1, pp. 6–36, 2008.

[11] F. Richardson, M. Ostendorf, and J. R. Rohlicek, "Lattice-based search strategies for large vocabulary speech recognition," in *International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 1995, pp. 576–579.

[12] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus among words: lattice-based word error minimization." in *Eurospeech*, 1999.

[13] E. K. Ringger and J. F. Allen, "Error correction via a post-processor for continuous speech recognition," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 1, May 1996, pp. 427–430 vol. 1.

[14] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," in *Proceedings of the Automatic Speech Recognition and Understanding*, Dec 1997, pp. 347–354.

[15] B. Roark, M. Saraclar, and M. Collins, "Discriminative n-gram language modeling," *Computer Speech & Language*, vol. 21, no. 2, pp. 373–392, 2007.

[16] M. S. Brian Roark and M. Collins, "Corrective language modeling for large vocabulary asr with the perceptron algorithm." in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 1. IEEE, 2004, pp. I–749.

[17] T. Oba, T. Hori, and A. Nakamura, "A comparative study on methods of weighted language model training for reranking lvcsr n-best hypotheses," in *In ICASSP'10*. IEEE, 2010, pp. 5126–5129.

[18] J. D. Williams and S. Balakrishnan, "Estimating probability of correctness for asr n-best lists," in *In Proc. SIGDIAL'09*. Association for Computational Linguistics, 2009, pp. 132–135.

[19] A. Stolcke, Y. Konig, and M. Weintraub, "Explicit word error minimization in n-best list rescoring." in *Eurospeech*, vol. 97. Citeseer, 1997, pp. 163–166.

[20] V. Goel, W. Byrne, and S. Khudanpur, "Lvcsr rescoring with modified loss functions: A decision theoretic perspective," in *In Proc. ICASSP'98*, vol. 1. IEEE, 1998, pp. 425–428.

[21] V. Goel, S. Kumar, and W. Byrne, "Segmental minimum bayes-risk asr voting strategies." in *INTERSPEECH*, 2000, pp. 139–142.

[22] V. Goel, S. Kumar, and W. J. Byrne, "Confidence based lattice segmentation and minimum bayes-risk decoding." in *INTERSPEECH*, 2001, pp. 2569–2572.

[23] V. Doumpiotis, S. Tsakalidis, and W. J. Byrne, "Lattice segmentation and minimum bayes risk discriminative training." in *INTERSPEECH*, 2003.

[24] W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf, and J. Woelfel, "Sphinx-4: A flexible open source framework for speech recognition," Mountain View, CA, USA, Tech. Rep., 2004.

[25] R. R. Kelly, J. A. Albertini, and N. B. Shannon, "Deaf college students' reading comprehension and strategy use," *American annals of the deaf*, vol. 146, no. 5, pp. 385–400, 2001.

[26] R. Baayen, R. Piepenbrock, and L. Gulikers, "Celex2," 1995.

[27] S. Bird, "Nltk: the natural language toolkit," in *In Proc. COLING'06*. Association for Computational Linguistics, 2006, pp. 69–72.

[28] F. Pedregosa, A. Varoquaux, V. Michel, and Thirion, "Scikit-learn: Machine learning in python," *The Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[29] M. Hart, *Project gutenberg*. Project Gutenberg, 2004.

[30] S. Loria, "Textblob: simplified text processing," *Secondary TextBlob: Simplified Text Processing*, 2014.

Towards graceful turn management in human-agent interaction for people with cognitive impairments

Ramin Yaghoubzadeh, Stefan Kopp

Social Cognitive Systems Group, CITEC, Bielefeld University, Germany
ryaghoubzadeh@uni-bielefeld.de, skopp@uni-bielefeld.de

Abstract

A conversational approach to spoken human-machine interaction, the primary and most stable mode of interaction for many people with cognitive impairments, can require proactive control of the interactive flow from the system side. While spoken technology has primarily focused on unimodal spoken interruptions to this end, we propose a multimodal embodied approach with a virtual agent, incorporating an increasingly salient superposition of gestural, facial and paraverbal cues, in order to more gracefully signal turn taking. We implemented and evaluated this in a pilot study with five people with cognitive impairments. We present initial statistical results and promising insights from qualitative analysis which indicate that the basic approach works.

Index Terms: human-computer interaction, virtual assistant, interruption, turn taking, gesture, cognitive impairment

1. Introduction

Spoken human-machine interaction has become a widely adopted paradigm in recent years. In addition to being a helpful technology to keep one's hands free in a variety of everyday contexts, spoken interaction also opens access to modern technology as a whole for certain groups of people, specifically those that cannot readily understand, learn, read, or manipulate interfaces employing other modalities. While graphical interfaces with flat hierarchies counteract some of these usability problems, the presentation and negotiation of information there does not always correspond well to those in spoken human-human interaction, which many of those people will be quite familiar with. However, off-the-shelf spoken language technology, which has become very good at recognizing words even when uttered by new users and answering common sets of – well-formed – questions, usually also lacks many of the aspects of this human-human mode of interaction, namely its conversational, incremental and reciprocal nature. Humans in interaction constantly exchange back-channel information relating to – possibly preliminary – evaluations of the unfolding stream of information. By attending to the other party, a speaker is aware of the back-channel feedback (paraverbal, facial, gestural) of a listener and will incrementally and smoothly incorporate it into their content selection and presentation. Likewise, the listener can tell when and where the speaker encounters a problem, and can intervene in a timely manner, if necessary. They might also be aware of possible points of misunderstanding, and either address them immediately – by barging in, often in a cooperative fashion – or queue them for later implicit or explicit resolution.

To achieve these capabilities, one crucial function in dialog systems is interrupting the user and taking the floor – but doing so in a cooperative manner that is consistently acceptable for users even over longer time spans and many repeated instances.

In the following sections, we will first provide an overview of the theoretical and analytical background relating to multimodal turn management signals, and look at related work on turn taking control in interactive systems. We will then present the scenario and user groups for our pilot study, and present the autonomous interruption controller that was run alongside a Wizard-of-Oz controlled main dialogue. After a description of the procedure and the interview structure for the assessment of subjective ratings, we will present initial statistical data, followed by a detailed analysis of one particularly informative interaction fragment, before concluding our presentation.

2. Background and related work

The different manifestations of turn-keeping and turn-grabbing signals were early described by Duncan and Fiske [1]; Bohle [2] provides a comprehensive overview and discussion. According to the latter source, one single characteristic, unimodal signal generally constructs a clear meaning in these situations, but the intensity can be increased by employing multimodal presentation. Addressing those behaviors in the listener role that do not generally have the effect of signaling a desire to obtain the floor, they list minimal acknowledgements, clarification requests, other-completions, short paraphrase, and head gestures – one should also add paraverbal back-channel feedback to the list [3]. As for the floor-asserting behaviors, more specifically floor grabs by the listener, they list head or gaze aversion from the speaker, as well as the initiation of gesticulation. Kaartinen [4] analyzed gestural behavior as turn-taking signals in news interviews, noting the role of adaptation of gestures in forming multifunctional constructs encompassing turn-taking information; and particularly highlighting (quasi-)deictic handshapes, first and foremost extended fingers.

The efficacy and acceptability of interruptive behavior on the part of dialog systems has been well researched.

Ter Maat et al. [5] investigated the effect of interruptive turn taking by an agent in a Wizard-of-Oz setup, comparing (unimodal spoken) early turn grabs, i.e. interruption-causing overlaps, to turn taking immediately at and slightly after appropriate points. They found that early barge-ins were perceived as more assertive, but also as significantly more disagreeable, rude and of lower conversational aptitude.

Cafaro et al. [6] examined ratings of simulated agent-agent interactions with comparable interruption types, but additionally manipulating the cooperativity of the interrupter's content selection strategy (e.g. elaboration vs. topic jump as a reply to a question). Strategy changes towards cooperativity in particular led to increased perceived friendliness and reduced dominance – but less so that the selection of interruption type, corroborating the findings of ter Maat et al.

SLPAT 2016, 7th Workshop on Speech and Language Processing for Assistive Technologies, pages 26–31,
San Francisco, USA, 13 September, 2016.

In our own work with autonomous dialog systems for older adults and people with cognitive impairments, we previously found that the spoken, conversational, paradigm of task-related interaction with a system transfers to both groups, both in terms of feasibility and acceptance [7]. In those studies, we strictly let the users control the pacing of the dialogue and the amount of transferred content, while priming for specific information presentation only when subjects yielded their turns spontaneously. While the performance and perception of the autonomous system was comparable to an earlier Wizard-of-Oz prototype [8] for most people, we found that a noticeable minority of participants from both groups were prone to excessively verbose or tangential presentation even after repeated instruction to the contrary (cf. Fig. 1) – which caused ASR and NLU to drastically decrease in performance, thus necessitating proactive, interruptive, system-side floor governance.

3. Pilot study

Considering this requirement for proactive floor management, and the aforementioned work on the reduced perceived cooperativity caused by pure verbal barge-ins, we constructed an autonomous prototype interruption controller (`flow_controller`) based on the research on the multimodal construction of turn-grabbing behavior. We employed it in a pilot study with five participants with cognitive impairments, engaging in a spoken human-agent interaction in a Wizard-of-Oz-controlled discussion game. These sessions were embedded in a larger study exploring the effects of agent body language on the persuasiveness and reception of system-generated argumentation for older adults as well as younger controls (n=40 each; analysis in progress), for which younger subjects with cognitive impairments were also recruited by our corporate partner, the large health and social care provider v. Bodelschwinghsche Stiftungen Bethel.

Since the participants with cognitive impairments were not expected to be able to fill out the required 90+-DOF questionnaire for the experiment proper, the interruption condition was piloted instead. Participants from all user groups were presented with the same scenario and task, described below.

3.1. Setup and participants

From the point of view of the participants, the setup consisted of a 27" touch screen, a microphone and an eye tracker, as well as cameras recording two angles (Fig. 2 depicts the view from the rear camera). The screen was able to show the game scene, showing the animated virtual agent, "Billie", as well as lists representing the game state. The agent was controlled by the ASAPRealizer software for behavior realization [9]; text-to-speech was provided by a CereVoice [10] component controlled by the realizer, which was able to provide some realizations of paraverbal signals. A directional microphone and a low-cost eyetracker were mounted below the screen. The system was primarily controlled by a Wizard-of-Oz console that interacted with a component managing the game state and graphical presentation. However, the agent's nonverbal and paraverbal behavior was controlled autonomously by the `flow_controller`, described below. This was contingent on the audio state as reported by a simple audio level detector, which was in turn inhibited by ongoing agent utterances (see Fig. 3 for an overview of components). The eye tracker was, in this incarnation of the system, employed as a source of data for qualitative analysis and as a basic functionality test for our user groups, although incorporation into the interruption controller is planned for the future.

Figure 2: *Overview of the setup (as seen from the rear camera). Touchscreen PC with eye tracker mounted below. High-fps face camera recording from below the screen. The physical item list and the items to allocate on it in preparation for the game are visible on the desk. The microphone is occluded by the participant (anonymized). Start of interaction.*

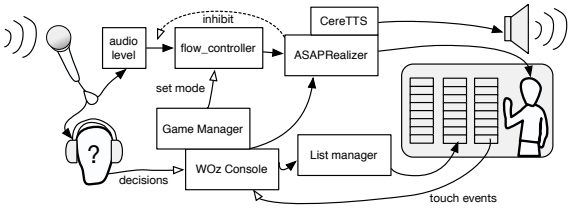


Figure 3: *Overview of components. "Audio level" and "Flow controller" constituted the autonomously acting subsystem. Discourse progress and contents were controlled by the Wizard (wearing headphones).*

Participants (n=5, 2 male, 3 female, ages 29–48) were recruited from a care institution providing support to people with cognitive impairments, both in communal and individual assisted living arrangements. Exact clinical diagnoses were not able to be divulged by the care providers. As with previous experiments, we asked that only subjects be recruited whose articulation was clear enough to be generally comprehensible by untrained, unfamiliar listeners.

3.2. Interruption controller

Animations for signaling turn grabs were first recorded using full-body motion capturing, then reduced to spinal and arm movements and preprocessed to obtain a chainable, smooth database for procedural animation. In lieu of a speech recognizer's voice activity detection module, a simple audio-level activated trigger was implemented using pyaudio that reported durations of ongoing and finished audio events. It was inhibited by open agent utterances, effectively removing cross-talk at the cost of ignoring overlapping speech. We were only interested in the duration of the user turn proper, and surmised from previous experiments that prolonged periods of overlap were unlikely to be frequently caused by the user group.

The `flow_controller` component, responsible for interruption generation, was able to be configured in four modes (0–3): modes 1, 2 and 3 started a slowly progressing three-state

```
a1  AGNT  Do you have another appointment?
    SUBJ                                    Yes. Then, I have yet another appointment ... on Friday

a2  AGNT  So, on Friday, right?        OK. At what time does it start?
    SUBJ                        Right.                                  Then I'll pick 3 PM again,

a3  AGNT                  So, at 3 PM, right?                 So, at 3 [interrupt]   Good.
    SUBJ  have ice cream.                      [hoarsely] Yah             Yes.

a4  AGNT  So, at that time, there is "Have ice cream", right?         Okay. Then I'll enter it as follows...
    SUBJ                                                    Right.


c1  AGNT Then tell me the next appointment, please.
    SUBJ                                          I have uhm (-) today shopping *thr 3 PM 3 PM *appoin

c2  AGNT
    SUBJ  appointment with <Name> (.) and then I also(?) later go shopping later *thr 3 PM with <Name>

c3  AGNT
    SUBJ  (.) and (-) then I also go shopping (-) later
```

Figure 1: Transcripts of interaction segments with different interaction styles observed in previous studies with an autonomous prototype system. The non-verbatim translation from German attempts to represent dysfluencies and errors intuitively. **Top:** older adult, brief but casual style; **bottom:** person with noticeable cognitive impairment, verbose turns, exacerbated by dysfluent and unclear articulation; this led to considerable ASR processing delays (the participant eventually entered the shopping appointment successfully).

Figure 4: *Four stages of nonverbal interruptive behavior. From left to right: Idle; first signal (reached after about 4s, shown with mouth half-open); second signal (reached after about 7s); final stage (held until user ends utterance or Wizard barges in).*

cascade of interrupting behaviors (cf. Table 1), varying slightly in surface form by mode, for all utterances above a threshold duration (set to 2s+). Behaviors included hand raises (half-open hand or pointing shape), gaze aversion, open mouth and paraverbals ("ah" and throat clearing). For short utterances, the agent would provide positive feedback by nodding. Mode 0 allowed for a non-interrupting state: the agent would nod at the defined transition points and then remain static in the idle position for the remainder of the user's turn.

3.3. Task and procedure

The task for the participants was a discussion game in the "desert survival" scenario. The premise was that the agent and the subject were stranded in a remote location, with their airplane destroyed and only a set of twelve items still intact. The task of participants was to order them, ranked by their perceived usefulness, and then engage in a discussion with the agent to find a consensus order.

A brief principal instruction was provided by the experimenter, then the interaction started. The Wizard controlling the agent would first greet the user and ask their name, then explain all stages of the game. The users were then asked to rank the list of items, for which we prepared a paper list with twelve empty item slots, and paper slips corresponding to the items to be placed on the list. All items were individually explained to the user, as was the meaning of list items "high" or "low" on the list to account for possible problems with abstract numbers. Subjects then had two minutes to decide on a preferred ranking. The agent would enter "hidden" rankings on his list on the screen. When the subject's list had been finished, the agent asked them to read them off in order. The agent entered those rankings in the leftmost (user) list. Thereafter, the agent "showed" its list – but instead silently generating a conflicting ranking according to a pre-defined permutation scheme.

The subsequent crucial discussion phase started, the agent presented the user's ranking along with its own and a relative statement (like: "You placed the Lighter on 1, I have it on 7. So, you rated it as more important than I did. Could you explain why you placed it there?") At this point in the discussion, the interruption controller, that was set to mode 0 (do not interrupt) in all other contexts, was set based on the index of the currently

Table 1: *Interruption controller: modes and phases, with respective generated actions. Modes were set according to the index of the discussed items (see text). Phases were successively entered during user turns that proceeded for long enough. *Note: all hand positions other than the idle position were combined with a slight gaze aversion (constant angle).*

Phase \ Mode	0	1	2	3
– (short)	nod	nod	nod	nod
1	nod	raise to mid*, index extended	raise to mid*, mouth open	raise to mid*, index ext'd, utter "ah"
2	nod	raise to high*, mouth open	raise to high*, mouth open	raise to high*, mouth open
3	nod	keep raising*, hand open, utter "ah"	keep raising*, hand open, clear throat	keep raising*, hand open, clear throat

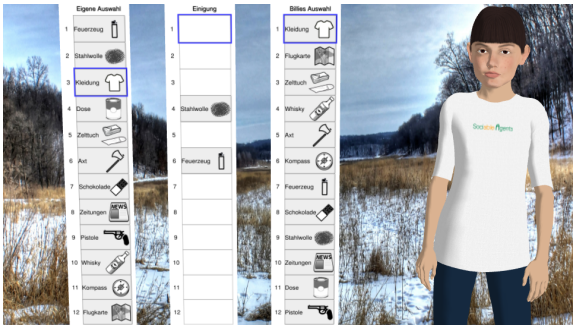


Figure 5: *Crashed in the taiga! Scene setup: left list: initial user choice; right list: agent 'choice'; center list: ranking made by user after the exchange of arguments. In this instance, the discussion has just taken place for the user's third most important item ("clothes", highlighted), and the user just selected a final slot for it, in this case following the agent's suggestion.*

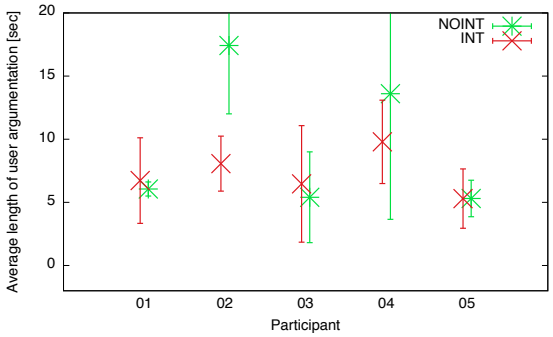


Figure 7: *Length statistics for discussion turns for each participant, non-interrupting (mode 0) vs. interrupting (mode 1–3) conditions. (Participant 03 was in the audio-only interaction.)*

discussed item (from first to twelfth: modes 0, 2, 0, 1, 2, 0, 1, 2, 3, 0, 2, 3). Therefore, at items 1, 3, 6, and 10 (a third of all items), the system would remain in non-interrupting mode as a reference for comparison.

After subjects presented their opinion about an item, the agent would invariably utter an argument from a precompiled list that contained one supportive and one dismissive argument for each item, selected depending on relative ranking. Then, the user was always given the choice to fix a position for the item on the common, central, list. Selections in the lists could be made either by speaking the rank number, or by touching the corresponding field on the screen (cf. Fig. 5). When all 12 items had been discussed and agreed upon, the user could modify the list one final time if they so wished, after which thanks and valedictions were presented by the agent, and the iteraction was over.

After the experiments, a simplified structured interview was conducted for each subject. A visual 5-point Likert rating aid (definitely yes – ... – definitely no) was employed to gain quantifiable ratings to ten questions, although the primary aim was to gather comments and qualitative information. The questions were (approximate correspondences in Simple English): Q1: Did the game with Billie go well? Q2: Was Billie nice to you? Q3: Did you understand what Billie said? Q4: Could Billie understand you as well? Q5: Did Billie listen when you wanted to say something? Q6: Did you find the game easy enough? Q7: Was the length of the game okay? Q8: Did you call the shots in the game? Q9: Did Billie butt in or interrupt you? Q10: Did you have fun playing the game? In particular, Q5 and Q9 were inserted as a pair of opposing valence, contingent on the experimental manipulation and its effectiveness and perceived intrusiveness. Subjects were also asked to report prior technical experience and answer the more general question "Do you enjoy talking to other people?".

3.4. Results and discussion

All five subjects were able to complete the whole task. Subject 2 mostly opted for the touch-based rank selection in the course of the discussion, while the other participants interacted using speech only. For subject 3, a technical problem led to the loss of the video signal on their screen after the introductory explanation; the subject accepted this silently and concentrated mostly on her physical list from then on. Since the task was completed successfully, we used this as an audio-only reference. The eye-tracker only reliably worked for subjects 1 and 4, thus final analysis of the gaze behavior can only be made after video-based annotation for the other subjects.

The scenario was not consistently conducive to long elaborations by all users, although two participants did produce them. While the sample size is much too small for robust statistical results, interesting trends can be gleaned from the graph in Fig. 7: for subject 2 and 4, who produced the most elaborate argumentations in the turn we focused on, a noticeable difference between the non-interrupting and interrupting items can be seen, indicating that the interruption strategy might have had an effect. This was most valid for the first three or four items, where all participants had a quite clear idea of their motivation to rank the items highest (note that subject 4, during item 6, where they were free to talk, just coughed, sighed and uttered "tough question", staring at the agent until the Wizard continued.) In the videos, no clear, hard 'breaking points' can be observed in any of these cases, indicating that the progressive nature of the signal might have progressively steered them to a smooth turn completion.


```
01  SUBJ        ja-aso meine überlegung war weil ö:    (0.6)     den ö:     (0.4)      ö:m   (0.3)
                yes-well my idea was since uh                    the uh               uhm
    SUBJ-gaze   @GUI                                                 @Table           @GUI
    AGNT-gest   | hold IDLE ------------------------------------------------------------| raise to MID ---

02  SUBJ        <<click>> der whiskey          ja    ess (.) essma so nich wichdich is ne den kam-man (0.8)
                          the whiskey          well fir...  firstly isn't that important right you can
    SUBJ-gaze                 FLICK@Agent @GUI                          @Agent          @GUI 0.6s    @Agent
    AGNT-gest   ---------------| hold MID -------------------------------| raise to HIGH, extend index
                                                                           Img#1      Img#2      Img#3

03  SUBJ        den |kamman weil wemman schmerzen hat oder so zum    (0.85)     <<click>>     (-)
                it   you can since when you are in pain or something to
    SUBJ-gaze   @GUI                                                 @Agent        @GUI
    AGNT             (ah)
                       ah
    AGNT-gest   ----| raise to APEX, open palm ----------------------| hold APEX ---------------

04  SUBJ         brauchen aba-ansonsten    (1.5)
                 use but otherwise
    SUBJ-gaze            @Table        @Agent
    AGNT                                        bei mir is-der whiskey an stelle fünf
                                                I've got the whiskey at position five
    AGNT-gest   | relax to IDLE -----------| hold IDLE ----------------------------------
```

Figure 6: Transcript and anonymized snapshots from an item argumentation by subject 1 (see 3.5 for discussion). Pause lengths in parentheses, short pauses given as (.), (-) [11]. Times of the three frames indicated by Img#x, in blue.

We surmise that an attenuation effect contributed to an observably reduced verbosity in later items: participants apparently ranked the most important and least important items with a clear idea of their merits or downsides, with the rest of ranks (around #7–#11) possibly filled rather indifferently with the unclear remainder.

The ratings of the interview questions relating to the experimental manipulation were rated equally by all participants (Q5 was rated "decidedly yes", Q9 as "decidedly no"). Agent niceness and enjoyment of the game were likewise rated with the most affirmative option by all subjects. Subject 2 noted in the free-form interview comments that she noticed the agent's gestures but found them slightly odd; she thought the agent looked like it "might want to say something".

3.5. Qualitative analysis

Even for those participants where no noticeable effect on utterance length could be observed, detailed analysis still indicates that the agent behavior modulated the subject's pacing and continued attention to the agent, indicating that timely and contingent content presentation, as afforded by an autonomous dialogue system as opposed to WOz, could allow for a cooperative takeover at these points.

Fig. 6 highlights one such situation. The participant is mainly focused on the GUI part of the screen (the agent's list) and his paper list, but regularly checks back with the agent while he is explaining his decision. The section highlighted in red shows a typical fragment of interaction where the system managed to capture the attention of the user. The user looks at the agent during his utterance – the agent has already performed the weakest interruption signal and is just about to generate the second animation (raise hand further with index finger extended and mouth slightly open). The user looks back to the left, but his gaze returns to the agent after merely 0.6 seconds. He then hestitates mid-sentence for 0.8 seconds, directing his attention immediately at the agent.

We argue that the interruption subsystem managed to construct a possible transition point here (which the Wizard did however not utilize before the user resumed) – with the short gaze shift to the lists either due to a delay in the user's reaction time, or else having seen the continuation of the signal in peripheral vision.

4. Conclusions

Our preliminary results indicate that the nonverbal agent behavior generated by the interruption controller did lead to graceful (self-)interruption in some of our participants with cognitive impairments, while others that did not noticeably vary in presentation length still reacted noticeably to the emitted signals. As for the ratings of intrusiveness and cooperativity, none of the participants judged this as interruptive behavior per se, and agent ratings were all maximally favorable. The practical efficacy of these interruption signals is certainly also dependent on proper contextual content selection at the very time of a generated transition opportunity or floor yield – which was not trivial to realize for the human Wizard. Depending on the scenario, this could work better with a spoken dialog system, which we will explore next. Another issue is the exact surface realization of the signals of increasing intensity. While we hand-crafted our signals based on literature on the topic, an evaluation of different versions of the signals (possibly using crowdsourcing with unimpaired users) could also be helpful.

Overall, we deem the further exploration of nonverbal control signals to govern the floor in conversational, spoken dialog systems a promising endeavor.

5. Acknowledgements

This research was partially supported by the German Federal Ministry of Education and Research (BMBF) in the project 'KOMPASS' (FKZ 16SV7271K) and by the Deutsche Forschungsgemeinschaft (DFG) in the Cluster of Excellence 'Cognitive Interaction Technology' (CITEC).

6. References

[1] S. Duncan and D. Fiske, *Face-to-face interaction: Research, methods, and theory*. Hillsdale, NJ: Erlbaum, 1977.

[2] U. Bohle, *Das Wort ergreifen – das Wort übergeben: Explorative Studie zur Rolle redebegleitender Gesten in der Organisation des Sprecherwechsels*. Berlin: Weidler, 2007.

[3] J. Allwood, J. Nivre, and E. Ahlsen, "On the semantics and pragmatics of linguistic feedback," *J Semantics*, vol. 9, pp. 1—-26, 1992.

[4] S. Kaartinen, "Multi-functional gestures in interruptions a look at news interview situations," Master's thesis, Oulu, Finland, 2013.

[5] M. ter Maat, K. P. Truong, and D. Heylen, *How Turn-Taking Strategies Influence Users' Impressions of an Agent*. Berlin, Heidelberg: Springer, 2010, pp. 441–453.

[6] A. Cafaro, N. Glas, and C. Pelachaud, "The effects of interrupting behavior on interpersonal attitude and engagement in dyadic interactions," in *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, ser. AAMAS '16. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2016, pp. 911–920.

[7] R. Yaghoubzadeh, K. Pitsch, and S. Kopp, "Adaptive grounding and dialogue management for autonomous conversational assistants for elderly users," in *Proceedings of the 15th International Conference on Intelligent Virtual Agents*, ser. LNCS (LNAI), vol. 9238, 2015, pp. 28–38.

[8] R. Yaghoubzadeh, M. Kramer, K. Pitsch, and S. Kopp, "Virtual agents as daily assistants for elderly or cognitively impaired people," in *Proceedings of the 13th International Conference on Intelligent Virtual Agents*, ser. LNCS (LNAI), vol. 8108, 2013, pp. 79–91.

[9] H. van Welbergen, D. Reidsma, and S. Kopp, "An incremental multimodal realizer for behavior co-articulation and coordination," in *Proceedings of the 12th International Conference on Intelligent Virtual Agents*, ser. LNCS (LNAI), vol. 7502, 2012, pp. 175–188.

[10] M. P. Aylett and C. J. Pidcock, *The CereVoice Characterful Speech Synthesiser SDK*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 413–414.

[11] M. Selting et al., "Gesprächsanalytisches Transkriptionssystem 2 (GAT 2)," *Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion*, vol. 10, pp. 353–402, 2009.

Simple and robust audio-based detection of biomarkers for Alzheimer's disease

Sabah Al-Hameed [1], *Mohammed Benaissa* [1], *Heidi Christensen* [2]

[1] Department of Electronic and Electrical Engineering, University of Sheffield, United Kingdom
[2]Department of Computer Science, University of Sheffield, United Kingdom
{ssaal-hammed1, m.benaissa, heidi.christensen}@sheffield.ac.uk

Abstract

This paper demonstrates the feasibility of using a simple and robust automatic method based solely on acoustic features to identify Alzheimer's disease (AD) with the objective of ultimately developing a low-cost home monitoring system for detecting early signs of AD. Different acoustic features, automatically extracted from speech recordings, are explored. Four different machine learning algorithms are used to calculate the classification accuracy between people with AD and a healthy control (HC) group. Feature selection and ranking is investigated resulting in increased accuracy and a decrease in the complexity of the method. Further improvements have been obtained by mitigating the effect of the background noise via pre-processing. Using DementiaBank data, we achieve a classification accuracy of 94.7% with sensitivity and specificity levels at 97% and 91% respectively. This is an improvement on previous published results whilst being solely audio-based and not requiring speech recognition for automatic transcription.

Index Terms: Dementia, feature extraction, feature selection procedure, de-noising, classification.

1. Introduction

Recent statistics show an increase of the elderly population around the world according to Alzheimer's Disease International [1] and a relatively high percentage of those will go on to develop dementia [2], [3]. Dementia is used as an umbrella term to describe symptoms of brain disease damaging the cells and neuron synapses caused by e.g. Alzheimer's disease. Dementia symptoms include cognitive decline (affecting amongst other things memory and the person's speech and language), limited motor control, abnormal behavior, loss of memory and judgment, apathy and at a late stage losing the ability to speak [2].

Currently, there is no powerful tool that gives a reliable diagnosis of dementia; rather, the patient has to go through a series of cognitive tests conducted by a professional neurologist for assessments. This process can be very challenging for the patient and involves a certain amount of anxiety and stress. Especially in the case of the early stage detection, complementary tests include the analysis of samples of cerebrospinal fluid taken from the brain and a magnetic resonance brain imaging test [4], [5]. Such methods are invasive, bring discomfort to the patients, are relatively costly and require a significant amount of effort and time.

Finding lightweight, noninvasive diagnostic and/or screening tools, that can be used in the comfort of peoples' homes and inform this process, is therefore of interest. This could be in the form of wearable sensors or incorporated in existing intelligent home technology. This paper describes a relatively simple audio-based tool for detecting biomarkers of dementia in a person's speech.

Changes in speech and language patterns offer valuable clues to the detection of dementia as the speech production process starts in the left hemisphere of the brain [6] and any decline in speech capabilities might indicate the presence of e.g. Alzheimer's disease. Several studies investigated the use of speech-based features for the detection of dementia providing a noninvasive and inexpensive tool that does not require extensive infrastructure or the presence of medical equipment [7], [8]. Automated speech and language analysis methods are potentially powerful tools, especially when using machine learning algorithms capabilities to evaluate the features extracted from the speech. Many methods rely on relatively computationally heavy processing involving speech recognition and the use of natural language processing techniques to achieve some degree of speech understanding at the linguistic level [9]. This makes them unsuitable as low-cost home-based solution and means they are expensive to port to new languages. The alternative solution presented in this paper investigates audio-only processing to address this challenge.

We propose a simple automated method for detecting/screening AD at an early stage. The proposed method is solely based on acoustic features and therefore would only require simple readily available audio technology that can be adapted to suit patient requirement either in terms of being portable or/and wearable. We also explore the performance of different classification techniques applied to a number of acoustic features automatically extracted from the speech recordings obtained from DementiaBank [10]. Finally, we investigate the effect of pre-processing and noise reduction on the performance of the proposed method.

The rest of the paper is organized as follows. Section 2 describes the background. Section 3 describes the experiment setup. Section 4 explores the machine learning. Section 5 presents the results. Finally, section 6 presents the conclusions and future work.

2. Background

Several publications have demonstrated the potential of speech based approaches to identifying dementia. Jarrold et al [11] distinguish between different types of dementia by combining two profiles of features related to acoustic and lexical features collected from 9 controls and 39 patients who have been diagnosed with different types of dementia. Features-based profiles were extracted from structured interviews and used as

SLPAT 2016, 7th Workshop on Speech and Language Processing for Assistive Technologies, pages 32–36,
San Francisco, USA, 13 September, 2016.

input to a machine learning algorithm. A score of 88% was achieved by using a multi-layer perceptron algorithm.

Orimaye et al [12] proposed a diagnostic method to identify people with AD using nine syntactic and eleven lexical features extracted from transcribed audio files from the DementiaBank dataset. They used a sample size of 242 files for both healthy older people and people with AD. They explored four different machine learning classification algorithms, achieving a 74% classification accuracy using a support vector machine (SVM) classifier with 10% cross-validation.

López et al [8], [13] investigated using features called Emotional Temperature derived from the speech along with acoustic features from 20 healthy subjects and 20 people suffering from dementia. This was done in an attempt to evaluate the importance of the emotions encapsulated in the spontaneous speech and they showed promising results when attempting to differentiate different stages of the disease.

Furthermore König et al [14] conducted an experiment of using four short cognitive vocal tasks with a number of participants divided into three groups: healthy control (HC), people with Mild Cognitive Impairment (MCI) and people with Alzheimer (AD). Their method included pre-processing, analyzing the data and feature extraction from the speech recordings. They were able to distinguish between HC and MCI with an accuracy of 79%, between HC and AD patients with an accuracy of 87%, and between those with MCI and AD with 80% accuracy.

Recently and similar to our work, Fraser et al [15] studied the potential of using linguistic features to identify Alzheimer's disease. They used speech recordings along with their manually transcribed files derived from the DementiaBank data set. They chose 240 speech recordings belonging to a group of 167 people identified as probably or possibly having AD and 233 samples from 97 subjects with no memory complaint. In total, a set of 370 acoustic, lexical and semantic features were extracted and they then applied two machine learning classification algorithms and obtained a highest accuracy of 92% in distinguishing between HC subjects and AD patients using the top 25 ranked features. Although they obtained promising results, their method relies on the accuracy of manually transcribed files, whereas a real system would need the added complexity of a speech recognizer to compute all three types of features. It is unclear how the results of Fraser's system are affected when having to rely on erroneous transcripts from an automatic speech recognizer.

Key aspects of our proposed method compared to state of art are listed as follows:

- Feasible for application in real time and in a range of environments (home/clinic) since our results have been evaluated in the presence of high levels of background noise.
- Higher classification accuracy; outperforming the recent highest score in [15] using the top 20 ranked features
- Robustness: as high classification accuracy is maintained when using higher numbers of features and even when using all 263 features.
- Not reliant on speech recognition to transcribe the audio, so the method could potentially be language independent.

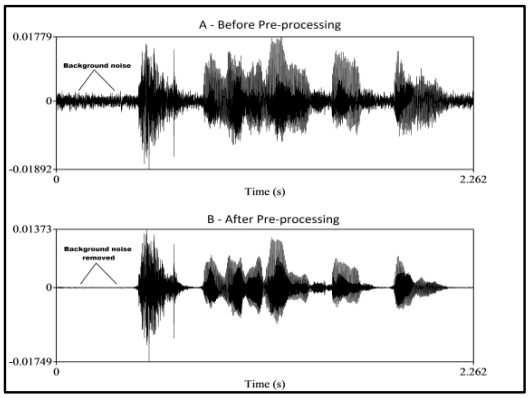


Figure 1. *Speech sample before (A) and after (B) the pre-processing step.*

3. Experimental setup

3.1. **Data set**

We utilized the DementiaBank data set [10], a free access large existing database for Alzheimer's and related dementia diseases collected during longitudinal study conducted by the University of Pittsburgh School of Medicine. A verbal description of the Boston Cookie Theft picture was recorded from people with different types of dementia with an age span from 49 to 90 years as well as from elderly HC subjects with an age range from 46 to 81 years. During the interviews, patients were given the picture and were told to discuss everything they could see happening in the picture. The speech samples were collected through annual visits from the majority of the participants and were transcribed using the CHAT transcription format Mac Whinney [16]. We consider the same sample size used by Fraser et al [15] with a total of 473 recording from 97 HC controls having 233 speech samples and the rest from 167 AD patients diagnosed as possible or probable AD.

3.2. **Pre-Processing**

The first step of the pre-processing is background noise reduction, as the DementiaBank data contains a high level of background noise. Effective de-noising is important to enable accurate features extraction. The spectral noise gating method using version 2.1.1 of the Audacity(R) recording and editing software [17] was applied to the audio without sacrificing the overall speech quality. Initial experimentation was carried out to examine the overall performance with and without the presence of the background noise as shown in Figure (1).

Next, using Praat [18], the instructor utterance was removed from the recordings and the audio files were converted from MP3 to mono wave; the sampling frequencies were kept unchanged.

3.3. **Features extraction**

In our study we focused on extracting acoustic features only and investigating the effectiveness of these features in detecting dementia at an early stage. This would avoid relying on the need for manually transcribed files or indeed the problems around achieving reliable speech recognition results in challenging far-field acoustic conditions.

Table 1. *Summary of all the features*.

#	Features set	Description
1	First Group (24) features	Task completion time
		Pitch variation features (mean, median, STD, Min and Max)
		Mean periods and STD periods
		Fraction of locally unvoiced frames and degree of voice breaks
		Jitter: (local, local-absolute, the relative average perturbation (rap), five-point perturbation quotient (ppq5) and the average absolute difference (ddp).
		Shimmer: (local, local-dB, three-point amplitude perturbation (apq3), five-point amplitude perturbation quotient (apq5), eleven-point amplitude perturbation quotient (apq11) and the average absolute difference (dda).
		Mean of autocorrelation
		Mean noise-to-harmonics ratio
		Mean harmonics-to-noise ratio
2	Second Group (17) features	Max, mean, median and STD of speech segment length >=0.4 sec
		No. of pauses (pause length of >=1ms are considered)
		Total speech & silent durations for the segments >= 0.4 sec
		Max, mean, median and STD of silent segment length >=0.4 sec
		Total silent length >=0.4 sec. including the pauses
		Number of speech and silent segments >=0.4 sec.
		Mean and STD of pauses and total duration of the pauses
3	Third Group (222) features	26 Spectral centroid coefficients
		26 Filter bank energy coefficients
		First 42 MFCC coefficients and their skewness, kurtosis, mean with kurtosis and skewness of the mean

Table (1) summarizes all 263 features extracted. The first group of features includes the pitch statistics, mean and standard deviation of periods, degree of voice breaks, fraction of locally unvoiced frames, and the voice quality measures including harmonic-to-noise ratio, mean of autocorrelation and noise-to-harmonic ratio. Various features related to jitter and shimmer scales were also extracted in accordance with [19], [8] using Praat [18].

The second group of features was derived by applying machine classification algorithms to identify speech/non-speech segments. This is done by windowing the audio files into 40ms frames with 50% overlapping window. For each frame we calculate the short time energy, zero crossing rate and the correlation coefficients. The three measures with labeled frames are used to train and build a voice activity detection (VAD) classifier using predefined frame samples randomly selected from the data. Next we used the VAD to label each frame for the rest of the audio files. The results from the VAD classifier gives us duration statistics for speech/silent regions with the amount of pauses presented in the recordings [20].

The last group of features includes the Mel Frequency Cepstral Coefficients extracted using the method mentioned by [21], including: the first 42 MFCC coefficients and their skewness, kurtosis, means and kurtosis and skewness of the means) previously used by [15] in addition to the first 26 coefficients for both filter bank energies and spectral centroid.

Table 2. *Top 20 rank features as automatically selected by the Weka attribute selection function*.

#	Features	Rank – Weight
1.	MFCC2	82.241
2.	Kurtosis -MFCC30	81.606
3.	Mean-MFCC30	81.606
4.	Skewness - MFCC2	80.972
5.	Mean-MFCC16	80.126
6.	Filter bank energy 22	79.069
7.	Spectral centroid -C14	79.069
8.	MFCC30	77.801
9.	Kurtosis -MFCC16	77.589
10.	Filter bank energy 2	77.589
11.	Filter bank energy 24	77.167
12.	MFCC1	76.532
13.	Filter bank energy 15	76.052
14.	Kurtosis -MFCC2	73.995
15.	Filter bank energy 20	72.304
16.	Filter bank energy 13	65.961
17.	No. of silent segments	61.522
18.	Fraction of locally unvoiced frames	59.830
19.	Minimum silent segments length	57.928
20.	Median pitch	49.48

4. Classification

4.1. **Automatic classification**

We used the capability and accuracy of the automated machine learning algorithms to measure the potential of the acoustic features to distinguish between AD patients and HC subjects. We applied four different classifiers: Bayesian Networks (BN), Trees-Random Forest (RF), AdaboostM1 (AB) and Meta- Bagging (MB). We used the Weka [22] software for running the experiments, with k-fold cross validation, in which we randomly divide the data into K equal-sized parts. We leave out part k, fit the model to the other K-1 parts (combined), and then obtain predictions for the left-out kth part. This is done in turn for each part k= 1, 2,..K [23], and then the results were averaged to obtain the final result. In our study we used k=10 as a cross validation.

4.2. **Feature selection**

Due to the variety and high number of features extracted as well as supporting the idea of simplicity, we applied a feature

Table 3. *Shows the performance under different running configurations*.

#	Machine Learning Algorithm	1st Configuration: 263 features	2nd Configuration: Top 22 features	3rd Configuration: Pre-processing with 263 features	4th Configuration: Pre-processing with top 20 features
		263 features	Accuracy %	Accuracy %	Accuracy %
1.	Bayes Net	89.64	91.75	**93.66**	**94.71**
2.	Meta-Bagging	90.27	**92.38**	93.65	92.6
3.	Random forest	**90.90**	91.96	91.96	92.8
4.	AdaBoost M1	82.87	85.83	91.96	91.75

selection technique. This is used to rank the features, to explore the lowest number of features that provides the best classification accuracy, and to avoid overfitting the data. For the unprocessed data (with the presence of background noise), this function automatically selected the top 22 features based on their ranks, whilst 20 features were selected when working with the files that had been pre-processed. Table (2) lists the top (20) features automatically selected by Weka using the built-in attribute selection technique function.

5. Results

We used the four machine learning algorithms stated in section 4.1 to achieve the final results in four different configurations resulting from using pre-processing or not, and using the full (263) or the reduced features sets we also calculate the sensitivity and specificity for the highest score achieved for the 2nd and the 4th configurations.

Table (3) lists the accuracies obtained for the four different configurations. The highest classification accuracy achieved was 94.71% using the (BN) classifier, running under the fourth configuration followed by configuration three with 93.66% using (BN) classifier, while configurations two and one score 92.38% and 90.90% using (MB) and (RF) classifiers respectively.

By adopting a pre-processing step and extracting fewer, better quality features for the classifiers, the highest accuracy was achieved.

The sensitivity and specificity for the 2nd configuration was 92.00%. Only 19 patients from 240 and 17 HC subjects from 233 were incorrectly classified, but when comparing with the 4th configuration, only 7 AD patients were incorrectly classified making the sensitivity level at 97.00%. However the specificity of the 4th configuration was slightly reduced to 91.00% (only 21 HC were misclassified)

Our results reveal two important facts: first, the majority of the features have the potential to identify dementia even when all the features have been utilized by the classifiers (93.66% classification accuracy using the full 263 features compared to 94.71% when feature selection is used). This is in contrast to what had been reported by [15], as their results showed a sharp drop off in the case of using all of the features (from 92.01% classification accuracy with feature selection down to 79% without).

The first group of features measures the perturbation of the fundamental frequency reflecting the defects on vocal folds closing and opening times. This is, captured by the shimmer and jitter parameters as they measure the differences of amplitude and cycles of consecutive periods. Also it is known that AD patients produce more noise in their speech due to the fluctuations in the airflow, caused by incomplete vocal fold closure than do healthy subjects. This is measured by the harmonics to noise ratio (HNR) feature, previously demonstrated by [24], [14]. Pauses and number of silent segments are more prevalent in AD patients as they tend to shorten the speech segments in contrast to HC subjects. This is because the AD patients most of the time find that talking requires much effort and concentration. The MFCC features, although they are well-known as standards in speech recognition systems, capture important separation between the two groups as they relate to the articulators (libs and tongue) control ability, that is decreased in AD patients [25].

Secondly our proposed method is robust and very capable of identifying dementia patients from healthy subjects even in the presence of significant background noise. These facts support our proposition for using only acoustic features for automatic detection and/or screening of AD at a low cost and within the home environment.

6. Conclusions and future work

Speech and language impairment serve as a strong evidence for Alzheimer's disease detection and it can be used to indicate its severity over the time [26].

In our study, for the same data set (based on short speech recordings from a picture description task.), but using only acoustic features, higher accuracy results were obtained, in distinguishing between HC subjects and AD patients, than those reported in the most recent state of the art [15].

Furthermore, we used acoustic features derived automatically from the speech recordings without the addition of any lexical or syntactic features that rely on complex speech recognition technology as in [9].

In this paper, we proposed a simple high accuracy automated method that can be used in the clinic and/or at home to guide the diagnosing and/or screening of dementia by using just speech. In the future, we plan to investigate more features and to test the performance of our method with different datasets to classify between neurodegenerative dementia patients and people with functional memory disorders. The analysis will be applied to the conversations between the neurologists and patients during their visit to the memory clinic.

7. References

[1] C. F. Martin Prince, Renata Bryce, "World Alzheimer Report - The benefits of early diagnosis and intervention World Alzheimer Report," *Alzeheimer's Dis. Int.*, p. 72, 2011.

[2] J. Berger, "The age of biomedicine: current trends in traditional subjects," *J. Appl. Biomed.*, vol. 9, no. 2, pp. 57–61, 2011.

[3] V. Taler and N. a Phillips, "Language performance in Alzheimer's disease and mild cognitive impairment: a comparative review.," *J. Clin. Exp. Neuropsychol.*, vol. 30, no. 5, pp. 501–556, 2008.

[4] C. Laske, H. R. Sohrabi, S. M. Frost, K. López-De-Ipiña, P. Garrard, M. Buscema, J. Dauwels, S. R. Soekadar, S. Mueller, C. Linnemann, S. A. Bridenbaugh, Y. Kanagasingam, R. N. Martins, and S. E. O'bryant, "Innovative diagnostic tools for early detection of Alzheimer's disease," *Alzheimer's Dement.*, vol. 11, no. 5, pp. 561–578, 2015.

[5] K. S. Santacruz and D. Swagerty, "Early diagnosis of dementia," *Am Fam Physician*, vol. 63, no. 4, pp. 703–713,717–718, 2001.

[6] B. Klimova Q1 and K. Kuca, "Speech and language impairments in dementia – a mini review," *J. Econ. Financ. Adm. Sci.*, pp. 1–7, 2016.

[7] R. Bucks, S. Singh, J. Cuerden, and G. Wilcock, "Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance," *Aphasiology*, vol. 14, no. July 2015, pp. 71–91, 2000.

[8] K. López-de-ipi, M. Ecay-torres, P. Martinez-lage, and B. Beitia, "Feature selection for spontaneous speech analysis to aid in Alzheimer ' s disease diagnosis : A fractal dimension approach," vol. 30, pp. 43–60, 2014.

[9] D. Hakkani-Tur, D. Vergyri, and G. Tur, "Speech-based automated cognitive status assessment," *Proc. Interspeech 2010*, no. September, pp. 258–261, 2010.

[10] "Dementia Bank." [Online]. Available: https://talkbank.org/DementiaBank/. [Accessed: 10-Dec-2015].

[11] W. Jarrold, B. Peintner, D. Wilkins, D. Vergryi, C. Richey, M. L. Gorno-Tempini, and J. Ogar, "Aided diagnosis of dementia type through computer-based analysis of spontaneous speech," *Proc. Work. Comput. Linguist. Clin. Psychol. From Linguist. Signal to Clin. Real.*, pp. 27–37, 2014.

[12] S. O. Orimaye and K. J. Golden, "Learning Predictive Linguistic Features for Alzheimer ' s Disease and related Dementias using Verbal Utterances," *Proc. Work. Comput. Linguist. Clin. Psychol. From Linguist. Signal to Clin. Real.*, pp. 78–87, 2014.

[13] K. López-de-Ipiña, J. B. Alonso, N. Barroso, M. Faundez-Zanuy, M. Ecay, J. Solé-Casals, C. M. Travieso, A. Estanga, and A. Ezeiza, "New approaches for Alzheimer's disease diagnosis based on automatic spontaneous speech analysis and emotional temperature," *Ambient Assist. Living Home Care. Lect. Notes Comput. Sci.*, vol. 7657, pp. 407–414, 2012.

[14] A. König, A. Satt, A. Sorin, R. Hoory, O. Toledo-Ronen, A. Derreumaux, V. Manera, F. Verhey, P. Aalten, P. H. Robert, and R. David, "Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease," *Alzheimer's Dement. Diagnosis, Assess. Dis. Monit.*, vol. 1, no. 1, pp. 112–124, 2015.

[15] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify Alzheimer's disease in narrative speech," *J. Alzheimer's Dis.*, vol. 49, no. 2, pp. 407–422, 2015.

[16] J. R. Booth, B. Mac Whinney, and Y. Harasaki, "Developmental differences in visual and auditory processing of complex sentences," *Child Dev.*, vol. 71, no. 4, pp. 981–1003, 2000.

[17] "Audacity® is free, open source, cross-platform software for recording and editing sounds." [Online]. Available: http://www.audacityteam.org/. [Accessed: 15-Jan-2016].

[18] P. Boersma and D. Weenink, "Praat: doing phonetics by computer." [Online]. Available: http://www.fon.hum.uva.nl/praat/. [Accessed: 05-Jan-2016].

[19] J. J. G. Meilán, F. Martínez-Sánchez, J. Carro, D. E. López, L. Millian-Morell, and J. M. Arana, "Speech in Alzheimer's disease: can temporal and acoustic parameters discriminate dementia?," *Dement. Geriatr. Cogn. Disord.*, vol. 37, no. 5–6, pp. 327–334, 2014.

[20] B. Roark, M. Mitchell, J. P. Hosom, K. Hollingshead, and J. Kaye, "Spoken language derived measures for detecting mild cognitive impairment," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 19, no. 7, pp. 2081–2090, 2011.

[21] J. I. Godino-Llorente, P. Gómez-Vilda, and M. Blanco-Velasco, "Dimensionality reduction of a pathological voice quality assessment system based on gaussian mixture models and short-term cepstral parameters," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 10, pp. 1943–1953, 2006.

[22] "Weka 3: Data Mining Software in Java." [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/. [Accessed: 01-Feb-2016].

[23] P. Taylor, R. R. Picard, and R. D. Cook, "Cross-Validation of Regression Models Cross-Validation of Regression Models," vol. 79, no. April 2013, pp. 37–41, 2012.

[24] D. E. L. Juan José G. Meilán, Francisco Martínez-Sánchez, Juan Carro, "Speech in alzheimer's disease: Can temporal and acoustic parameters discriminate dementia?," *Dement. Geriatr. Cogn. Disord.*, vol. 37, no. 5–6, pp. 327–334, 2014.

[25] A. Tsanas, M. a. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of Parkinsons disease," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 5, pp. 1264–1271, 2012.

[26] M. Yancheva, K. Fraser, and F. Rudzicz, "Using linguistic features longitudinally to predict clinical scores for Alzheimers disease and related dementias." in *6th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, 2015, p. 134.

Discriminating the Infant Cry Sounds Due to Pain vs. Discomfort Towards Assisted Clinical Diagnosis

Vinay Kumar Mittal

Indian Institute of Information Technology Chittoor, Sri City, India

vkmittal@iiits.in

Abstract

Cry is a means of communication for an infant. Infant cry signal is usually perceived as a high-pitched sound. Intuitively, significant changes seem to occur in the production source characteristics of cry sounds. Since the instantaneous fundamental frequency (F_0) of infant cry is much higher than for adults and changes rapidly, the signal processing methods that work well for adults may fail in analyzing these signals. Hence, in this paper, we derive the excitation source features F_0 and strength of excitation (SoE) using a recently proposed *modified zero-frequency filtering method*. Changes in the production characteristics of acoustic signals of infant cries due to *pain* and *discomfort* are examined using the features F_0, SoE and signal energy. These changes are validated by visually comparing their spectrograms with the spectrograms of the acoustic signals. Effectiveness of these discriminating features is examined for different pain/discomfort cry sounds pairs in an 'Infant Cry Signals Database (IIIT-S ICSD)', especially collected for this study. Fluctuations in the features F_0, SoE and energy are observed to be larger in the case of infant cry due to *pain*, than for *discomfort*. These features can help in developing further the clinical assistive technologies for discriminating different infant cry types and initiating the remedial measures automatically.

Index Terms: Infant cry signals, modified zero-frequency filtering, pain cry, discomfort cry, excitation source characteristics

1. Introduction

Whenever an infant cries, his/her mother invariably knows the reason why her baby is crying. Infants *cry* to communicate either their some need or condition that could be physiological, pathological or environmental. In today's fast paced life, there are situations routinely, when the infants are under the physical/medical care of people other than the parents. For such conditions, if the assistive technologies could be developed that can help the parents and other care-taking people know the cause of an infant's crying, several diverse applications can then be evolved. It could also assist clinical diagnosis of any medical condition that an infant may be suffering from. But, this necessitates characterising the changes in the acoustic signal of infant cry, and also possibly understand the differences in the features of infant cry signal due to different causes. This paper aims at exploring few such discriminating features of infant cry signal.

Infant cry is a combination of vocalization, constrictive silence, coughing, choking and interruptions [1]. It provides information about the health, gender, disease and emotions etc. of the infant. The first cry of an infant is an important parameter in determining the Apgar count, which can be used to classify neonates into healthy and unhealthy (or weak) [1]. This is the first tool of communication and the sign of life at birth. Various characteristics of the first cry are vocalizations, facial expressions and limb movements, all of which change over time. Infants cry to let others know about their problems or needs, just like adults do by talking. Thus infant cry falls in the most sensitive range of the human auditory perception [2]. Physiological variables such as, facial expressions, muscular tonus, sleep and suction abilities have been studied as parameters to estimate the needs of an infant [3]. The study of infant cry has gained significance over the years, for diverse applications including early detection of the cause of cry and the possible ailment.

Infant cry signal, produced in response to a stimuli, involves a rhythmic pattern of cry sounds and inhalation. An important feature used for the analysis of infant cries in most studies is the instantaneous fundamental frequency (F_0). The fundamental frequency and its first three harmonics were studied [4, 5]. Infant cries were attempted to be classified on the basis of pain, sadness, hunger, fear and other few causes [6, 7, 8, 9]. Pitch characteristics of infant cries were categorized into urgent, arousing, distressing or sick, using linear predictive (LP) coding [10]. Pitch measurements at every epoch were taken, using a time-domain based cross-correlation [11]. The start and end time of each cry segment were detected, using short-time energy function and zero-crossing rate [8, 12].

Spectrographic analysis of cry signal was carried out to characterize pitch and its harmonics [1]. First three formants along with fundamental frequency were used for the analysis in [7, 13]. Cepstrum analysis was used for extracting the fundamental frequency, along with LP analysis for extracting the first three formants [13]. Features such as short-time energy, zero-crossing rate and linear prediction cepstral coefficients (LPCCs) were used for the analysis of cry signals [8]. Parameters such as mean, standard deviation and peak value of the fundamental frequency were used to examine hyper-phonation [10]. Parameters such as segment density, segment length and pause length were used to study their relation with the gender of the baby [12]. Analyses of cries of infants with different heart disorders were carried out by comparing frame-wise mean F_0, and minimum and maximum F_0 values [13]. The parameters such as duration, fundamental frequency and the shape of F_0 contour were also explored to describe a cry [14].

In infant cry, the harmonic structure was observed to be related to abnormalities in larynx, and dysphonation due to muscle pain or discomfort [1]. The LPCC magnitudes for cries due to similar causes were observed to be similar, and different for cries due to different causes [8]. Typical characteristics from the shape of power spectra of signals were obtained for cries due to hunger, sleepiness and discomfort, with classification accuracy of 85% [9]. Segment length analysis of cry signals had shown similarity among normal infants as compared to those with hearing disorder [12]. Cries of infants were divided into normal,

SLPAT 2016, 7th Workshop on Speech and Language Processing for Assistive Technologies, pages 37–42,
San Francisco, USA, 13 September, 2016.

and due to disorders such as Tetralogy of Fallot, Ventricular Septal Defect, Atrial Septal Defect, and Patent Ductus Arteriosus [13]. However, the production characteristics of the acoustic signals of infant cries due to different causes have not been studied much. Our preliminary study of infant cry signals [15] also indicated the need for examining in detail their production characteristics, where intuitively larger changes seem to occur. However, evolving the signal processing methods for reliable F_0 extraction from the infant cry signal remains a challenge.

This paper focuses on examining changes in the acoustic signals of infant cries due to different causes, from their production point of view. Especially, the excitation source characteristics derived from the acoustic signals of infant cries due to *pain* and *discomfort* are analysed. An *Infant Cry Signals Database (IIIT-S ICSD)* [16] is used. The database consists of infant cries due to six different causes. But, due to better availability of multiple pain/discomfort cry sound pairs produced by same infant, the discriminating features are examined for infant cry sounds due to *pain* and *discomfort* categories of cry-causes. Changes in the signals are examined using three production features, namely, F_0, strength of excitation (SoE) and signal energy. The excitation source features F_0 and SoE are extracted using the *modified zero-frequency filtering (modZFF)* method [17, 18, 19]. Effectiveness of the discriminating features derived using the *modZFF method* is validated by visually comparing the spectrograms of an excitation source feature, the SoE impulse sequence, with that of the acoustic signal. Results indicate larger changes in the production features of infant cry due to *pain*, than for *discomfort*. These discriminating features can be used in developing further the assistive systems for clinical diagnosis of infant cry sounds, with wide ranging applications.

This paper is organized as follows. In Section 2, the details of the data collected are discussed. The *modZFF method* used for extracting the excitation source features from the infant cry signals, is described briefly in Section 3. Section 4 analyses the production characteristics of acoustic signals of infant cries due to pain and discomfort. Results are discussed in Section 5. Section 6 gives a summary, along with scope of further work.

2. Data for the study

The data of acoustic signals of infant cries due to different causes was especially collected for this study. This database is named as the *Infant Cry Signals Database (IIIT-S ICSD)* [16]. The infant cry signals in the ICSD were recorded in a private paediatric hospital, under the supervision of two medical experts (paediatricians). The infant cry data was recorded for infants needing routine check-up, vaccination or cure to some ailment. The recordings were made in multiple sessions in the doctor's room whenever an infant was brought-in for regular check-up, vaccination or for cure of some ailment, and the infant cried because of pain, ailment, discomfort, emotional need, change of environment or hunger/thirst etc. Infant cries were categorized as per the doctors and parents, into six classes of cause factors [16], as elaborated in Table 1. The cry data was recorded for infants in the age group of 3 months to 2 years.

The acoustic signals' data in the ICSD was recorded using a Roland Edirol R-09 Wave/MP3 recorder, placed at around 10-20 cm from infant's mouth. The data was recorded in stereo mode, at a sampling rate of 48 kHz and 24 bit/sample coding. People in the doctor's room were requested not to speak (for few secs) during recording of the cry data. Parents were requested not to make any efforts to calm the baby for a short period of time, so as to record the clean signals. Data was further pre-processed manually by listening carefully, and using the software tools such as Wavesurfer, Audacity and MATLAB to make it free from any noise or any overlapping speech sounds. The *IIIT-S ICSD* consists of 693 infant cry samples of 33 speakers (infants), recorded for total about 670 sec, that are stored in 76 files [16]. This *IIIT-S ICSD* database collected for the purpose of research towards evolving the speech sounds based assistive technologies, can be made available on request.

Table 1: *Categories of Possible Causes of Infant Cry*

(a) *Causes*	(b) *Description of Causes of Infant Cry*
1. Pain	Cry due to internal pain, or external pain caused by vaccination or any physical hurt on the body
2. Ailment	Cries due to any ailment such as cold, cough or fever etc.
3. Discomfort	Cry due to irritation caused by the external factors, e.g., the doctor opening baby's mouth (investigation) or nurse holding the baby (vaccination)
4. Emotional need for attention	Cry when the baby has emotional need to go back into parents arms and feel their touch, or need cuddling
5. Environmental factors	Cry due to fear of the surroundings or need a change in the environment (e.g., need for changing the diapers)
6. Hunger/thirst	Cries due to hunger or thirst

The study of relative changes in the production characteristics of infant cry sounds due to different causes requires availability of acoustic signals data of cry sounds due to different causes, produced by the same speaker, i.e., an infant. Since, multiple pairs of acoustic signals of pain/discomfort cry sounds produced by same infant are better available in the ICSD, the discriminating features are examined in this paper for *pain* and *discomfort* categories of infant cry sounds. Though, the *IIIT-S ICSD* consists of acoustic signals of infant cries due to six categories of different causes, as diagnosed by medical experts and the parents (see Table 1). The data-pairs of *pain* and *discomfort* cry signals both produced by same infant are chosen from the *IIIT-S ICSD* [16]. These are examined using a recently proposed signal processing method, the *modZFF method*, discussed in the next section.

3. Deriving the excitation source features using Modified Zero-Frequency Filtering

Changes in the acoustic signals of infant cry sounds are examined in this study using three production features, namely, the instantaneous fundamental frequency (F_0), the strength of excitation (SoE) and frame-wise signal energy (E). Production of acoustic signal of infant cry apparently involves significant changes in the glottal excitation source characteristics [15]. Hence, the excitation source features F_0 and SoE are focused in this study. Though the feature F_0 can be derived from the acoustic signal using the autocorrelation or linear prediction residual methods, but those methods give only the indicative results, good for preliminary investigation [15]. In this study, the excitation source characteristics F_0 and SoE are extracted from the acoustic signal of infant cry using the *modified zero-frequency filtering (modZFF)* method [17, 18, 19]. Differences

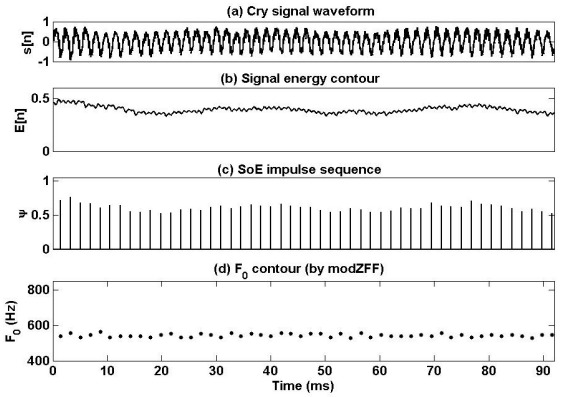


Figure 1: Illustration of features for *discomfort cry signal*: (a) cry signal waveform, (b) signal energy contour, and (c) SoE impulse sequence and (d) $F0$ contour derived using the *modZFF* method, for (discomfort) cry signal of infant #S02 (male).

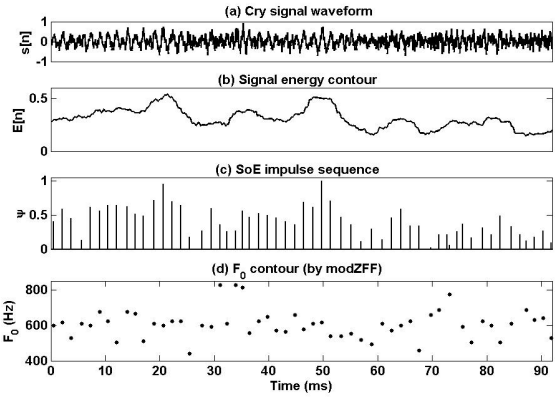


Figure 2: Illustration of features for *pain cry signal*: (a) cry signal waveform, (b) signal energy contour, and (c) SoE impulse sequence and (d) $F0$ contour derived using the *modZFF* method, for (pain) cry signal of infant #S02 (male).

of changes in these features are examined using their mean and standard deviation. The infant cry cause assessed by the doctor and parents is used as base reference. Effectiveness of the discriminating features, mainly the excitation source features, is validated by visually comparing the spectrograms of the SoE impulse sequence and the acoustic signal.

The excitation source characteristics can be derived from the normal speech signals, using the zero-frequency filtering (ZFF) method [17, 18]. But use of the ZFF method to derive the excitation source features from the acoustic signal of *nonverbal sounds* such as *infant cry*, that have significant changes in their source characteristics, pose two limitations [19]. (i) Shorter window length would be required for trend removal operation. (ii) Impulse sequence for aperiodic signals may be affected by the choice of shorter window length. Both of these limitations are addressed in the *modified zero-frequency filtering (modZFF)* method [19, 20], that uses gradually reducing window lengths instead of a fixed window length, for the trend removal operation. Key steps involved in the *modZFF* method are as follows:

1. Pre-process the input cry signal $(s[n])$ by downsampling to 8 kHz, smoothen it over m sample points to obtain an equivalent effect of low-pass filtering, and then upsample it back to the sampling frequency of the original signal [19]. The resultant pre-processed signal is $(s_p[n])$.
2. Get the differenced signal $(\tilde{x}[n])$ from the pre-processed signal $(s_p[n])$, using:

$$\tilde{x}[n] = s_p[n] - c\, s_p[n-1] \tag{1}$$

   where c is a constant (usually value of c = 0.9 to 1.0 is chosen) and $n = 1, 2, 3, \ldots$. The differenced signal $(\tilde{x}[n])$ gives a zero-mean signal $(\hat{x}[n])$.
3. Pass the zero-mean signal $\hat{x}[n]$ through a cascade of two zero-frequency resonators (ZFRs), i.e., two ideal digital resonators at 0 Hz, to get the ZFR output signal $\tilde{y_1}[n]$.

$$\tilde{y_1}[n] = \sum_{k=1}^{4} a_k \tilde{y_1}[n-k] + \hat{x}[n], \tag{2}$$

   where, $a_1 = +4$, $a_2 = -6$, $a_3 = +4$, $a_4 = -1$.
4. Remove the trend built-up in the cascaded ZFRs' output $(\tilde{y_1}[n])$ by successive integration operations, i.e., by subtracting the local mean computed over a window. Gradually reducing window lengths of 20 ms, 10 ms, 5 ms, 3 ms, 2 ms and 1 ms are used in the successive trend removal stages, to highlight the excitation source information better. Output of each trend removal stage $(\hat{y_2}[n])$ is given by below equations:

$$\hat{y_2}[n] = \tilde{y_1}[n] - \bar{y}[n] \tag{3}$$

$$\bar{y}[n] = \frac{1}{2N+1} \sum_{n=-N}^{N} \tilde{y_1}[n] \tag{4}$$

   where, $2N + 1$ is window length in number of samples. Here, $\bar{y}[n]$ represents the local mean computed over each successive window. The final trend removed output $(\hat{y_2}[n])$ is called the *modified zero-frequency filtered (modZFF)* signal, i.e. $z_m[n]$ [19].
5. The positive to negative going zero-crossings of the *modZFF* signal $(z_m[n])$ give locations of impulses.
6. The slope of the *modZFF* signal $(z_m[n])$ around each impulse location indicates the relative *strength of excitation (SoE)* there. The *SoE* is denoted as ψ in this paper.

An illustration of SoE impulse sequence and F_0 contour derived from the acoustic signal of infant cry, using the *modZFF method* is shown in Fig. 1(c) and Fig. 1(d), respectively. The *modZFF method* helps deriving an SoE impulse sequence (as the excitation source characteristics) from the acoustic signal of infant cry. The *SoE*, i.e., the amplitudes of impulses, indicate the strength of excitation around the respective impulse locations. Using these excitation source features (F_0 and SoE), derived using the *modZFF method* larger changes are observed in the production features of infant cry due to *pain*, than for *discomfort*. Analysis details are discussed in the next section.

4. Discriminative analysis of Pain vs. Discomfort infant cry sounds

In this paper, the production characteristics of acoustic signals of infants cries data in the *IIIT-S ICSD* are examined under two

Table 2: *Changes in F_0 for pain vs. discomfort cry signals:* (a) speaker #, the (b) mean (μ_{F0_D}) and (c) std. dev. (σ_{F0_D}) for *discomfort* cry, the (d) mean (μ_{F0_P}) and (e) std. dev. (σ_{F0_P}) for *pain* cry, and the changes (Δ %) in (f) mean (μ_{F0}), (g) std. dev. (σ_{F0}) and (h) normalized std. dev. ($\sigma_{N_{F0}}$) in F_0 for *pain* cry from *discomfort* cry signals. Note: M indicates Male and F indicates Female infant.

(a) Speaker (Infant) #	(b) μ_{F0_D} (Hz)	(c) σ_{F0_D} (Hz)	(d) μ_{F0_P} (Hz)	(e) σ_{F0_P} (Hz)	(f) $\Delta\mu_{F0}$ (%)	(g) $\Delta\sigma_{F0}$ (%)	(h) $\Delta\sigma_{N_{F0}}$ (%)
S02 (M)	667.9	184.6	737.9	272.5	10.48	47.60	33.60
S03 (M)	705.9	228.5	620.1	268.3	-12.16	17.42	33.68
S04 (F)	676.9	311.8	754.3	380.0	11.43	21.88	9.38
S05 (M)	654.9	259.9	707.8	306.3	8.10	17.83	9.00
S06 (M)	734.7	232.7	647.9	277.5	-11.81	19.21	35.18
S07 (F)	581.8	219.7	619.8	252.9	6.53	15.12	8.07
S09 (M)	620.7	200.4	664.4	308.8	7.04	54.06	43.94
S10 (F)	667.8	239.8	658.6	283.2	-1.37	18.05	19.69
S11 (F)	579.3	152.2	664.0	211.1	14.63	38.70	20.99
S12 (M)	684.1	224.9	685.4	294.8	0.19	31.09	30.85
S13 (F)	523.6	132.9	530.1	150.4	1.24	13.18	11.79
Average	*645.2*	*217.1*	*662.8*	*273.3*	*3.12*	*27.74*	*23.29*

Table 3: *Changes in SoE (ψ) for pain vs. discomfort cry signals:* (a) speaker #, the (b) mean (μ_{ψ_D}) and (c) std. dev. (σ_{ψ_D}) for *discomfort* cry, the (d) mean (μ_{ψ_P}) and (e) std. dev. (σ_{ψ_P}) for *pain* cry, and the changes (Δ %) in (f) mean (μ_{ψ}), (g) std. dev. (σ_{ψ}) and (h) normalized std. dev. ($\sigma_{N_{\psi}}$) in SoE (ψ) for *pain* cry from *discomfort* cry signals. Note: M/F indicates Male/Female infant.

(a) Speaker (Infant)#	(b) μ_{ψ_D}	(c) σ_{ψ_D}	(d) μ_{ψ_P}	(e) σ_{ψ_P}	(f) $\Delta\mu_{\psi}$ (%)	(g) $\Delta\sigma_{\psi}$ (%)	(h) $\Delta\sigma_{N_{\psi}}$ (%)
S02 (M)	.3253	.2132	.2056	.1374	-36.80	-35.55	1.97
S03 (M)	.1613	.1247	.2640	.1776	63.37	42.42	-12.98
S04 (F)	.2454	.1880	.2185	.1556	-10.96	-17.23	-7.04
S05 (M)	.2986	.1860	.2149	.1815	-28.02	-2.42	35.56
S06 (M)	.2782	.1950	.2098	.1674	-24.59	-14.15	13.83
S07 (F)	.2321	.1770	.2955	.2292	27.32	29.49	1.71
S09 (M)	.2305	.1830	.1713	.1528	-25.68	-16.50	12.35
S10 (F)	.1582	.1305	.2262	.1637	42.98	25.44	-12.27
S11 (F)	.2745	.1880	.2212	.1676	-19.42	-10.85	10.63
S12 (M)	.1447	.1307	.1453	.1433	0.41	9.64	9.19
S13 (F)	.3542	.2220	.3127	.2407	11.72	8.42	22.81
Average	*.2457*	*.1762*	*.2259*	*.1743*	*0.03*	*1.70*	*6.89*

categories, namely, *pain* and *discomfort*. Since the data available in other categories is less for same speaker (i.e., same infant), these cry categories may be analysed in future after extending the database. The excitation source features F_0 and SoE are derived using the modZFF method, for each infant cry signal. The signal energy E, a production feature, represents the combined effect of the excitation source and the vocal tract filter. It is computed for frame size of 5 ms at each time-instant.

Relative changes are examined in the production features (F_0, SoE and E) of the acoustic signals of infant cries due to *pain* and *discomfort*. An illustration of energy (E) contour, the SoE impulse sequence and the F_0 contour for *discomfort cry* is shown in Fig. 1(b), (c) and (d), respectively. Changes in these features for *pain cry* are illustrated in Fig. 2, in a similar way. Some patterns can be observed. In Fig. 2, for infant cry due to *pain*, the F_0 contour has near *cyclic* changes with larger fluctuations, that could be due to physiological conditions during pain. In Fig. 1, for infant cry due to *discomfort*, the F_0 contour is relatively *flat*, with changes at larger intervals and shorter fluctuations. Similar changes are observed in the SoE and E also.

Quantitative analysis is carried out by measuring the statistical parameters *mean* (μ), *standard deviation* (σ or std dev) and *normalized standard deviation* ($\sigma_N = \sigma/\mu$) in the production features F_0, SoE and E. Changes in these parameters for *pain* cry vs. *discomfort* cry are compared using the percentage difference, e.g., $\Delta\mu_{F_0} = (\mu_{F_{0_P}} - \mu_{F_{0_D}})/\mu_{F_{0_D}} \times 100(\%)$. Likewise, changes in the features SoE and E are compared using the $\Delta\mu_{\psi}(\%)$ and $\Delta\mu_{E}(\%)$, respectively. Changes in fluctuations in the features F_0, SoE and E are given in Table 2, Table 3 and Table 4, respectively, for *pain* vs. *discomfort* cry signals of first 11 infants (6 males, 5 females). Data of two speakers (S01 and S08) is discarded due to overlapping speech present in the signal. Similar changes in the production features are observed for pain/discomfort across speakers in the database.

In Table 2, the *fluctuations* in F_0 ($\Delta\sigma(\%)$ in column (g)) are larger for *pain* cry in comparison to *discomfort* cry. *Normalized fluctuations* ($\Delta\sigma_N(\%)$ in column (h)) are also larger for *pain* cry. In general, the average F_0 values increase for *pain* cry w.r.t. *discomfort* cry ($\Delta\mu_{F_0}(\%)$ in column (f)). But, in few cases (S03, S06 and S10), the F_0 for *discomfort* cry (if high) may reduce for *pain* cry.

In Table 3, the average SoE values ($\Delta\mu_{\psi}(\%)$ in column (f)) reduce in general for *pain* cry w.r.t. *discomfort* cry. This reduction in the SoE with increase in F_0, is in line with earlier observation of relative changes (in opposite direction) in the F_0 and SoE, for normal and shouted speech of adults [21, 22]. Interestingly, in the cases (S03, S10) where the mean F_0 (μ_{F_0}) decreases for *pain* cry w.r.t. *discomfort* cry, the mean SoE

Table 4: *Changes in signal energy (E) for pain vs. discomfort cry signals:* (a) speaker #, the (b) mean (μ_{E_D}) and (c) std. dev. (σ_{E_D}) for *discomfort* cry, the (d) mean (μ_{E_P}) and (e) std. dev. (σ_{E_P}) for *pain* cry, and the changes (Δ %) in (f) std. dev. (σ_E) and (g) normalized std. dev. (σ_{N_E}) in E for *pain* cry from *discomfort* cry signals. Note: M indicates Male and F indicates Female infant.

(a) Speaker (Infant)#	(b) μ_{E_D}	(c) σ_{E_D}	(d) μ_{E_P}	(e) σ_{E_P}	(f) $\Delta\sigma_E$ (%)	(g) $\Delta\sigma_{N_E}$ (%)
S02 (M)	.2351	.1361	.1904	.1042	-23.44	-5.46
S03 (M)	.1370	.1025	.1625	.0905	-11.71	-25.56
S04 (F)	.1474	.1070	.1525	.0893	-16.54	-19.34
S05 (M)	.1818	.1080	.1386	.0882	-18.33	7.12
S06 (M)	.1756	.1170	.1561	.1037	-11.37	-0.28
S07 (F)	.1530	.0960	.1917	.1049	9.27	-12.77
S09 (M)	.1617	.1120	.1303	.0874	-21.96	-3.14
S10 (F)	.1444	.0930	.1346	.0739	-20.54	-14.75
S11 (F)	.1602	.1120	.1502	.1082	-3.39	3.04
S12 (M)	.0829	.0746	.1073	.0981	31.50	1.60
S13 (F)	.1922	.1020	.1759	.1243	21.86	33.16
Average	*.1610*	*.1055*	*.1536*	*.0975*	*-5.88*	*-3.31*

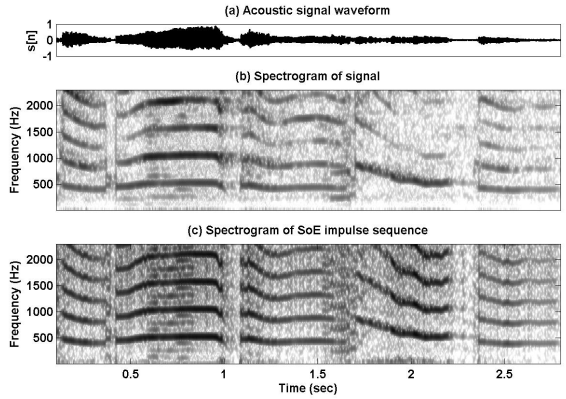


Figure 3: Validation of the excitatation source characteristics of acoustic signal of *discomfort* cry, using spectrograms: (a) acoustic signal, and spectrograms of (b) the acoustic signal and (c) the SoE impulse sequence derived using the modZFF method, for *discomfort* cry signal of infant #S13 (female).

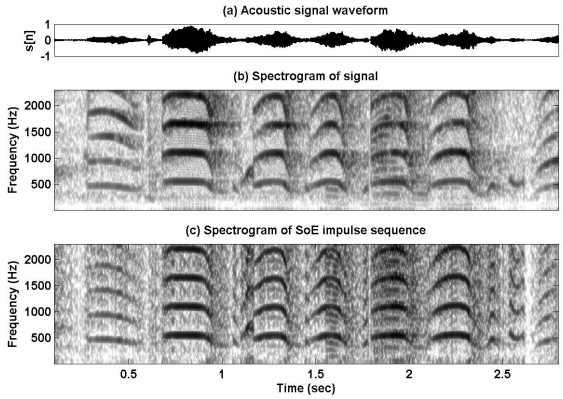


Figure 4: Validation of the excitatation source characteristics of acoustic signal of *pain* cry, using spectrograms: (a) acoustic signal, and spectrograms of (b) the acoustic signal and (c) the SoE impulse sequence derived using the modZFF method, for *pain* cry signal of infant #S13 (female).

(μ_ψ) increases. The fluctuations in the SoE ($|\Delta\sigma_\psi|(\%)$ in column (g)) and normalized fluctuations ($|\Delta\sigma_{N_\psi}|(\%)$ in column (h)) are larger for *pain* cry w.r.t. *discomfort* cry of most infants. Similar trend of changes in the fluctuations in signal energy (E) are observed in Table 4, in columns (f) and (g).

5. Discussion on results

Changes in the excitation source feature SoE (in Table 3) appear to be more prominent than changes in the signal energy E (in Table 4), as also observed in Fig. 2. Prominence of changes in the source characteristics in comparison to acoustic signal is validated by visual inspection of the spectrograms ($|X(\tau,\omega)|^2$), used as the ground truth. These spectrograms are obtained using the *short-time Fourier transform* $X(\tau,\omega) = \sum_{n=-\infty}^{n=\infty} x[n]\, w[n-m]e^{-jwn}$ [23, 24], for signal frames of size 20 ms, with a frame shift of 1 ms. Spectrograms for the SoE impulse sequences (Fig. 3(c) and Fig. 4(c)), reveal the excitation source characteristics better than the spectrograms of the acoustic signals (Fig. 3(b) and Fig. 4(b)). These spectrograms also indicate the nature of inter-cry changes in the excitation source features, that are similar to those illustrated in Fig. 1(d) and Fig. 2(d).

For *pain* cry signal, the contours of F_0 and harmonics have *cyclic* changes with larger fluctuations (region 1.2 sec to 2.4 sec in Fig. 4). But, for cries due to *discomfort* these contours are relatively *flat* with fewer fluctuations (region 1.1 sec to 2.3 sec in Fig. 3). The cyclic changes with larger fluctuations for *pain* cry are possibly due to significant changes in the excitation source characteristics during the production of cry signal in shorter and louder bursts. Whereas, relatively flat contours with lesser fluctuations for *discomfort* cry may be due to slow moaning, with related smaller changes in the excitation source characteristics.

6. Summary and conclusion

In this paper, the production characteristics are examined for acoustic signals of the infant cries due to *pain* and *discomfort*. Aim is to characterise the infant cry signal and identify features that help distinguishing the causes of infant cries. Acoustic signals of infant cries due to *pain* or *discomfort* are examined us-

ing three production features F_0, SoE and signal energy. The excitation source features F_0 and SoE are derived using the *modified zero-frequency filtering* method. The F_0 contour for *pain* cry has *cyclic* changes with larger fluctuations, but it is relatively *flat* with lesser fluctuations for *discomfort* cry. This observation is validated quantitatively by comparing the relative fluctuations and normalized fluctuations in F_0 and SoE for *pain vs. discomfort* cry. Significance of changes in the source characteristics is validated using spectrograms of the SoE impulse sequence and the acoustic signal. The results are consistent across cry signals of speakers (infants) in the database.

In future, changes may be examined in the production characteristics of acoustic signals of infant cries due to remaining categories other than pain and discomfort. This author is working towards developing the systems where automated detection of cause of infant cry may help the parents as well the doctors towards assisted clinical diagnosis of an infant's ailment. Details of these attempts may be expected to appear in future publications of the author.

However, this study highlights the importance of examining changes in the excitation source characteristics of acoustic signals of the paralinguistic sounds such as infant cry. The study should also be helpful towards developing the assistive technologies and systems that may help early diagnosis of the ailment and medical care to an infant by identifying the cause of cry, from the acoustic signal. It could be immensely useful in the cases of ailments where the reaction-time of ailment detection and remedial measures could be of critical importance for an infant's life.

7. Acknowledgements

The author would like to thank paediatricians Dr. Manish Gour (M.B.B.S, DCH) and Dr. Nizam (M.B.B.S) from Pranaam Hospital, Hyderabad for their help in collecting the infant cry data. The author is also grateful to all parents of the infants, for willingly giving the details about themselves and their infants, and cooperate in the data collection. The author would also like to thank the anonymous reviewers for their insightful suggestions.

8. References

[1] A. Neustein, *Advances in speech recognition: mobile environments, call centers and clinics*. Springer, 2010, ch. fourteenth, pp. 324–345.

[2] G. Varallyay Jr, Z. Benyo, A. Illenyi, Z. Farkas, and G. Katona, "The speech-chorus method at the analysis of the infant cry," *Acoustic Review*, vol. 6, no. 2, pp. 9–15, 2005.

[3] Y. Skogsdal, M. Eriksson, and J. Schollin, "Analgesia in newborns given oral glucose," *ACTA Paediatrica*, vol. 86, no. 2, pp. 217–220, 1997.

[4] H. E. Baeck and M. N. Souza, "Study of acoustic features of newborn cries that correlate with the context," in *Proc. 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 3, 2001, pp. 2174–2177.

[5] R. P. Daga and A. M. Panditrao, "Acoustical Analysis of Pain Cries in Neonates: Fundamental Frequency," *IJCA Special Issue on Electronics, Information and Communication Engineering (ICEICE)*, vol. 3, pp. 18–21, Dec. 2011.

[6] Y. Abdulaziz and S. M. Syed Ahmad, "Infant cry recognition system: A comparison of system performance based on mel frequency and linear prediction cepstral coefficients," in *Proc. IEEE International Conference on Information Retrieval & Knowledge Management (CAMP)*, 2010, pp. 260–263.

[7] R. Hidayati, I. K. E. Purnama, and M. H. Purnomo, "The extraction of acoustic features of infant cry for emotion detection based on pitch and formants," in *Proc. IEEE International Conference on Instrumentation, Communications, Information Technology, and Biomedical Engineering (ICICI-BME)*, 2009, pp. 1–5.

[8] K. Kuo, "Feature extraction and recognition of infant cries," in *Proc. IEEE International Conference on Electro/Information Technology (EIT 2010)*, 2010, pp. 1–5.

[9] Y. Mima and K. Arakawa, "Cause Estimation of Younger Babies' Cries from the Frequency Analyses of the Voice-Classification of Hunger, Sleepiness, and Discomfort," in *Proc. IEEE International Symposium on Intelligent Signal Processing and Communications (ISPACS'06)*, 2006, pp. 29–32.

[10] P. S. Zeskind and T. R. Marshall, "The relation between variations in pitch and maternal perceptions of infant crying," *Child Development*, pp. 193–196, 1988.

[11] M. Petroni, M. E. Malowany, C. C. Johnston, and B. J. Stevens, "A new, robust vocal fundamental frequency (F_0) determination method for the analysis of infant cries," in *Proc. IEEE Seventh Symposium on Computer-Based Medical Systems*, 1994, pp. 223–228.

[12] G. Várallyay, "Future prospects of the application of the infant cry in the medicine," *Electrical Engineering*, vol. 50, no. 1-2, pp. 47–62, 2006.

[13] S. Chandralingam, T. Anjaneyulu, and K. Satyanarayana, "Estimation of Fundamental and Formant frequencies of infants cries; a study of Infants with congenital Heart disorder," *Indian Journal of Computer Science and Engineering*, vol. 3, no. 4, pp. 574–582, 2012.

[14] G. Várallyay Jr., "The melody of crying," *Elsevier International Journal of Pediatric Otorhinolaryngology*, vol. 71, no. 11, pp. 1699–1708, 2007.

[15] S. Asthana, N. Varma, and V. K. Mittal, "Preliminary Analysis of Causes of Infant Cry," in *Proc. IEEE 14th International Symposium on Signal Processing and Information Technology (ISSPIT 2014)*, Dec. 15-17 2014.

[16] S. Sharma, S. Asthana, and V. K. Mittal, "A Database of Infant Cry Sounds to Study the Likely Cause of Cry," in *Proc. ACL Anthology 12th International Conference on Natural Language Processing (ICON-2015)*, IIITM-K, Trivendram, Dec. 11-14 2015.

[17] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1602–1613, 2008.

[18] B. Yegnanarayana and K. S. R. Murty, "Event-based instantaneous fundamental frequency estimation from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 614–624, 2009.

[19] V. K. Mittal and B. Yegnanarayana, "Study of characteristics of aperiodicity in Noh voices," *The Journal of the Acoustical Society of America*, vol. 137, no. 6, pp. 3411–3421, 2015.

[20] V. K.Mittal and B. Yegnanarayana, "Analysis of production characteristics of laughter," *Elsevier Computer Speech & Language*, vol. 30, no. 1, pp. 99–115, 2015.

[21] V. K. Mittal and B. Yegnanarayana, "Production features for detection of shouted speech," in *Proc. 10th Annual IEEE Consumer Communications and Networking Conference, 2013 (CCNC'13)*, Jan. 11-14, 2013, pp. 106–111.

[22] V. K. Mittal and B. Yegnanarayana, "An automatic shout detection system using speech production features," in *Proc. Second International Workshop on Multimodal Analyses enabling Artificial Agents in Human-Machine Interaction (MA3HMI 2014), Held in Conjunction with INTERSPEECH 2014*, Singapore, Sep. 2014, pp. 88–98, published as Lecture Notes in Artificial Intelligence (LNAI 8757) by Springer.

[23] A. V. Oppenheim, R. W. Schafer, J. R. Buck *et al.*, *Discrete-time signal processing*. Prentice-hall Englewood Cliffs, 1989, vol. 2, ch. third.

[24] L. Rabiner, B. H. Juang, and B. Yegnanarayana, *Fundamentals of Speech Recognition*, Indian subcontinent adaptation, first ed. New Delhi, India: Pearson Education Inc., 2009, ch. third, pp. 88–113.

Improvement of Continuous Dysarthric Speech Quality

Anusha Prakash[1], *M. Ramasubba Reddy*[1], *Hema A Murthy*[2]

[1]Dept of Applied Mechanics, Indian Institute of Technology Madras, India
[2]Dept of Computer Science & Engineering, Indian Institute of Technology Madras, India
am13s002@smail.iitm.ac.in, hema@cse.iitm.ac.in

Abstract

Dysarthria refers to a group of motor speech disorders as the result of any neurological injury to the speech production system. Dysarthric speech is characterised by poor speech articulation, resulting in degradation in speech quality. Hence, it is important to correct or improve dysarthric speech so as to enable people having dysarthria to communicate better.

The aim of this paper is to improve the quality of continuous speech of several people suffering from dysarthria. Experiments in the current work use two databases- Nemours database and speech data collected from a dysarthric speaker of Indian origin. Durational analysis of dysarthric speech versus normal speech is performed. Based on the analysis, manual modifications are made directly to the speech waveforms and an automatic technique is developed for the same. Evaluation tests indicate an average preference of 78.44% and 67.04% for the manually and automatically altered speech over the original dysarthric speech, thus emphasising the effect of durational modifications on the perception of speech quality. Intelligibility of speech generated by three techniques, namely, proposed automatic modification technique, a formant re-synthesis technique, and an HMM-based adaptive system, is compared.

Index Terms: continuous dysarthric speech, Indian dysarthric speaker, durational modifications, formant re-synthesis, HMM-based adaptive system

1. Introduction

The word dysarthria, originating from *dys* and *arthrosis*, means difficult or imperfect articulation. Speech of a person suffering from dysarthria is affected due to a neurological defect in the speech production system [1]. There is a lack of coordination amongst the various parts involved in speech production to produce understandable speech. Dysarthric speech is characterised by the poor articulation of phonemes, problems with speech rate, incorrect pitch trajectory, swallowing or drooling while speaking. As a result, people with dysarthria have problems with speaking most often. The main aim of the paper is to improve the speech quality of continuous dysarthric speech.

Several efforts on correcting dysarthric speech to make it more intelligible are available in the literature. In [2], dynamic time warping (DTW) is first performed across dysarthric and normal phoneme feature vectors for each utterance, and then a transformation function is determined to correct dysarthric speech. In [3] and [4], the intelligibility of vowels in isolated words spoken by a dysarthric person is improved by formant re-synthesis of transformed formants, smoothened energy and synthetic pitch contours. In [5] and [6], dysarthric speech is improved by correcting pronunciation errors based on given transcriptions and by morphing the waveform in time and frequency. The authors report that the morphing doesn't increase intelligibility of the dysarthric speech. Some corrections are made by using an HMM-based speech recogniser followed by a concatenation algorithm and grafting technique to correct wrongly uttered units [7], or by synthesising speech using HMM-based adaptation [8]. In [9], poorly uttered phonemes are replaced by phonemes from normal speech with discontinuities in short term energy, pitch and formant contours at concatenation points addressed.

The work carried out in this paper focuses on continuous speech and also unstructured text. A durational analysis is carried out across dysarthric and normal speech. Though dysarthria is mostly characterised by slow speech, there are studies reporting rapid rate of speech [1], [10]. Based on the analysis for every dysarthric speaker, manual modifications are made directly to the speech waveforms. An automatic technique is proposed to achieve the same. The effect of these durational modifications on the perceptual quality of speech is studied.

Nemours database [11], a standard database for dysarthric speech, is used in the experiments. Additionally, a dysarthric speech dataset collected from an Indian speaker is also used. Unlike the text in Nemours database, the text in the Indian speech data does not conform to any particular structure. Analysis and modifications are made to speech data of different speakers in the Nemours database and the Indian English dysarthric dataset. Results of subjective evaluation, comparing modified and original dysarthric speech are then presented.

Additionally, two other systems are developed to improve the intelligibility of dysarthric speech. The first is a formant re-synthesis method based on an earlier work [4]. The second is an HMM-based text-to-speech (TTS) synthesis system adapted to the dysarthric person's voice [8]. We assume that a recognition system having 100% recognition accuracy is already available to transcribe speech for synthesis. A word error rate test is conducted to assess the intelligibility of the speech produced by these two systems along with the proposed automatic technique.

The rest of the paper is organised as follows. Section 2 describes the databases used in the experiments. The formant re-synthesis method is described in Section 3 followed by the HMM-based speech synthesis system using adaptation in Section 4. Durational analysis performed on the data along with the proposed modifications are detailed in Section 5. Evaluation results are presented in Section 6. The work is concluded in Section 7.

SLPAT 2016, 7th Workshop on Speech and Language Processing for Assistive Technologies, pages 43–49,
San Francisco, USA, 13 September, 2016.

2. Speech databases used

Standard databases available for dysarthric speech are Universal Access, TORGO and Nemours [11–13]. Universal Access database contains audiovisual isolated word recordings and is hence not suitable for our purpose. TORGO database consists of acoustic and articulatory data of non-words, short words, and complete sentences. However, complete sentences are fewer in number and they account for low phone coverage. Nemours database consists of 74 sentences for each dysarthric speaker. Experiments are therefore performed with Nemours database and Indian English dysarthric speech dataset [1]. The Indian English dysarthric speech data will be referred to as "IE" in this paper.

2.1. Nemours database

Nemours database [11] consists of dysarthric speech data of 11 male North American speakers. The degree of severity of dysarthria varies across speakers: mild (BB, FB, LL, MH), moderate (JF, RK, RL) and severe (BK, BV, SC). The speech data consists of 74 nonsense sentences for each speaker. The sentences follow the same format: "The X is Y'ing the Z", where X and Z are monosyllabic nouns and Y'ing is selected from a set of bisyllabic verbs. Along with the recording of each dysarthric speaker, the corresponding speech by a normal speaker is recorded. The normal speakers are appended with the prefix "JP". Transcriptions are available in terms of Arpabet labels [14].

Phone level segmentation is available for dysarthric speech while word level segmentation is available for normal speech. The procedure to obtain phone level segmentation for normal speech is described in the following section. Pauses within an utterance were already marked for speaker RK in the database but were not available for speakers BK, RL and SC. Hence for these three speakers, pauses were marked manually. Significant intra-utterance pauses are not present in the speech of other dysarthric speakers. For speaker KS, phonemic labeling is not provided. Hence, it is excluded from the experiments.

2.1.1. Segmentation of normal speech data at the phone level

Hidden Markov models (HMM) are used to segment normal speech data at the phone level. Word level boundaries and phone transcriptions for each word are available in the database. HMMs are used to model monophones in the data. Source and system parameters of speech are modeled by these HMMs. The source features are $logf0$ (pitch) values, along with their velocity and acceleration values. The system parameters are mel frequency cepstral coefficients (MFCC), along with their velocity and acceleration values. Instead of embedded training of HMM parameters at the sentence level, embedded re-estimation is restricted to the word boundary. This is inspired by [15], where phone level alignment is obtained from embedded training within syllable boundaries.

HMMs built using Carnegie Mellon University (CMU) corpus [16] were used as initial monophone HMMs instead of using the conventional flat start method to build HMMs, where the models were initialised with global mean and variance. This resulted in better phone boundaries. Data of American speaker referred to as "rms" in CMU corpus was used for this purpose.

[1]The Indian English dysarthric speech data can be found at the link: `www.iitm.ac.in/donlab/website_files/resources/IEDysarthria.zip`

2.2. Indian English dataset

2.2.1. Text selection

The text was chosen from CMU corpus [16]. 73 sentences were selected such that they ensured enough phone coverage. The phoneme transcriptions of the text were obtained from CMU pronunciation dictionary [17] and were later manually corrected when the word pronunciation varied. An additional label "pau" was added to account for pauses or silences.

2.2.2. Speech recording

The speech of an Indian male suffering from cerebral palsy, who is mildly dysarthric, was recorded. The speech was recorded in a low-noise environment and sampled at 16 kHz, with 16 significant bits. The recording was performed over several sessions, each session not exceeding half-an-hour. Frequent breaks were given during the sessions as per the convenience of the speaker so that fatigue didn't affect the quality of speech. About 11 minutes of speech data was collected. Frenchay dysarthria assessment (FDA) [18] was not performed due to unavailability of a speech pathologist.

2.2.3. Segmentation at the phone level

Before segmenting the dysarthric speech data, long silence regions (more than 100 ms) were removed from the speech waveforms by voice activity detection (VAD). 11 minutes of data then reduced to about 8.5 minutes. Segmentation was performed semi-automatically. HMMs were built from already available normal English speech data of an Indian (Malayalam) speaker "IEm" [19], as speaker IE is a native Malayalam speaker. These HMMs were used as initial HMMs to segment dysarthric speech data at the phoneme level. Segmentation was then manually inspected and corrected.

3. Formant re-synthesis technique

In reference [4], the intelligibility of dysarthric vowels in isolated words of CVC type (C- consonant, V- vowel) is improved. Borrowing from this work, a similar approach is adopted to improve intelligibility of continuous dysarthric speech in this paper. Formants F1-F4, pitch and short-term energy values are extracted from dysarthric and normal speech. Frame length of 25 ms and frame shift of 10 ms are considered. Formant transformation from dysarthric space to normal space is only carried out in the vowel regions. For this purpose, utterances segmented at the phone level are required. The transformation makes use of vowel boundaries and vowel identities. Then, formant values at the stable point of the vowels are determined [4]. The stable point (or region) is the vowel point (or region) that is least affected by context. A 4-dimensional feature vector represents each instance of a vowel- F1stable, F2stable, F3stable and vowel duration. In [4], formant transformation is achieved by training Gaussian mixture model (GMM) parameters using joint density estimation (JDE). This works well for data that

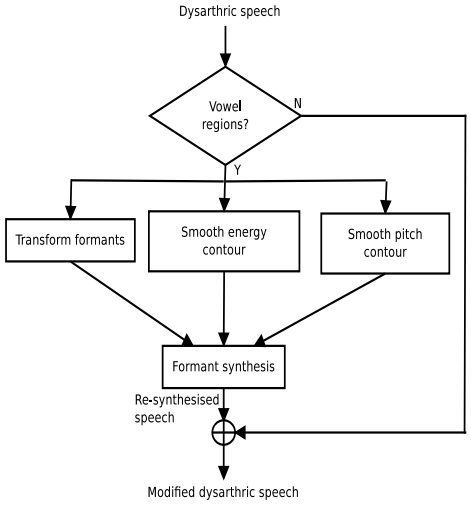


Figure 1: Formant re-synthesis of dysarthric speech

is phonetically balanced. The data used in the experiments in this paper suffers from data imbalance as the frequency of individual vowels in the database varies. To overcome this problem, a universal background model-GMM (UBM-GMM) [20] is trained and adapted to individual vowels of dysarthric and normal speech. Maximum a posteriori (MAP) is the adaptation algorithm used. The procedure to obtain adapted models is as follows:

- Each frame in a vowel region is represented by a 4-dimensional feature vector- formants F1-F3, and vowel duration. All the feature vectors, irrespective of stable points, are pooled together for all vowel instances across dysarthric and normal speech to train the UBM-GMM.
- The adaptation data for a vowel of dysarthric or normal speech is a 4-dimensional feature set (F1stable, F2stable, F3stable, vowel duration) across all instances of that vowel.
- A set of $(2 * number\_of\_vowels)$ models is obtained by adapting only the means of the UBM-GMM. This is a codebook of means for the same vowel across dysarthric and normal speech.

The dysarthric speech data is initially split into train (80%) and test data (20%). Normal speech corresponding to the dysarthric speech in Nemours database is used for obtaining the codebook. For the Indian dysarthric speech, speech of speaker "ksp" from CMU corpus [16] is used as normal speech. Adapted models are built using the train data. The codebook size of the UBM-GMM is 64. The procedure to re-synthesise dysarthric speech is shown in Figure 1. For test data, pitch ($F0$) and energy contours are smoothened. Smoothening is performed by using a median filter of order 3 and then low-passing using a Hanning window. This approach differs from the work carried out in reference [4], where a synthetic F0 contour is used for the dysarthric speech. Using the vowel boundaries, every vowel in the test utterance is represented by a 4-dimensional feature vector (stable F1-F3+vowel duration). Using the codebook of means for vowels across dysarthric and normal speech, the features of the dysarthric vowels are replaced by the means of their normal counterpart. The replaced or transformed stable point formants represent the entire vowel. Hence, the same stable point formant value is repeated across the duration of the vowel. Using the transformed formant contours, smoothened pitch and energy contours, speech is synthesised using a formant vocoder [21]. The modified dysarthric speech is then obtained by replacing non-vowel regions in the re-synthesised dysarthric speech by the original dysarthric speech.

4. HMM-based synthesiser using adaptation

An HMM-based TTS synthesiser (HTS) adapted to the dysarthric person's voice is developed [22], [8]. This is to evaluate the maximum intelligibility of synthesised speech that can be obtained given a recognition system for dysarthric speech that is 100% accurate. The purpose of using an HMM-based adaptive TTS synthesiser is two-fold: (1) not enough data to build a speaker-dependent system for every dysarthric speaker, and (2) to correct the pronunciation of the dysarthric speaker.

The HMM-based adaptive TTS can be divided into three phases- training, adaptation and synthesis. Audio files and corresponding transcriptions are available for training and adaptation data. In the training phase, mel-generalized cepstral (MGC) coefficients and $logf0$ values, along with their velocity and acceleration values are extracted from the audio files. Average voice models are then trained from speech features corresponding to the training data. In the adaptation phase, CSMAPLR+MAP adaptation (CSMAPLR- constrained structural maximum a posteriori linear regression) is performed to adapt the average voice models to the adaptation features. Speaker adaptive training (SAT) is performed to reduce the influence of speaker differences in the training data. In the synthesis phase, the test sentence is broken down into phones. Phone HMMs are chosen based on the context and concatenated to form the sentence HMM. MGC coefficients and $f0$ values are generated from the sentence HMM, and speech is synthesised using mel log spectrum approximation (MLSA) filter.

To build an adaptive TTS system for speakers in Nemours database, speech of two normal American male speakers, "bdl" and "rms" from the CMU corpus, is used as the training data. For the Indian English dysarthric data, the training data is speech of an Indian speaker "ksp" from the CMU corpus. 1 hour of speech data is available for every speaker in the CMU corpus. Dysarthric speech data is split into adaptation data (80%) and test data (20%). Synthesised speech of the sentences in the test data is used in the subjective evaluation. For developing the HMM-based adaptive TTS sythesiser, HTS version 2.3 software is used.

5. Proposed modifications to dysarthric speech

5.1. Durational analysis

A durational analysis across dysarthric and normal speech is performed. The following observations with respect to dysarthric speech are made:

- The average phone durations of dysarthric speech in the

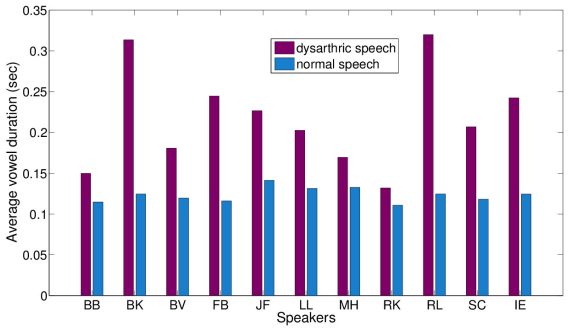


Figure 2: Average vowel durations across dysarthric and normal speakers

databases are longer than their normal speech counterparts [13, 23, 24]. As an example, average vowel durations are plotted in Figure 2.

- Standard deviations of vowel durations of dysarthric speakers are also longer (Figures 3 and 4), indicating that either the vowel is sustained for a longer duration or is hardly uttered.
- Speech data of the Indian dysarthric speaker IE is compared with the speech of different normal speakers. Four different nativities of Indian English (Hindi, Tamil, Telugu and Malayalam) in the Indic TTS corpus [19], and speech of an American speaker "rms" from CMU corpus [16] are the normal speech data considered. It is observed that the duration plot of speaker IE is clearly shifted with respect to that of normal speakers (Figure 5).
- For the same set of sentences spoken by dysarthric and normal speakers, the total utterance duration is longer for the dysarthric speaker. This indicates insertion of phones, intra-utterance pauses, etc. while speaking.

Based on the above analysis, if the duration is reduced closer to that of normal speech, the quality of dysarthric speech may improve. Reference [25] observes that as phone durations of dysarthric speech increase, the intelligibility of speech in terms of FDA score comes down. Taking this observation forward, in this paper, dysarthric speech is modified both manually and automatically to achieve this durational reduction.

5.2. Manual Modifications

The increase in phone duration is due to elongation of vowels, artifacts while producing sounds, or significant pauses within words. Hence, randomly increasing the speech rate of the utterance won't be useful, specific corrections are required. Each phoneme segment of the dysarthric speech is compared with its counterpart in normal speech. Elongations and artifacts are manually removed, keeping in mind not to degrade the intelligibility of speech. Steady regions of elongated vowels are spliced out. Segments are carefully deleted so as to not cause a sudden change in spectral content. For speaker IE, the recorded speech of Malayalam speaker "IEm" is considered as the reference. Original and corresponding manually modified waveforms are used in the subjective evaluation.

5.3. Proposed automatic method

A Dynamic Time Warping (DTW) algorithm is used to compare the similarity between MFCC features of dysarthric (test) and corresponding normal speech (reference). 39-dimensional MFCC features, including velocity and acceleration values are used. Wherever the slope of the DTW path is zero for a minimum number of frames, termed as *frameThres*, those frames are considered for deletion. When deleting frames, it is important to ensure that there is no sudden change in energy at the points of join, i.e., the energies between frames before and after deletion. It is observed that artifacts are introduced in places where the energy difference between frames at concatenation points is high. Therefore, the short-term energy (STE) difference is considered as an additional criterion for deletion. Whenever STE difference is less than a certain limit, *STEThres*, frames are deleted. In the experiments, *frameThres* and *STEThres* are set to 6 and 0.5 respectively. These thresholds are obtained empirically after testing with *frameThres* ranging from 4 to 10 and *STEThres* ranging from 0.3 to 2.5. This automatic procedure of deletion is illustrated in Figure 6.

The DTW paths of a sample utterance of dysarthric speaker RL before and after automatic modifications compared with respect to the same utterance of normal speaker JPRL is shown in Figure 7. It is observed that the DTW path is more diagonal in Figure 7b compared to Figure 7a, indicating that the modified dysarthric utterance is more similar to the normal utterance. It also results in a considerable reduction in number of frames or duration of the utterance. This method is referred to as the

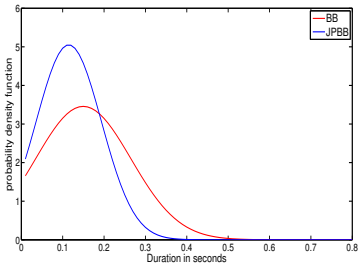


Figure 3: Duration plot for vowels of dysarthric speech BB and normal speech JPBB

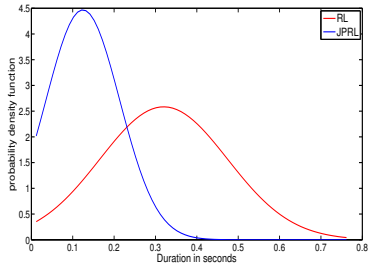


Figure 4: Duration plot for vowels of dysarthric speech RL and normal speech JPRL

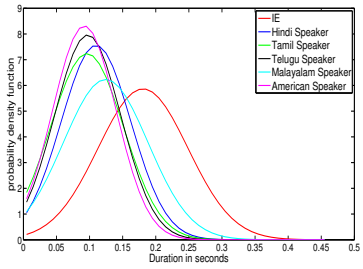


Figure 5: Duration plot for vowels of dysarthric speech IE and normal speech of other speakers

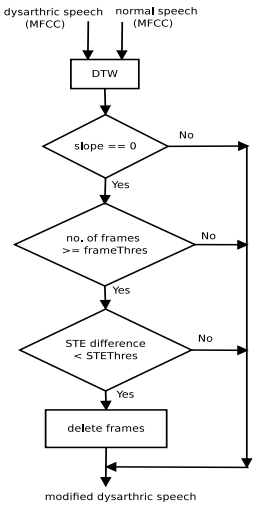


Figure 6: Flowchart of automatic (DTW+STE) method to modify dysarthric speech

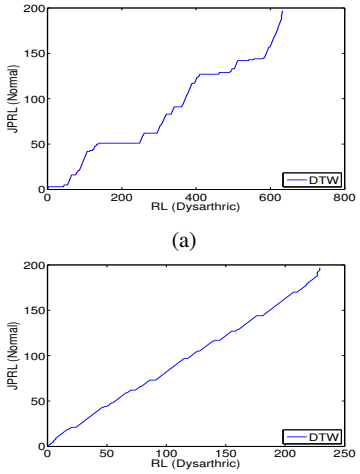


Figure 7: DTW paths of an utterance of speaker RL between: (a) original dysarthric speech and normal speech, and (b) modified dysarthric speech and normal speech

DTW+STE modification method.

6. Performance evaluation

Subjective evaluation is conducted to evaluate the techniques used. A pairwise comparison test is performed to assess the proposed modification techniques and a word error rate test to compare intelligibility across different methods. Naive listeners are used in the subjective tests rather than expert listeners in order to assess how a naive listener, who has little or no interaction with dysarthric speakers, evaluates the quality of dysarthric speech. Tests are conducted in a noise-free environment.

6.1. Pairwise comparison tests

A pairwise comparison test is conducted to compare the quality of speech modified by the proposed techniques and original dysarthric speech [26]. In the "A-B" test, A is played first and then B, and vice-versa in the "B-A" test to remove the bias in listening. "A" is the modified speech and "B" is the original speech in both the tests. Preference is always calculated in terms of the audio sample played first. The score "A-B+B-A" gives an overall preference for system A against system B and is calculated by the following formula:

$$\text{“}A-B+B-A'' = \frac{\text{“}A-B'' + (100 - \text{“}B-A'')}{2}$$

About 11 listeners evaluated a set of 8 sentences for each speaker. Results of the evaluation are shown in Figure 8. Results indicate a preference for the modified versions over original dysarthric speech in almost all cases. From Figure 8, it is evident that the manual method out-performs the DTW+STE (automatic) method. This is because manual modifications are hand-crafted carefully so as to produce better-sounding speech. For speakers BB and IE, who are mildly dysarthric, the performance of the DTW+STE method drops drastically due to artifacts introduced in the modified speech. This is true for speakers

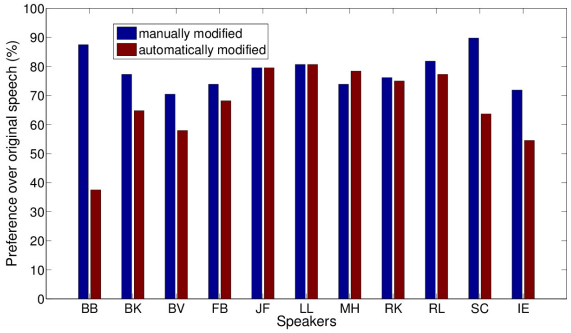


Figure 8: Preference for manually and automatically (DTW+STE) modified speech over original dysarthric speech of different speakers

BK and BV, where artifacts in the original speech are not eliminated by the DTW+STE technique. The drop in performance from the manual to the automatic technique is quite high for speaker SC because of the slurry nature of speech. Hence, in such cases identifying the specifics of dysarthria for individual speakers is vital to improving speech quality. Nonetheless, the performance of both methods is almost on par for speakers JF, LL, FB, MH, RL, RK who are mild to severely dysarthric.

Pairwise comparison tests were also conducted between original and formant re-synthesised speech, and between automatically modified and formant re-synthesised speech. About 10 listeners evaluated a set of 8 sentences for each speaker in each test. Preference was individually over 82% for the original dysarthric speech and DTW+STE method over the formant re-synthesis method.

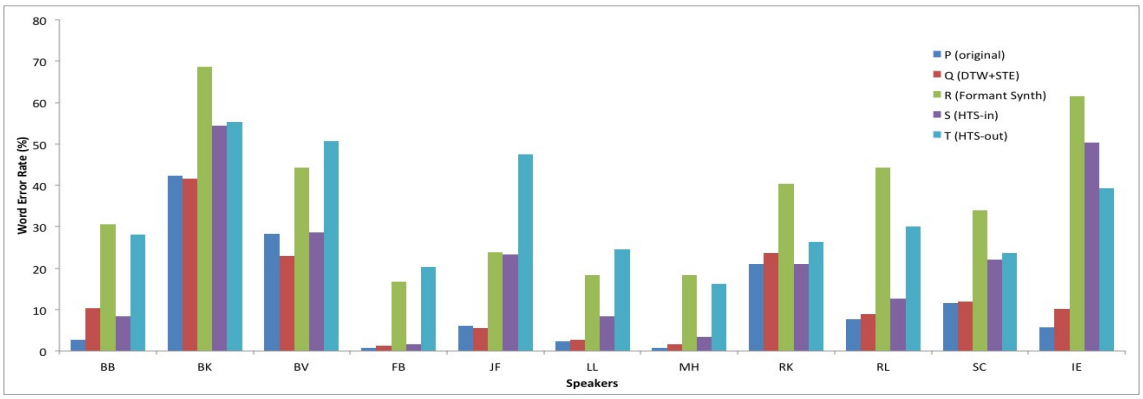


Figure 9: Word error rates for different types of speech across dysarthric speakers

6.2. Intelligibility tests

To evaluate intelligibility across different systems, a word error rate (WER) test was conducted. Based on the feedback on the pairwise comparison tests and the text in Nemours database containing nonsensical sentences, it is difficult to recognise words in dysarthric speech. Hence, given the text, listeners were asked to enter the number of words that was totally unintelligible. Though the knowledge of the pronounced word may have an influence on its recognition, this is a uniform bias that is present when evaluating all systems. About 10 listeners participated in the evaluation. The following types of speech were used in the listening tests:

P (original): original dysarthric speech
Q (DTW+STE): dysarthric speech modified using the DTW+STE method
R (Formant Synth): output speech of the formant re-synthesis technique
S (HTS-in): speech synthesised using the HMM-based adapted TTS for text in the database not used for training (held-out sentences)
T (HTS-out): speech synthesised using the HMM-based adapted TTS for text from the web

The results of the WER test are presented in Figure 9. It can be seen that the intelligibility of formant re-synthesis technique is poor for all speakers. For the DTW+STE method, WER is higher compared to original dysarthric speech for a majority of speakers. WER of HMM-based adaptive synthesiser on held out-sentences, i.e., sentences not used during training is high compared to original dysarthric speech in almost all cases. Intelligibility of sentences synthesised from the web is quite poor compared to that of held-out sentences for speakers in Nemours database. This is the opposite for Indian dysarthric speaker IE. This is due to the similar structure of held-out sentences and sentences used in training the HMM-based synthesiser in Nemours database, unlike the sentences in the Indian dysarthric dataset that are unstructured. Overall, the intelligibility of original dysarthric speech does not increase. However, for speakers BK, BV and JF, DTW+STE modified speech has the lowest WER. For speaker RK, the intelligibility of HMM-based adaptive synthesised speech is on par with that of original dysarthric speech. By informal listening, it is noted that some pronunciations of the dysarthric speaker do get corrected in the sentences synthesised using the HMM-based adaptive TTS system. This indicates that the technique used to increase intelligibility largely depends on the type and severity of dysarthria.

While the DTW+STE does not need segmented boundaries, it makes use of a reference for comparison. Only insertion of sounds are taken care of, deletion and substitution of phonemes are not addressed. Though this technique does not increase intelligibility for most speakers, the overall perceptual quality of the modified dysarthric speech is improved.

In the speech synthesis domain, the HMM-based adaptive synthesiser is a statistical parametric speech synthesiser (SPSS) and the DTW+STE technique is analogous to a unit selection speech (USS) synthesiser. The synthesised speech of the HMM-based synthesiser lacks the voice quality of the dysarthric speaker. Similar to the USS system, the speech output of the DTW+STE method has discontinuities but preserves the voice characteristics of the dysarthric speaker.

7. Conclusions

Continuous dysarthric speech quality is improved upon in the work. A durational analysis is performed by comparing dysarthric and normal speech for speakers in Nemours database and an Indian English speaker having dysarthria. Based on the analysis, dysarthric speech is directly modified manually, and an automatic method is developed to do the same. The intelligibility of dysarthric speech modified using different techniques is studied. Evaluations indicate an improvement in speech quality using the STE+DTW method. This emphasises the importance of duration in perceptual speech quality, indicating that this kind of modification may be used as a pre-processing step for improving dysarthric speech quality. Only durational attributes are analysed in this work, this can be extended to analyse other attributes that affect the speech of a dysarthric person.

8. Acknowledgements

The authors would like to thank Rajiv Rajan, Kalpana Rao and Namita Jacob of Vidyasagar, Chennai, with whose help, collecting dysarthric speech data was possible. The authors would also like to thank Jom, Shreya and other colleagues in the lab for their inputs and assistance in conducting the evaluations.

9. References

[1] American Speech Language Hearing Association, "Dysarthria," http://www.asha.org/public/speech/disorders/dysarthria/.

[2] J. P. Hosom, A. B. Kain, T. Mishra, J. P. H. van Santen, M. Fried-Oken, and J. Staehely, "Intelligibility of modifications to dysarthric speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, April 2003, pp. 924 – 927.

[3] A. Kain, X. Niu, J. Hosom, Q. Miao, and J. P. H. van Santen, "Formant re-synthesis of dysarthric speech," in *Fifth ISCA ITRW on Speech Synthesis*, June 2004, pp. 25–30.

[4] A. B. Kain, J.-P. Hosom, X. Niu, J. P. van Santen, M. Fried-Oken, and J. Staehely, "Improving the intelligibility of dysarthric speech," *Speech Communication*, vol. 49, no. 9, pp. 743 – 759, 2007.

[5] F. Rudzicz, "Acoustic transformations to improve the intelligibility of dysarthric speech," in *2nd Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, July 2011, p. 11 21.

[6] ——, "Adjusting dysarthric speech signals to be more intelligible," *Computer Speech & Language*, vol. 27, no. 6, pp. 1163–1177, 2013.

[7] M. S. Yakcoub, S. A. Selouani, and D. O'Shaughnessy, "Speech assistive technology to improve the interaction of dysarthric speakers with machines," in *3rd International Symposium on Communications, Control and Signal Processing (ISCCSP)*, March 2008, pp. 1150–1154.

[8] M. Dhanalakshmi and P. Vijayalakshmi, "Intelligibility modification of dysarthric speech using HMM-based adaptive synthesis system," in *2nd International Conference on Biomedical Engineering (ICoBE)*, March 2015, pp. 1–5.

[9] M. Saranya, P. Vijayalakshmi, and N. Thangavelu, "Improving the intelligibility of dysarthric speech by modifying system parameters, retaining speaker's identity," in *International Conference on Recent Trends In Information Technology (ICRTIT)*, April 2012, pp. 60–65.

[10] G. L. Dorze, L. Ouellet, and J. Ryalls, "Intonation and speech rate in dysarthric speech," *Journal of Communication Disorders*, vol. 27, no. 1, pp. 1 – 18, 1994.

[11] X. Menendez-Pidal, J. Polikoff, S. Peters, J. Leonzio, and H. Bunnell, "The Nemours database of dysarthric speech," in *Fourth International Conference on Spoken Language (ICSLP)*, vol. 3, October 1996, pp. 1962–1965.

[12] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. S. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research." in *Annual Conference of the International Speech Communication Association, INTERSPEECH*, September 2008, pp. 1741–1744.

[13] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The TORGO database of acoustic and articulatory speech from speakers with dysarthria." *Language Resources and Evaluation*, vol. 46, no. 4, pp. 523–541, 2012.

[14] Wikipedia, "Arpabet," https://en.wikipedia.org/wiki/Arpabet.

[15] S. A. Shanmugam and H. Murthy, "A hybrid approach to segmentation of speech using group delay processing and HMM based embedded reestimation," in *Annual Conference of the International Speech Communication Association, INTERSPEECH*, Singapore, September 2014, pp. 1648–1652.

[16] J. Kominek and A. W. Black, "The CMU arctic speech databases," in *5th ISCA Speech Synthesis Workshop*, June 2004, pp. 223–224.

[17] Carnegie Mellon University, "The CMU pronounciation dictionary," http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[18] P. Enderby, "Frenchay Dysarthria Assessment," *International Journal of Language & Communication Disorders*, vol. 15, no. 3, pp. 165–173, December 2010.

[19] TTS Consortium, DeitY, Government of India, "Indic TTS," https://www.iitm.ac.in/donlab/tts/.

[20] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, January 2000.

[21] L. Rabiner and R. Schafer, *Theory and Applications of Digital Speech Processing*, 1st ed. Upper Saddle River, NJ, USA: Prentice Hall Press, 2010.

[22] J. Yamagishi, T. Nose, H. Zen, Z. H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, "Robust speaker-adaptive HMM-based text-to-speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1208–1230, August 2009.

[23] V. Surabhi, P. Vijayalakshmi, T. S. Lily, and R. V. Jayanthan, "Assessment of laryngeal dysfunctions of dysarthric speakers," in *IEEE Engineering in Medicine and Biology Society*, Minnesota, September 2009, p. 2908 2911.

[24] H. Ackermann and I. Hertrich, "Speech rate and rhythm in cerebellar dysarthria: An acoustic analysis of syllabic timing," *Folia Phoniatrica et Logopaedica*, vol. 46, no. 2, pp. 70–78, 1994.

[25] P. Vijayalakshmi and M. Reddy, "Assessment of dysarthric speech and analysis on velopharyngeal incompetence," in *28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*, August 2006, pp. 3759–3762.

[26] P. Salza, E. Foti, L. Nebbia, and M. Oreglia, "MOS and pair comparison combined methods for quality evaluation of text to speech systems," in *Acta Acustica*, vol. 82, 1996, pp. 650–656.

PAoS Markers: Trajectory Analysis of Selective Phonological Posteriors for Assessment of Progressive Apraxia of Speech

Afsaneh Asaei[1]*, Milos Cernak*[1]*, Marina Laganaro*[2]

[1]Idiap Research Institute, Martigny, Switzerland
[2]University of Geneva, Switzerland
{aasaei,mcernak}@idiap.ch, marina.laganaro@unige.ch

Abstract

Progressive apraxia of Speech (PAoS) is a progressive motor speech disorder associated with neurodegenerative disease causing impairment of phonetic encoding and motor speech planning. Clinical observation and acoustic studies show that duration analysis provides reliable cues for diagnosis of the disease progression and severity of articulatory disruption. The goal of this paper is to develop computational methods for objective evaluation of duration and trajectory of speech articulation. We use phonological posteriors as speech features. Phonological posteriors consist of probabilities of phonological classes estimated for every short segment of the speech signal.

PAoS encompasses lengthening of duration which is more pronounced in vowels [1, 2]; we thus hypothesize that a small subset of phonological classes provide stronger evidence for duration and trajectory analysis. These classes are determined through analysis of linear prediction coefficients (LPC). To enable trajectory analysis without phonetic alignment, we exploit phonological structures defined through quantization of phonological posteriors. Duration and trajectory analysis are conducted on blocks of multiple consecutive segments possessing similar phonological structures. Moreover, unique phonological structures are identified for every severity condition.

Index Terms: Progressive apraxia of speech (PAoS), Phonological posterior features, Phonological structures, Linear prediction coefficient (LPC).

1. Introduction

Dysarthria and Progressive Apraxia of Speech (PAoS) are two common speech motor disorders observed in neurodegenerative diseases. While automatic processing (for assessment and assistive applications) of dysarthric speech is getting considerable attention in the speech community [3, 4, 5, 6], acoustic and automatic processing studies of PAoS are rather rare. It might be due to increased complexity of speech degradations of patients with PAoS, where production errors are more inconsistent and unpredictable [7].

PAoS is a speech motor disorder associated to several neuropathological conditions, which causes progressive degradation of the main speech characteristic and of speech intelligibility. The main symptoms of PAoS are phonetic distortions and phonemic errors, groping and effortful speech initiation with successive approximations, changes in inter- and intra-syllabic transitions, increased syllabic duration and decreased speech rate [6, 7].

PAoS has been associated with impaired phonetic encoding (planning of speech gestures) rather than to impaired motor execution [7, 2, 8]. Here we hypothesize that analysis of phonological features extracted from the degraded speech signal could contain clues for assessment of the progressive disruption and severity levels of PAoS.

One particular contribution of this paper is selection of phonological classes using linear prediction analysis of phonological posteriors. In addition, we exploit phonological structures [9] to enable automatic analysis of duration and trajectory without any need for automatic alignment. Prior work on phonological structures demonstrate their relation to articulatory postures [9], thus considering the structure of multiple consecutive segments enables quantification of the dynamic and trajectory of articulatory movements and co-articulation. The studies presented in this paper exploit this structural property of phonological posteriors to obtain speech-based markers of PAoS severity.

The results obtained in Section 5 demonstrate a significant increase in duration and less consistency in articulatory movements as the neuro-degeneration thrives. Furthermore, we identify unique structures per severity condition which indicates that certain articulatory postures disappear in AoS progression and are replaced by new postures and trajectory of movements. This observation can lead to development of a novel automatic assessment method relying on the nearest neighbor rule of classification. Preliminary studies show the potential of this method. However, recordings of many patients is required to validate/endorse its usefulness for clinical applications.

To the best of authors' knowledge, prior work on application of phonological features for objective intelligibility prediction of pathological speech considered statistical measures of independent processing of segments of speech [10]. In contrast, we propose ranking ad selection of phonological classes, and we study the relation between adjacent segments, or trajectory of articulation. The proposed approach provides simple objective tools that correlate with higher level speech production behaviors such as speaking rate and co-articulation without any requirement for speech alignment.

The rest of the paper is organized as follows. Section 2 describes the data available from 3 assessment sessions of a patient diagnosed with isolated PAoS. This data is used for estimation of phonological posteriors through the procedure explained in Section 3. We used linear prediction analysis to rank the phonological classes for trajectory analysis in Section 4. The trajectory analysis methods are explained in Section 5 where selection of phonological classes is found an effective approach to obtain more distinct markers of severity. Moreover, distinct structural patterns are observed for every severity condition. This observation leads to devising a classifier for detection of the level of severity which is elaborated in Section 6. The conclusions are drawn in Section 7.

SLPAT 2016, 7th Workshop on Speech and Language Processing for Assistive Technologies, pages 50–55,
San Francisco, USA, 13 September, 2016.

2. PAoS Data

The data used for evaluation of the methods proposed in this paper consist in 3 recordings over A 28-months period of a 67 year-old french speaking woman diagnosed with isolated PAoS. The patient has been recorded for about 2 minutes while reading the same text ("La bise et le soleil" [11]). The total duration is thus about 7 minutes. Across the 3 sessions the severity of speech disruption progresses from mild, to medium and to severe impairment according to clinical assessment by speech and language therapists and to normative acoustic data. Diadochokinetic rate assessed with standard diadochokinetic tasks [12] and articulation rate are reported in Table 1.

Table 1: *Clinical PAoS pattern: speech rate and diadochokinesic rate (syll/sec) across the assessment sessions. The numbers in parenthesis shows the relative reduction in rates with respect to the mild condition. The patient's production impairment in medium condition is assessed after 16 month from the mild condition; the severe impairment is evaluated after 12 month from the medium condition.*

Condition	**Mild**	**Medium**	**Severe**
Articulation rate	2.73	2.39 (13%)	2.06 (25%)
Diadochokinesis rate	2.85	2.22 (22%)	1.58 (45%)

The clinical and acoustic durational measurements show that this patient after 16 months from its initial mild AoS, exhibits increased impairment where the articulation rate is decreased by 13% and diadochokinetic rate is decreased by 22%. In the follow up assessment session after 28 months from the diagnosis of mild AoS, the patient reaches more severe impairment manifested in 25% reduction in articulation rate and 45% reduction in diadochokinetic rate.

In the rest of the paper, our goal is to quantify speech markers that correlate with the clinical markers. Motivated from the intuitive effects of PAoS on articulatory disruptions, and how the clinical assessments quantify this impairment, we focus our work on ranking and selection of speech representations, and their evolution through time in trajectory analysis.

3. Phonological Structures

We use deep neural network (DNNs) to estimate the phonological posterior features. As we have already seen in Section 1, PAoS affects phonetic planning. Hence, phonological posteriors are suitable representation of speech to enable assessment of these patients. Moreover, phonological posteriors exhibit highly constrained structures that are consistent for adjacent segments and change according to the speaking rate. In the next Section 3.1, we explain the framework for estimation of phonological posteriors.

3.1. Phonological Posteriors

Figure 1 illustrates the process of the phonological analysis [13, 14]. This process starts by converting a segment of speech samples into a sequence of acoustic features $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \ldots, \boldsymbol{x}_N\}$ where N denotes the number of segments in the utterance. Conventional cepstral coefficients can be used as acoustic features. Then, a bank of phonological class analyzers realized via neural network classifiers converts the acoustic feature observation sequence X into a sequence of phonological posterior probabilities $Z = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n, \ldots, \boldsymbol{z}_N\}$; a posterior probability

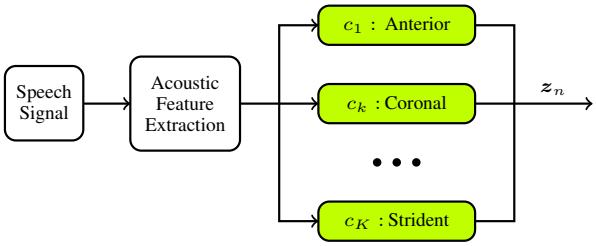


Figure 1: *The process of phonological analysis. Each segment of speech signal is represented by phonological posterior probabilities $\boldsymbol{z}_n$ that consist of K class-conditional posterior probabilities. For each phonological class, a DNN is trained to estimate its posterior probability given the input acoustic features.*

$$\boldsymbol{z}_n = [p(c_1|x_n), \ldots, p(c_k|x_n), \ldots, p(c_K|x_n)]^\top \quad (1)$$

consists of K phonological class-conditional posterior probabilities where c_k denotes the phonological class and $.^\top$ stands for the transpose operator. The phonological posteriors Z yield a parametric speech representation.

Phonological analysis was performed with the PhonVoc: phonetic and phonological toolkit [16]. Probabilities of $K = 24$ phonological classes corresponding to the French version of the Sound Pattern of English [14] were extracted from 25 ms speech segments, using 10 ms steps [16].

3.2. Structured Sparsity

Phonological posteriors are indicators of the physiological posture of human articulation machinery. Due to the physical constraints, only few combinations can be realized in our vocalization. This physical limitation leads to a small number of unique patterns exhibited over the entire speech corpora [17]. We refer to this structure as *first-order structure* which is exhibited at *segmental* level. These structures can be quantified using binary (1-bit) quantization or finer quantization levels. We will compare both binary (Q1) and 2-bit (Q2) quantization levels to perform trajectory analysis in Section 5. We will see that binary structures are the best level of structural definition.

In addition to the first-order structures, the dynamic of the phonological posteriors can be quantified considering the higher-order structure underlying a sequence (trajectory) of phonological posteriors. This structure is exhibited at *supra-segmental* level which is associated to the syllabic information or more abstract linguistic attributes. We refer to this structure as *high-order structure*.

Previously we have shown that the trajectories of the articulatory-bound phonological posteriors correspond to the distal representation of the gestures in the gestural model of speech production (and perception) [9]. In this paper, we exploit these structures as markers for objective evaluation and assessment of the level of severity of speech motor disorder in patients diagnosed with progressive apraxia of speech. The details of our analysis are explained in Sections 5.

Unlike previous work on application of phonological posteriors for assessment of pathological speech, we hypothesize that not all phonological classes are equally important. In other words, a small subset of phonological classes may provide stronger cues for diagnosis of the level of PAoS. This hypothesis is supported by clinical investigations confined to the

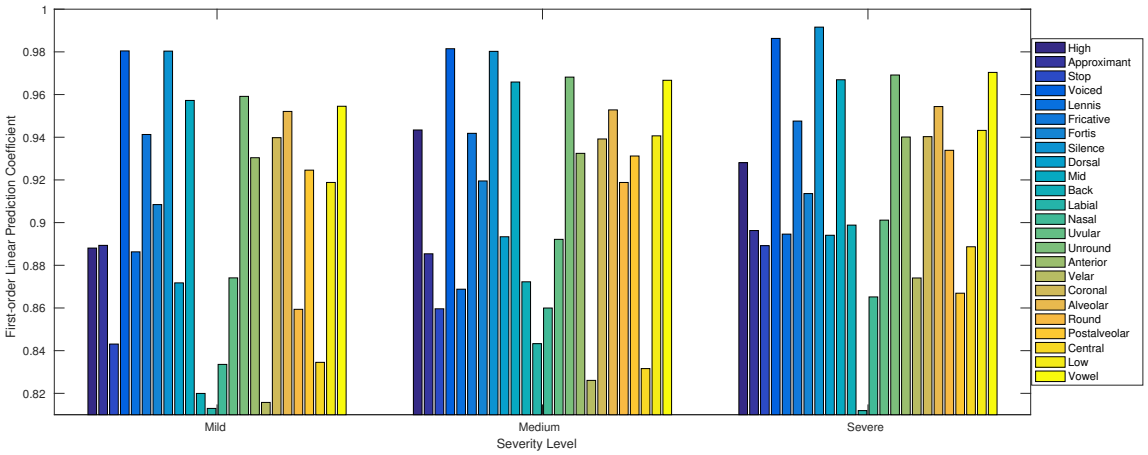


Figure 2: *Ranking and selection of important phonological classes: The first coefficient of the linear prediction analysis [15] is computed for the blocks of posteriors consisted on adjacent segments with similar binary structures. The average of the first LPC for each phonological classes is calculated per recording session. Finally, the mean of class-specific LPCs are computed for all the session and sorted. The classes with the highest mean values are considered as those contributing the most to the assessment of speech rate and trajectory in PAoS. The top-4 classes are "Voiced", "Vowel", "Unround" and "Mid". Considering the phonological-phonetic mapping used for the neural network training [14], these classes capture phonological variation of vowel-like sounds (c.f. Table 2).*

vowels and in particular lengthening of the vowels as the PAoS thrives [1, 2]. In the next section, we elaborate on a method for ranking and selection of phonological classes using linear prediction analysis [15].

4. Selection of Phonological Classes

We hypothesize that phonological classes are not equally important for assessment of PAoS. The focus of the present work is on trajectory analysis of phonological posteriors; hence, we rely on linear prediction coefficients (LPC) to measure the dependency and predictability of consecutive phonological posteriors.

4.1. Linear Prediction Analysis

The goal of linear prediction analysis is to minimize prediction error of the current segment using the values of the posteriors from the past consecutive segments. The predicted posterior at segment n is thus obtained as

$$\hat{\boldsymbol{z}}_n = \sum_{p=1}^{P} \boldsymbol{\alpha}_p \odot \boldsymbol{z}_{n-p} \tag{2}$$

where $\boldsymbol{\alpha}_p = \left[\alpha_p^1 \dots \alpha_p^k\right]^\top$ is a K dimensional vector, and $\odot$ stands for element-wise product. The LPCs are estimated to minimize the reconstruction error in mean square sense, i.e. $||\hat{\boldsymbol{z}}_n - \boldsymbol{z}_n||_2$.

The procedure for LPC analysis of phonological posteriors and selection of the most important classes is as follows:

1. *Blocking:* The posteriors are analyzed in blocks of consecutive segments which possess similar structures after quantization.
2. *Class-specific LPC:* The high-order LPC analysis is performed where the LPC order is chosen as the block length, i.e. number of segments.
3. *Ranking:* Means of the first LPC coefficient α_p^1 for all blocks and recording conditions are computed for every phonological class. The means are then sorted and the classes which exhibit largest means are considered as the most informative classes for trajectory analysis.

Figure 2 illustrates the average value of the first LPCs for each phonological posterior per recording sessions corresponding to mild, medium and severe condition. The top-4 most important classes are identified as "Voiced", "Vowel", "Unround" and "Mid". We can see the consistent high value of LPCs estimated for these classes throughout progression of AoS.

4.2. Phonological-Phonetic Importance of Vowels

To have a better understanding of what the selection of phonological classes may imply, we refer to the phonological-phonetic mapping used for neural network training in posterior estimation [14]. Table 2 shows the mapping between selected classes and phonemes associated with these classes.

Table 2: *Association of selected French phonological classes and phonemes. The French phoneme set is taken from BDLex [18].*

Class c_k	Phonemes
Voiced	a ɑ̃ ə i y u e ɛ̃ ø o õ ɔ ə ɛ œ œ̃ j l m ɥ w b ŋ ɲ n v ɡ ʁ d z ʒ
Vowel	i y u e ɛ̃ ø o õ ə ɛ œ ɔ a ɑ̃ œ̃ ə
Unround	a ɑ̃ i e ɛ̃ ɛ ə
Mid	ø ɛ̃ e œ ɛ

We can see that the selected top-4 classes capture phonological variability of all vowel-like speech sounds (all vowels and voiced consonants). This might indicate that vowel analysis is more important in PAoS than the consonant analysis. This observation is inline with the clinical assessment of PAoS [8].

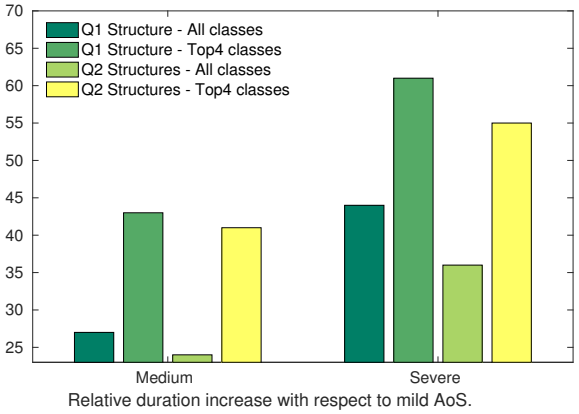


Figure 3: *Relative increase in duration illustrated for (i) Q1 and Q2 quantizations to obtain phonological structures, and (ii) when phonological structures are obtained from all (24) classes vs. the case that only top-4 classes are used for duration modeling. We see clear benefit when Q1 binary structures are used for duration analysis, and when we use the selective top-4 classes. In the most distinctive case, using Q1 structure of top-4 classes, we observe 43% increase in duration in medium AoS and 61% increase in duration in severe AoS with respect to mild AoS (c.f. Table 3).*

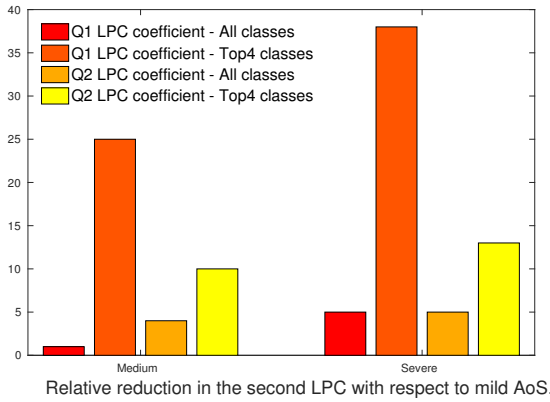


Figure 4: *Relative reduction in the second coefficient of high-order LPC analysis for (i) Q1 and Q2 quantizations to obtain phonological structures, and (ii) when phonological structures are obtained from all 24 classes vs. the case that only top-4 classes are used for duration modeling. In the most distinctive case, using Q1 structure of top-4 classes, we observe 25% reduction in the second LPC in medium AoS and 38% reduction in the second LPC in severe AoS with respect to mild AoS (c.f. Table 4).*

5. Trajectory Analysis

Analysis of trajectory of phonological posteriors is performed using three metrics defined through (5.1) Duration of phonological structures, (5.2) Predictability of phonological classes from the previous segments, and (5.3) Dynamic of posteriors quantified through high-order structures underlying consecutive segments.

Clinical assessment for diagnosis of PAoS asserts that speaking rate is reduced and the control over muscle movements is less consistent as the disease thrives [2, 8]. Accordingly, we expect to see an increase in duration and less predictability from mild to severe condition. Complying to the clinical emphasis put on vowel analysis, focusing our analysis on top-4 classes (c.f Table 2) is expected to be advantageous in distinction of the severity conditions.

5.1. Structural Duration

Typically, duration analysis requires automatic alignment of speech with the actual transcription using automatic speech recognition (ASR). However, this can be a cumbersome method that requires ASR resources and expertise. Moreover, automatic alignment is affected by the ASR errors due to *progressive* mismatch between the training and testing conditions.

To alleviate this limitation, we propose to use the phonological structures for duration analysis. The structures are obtained through quantization of posteriors, and they are often similar for adjacent segments. As the phonological structures can be related to the articulatory postures of speech production [9], slower speaking rate indicates a slower dynamic in the structural changes.

Applying the same blocking procedure as explained in Section 4, we quantify the *structural duration* as the average number of the segments in one block.

Different level of quantization can be applied to obtain the structures. We compare 1-bit (Q1) and 2-bit (Q2) quantization in our studies. Furthermore, to see the effectiveness of working on a small subset of phonological classes, we compare the results when the structures are obtained from all classes or only from the top-4 most important classes.

The results are illustrated in Figure 3. More details are listed in Table 3. We can see a clear benefit of binary structures over the 2-bit quantization. Moreover, obtaining the structural duration using a subset of most indicative phonological classes leads to higher distinction across different PAoS conditions.

Table 3: *Structural duration measured in terms of the average number of segments in all blocks of similar structures. The numbers in parenthesis show the relative increase in duration with respect to the mild condition.*

Condition	**Mild**	**Medium**	**Severe**
Q1 duration (all)	2.3	3.1 (27%)	3.6 (44%)
Q1 duration (top-4)	4.7	6.7 (43%)	7.6 (61%)
Q2 duration (all)	1.5	1.97 (24%)	2.2 (36%)
Q2 duration (top-4)	2.3	3.3 (41%)	3.6 (55%)

5.2. Long Term Dependency

Similar to the method explained in Section 4, the linear prediction analysis is conducted on blocks of the same phonological structures. We perform high-order LPC analysis where the order is determined from the length of the block, i.e. number of segments. We measure the mean of the second LPC for every recording session. Less control over the muscle movement leads to less consistency of the articulation trajectories. Figure 4 illustrates the relative reduction of the second LPC with respect to the mild condition. The details of the results are listed in Table 4.

It is evident that high-order dependencies are reduced. This

Table 4: *High-order linear prediction analysis: The values for the second coefficient averaged for all segments are listed at different recording sessions. The numbers in parenthesis show the relative decrease in high-order dependency with respect to the mild condition.*

Condition	**Mild**	**Medium**	**Severe**
Q1 LPC (all)	0.53	0.51 (1%)	0.48 (5%)
Q1 LPC (top-4)	0.22	0.17 (25%)	0.14 (38%)
Q2 LPC (all)	0.78	0.74 (4%)	0.75 (5%)
Q2 LPC (top-4)	0.61	0.54 (10%)	0.52 (13%)

effect is much more pronounced if we only consider top-4 most important phonological classes (c.f. Table 2). Similar studies on other LPCs larger than the second coefficient shows that those values are very small and their changes are less distinctive for PAoS objective evaluation.

5.3. High-order Structures

Relying on the relation between phonological structures and articulatory postures, the dynamic/trajectory of articulation or co-articulation can be quantified considering high-order structures [9]. To that end, we append consecutive phonological posteriors to define the trajectories through quantization of augmented posteriors. The number of consecutive posteriors determine the level (order) of trajectory structures. More specifically, C adjacent posterior vectors are appended to define a new posterior which encode C-order dynamic of features as

$$\boldsymbol{z}_n^C = \left[\boldsymbol{z}_n^\top \dots \boldsymbol{z}_{n+C}^\top\right]^\top. \tag{3}$$

As the phonetic planning is disrupted in PAoS [8], we expect to see unique structures for distinct levels of severity. More intuitively, certain articulatory postures can only occur at an specific level of neurodegeneration.

The percentage of unique structures (number of unique structures / number of all segments) is listed in Table 5. Furthermore, we quantify the percentage of structures that only occur in *one* severity condition.

Table 5: *Ratio of unique structures (%) per condition. The numbers in parenthesis show the ratio of the structures that only occur in one particular condition, thus indicative of particular articulatory posture that may only occur at a specific level of impairment.*

Condition	**Mild**	**Medium**	**Severe**
Q1 structures 1-order	7.1 (32)	5.9 (29)	4 (36)
Q1 structures 2-order	28 (60)	21 (59)	17 (63)
Q1 Structures 3-order	44 (76)	34 (74)	28 (74)
Q2 structures 1-order	62 (90)	47 (88)	42 (90)

We can see that the number of distinct structures grow rapidly as the order is increased. This demonstrates that the trajectories of phonological posteriors exhibit more distinct properties as we consider larger context (C) or finer structures (Q2).

Nevertheless, even at a segment level (no augmentation, $C = 1$), we can see a significant number of structures that only occur in one specific severity condition: nearly 30% of structures are unique. This observation on structural differences motivates us to perform automatic assessment using nearest neighbor rule of classification. The procedure is explained in the following Section 6.

6. Preliminary Automatic Assessment

To visualize the structural differences between phonological posteriors across recording sessions, we used the t-distributed stochastic neighbor embedding (tSNE) method [19] for visualization of high-dimensional (posterior) features. Figure 5.2 illustrates the results. Phonological posteriors without augmentation ($C = 1$ in (3)) are used for visualization. We contrast the visualization of posteriors where all 24 classes are used vs. only top-4 selective classes are considered. We can see that the distinction in posterior distribution is well preserved.

Exploiting the structural differences enables us to perform automatic assessment via classification. We consider the nearest neighbor classification rule for this purpose. To that end, we divide the data in two training and testing splits. Each testing segment is independently labeled based on the label of its nearest neighbor phonological posterior in the training set.

We use segmental posteriors without augmentation, i.e. $C = 1$. The *cosine* similarity metric is used to find the nearest neighboring vector which is defined as one minus cosine of the angle between two vectors; mathematically, it is expressed as

$$S_{\text{cosine}}(\boldsymbol{z}_1, \boldsymbol{z}_2) = 1 - \frac{\sum_k z_1^k z_2^k}{\sqrt{\sum_k \left(z_1^k\right)^2 \sum_k \left(z_2^k\right)^2}}. \tag{4}$$

where z_1^k denotes the k^{th} element of posterior vector $\boldsymbol{z}_1$. This metric has been found a suitable choice when comparing the similarity of two posteriors vectors [20].

The segment-level labels are then pulled to make a decision about the severity of articulatory disruption based on majority voting. We observe that exploiting about 5 seconds for training and testing data, enables us to perfectly classify the session of recording. The training size in all severity conditions is equal.

This suggests that inter-patient PAoS severity might be automatically assessed using nearest neighbor classifier and the phonological posteriors as speech features. Of course, this can not be validated or endorsed clinically unless a sufficiently large number of patients are recorded and used for exhaustive evaluation of this method.

7. Conclusions

Trajectory analysis of phonological posteriors enables objective assessment of progressive apraxia of speech. We demonstrated that a selected set of phonological classes can be considered as strong indicators of PAoS. In this paper, we performed linear prediction analysis to select the most important classes for trajectory analysis. Interestingly, these classes highly correlate with the clinical observation on importance of vowels in PAoS diagnosis.

To enable trajectory analysis without any need for automatic alignment of speech, we build on our previous work on phonological structures obtained through quantization as a method for quantifying the articulatory postures [9]. Our investigations on structural duration shows a significant increase in duration if we consider mainly the top-4 important classes. This observation is inline with the clinical evidence of speech rate reduction more pronounced in production and lengthening of vowels. This has been also verified in scientific studies using acoustic analyses.

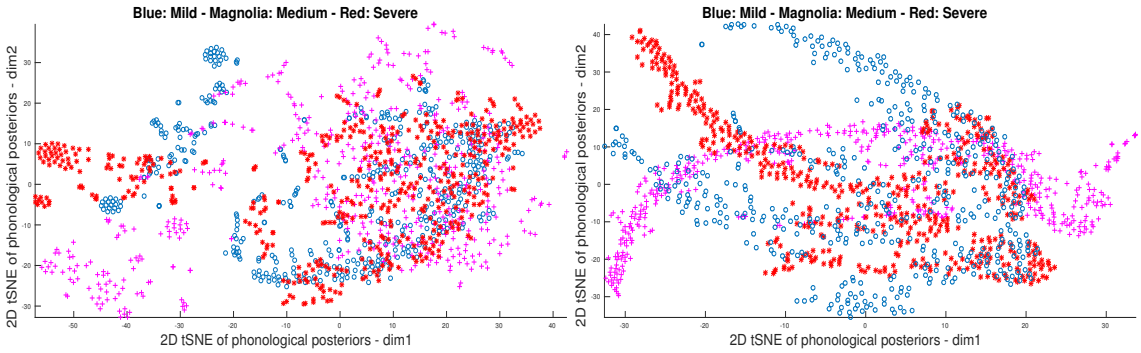


Figure 5: *tSNE visualization of (left) all phonological posteriors and (right) selected top-4 classes (c.f. Table 2).*

Furthermore, high-order LPC analysis demonstrate a significant decrease in consistency of phonological trajectories. The dynamic of phonological posteriors can be quantified by appending multiple adjacent phonological posteriors to form a super-vector of multiple phonetic classes. This method enables us to quantify the transitions or co-articulation among articulatory postures. The studies presented in this paper confirmed that unique phonological structures are exhibited for every severity condition. Exploiting this property can potentially enable us to perform automatic assessment of the level of severity in PAoS. Preliminary studies motivate us to explore this direction further.

8. Acknowledgments

The research leading to these results has received funding from by Swiss NSF project on "Parsimonious Hierarchical Automatic Speech Recognition (PHASER)" grant agreement number 200021-153507. We also thank the support of Swiss NSF projects on "Adaptive Multilingual Speech Processing (AMUSE)" grant agreement number 200020-144281.

9. References

[1] K. J. Ballard, S. Savage, C. E. Leyton, A. P. Vogel, M. Hornberger, and J. R. Hodges, "Logopenic and nonfluent variants of primary progressive aphasia are differentiated by acoustic measures of speech production." *PloS one*, vol. 9, no. 2, 2014. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/24587083

[2] M. Laganaro, M. Croisier, O. Bagou, and F. Assal, "Progressive apraxia of speech as a window into the study of speech planning processes," *cortex*, vol. 48, no. 8, pp. 963–971, 2012.

[3] A. Rosenberg, "Introducing Objective Acoustic Metrics for the Frenchay Dysarthria Assessment Procedure," Ph.D. dissertation, University of Sheffield, Sheffield, UK, 2007.

[4] D. Martínez, P. Green, and H. Christensen, "Dysarthria Intelligibility Assessment in a Factor Analysis Total Variability Space," in *Proc. of Interspeech*, 2013, pp. 2133–2137.

[5] K. H. Wong, Y. T. Yeung, P. C. M. Wong, G. Levow, and H. Meng, "Analysis of Dysarthric Speech using Distinctive Feature Recognition ," in *Proc. of 6th Workshop on Speech and Language Processing for Assistive Technologies*, 2015.

[6] L. Baghai-Ravary and S. W. Beet, *Automatic Speech Signal Analysis For Clinical Diagnosis And Assessment of Speech Disorders*. Springer, 2013.

[7] M. R. McNeil, S. R. Pratt, and T. R. D. Fossett, "Speech motor control in normal and disordered speech, B. Maassen, R. D. Kent, H. Peters, P. H. H. M. Van Lieshout and W. Hulstijn (Eds.)," *Oxford University Press*, pp. 389–414, 2004.

[8] M. Laganaro, "Patterns of impairments in aos and mechanisms of interaction between phonological and phonetic encoding," *Journal of Speech, Language, and Hearing Research*, vol. 55, no. 5, pp. S1535–S1543, 2012.

[9] M. Cernak, A. Asaei, and H. Bourlard, "On Structured Sparsity of Phonological Posteriors for Linguistic Parsing," *Speech Communication (to appear)*, 2016.

[10] C. Middag, Y. Saeys, and J.-P. Martens, "Towards an asr-free objective analysis of pathological speech," in *INTERSPEECH*. International Speech Communication Association (ISCA), 2010, pp. 294–297.

[11] Fougeron and S.-A. Jun, "Rate effects on French intonation: prosodic organization and phonetic realization," *Journal of Phonetics*, vol. 26, no. 1, pp. 45–69, Jan. 1998. [Online]. Available: http://dx.doi.org/10.1006/jpho.1997.0062

[12] G. Python, P. Pellet Cheneval, and M. Laganaro, "Dpistage norm des troubles de parole : apport des diadococinsies," *Aphasie et domaines associs*, vol. 1, 2015.

[13] D. Yu, S. Siniscalchi, L. Deng, and C.-H. Lee, "Boosting attribute and phone estimation accuracies with deep neural networks for detection-based speech recognition," in *Proc. of ICASSP*. IEEE SPS, March 2012. [Online]. Available: http://research.microsoft.com/apps/pubs/default.aspx?id=157585

[14] M. Cernak, B. Potard, and P. N. Garner, "Phonological vocoding using artificial neural networks," in *Proc. of ICASSP*. IEEE, Apr. 2015, pp. 4844–4848.

[15] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[16] M. Cernak and P. N. Garner, "PhonVoc: A Phonetic and Phonological Vocoding Toolkit," in *Proc. of Interspeech*, 2016.

[17] A. Asaei, M. Cernak, and H. Bourlard, "On Compressibility of Neural Network Phonological Features for Low Bit Rate Speech Coding," in *Proc. of Interspeech*, Sep. 2015, pp. 418–422.

[18] G. Perennou, "B.D.L.E.X. : A data and cognition base of spoken French," in *Proc. of ICASSP*, vol. 11, 1986, pp. 325–328.

[19] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[20] A. Asaei, H. Bourlard, and B. Picart, "Investigation of kNN Classifier on Posterior Features Towards Application in Automatic Speech Recognition," Tech. Rep. Idiap-RR-11-2010, 2010, [online] http://publications.idiap.ch/index.php/publications/show/1671.

The Effect of Semantic Difference on Non-expert Judgments of Simplified Sentences

Sven Anderson, S. Rebecca Thomas, Ki Won Kwon, Wayne Zhang

Bard College
Annandale-on-Hudson, NY 12504 USA
sanderson@bard.edu,thomas@bard.edu

Abstract

Advances in text simplification depend on reliable judgments of sentence difficulty. The ability of untrained native English speakers to judge sentence difficulty in the presence of variation in semantic similarity is examined using cloze tests and a forced-choice comparison task. Judgments from participants in web-based experiments demonstrate ability to assess sentence difficulty of professionally leveled sentence pairs with 84% accuracy. The comparison task results suggest that participants' ability to judge comparative sentence difficulty is inversely related to semantic similarity; that is, contrary to our intuition, speakers appear more accurate at judging sentence difficulty for sentences that are dissimilar than for those that are similar.

1. Introduction

Text simplification aims to expand access to textual information by algorithmically reducing the reading level of text – ideally without changing its meaning. Understanding text simplification can help us to design or customize tools for Augmentative and Alternative Communication (AAC) users, language learners, and other populations who need better access to information including the wealth of information available on the web.

Many approaches to simplification use statistical machine learning which depends on large corpora of text at different levels. Of particular use, though still quite limited in size, are bi-text corpora; these corpora present the same ideas in standard/simplified text pairs and have usually been created by aligning sentences from longer text, e.g. [1]. Unfortunately, the paucity of simplified text data has hindered most efforts to optimize machine learning approaches to the problem [2]. This led us to consider whether a necessary prerequisite for the identification of simplified texts, namely reading level judgments, might be more readily obtained from untrained native speakers using crowd-sourcing methods such as Amazon's Mechanical Turk.

Although professionals who create leveled readers are able to write sentences, paragraphs, and entire book collections at a pre-specified reader grade level, it is unclear whether untrained native speakers can meaningfully assess text difficulty, particularly of short texts. In addition, when text is simplified, its content is necessarily transformed, and judgments that compare the levels of two similar sentences may be confounded by differences that go beyond complexity differences, including substantial semantic and syntactic variation unrelated to simplification. Given the difficulty of the task, it appears doubtful that untrained readers can reliably assign grade levels to texts; however, such readers might be able to determine relative difficulty of two sentences. If so, a comparison-based task might enable a crowd-source approach much like what underlies test similarity tasks like SemEval [3].

In this paper we report on experiments that seek to discover whether native speakers can compare two sentences and judge when one is "more difficult" than another, despite other differences in syntax and semantics. We expected to find that more similar sentences, having fewer lexical and syntactic differences, thereby allow greater focus on the features relevant to the task and thus more accurate assessment of sentence difficulty. Leroy et al. [4] assume that explicit judgments of difficulty are possible, and refer to these explicit judgments, collected from a population of participants, as "perceived difficulty," which they distinguish from "actual difficulty." Like these researchers we use cloze measure scores to estimate "actual difficulty," but recognize that this simply measures the predictability of words given sentence contexts. We bring these two measures together by comparing cloze scores with comparative judgments.

2. Related Work

Text simplification improves accessibility of written or spoken language by transforming it to better meet the needs of the reader, including those with cognitive and/or linguistic challenges. The primary approaches to text simplification, all of which remain open research problems, include lexical and syntactic simplification, machine translation, and explanation generation [5]. For example, lexical simplification has been used to reduce the size of an

SLPAT 2016, 7th Workshop on Speech and Language Processing for Assistive Technologies, pages 56–62,
San Francisco, USA, 13 September, 2016.

article's vocabulary set to a smaller size that could determine which icons to use in communication boards [6]. Text simplification also applies to spoken language [7], and even non-linguistic media.

The present study focuses on the simplification of single sentences. It is not self-evident that this is the best level at which to study simplification: leveled readers, for example, are often simplified at the document level with transformations that restructure the linguistic presentation of ideas across paragraphs or sections. However, several considerations suggest the sentence as a viable unit of simplification. First, the sentence is a linguistic entity that conveys a complete proposition, and it is shared by all human languages. Sentence-level simplification involves both lexical and syntactic transformations, and even studies focused on word- or phrase-level transformations usually rely on delivery within a sentence context. Finally, simplification approaches derived from statistical machine translation depend on sentences aligned with their simplified counterparts, providing a basis for machine learning [1, 8].

A major challenge to the development of usable simplification is a clearer understanding of the factors that influence perception of text difficulty. That is, the creation of reliable automated text simplification algorithms depends on being able to detect whether text has been simplified! Lasecki [9] demonstrated that judgments from untrained native speakers can be used to assess sentence level along a 7-point Likert scale. In this case, the authors assumed that sentence simplicity is correlated with the number of simplifying transformations applied to a complex sentence. Many other studies also rely on comparing sentences that are semantically related, perhaps derived from the same original complex sentence. For example, [10] demonstrated a positive effect of lexical simplification using pairs of lexically simplified sentences. One objective of our study was to determine the degree to which semantic similarity affects simplicity judgments, since text simplification often introduces significant semantic changes.

3. Background

This section provides background concerning three topics central to our methods: measures of readability, empirical measures of difficulty, and algorithmic measures of sentence similarity. To contextualize this discussion, consider a sample task from our comparison experiment shown in Figure 1. In these sentence pairs the first pair is considered semantically similar and second pair dissimilar. (Additional sentence pairs from our experiments appear in the appendix.) Each pair of sentences is drawn from texts at different author-identified reading levels and cloze scores were used as an alternate measure of actual reading level.

3.1. Measures of Reading Level

Measuring the reading grade level and readability of text is important to writers of educational materials. Dubay [11] reports that from 1940 to the 1980's, approximately 200 readability formulas appeared in the literature. Many of these formulas use simple surface measures such as word count, word length, syllable count, average syllables per word, etc. to estimate sentence readability. As one example, the widely-used Flesch-Kincaid Grade level score is given by

$$\text{grade} = 0.39\frac{n_w}{n_s} + 11.8\frac{n_\sigma}{n_w} - 15.59$$

where n_s, n_w, and n_σ are the number of sentences, words, and syllables, respectively. The result is intended to be interpreted as a grade level; thus a text written for beginning readers would score roughly 1, with more complex texts assigned higher scores.

Most formulas are intended to measure the level of longer passages, not single sentences. Fry [12] notes that most formulas require at least 300 words. He proposes a formula for 40-99 words of three or more sentences, but observes that for shorter texts this formula is unreliable. The feasibility of determining readability based exclusively on surface level measures is limited, leading some researchers to explore models that incorporate semantic content.

3.2. The Cloze Measure as Actual Difficulty

The cloze measure estimates text difficulty by relating it to the ease with which a missing word can be guessed from its context [13], usually a text of several paragraphs. In most cloze experiments every Nth word – where N is generally 5-7 – is replaced by a blank. When human subjects guess the missing word with roughly 60% accuracy or greater, the text is considered relatively easy. Human prediction of a word based on surrounding context is reminiscent of the goal of language models based on n-grams: both human and algorithm rely on context to predict a most likely word. Smith and Levy [14] compared a 5-gram continuation (cloze) task in which participants completed a short phrase with corpus statistics derived from Web 1T and scanned books. They found that cloze scores varied substantially from corpus statistics.

One challenge in this project is how to assess the "true" difficulty of a sentence, against which to compare human perceptions or ratings. Leroy et al. [4] differentiates implicit tests of text difficulty based on cloze tests or tests of comprehension from explicit reports based on comparison or Likert-scale judgments by participants. The authors call the former *actual difficulty* and the latter *perceived difficulty*. We adopt the same terminology in this paper, acknowledging that the cloze score is merely our best approximation to actual difficulty.

Similar	S: North of Cairo, Egypt, the Nile enters the region called the delta. C: North of Cairo the Nile enters the delta region, a level triangular lowland. SEMILAR score: 0.883
Dissimilar	S: Fish come from the sea or from fish farms. C: Small clustered fishing villages are found along the coastline. SEMILAR score: 0.349

Figure 1: Example sentences; in each pair, the simpler sentence is marked with S. See appendix for more.

3.3. Semantic Similarity of Sentences

We employed the SEMILAR Toolkit [15] to measure the semantic similarity of sentence pairs. We measured the semantic similarity of each sentence pair using SEMILAR, with word similarity measured by LSA using the TASA model provided with the SEMILAR download. This variant of SEMILAR was selected based on empirical tests of several similarity measures as applied to three datasets: O'Shea [16], Sem* STS 2012 [17], and Sem* STS 2013 [3]; it scored the closest to human judgments for all three. The sentence pairs used in this study were selected to be of similar length. Thus sentence pairs with high semantic similarity are likely to share numerous similar word pairs, whereas sentences with low semantic similarity are likely to have a large number of word-to-word differences.

4. Methods

The sentences used for all experiments were drawn from Brittanica School articles. Topic-matched articles from Level 1 (elementary school level) and Level 3 (high school level) were mined to find sentence pairs differing markedly in difficulty, one drawn from each level. Aligned sentences extracted from these articles were retained only if their lengths (in words) differed by less than 20% to avoid confounding effects. For each topic, at most one sentence pair was retained in order to avoid semantic contamination across different sentence pairs. Sentence pairs were chosen to provide a group of highly similar and highly dissimilar sentences. The SEMILAR score falls in the range of 0.0 to 1.0, with 1.0 meaning identical sentences and 0.0 meaning no similarity. In order to facilitate comparing results on high similarity pairs vs. low similarity pairs, sentence pairs were ranked according to SEMILAR scores; the pairs falling in the 85-87.5th percentile and the 12.5-15th percentile were used in these experiments. The low-similarity pair SEMILAR scores ranged from 0.0 to 0.36; the high-similarity pair SEMILAR scores ranged from 0.83 to 0.97.

All experiments were run using a custom, web-based LimeSurvey [18] redirected from Amazon Mechanical Turk. Participants were paid a small amount for completing the experiment. The first two questions were "dummy" questions, with responses collected but not analyzed. The third question was designed with an unambiguous correct answer, and any participant who answered this incorrectly was not allowed to proceed with the experiment, to eliminate inattentive subjects. Demographic information was collected, and participants were told that only native English speakers could participate. In addition, each participant was asked about current English usage, e.g. whether they primarily or exclusively spoke English in their home.

5. Cloze Experiment

A total of three paired cloze experiments were run, making a total of six different sets of sentences, each of which was completed by twenty participants. Each experiment pair used the same set of 38 sentence pairs, which were taken from topic matched, level-differing articles as described above. The sentence pairs were divided into tasks A and B, such that for each sentence pair, the Level 1 sentence was randomly assigned to either task A or B, and the Level 3 sentence assigned to the other. Every seventh word in each sentence was deleted, starting at a specified position in the sentence: the fourth in the first experiment pair, the third in the second experiment pair, and the fifth in the third experiment pair. Participants provided responses to only one of the six different tasks and sentence order was randomized for each participant. A participant's response was considered correct only if it was a case-insensitive exact match. In total, participants produced 8,160 individual cloze responses, approximately 68 words per participant.

5.1. Cloze Results

Our sentences have an average of 12.8 words; a sentence will have one to four blank cloze positions on any trial. The proportion of a sentence's blanks correctly filled in by all participants, its cloze score, is highly variable, since scores depend largely on whether the cloze position happens to fall on an easily guessed word. To overcome this source of noise we average scores for each sentence over the three different blank positions and call this the *average cloze score*. The distribution of average cloze scores for all sentences shown in Figure 2 suggests that sentence difficulty is quite variable ($\overline{x} = 0.42$; $\sigma = 0.17$).

The original sentences were drawn from texts at elementary (Level 1) and high-school (Level 3) grade levels. The average cloze scores of sentences drawn from texts

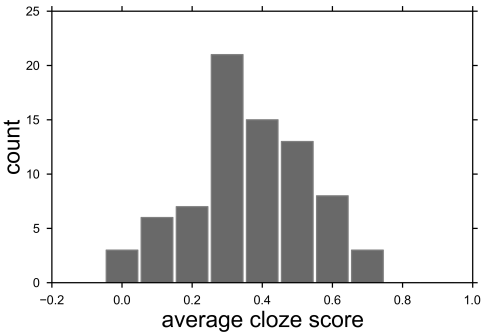


Figure 2: Histogram of average cloze scores for sentences. Averages are over three different initial positions in each sentence, with that and every seventh subsequent word deleted.

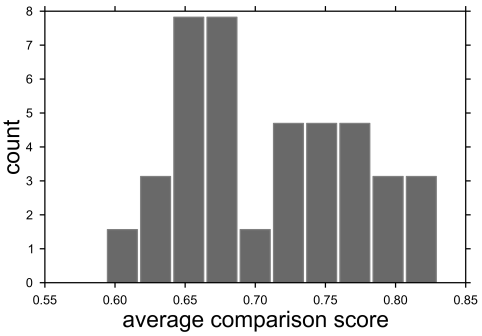


Figure 3: Average comparison accuracy per participant. Comparison is correct if Level 1 is judged simpler than Level 3.

at the elementary (49.3%) versus those at the high-school level (35.2%) are significantly different ($p < 0.0001$; $t = 4.05$), reflecting agreement between professional leveling of the articles and sentence difficulty as estimated by the cloze measure.

Given this result, we might ask whether other measures, such as unigram frequency or common readability measures show greater agreement with professional judgments of reading level than cloze scores do. To explore the utility of unigram frequency, word probability for each cloze word was estimated from the Web1T corpus. Unigram probabilities and average cloze scores have a correlation of $r = 0.506$, indicating some shared information. We then evaluated the ability of a logistic regression model to predict reading level based on both average cloze scores and unigram probabilities. Only the average cloze score is a significant ($p < 0.001$) predictor of level, suggesting that average cloze scores are superior predictors compared to unigram frequency.

The readability measures we examined, Flesch-Kincaid grade level and Lexile Score, do not generate significantly different values for Level 1 versus Level 3 sentences. The Flesch-Kincaid grade level has often been used to measure text difficulty, even in single sentences [4, 10]. Average Flesch-Kincaid grade levels are 7.2 and 8.6 for the Level 1 and 3 sentences, respectively; however, this difference does not reach significance ($p = 0.11$). This result is consistent with findings of Leroy et al. [10] who found that Flesch-Kincaid grade level did not differ significantly for sentence pairs that had been lexically simplified. We obtained similar results when using Lexile scores; we note that Lexile is intended for much longer texts. In this case average Lexile Scores [19] were 897.9 and 932.9 for Levels 1 and 3, respectively. The scores, grouped by reading level, were not significantly different ($p = 0.55$). Thus cloze scores better predict reading level than either unigram frequency or these reading level measures.

6. Comparison Experiment

The sentence comparison task tests whether participants are able to compare two sentences and determine which is simpler, despite variation in semantic similarity. Each of the 38 randomly ordered sentence pairs was presented in turn to the participant. The order in which the Level 1 and Level 3 sentences were presented was also randomized as the experiment was run. The forced-choice task requires participants to select the less difficult of two sentences.

6.1. Comparison Results

Using author-assigned text level to measure sentence difficulty, 27 of 29 participants responded above 60% correct. The scores of two participants were markedly lower outliers, within 5% of chance; thus, their results are removed from subsequent analysis. The mean score for the remaining 27 participants was 72.1% (Figure 3).

Using author-assigned text level as the measure of actual sentence difficulty, the per-sentence comparison accuracy results are shown in Figure 4. 84.2% (32/38) of sentence pairs had comparison judgments consistent with the author-assigned reading level; that is, the Level 1 sentence was more often labeled simpler than the Level 3 sentence. If instead we use the average cloze scores as the measure of actual sentence difficulty, only 76.3% (29 of 38) of comparison judgments match.

There are nine sentence pairs for which participants had better cloze accuracy on the Level 3 sentence than on the Level 1 sentence, which is not the expected result. This may, in part, be an artifact of having used only three of the possible seven initial positions for deleting words; in some sentences the deleted words may have been the hardest or easiest to guess. Another potential explanation is that the reading level of individual sentences may not always reflect the reading level of an entire text; a com-

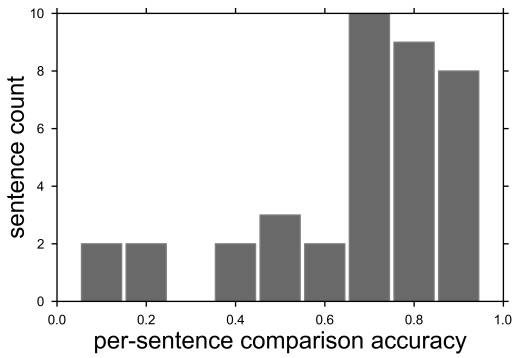


Figure 4: Histogram of per-sentence average comparison accuracy.

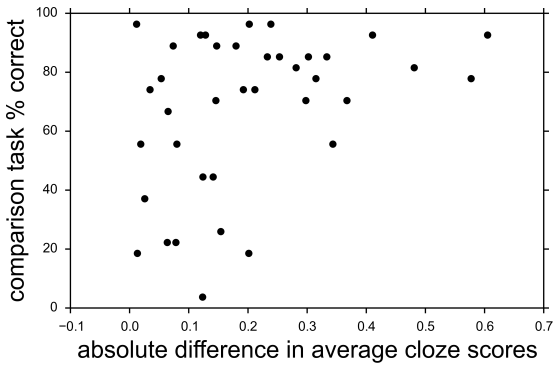


Figure 5: Scatter plot of percent correct perceived difficulty vs. absolute difference in average cloze.

plex text may contain relatively simple individual sentences. In selecting our experimental data, we assumed that sentences drawn from a text reflected the level of that text. The variation of average cloze scores and the variation in comparison scores both suggest that text level and sentence level are not in complete agreement.

6.2. Combined Results

The primary focus of this study was to determine whether the comparative difficulty of sentences can be assessed without regard to semantic or syntactic differences. The sentence pairs in this study belong to one of two groups: high similarity or low similarity. We compared the degree to which comparison scores agreed with author-assigned reading level for the high similarity (67.7%) versus low similarity (77.0%) groups; a t-test indicates that the comparison scores for these groups are not significantly different ($p = 0.158$; $t = -1.44$). That is, the ability of participants to compare sentence level was not significantly affected by the semantic and syntactic differences between those sentence pairs. However, when we compared the degree to which comparison scores agreed with comparative average cloze scores, the high similarity pairs average 58.4% versus 75.7% for low similar sentence pairs (one-tailed, $t = -2.12$, $p = 0.02$), reaching significance. Surprisingly, whether we assess actual sentence difficulty based on average cloze score or on author-assigned level, lower semantic similarity appears to aid the ability to judge comparative sentence difficulty.

Although not a simple linear relationship, the relation between perceived and actual difficulty shown in Figure 5 suggests that the ability to judge differences in sentence difficulty improves as cloze difference between two sentences increases.

Other factors that have been shown to be related to perceived difficulty include function word density, the occurrence of difficult words as measured by Dale-Chall, and noun-phrase complexity [4]. We measured noun-phrase complexity using the maximal depth of all noun phrases in a sentence. Function word density was based on English stopwords listed in NLTK. A multiple regression test including these variables, average cloze scores, and ratios of cloze scores was performed. The overall adjusted R-squared value was 0.17 ($p = 0.12$). The only significant variable was the Dale-Chall score of the difficult sentence ($p = 0.05$). The average cloze score of the Level 1 sentence had the next smallest p-value ($p = 0.10$). We have found no variables that are significant predictors of perceived difficulty scores.

7. Discussion

One goal of this study was to determine whether sentence level judgements of untrained native speakers could be used to replace professional judgments of sentence difficulty. Experience with simplifications in the Simple English Wikipedia suggests that non-professional simplification is much less reliable than that obtained from professionals [2]. However, [9] showed that crowd-source methods could accurately rate the number of simplifying transformations made to single complex sentences. These results, from a slightly smaller set of sentences, are most directly comparable to the results reported above. Note that in [9] sentence simplicity was based on a count of transformations to a base sentence and did not incorporate the wide variation in semantic similarity we investigate here. Our cloze results support the claim in [9] that untrained speakers can implicitly judge sentence difficulty in agreement with expert judgments, but imperfectly so. Moreover, perceived difficulty, as measured by the sentence comparison task somewhat similar to that in [9], is not clearly related to cloze scores on the same sentences. This suggests that actual and perceived difficulty do not measure identical sentence attributes. Additionally, neither directly represents the benefit in comprehension that such a simplification might afford the reader. The best metric for measuring text difficulty depends on

the task, and the field has not settled on a clear methodology for such measures.

Our unexpected finding that comparing sentence difficulty is improved by semantic dissimilarity is one that must be more carefully studied. Semantic priming presents one possible explanation. By this account, when participants are presented with a sentence pair, the second sentence may seem easier because the first sentence has semantically primed the participant for the second sentence. Of course, this is likely to be a much stronger effect for semantically similar sentences. If so, we would predict that there may be a consistent bias toward underestimating the complexity of the second sentence in high-similarity pairs. Unfortunately, our experimental design was such that the order in which the two sentences were presented was randomized for each participant at run time, and the order of presentation was not recorded. Thus the experimenters cannot assess whether such a bias actually was observed.

7.1. Crowd-sourced Effort to Judge Difficulty

One advantage of both the cloze procedure and the forced-choice comparison task is that they can be undertaken by readers of a language with very little training and are therefore suitable for crowd-sourced data collection. The comparison task requires a single decision that ranks one sentence relative to another. By contrast, the cloze procedure requires multiple lexical inputs, but it yields a percentage cloze score that provides a total order for all sentences. Which approach is best?

In cases where experimenters desire a total order over a particular set of sentences, a simple analysis provides a means to compare the effort required for each of these two methods. If we have n sentences of average length m, the comparison task is essentially sorting by binary decisions and will provide a total order of the sentences with $nlog(n)$ comparisons. Assuming we must provide blanks on approximately half of the words to estimate the cloze score, the cloze Procedure will require $\frac{m}{2} \times n$ blanks to be completed by a set of participants. Thus, the number of inputs provided by participants is numerically similar when $log(n) = \frac{m}{2}$. For sentences of about 20 words, this implies $n \approx 1024$. That is, the comparison procedure will require fewer human decisions until approximately 1000 sentences are to be completely ordered. Note that this ignores the relative difficulty of, and time required to make, comparison vs. word-choice decisions.

In this study, the cloze procedure provided a total order using about 210 word choices per participant, but took significantly more effort for both participants and the experimenters. In contrast, the comparison task required only 38 comparisons per participant but was not designed to provide a total order. If a total order is desired, further comparison tasks might be generated for additional test subjects, guided by an $O(nlog(n))$ sorting algorithm such as mergesort.

7.2. Conclusion

The methodology by which sentence difficulty is measured has direct consequences for the creation of corpora (e.g., [20, 8, 1, 10]) underlying future development of text simplification. The genesis of this project was the authors' awareness of the need for more bi-text data at different levels of reading complexity, which might hypothetically be used to train machine translation systems to perform text simplification, or to train systems that measure text readability. Naive users could presumably judge relative readability more easily and more quickly than they could perform cloze exercises. If their judgments were sound, then one might use crowdsourcing to efficiently sort sentences by readability. Our results indicate that there is no need to match the sentences by content or even by topic; in fact, it appears be an advantage not to do so. Future research should clarify whether this is more generally true.

8. Appendix: Sample sentence pairs

Sample high-similarity pairs
S: The two openings in the nose are called nostrils.
C: The external openings are known as nares or nostrils.
SEMILAR score: 0.828

S: Northern Ireland is often called Ulster because it includes six of the nine counties that made up the ancient kingdom of Ulster.
C: Northern Ireland is sometimes referred to as Ulster, although it includes only six of the nine counties which made up that historic Irish province.
SEMILAR score: 0.869

S: Four main aerodynamic forces act on an airplane in flight.
C: An aircraft in straight-and-level unaccelerated flight has four forces acting on it.
SEMILAR score: 0.843

Sample low-similarity pairs
S: Guglielmo Marconi was an Italian scientist and inventor.
C: Marconi's great triumph was, however, yet to come.
SEMILAR score: 0.010

S: Unlike many plants, cacti do not have deep roots.
C: The fruit is usually a berry and contains many seeds.
SEMILAR score: 0.013

S: In nearly all mammals, the female carries the developing young in her body after mating.
C: The winter dormancy of bears at high latitudes is an

analogous phenomenon and can not be considered true hibernation.
SEMILAR score: 0.113

9. References

[1] W. Coster and D. Kauchak, "Simple English wikipedia: a new text simplification task," in *Proc. of the 49th Association for Computational Linguistics*, 2011, pp. 665–669.

[2] W. Xu, C. Callison-burch, and C. Napoles, "Problems in current text simplification research : New data can help," *Transactions of the ACL*, vol. 3, pp. 283–297, 2015.

[3] E. Agirre, D. Cer, M. Diab, A. Gonzalez-Agirre, and W. Guo, "SEM 2013 shared task: Semantic textual similarity, including a pilot on typed-similarity," in *In *SEM 2013: The Second Joint Conference on Lexical and Computational Semantics. Association for Computational Linguistics*, 2013.

[4] G. Leroy, S. Helmreich, and J. R. Cowie, "The influence of text characteristics on perceived and actual difficulty of health information," *International journal of medical informatics*, vol. 79, no. 6, pp. 438–49, 2010.

[5] M. Shardlow, "A survey of automated text simplification," *International Journal of Advanced Computer Science and Applications*, vol. 4, no. 1, pp. 58–70, 2014.

[6] S. Anderson, S. Thomas, C. Segal, and Y. Wu, "Automatic reduction of a document-derived noun vocabulary," in *Twenty-Fourth International FLAIRS Conference*, 2011.

[7] D. J. Higginbotham, G. W. Lesher, B. J. Moulton, and B. Roark, "The application of natural language processing to augmentative and alternative communication," *Assistive Technology*, vol. 24, no. 1, pp. 14–24, 2012.

[8] Z. Zhu, D. Bernhard, and I. Gurevych, "A monolingual tree-based translation model for sentence simplification," in *Proceedings of the 23rd International Conference on Computational Linguistics*, 2010, pp. 1353–1361.

[9] W. S. Lasecki, L. Rello, and J. P. Bigham, "Measuring text simplification with the crowd," *Proceedings of the 12th Web for All Conference (W4A '15)*, pp. 4:1—4:9, 2015.

[10] G. Leroy, D. Kauchak, and O. Mouradi, "A user-study measuring the effects of lexical simplification and coherence enhancement on perceived and actual text difficulty," *International journal of medical informatics*, vol. 82, no. 8, pp. 717–30, Aug. 2013.

[11] W. DuBay, "The principles of readability," *Costa Mesa: Impact Information*, 2004.

[12] E. Fry, "A readability formula for short passages," *Journal of Reading*, vol. 33, no. 8, pp. 594–597, 1990.

[13] W. L. Taylor, "Cloze procedure: a new tool for measuring readability." *Journalism and Mass Communication Quarterly*, vol. 30, no. 4, p. 415, 1953.

[14] N. J. Smith and R. Levy, "Cloze but no cigar: The complex relationship between cloze, corpus, and subjective probabilities in language processing," *Proceedings of the 33rd Annual Meeting of the Cognitive Science Conference*, pp. 1637–1642, 2011.

[15] V. Rus, M. C. Lintean, R. Banjade, N. B. Niraula, and D. Stefanescu, "Semilar: The semantic similarity toolkit." in *ACL*, 2013, pp. 163–168.

[16] J. O'Shea, Z. Bandar, K. Crockett, and D. McLean, "Benchmarking short text semantic similarity," *International Journal of Intelligent Information and Database Systems*, vol. 4, no. 2, p. 103, 2010.

[17] E. Agirre, M. Diab, D. Cer, and A. Gonzalez-Agirre, "Semeval-2012 task 6: A pilot on semantic textual similarity," in *Proc. First Joint Conference on Lexical and Computational Semantics*, 2012, pp. 385–393.

[18] C. Schmitz, *LimeSurvey:An Open Source Survey Tool*, 2015, http://www.limesurvey.org.

[19] A. J. Stenner, H. Burdick, E. E. Sanford, and D. S. Burdick, "The Lexile framework," Durham, NC: MetaMetrics, Tech. Rep., 2007.

[20] R. Chandrasekar and B. Srinivas, "Automatic induction of rules for text simplification," *Knowledge-Based Systems*, vol. 10, no. 3, pp. 183–190, 1997.

An ASR-Based Interactive Game for Speech Therapy

Mario Ganzeboom[1]*, Emre Yılmaz*[1]*, Catia Cucchiarini*[1] *and Helmer Strik*[1]

[1]CLS/CLST, Radboud University, Nijmegen, The Netherlands

{m.ganzeboom,e.yilmaz,c.cucchiarini,w.strik}@let.ru.nl

Abstract

The demand for intensive and costly speech therapy to patients impaired by communicative disorders can potentially be alleviated by developing computer-based systems that provide automatized speech therapy in the patient's home environment. In this paper we report on research aimed at developing such a system that combines serious gaming with automatic speech recognition (ASR) technology to provide computer-based therapy to dysarthric patients. The aim of the serious gaming environment is to increase the patients' motivation to practice, which tends to decrease over time with conventional speech therapy, as progress in dysarthric patients is often slow. Additionally, some speech exercises (e.g. drills) are not particularly motivating due to their repetitive nature. The ASR technology is aimed at providing feedback on speech quality during training to improve speech intelligibility. Different types of acoustic models were trained on normal speech of adults and elderly people, and tested on dysarthric speech. The results show that speaker-adaptive training and Deep Neural Networks (DNN)-based acoustic models substantially improve the performance of ASR in comparison to traditional GMM-HMM-based methods. In this specific case, the ASR-based game is developed to provide speech therapy to dysarthric patients, but this approach can be adapted for use in other types of communicative disorders.

Index Terms: communicative disorders, speech therapy, serious gaming, ASR.

1. Introduction

Among the problems that are likely to be associated with an increasingly ageing population worldwide is a growing incidence of neurological disorders such as Parkinson's Disease (PD), Cerebral Vascular Accident (CVA or stroke) and Traumatic Brain Injury (TBI). Possible consequences of such diseases are communicative disorders. One of them is dysarthria, a motor speech disorder that affects speech intelligibility and causes communication problems [1]. Face-to-face speech therapy has proven beneficial for improving speech intelligibility in dysarthric patients, but to be effective therapy should be intensive [2, 3, 4, 5]. Owing to the increasing number of patients and the related high expenses, it may become difficult to provide intensive care in the future. As a result, attempts are being made at finding alternative, sustainable solutions that can guarantee the amount of care that is required for dysarthria patients in addition to or even without face-to-face sessions [6, 7]. In analogy to applications for pronunciation improvement in second language learning, [8, 9, 10], computer-based systems that employ ASR technology can be used to provide dysarthric patients with more robust and focused practice. A compounding problem however is that progress in these patients is slow, which is likely to reduce their motivation to practice. So one of the challenges is to develop systems that can motivate patients to get the necessary amount of practice. This can be achieved by resorting to games, which are known to increase motivation in learners and patients [11]. This aim is pursued in the CHASING project[1], in which a serious game employing ASR is being developed and evaluated to provide additional speech therapy to dysarthric patients.

In this paper we report on research that we conducted to develop and optimize this ASR-based game. Although we briefly refer to the process of game development and optimization, the emphasis is mainly on developing the ASR technology to be integrated in this game. The remainder of the paper is organized as follows. Section 2 briefly summarizes related work on game-based and ASR-based speech therapy. Section 3 presents the architecture of the ASR-based game. Section 4 describes the methodology adopted in experiments aimed at investigating how ASR can be improved to be incorporated in the game. Section 5 reports on the results of these experiments, while Section 6 presents a discussion of the results and the conclusions

2. Games and ASR for speech therapy

Neurological disorders like PD, stroke or TBI manifest more frequently at later ages (e.g. 55 or above), although these disorders sometimes may also occur at a younger age. Our research focuses on developing and evaluating an ASR-based serious game for providing speech therapy to elderly individuals with dysarthria, because these constitute the majority of the patients' group. Krause et al. [12] reported of work in this direction. They developed a game that challenged patients with PD to break glasses and vases by producing sufficiently loud and long /a/-phonemes. The game aimed to improve the reduced voice intensity of the patients that were reported to have a mild form of dysarthria. Speech processing algorithms sufficed to provide real-time feedback on the current and desired intensity only. An initial evaluation of the game was conducted with eight patients. Significant improvements of average peak voice loudness were observed in comparison with previously calibrated limits to those measured during game play.

Outside of academic research, similar types of games already existed: Dr. Speech[2] and VoxGames [3]. These games targeted children with varying speech disorders. RodrĂguez et al. [13] describe a set of small games named 'PreLingua'. These games were intended to assist the work in speech therapy focussing on deviations in phonatory related speech dimensions. The aim was to improve the use of voice activity, intensity, breathing, tone and vocalization of children with develop-

[1]See http://hstrik.ruhosting.nl/chasing/. Last retrieved on June 24, 2016.

[2]See http://www.drspeech.com/SpeechTherapy5.html, last retrieved on March 25th, 2016.

[3]See http://www.ctsinf.com/english/#voxGames.html, last accessed on March 25th, 2016.

SLPAT 2016, 7th Workshop on Speech and Language Processing for Assistive Technologies, pages 63–68,
San Francisco, USA, 13 September, 2016.

mental disorders. Speech processing algorithms were utilized to analyse children's voices and control interactive elements in the games. Although the games were in use at a school for special education and were positively evaluated by a group of speech therapists, no evidence related to the efficacy of the games was reported.

Bunnell et al. [14] described work that included games and ASR technology. Their STAR system was intended for children with articulation disorders who practiced speech production starting with CV syllables progressing to words/phrases. Hidden Markov Models (HMMs) were trained on children's speech and evaluated on speech of children substituting /w/ for /r/. The results reported in their research showed that the log likelihoods from the HMMs correlated well with the perceptual ratings collected for utterances that contained substitutions, but poorly for correctly pronounced examples.

Vaquero et al. [15] introduced a set of small games named 'Vocaliza' as an addition to the previously described PreLingua. ASR technology was used to recognize the words children spoke while completing speech exercises. The novel addition was the utilization of ASR-based utterance verification (UV) technology to detect mispronunciations in a child's speech on a word level by calculating a confidence measure based on likelihood ratios. Similar to their research describing PreLingua, no results on performance evaluation were reported.

In a collaboration with Yin [16], they introduced mispronunciation detection at a phoneme level and obtained 6.7% absolute improvement in Equal Error Rate (EER) when replacing their baseline speaker-independent acoustic models with speaker-adapted models. In [17] they reported results on mispronunciation detection on both word and phoneme levels. They showed that their word-level pronunciation verification system in Vocaliza was rather unreliable for speech therapy due to a trade-off lowering the False Rejection Rate (FRR), at the cost of increasing the False Acceptance Rate (FAR). This limited frustration to the user, but at the same time accepted many mispronounced words. Their phoneme-level mispronunciation detection method did not show this trade-off and obtained considerable improvements in EER (i.e. equal FRR and FAR percentage). Additionally, further improvements in detecting mispronunciations were reported by fusing prior knowledge of the target word, target phoneme and its position in the word with the obtained posterior probabilities using MultiLayer Perceptron Neural Networks (NN-MLP).

Notable is also the study by Tan et al. [18] in which they combine word-level articulation exercises with popular gameplay (i.e. Pac-Man) employing an off-the-shelve ASR package. Feedback on the user's pronunciation was provided by triggering a predefined action in the game if a word was recognized and by displaying the recognized word and its corresponding confidence score at the top of the game screen. Evaluation took place using an informal play test with two children. Noteworthy observations are that the ASR package, not adapted to speech of children, frequently did not recognize their correct pronunciations of the target words. However, the children appeared to stay engaged and interested and continued playing. This may point to a certain level of tolerance for recognition errors, before a user gets frustrated.

As the previous paragraphs show, most research involving ASR-based games for speech therapy was aimed at disordered speech of children. In our research we developed a game aimed at elderly patients with dysarthric speech. To the best of our knowledge, this has not been done before. In addition to voice intensity in Krause et al. [12], our game also employs speech processing algorithms to provide real-time feedback on fundamental frequency (F0). We are currently researching strategies to also add feedback on pronunciation to the game. This strategy potentially consists of two phases that both employ ASR technology for pronunciation evaluation: utterance verification and pronunciation error detection. For the utterance verification phase we need to recognize the user's utterance. Mustafa et al. [19] provide an interesting overview of previous research on ASR for dysarthric speech. In section 4 we report on our initial recognition experiments that include dysarthric speech from our target group. Some previous research on pronunciation error detection in elderly dysarthric speech has been conducted [20, 21], but this addressed dysarthric speech due to different etiologies.

In the introduction we argued that serious games can increase patients' motivation for speech therapy. It is therefore important to limit potential sources of frustration in the game. Utterance verification and pronunciation error detection technology could be a source of such frustration, because it is not guaranteed that the patient's utterance is always recognized. In previously described research, patients potentially had to say the same utterance multiple times if it was not recognized by the system, before they were able to continue. This causes frustration to the patient and potentially lowers motivation. Elements in the gameplay should therefore not completely rely on the outcome of ASR technologies. A nice example is perhaps the game in Tan et al. [18]. The patient was only rewarded with a 'power up' if the associated word was recognized, but could still continue playing the game normally if this was not the case.

In the next section we outline our current game and briefly describe our ideas for integrating gameplay elements that do not fully rely on the outcomes of ASR technology.

3. The CHASING game for ASR-based speech therapy to dysarthric patients

In the CHASING project a serious game has been developed that Dutch-speaking dysarthric patients can play with their friends and relatives. The choice of the game was based on user tests in which several game concepts had been proposed and evaluated. An important aspect in this respect was whether the game should be a single player game or a multiplayer game. A single player game has the advantage that it can be played independently by a patient, without having to rely on other participants. On the other hand, multiplayer games are generally more engaging and motivating and are therefore likely to be played more frequently, which is of course very important for the therapy to be effective. The patients in our focus group showed a clear preference for multiplayer games and indicated that finding players would not be a problem as these could be their friends and relatives. Further tests with initial versions of the game revealed that additional considerations had to be made for the intended target group of elderly people who are no experienced gamers. For instance, it turned out that it was necessary to proceed more gradually both in introducing new game elements and in advancing to higher levels of difficulty. Moreover, the use of direct visual feedback was found easier to interpret as opposed to indirect feedback integrated into the gameplay. This is potentially due to diminished cognitive skills as an additional consequence of their neurological disorder. The game that was eventually selected and developed is called 'Schatzoekers' (i.e. 'Treasure hunters'). It is a two-player cooperative game in which players talk to each other through an audio con-

nection and have to help each other in finding the treasure and the key to open it. One player, the 'digger', can dig up the treasure on land and the other, the 'diver', dives in various rivers and canals in search for the key to open it. The locations of both treasure and key are marked on the map, but only the 'digger' can see the location of the key (where the 'diver' should go) and only the 'diver' can see the location of the treasure (where the 'digger' should go). The players thus have to explain to each other where to go. This way, players are encouraged to keep speaking to each other to describe where they are on the map and giving directions where the other should go. Figure 1 shows the tablet set up for which the game was developed. Figure 2 displays a screenshot of the game.

Figure 1: The tablet set up displaying the start screen of the game.

Every map in the game is a different level. Levels of difficulty are influenced by the size and layout of the map, the complexity of street names and icons to describe one's location, the availability of an overview map and its level of detail. In the initial levels, your location is also visible to your co-player, in addition to the location of the item you need to find. That visibility is removed in later levels. Players talk to each other using the headset and get feedback on their loudness of voice and pitch from the game, especially when they are above or below specified thresholds. This is indicated by the horizontal green bar shown in Figure 2, which provides real-time feedback while the patient is speaking and shows a green, orange or red color when the loudness of the patient's speech is within, near or below the threshold, respectively. When the pitch is too high a notification slides down from underneath the bar instructing to 'speak loud and low'. The therapeutic goals of the game are to motivate dysarthric individuals in using continuous speech, and to speak up and maintain predefined levels of pitch and loudness.

In addition to feedback on loudness and pitch, our idea is to incorporate ASR technology in the game to be able to automatically provide more robust and focused feedback on speech quality. An initial idea to integrate this into the gameplay is to present the user with passphrases that have to be uttered, before the treasure is opened in the game. The user is always rewarded with treasure, but potentially depending on the level of speech quality determined by the ASR, the user may be rewarded with different kinds and/or amounts of treasure.

In the preparations for providing this type of feedback using ASR, we developed an ASR architecture that runs on a server in the cloud. Every time the game authenticates to this server, a separate ASR session is initialized which is only available to the

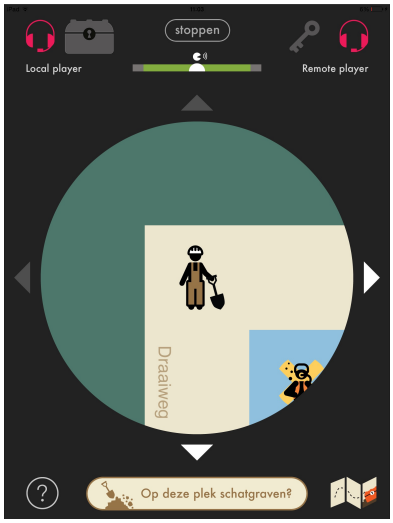


Figure 2: An in-game screenshot displaying the game from the perspective of the player 'digger'. In the partially blue square it can be observed that the, 'diver' already reached the correct location of the key.

user who requested it. The audio containing the player's speech is then continuously streamed to the server for offline analysis later on. As we have to handle privacy sensitive data, all communication with the server happens over secured connections.

We are currently developing and optimizing the ASR technology for the game and this work has to be done while the game is still being developed, improved and finalized. This means, among other things, that we cannot test the technology on the actual speech that will be produced in the game. A compounding problem in developing ASR applications for pathological speech is the limited availability of sufficient representative data. Since this was all anticipated, we started experimenting with already available speech data that can be considered representative for the type of speech that will have to be dealt with in the game (see section 4.1). Initial experiments were run to investigate to what extent ASR performance can be improved by speaker-independent Subspace Gaussian Mixture Model-Hidden Markov Models (SGMM-HMMs) and speaker-adaptive Deep Neural Networks (DNNs) in comparison to the traditional system using speaker-adaptive GMM-HMMs.

4. ASR experiments for the CHASING game

The ASR module in the CHASING game has to process dysarthric speech which is notoriously more difficult to recognize than normal speech. One of the obstacles in developing ASR technology that can handle dysarthric speech is the limited amount of dysarthric speech data available for training and testing the ASR algorithms. To partly circumvent this problem experiments were conducted in which maximum use was made of existing databases.

To investigate the baseline performance of the deep neural network-based acoustic models on dysarthric speech, we performed recognition experiments on a similar type of speech in Flemish which is a variety of the Dutch language spoken in Flanders. This choice is motivated by the availability of a prin-

cipled Flemish pathological speech database, namely the COPAS database [22], and the phonetic similarity between the two varieties Flemish and Dutch.

4.1. Speech databases

Within the framework of the CHASING project, a database of dysarthric speech is being collected [23], but this corpus is still relatively limited. For Dutch another, larger corpus of pathological speech has been collected [22], which also contains a considerable number of recordings of dysarthric speech. Although this database was compiled in Flanders and contains speech of patients who speak the Southern variety of Dutch (Flemish), it can be useful to investigate the baseline performance of the deep neural network-based acoustic models on dysarthric speech. First, the two varieties of Dutch spoken in the Netherlands and in Flanders are mutually intelligible and the most important phonological and phonetic differences are well known. Second, developing and testing the ASR on Flemish speech material makes it more feasible to adapt the CHASING game for Flemish patients at a later stage.

4.1.1. Training Data

Since the idea was to investigate the performance of the deep neural network-based acoustic models on Flemish dysarthric speech, Flemish speech data were used for training. These were obtained from the Flemish components of two Dutch and Flemish speech databases, i.e. CGN [24] and JASMIN-CGN [25]. The Flemish CGN component contains recordings of standard Flemish as spoken by adults in different regions of Flanders. The components with read speech, spontaneous conversations, interviews and discussions were used for training the acoustic models. The total duration of the normal Flemish speech (FLN) used in the recognition experiments is 186.5 hours. Additional speech material was taken from the Flemish component of the JASMIN-CGN corpus, which is an extension of the CGN database with speech of children non-natives and elderly people. The elderly speech component, with a total duration of approximately 5 hours, was employed in our experiments.

4.1.2. Testing Data

The COPAS pathological Flemish speech database [22] was used for testing the acoustic models trained on various speech types described in the previous section. The COPAS database has been collected within the framework of the SPACE project which was aimed at developing a reliable ASR-based speech assessment tool for pathological speech. This speech database contains recordings of 122 normal speakers as a control group and 197 speakers with speech disorders such as dysarthria, cleft, voice disorders, laryngectomy and glossectomy. The speech material includes not only word reading tasks, but also isolated sentence and short passage reading tasks.

The word reading tasks used in this paper is the Dutch Intelligibility Assessment (DIA) [26] material which contains 35 versions of 50 consonant-vowel-consonant (CVC) words and pseudowords organized in 3 subgroups. Moreover, we added all sentence reading tasks with annotations. These include 2 isolated sentence reading tasks (S1 and S2), 11 text passages (S) of reading difficulty levels AVI 7 and AVI 8 according to a system adopted in the Dutch language area that indicates reading difficulty based on text structure, vocabulary and length of words and sentences, and varies from AVI 1 up to AVI 9, and a phonetically balanced text known as Text Marloes (TM) [27].

For the recognition experiments, we classified the aforementioned material based on the type of speaker (normal vs. pathological) and speech material (word vs. sentence) resulting in 4 test sets. The speech segments in which the speaker does not utter the target word are discarded to be able to evaluate the recognizer errors only. There are 687 different words and 212 different sentences in the test data. The test set containing the word tasks uttered by normal speakers (WN) and speakers with disorders (WD) consists of 6154 and 8648 utterances with a total duration of 1.5 and 2 hours, respectively. The test set containing the sentence tasks uttered by normal speakers (SN) and speakers with disorders (SD) consists of 1918 (15,149) and 1034 (8287) sentences (words) with a total duration of 1.5 and 1 hour, respectively.

4.2. Implementation Details

The recognition experiments are performed using the Kaldi ASR toolkit [28]. The standard training recipe provided for multiple databases is applied to train a conventional context-dependent GMM-HMM on MFCC, LDA-MLLT and FMLLR-adapted features. Then, a system using an SGMM-based [29] acoustic model is also trained with a universal background model having 800 Gaussians and substate phone-specific vector size of 40. Providing the best performance among the aforementioned recognizers, this system is used to obtain the state alignments required for DNN training.

For DNN training, a standard feature extraction scheme is used by applying Hamming windowing with a frame length of 25 ms and frame shift of 10 ms. The DNNs with 6 hidden layers and 2048 sigmoid hidden units at each hidden layer were trained on the FMLLR-adapted features. The DNN training is done by mini-batch Stochastic Gradient Descent with an initial learning rate of 0.008 and a minibatch size of 256. The time context size is 11 frames achieved by concatenating ± 5 frames. Unigram language models were trained on the target word transcriptions and used in the word recognition tasks. For the sentence recognition tasks, trigram language models were trained on the target sentence transcriptions.

5. Results and Discussion

We performed ASR experiments using the speech data described in Section 4.1. The recognition results obtained on the word and sentence tasks uttered by normal and pathological speakers from the COPAS database are presented in Table 1. For each column, the best results are marked in bold. In the context of the proposed serious game, sentence recognition is a more relevant task compared to isolated word recognition. For completeness, we present both word and sentence task results in this section.

Table 1: Word error rates in % obtained on the word and sentence COPAS test sets

Acoustic models	WordDys	WordNor	SentDys	SentNor
GMM+MFCC	76.2	55.0	37.3	13.3
GMM+LDA-MLLT	73.8	51.6	36.7	11.7
GMM+FMLLR	66.2	41.0	27.8	7.8
SGMM	59.2	34.0	**23.6**	5.7
DNN+FMLLR	**56.2**	**30.2**	**23.6**	**4.2**

The conventional GMM-HMM trained on Mel Frequency Cepstral Coefficients (MFCCs) provides a WER of 37.3% on

the dysarthric sentence utterances and a WER of 76.2% on the dysarthric word tasks. The WER difference between the normal and dysarthric speakers on the two tasks is larger than 20% for this system. The high WERs on the word tasks were due to the challenging recognition of one-syllable words and phonetically similar pseudowords. By using LDA-MLLT the WERs were reduced slightly as the second row in Table 1 shows.

Using discriminately trained features and including speaker information by applying speaker adaptive training (SAT) further reduced the WERs to 27.8% and 66.2%. Compared to GMM+LDA-MLLT, this is an absolute improvement of 8.9% and 7.6% on sentence and word tasks, respectively. The dysarthric speech recognizer benefits considerably from the speaker adaptive training.

Training SGMM-based acoustic models improves the recognition accuracy on the two tasks to a WER of 23.6% for sentence recognition and 59.2% on word recognition tasks. The DNN-based recognizer provides a similar performance with the SGMM-based recognizer on the sentence recognition task. An absolute improvement by 3% is obtained using DNNs on the word recognition task.

By using state-of-the-art DNN-based acoustic models it was possible to substantially lower the WERs, especially for the sentence task. In practice, we could lower the WERs even more, by using simpler tasks (exercises) with less complex language models. For instance, we could elicit speech in such a way that the number of possible correct answers is low, and then the ASR only has to determine whether one of these answers was spoken.

6. Conclusions

In this paper we have reported on our research aimed at developing an ASR-based game that can provide speech therapy to dysarthric patients with Dutch as their mother tongue. In particular, we have described experiments in which different types of acoustic models trained on normal speech were tested on dysarthric speech. The results show that speaker-adaptive training and Deep Neural Networks (DNN)-based acoustic models substantially improve the performance of ASR in comparison to traditional GMM-HMM-based methods.

Considering that the performance can further be improved by adopting more specific tasks in the game, as discussed in the previous section, we can conclude that the levels of accuracy obtained in these experiments bode well for the deployment of ASR in speech therapy applications. The experiments reported on in this paper were conducted on dysarthric speech of a close variety of Dutch, i.e. Flemish. Given that data sparsity is one of the major obstacles in developing ASR-based speech therapy applications, employing speech data of a closely related language variety is a possible way of approaching this problem. In the near future we intend to conduct similar experiments with dysarthric speech of the Northern variety of Dutch. However, since developing ASR-based speech therapy applications is very costly, it is important to know to what extent they are portable to other language varieties, so that more patients can profit from them.

Finally, we would like to underline that although in this specific case the ASR-based game has been developed to provide speech therapy to dysarthric patients, this approach can be easily adapted for use in other types of communicative disorders. To conclude, these results thus indicate that employing ASR technology for speech therapy to patients with communicative disorders is becoming more viable. In turn, this will allow them to get more intensive therapy, also in their home environment.

7. Acknowledgements

This research is funded by the NWO research grant with ref. no. 314-99-101 (CHASING). We would like to thank all members of the chasing team for their contribution: Marjoke Bakker, Lilian Beijer, Douwe-Sjoerd Boschman, Lodewijk Loos, Paulien Melis, Jurre Ongering, Toni Rietveld and Sabine Wildevuur.

8. References

[1] J. R. Duffy, *Motor Speech Disorders: Substrates, Differential Diagnosis, and Management*, 1st ed. St Louis, MO: Mosby, Jan. 1995.

[2] L. Ramig, S. Sapir, S. Countryman, A. Pawlas, C. O'Brien, M. Hoehn, and L. Thompson, "Intensive voice treatment (LSVT®) for patients with parkinson's disease: a 2 year follow up," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 71, no. 4, pp. 493–498, 2001.

[3] S. K. Bhogal, R. Teasell, and M. Speechley, "Intensity of aphasia therapy, impact on recovery," *Stroke*, vol. 34, no. 4, pp. 987–993, 2003.

[4] G. Kwakkel, "Impact of intensity of practice after stroke: Issues for consideration," *Disability and Rehabilitation*, vol. 28, no. 13-14, pp. 823–830, 2006.

[5] M. Rijntjes, K. Haevernick, A. Barzel, H. van den Bussche, G. Ketels, and C. Weiller, "Repeat therapy for chronic motor stroke: A pilot study for feasibility and efficacy," *Neurorehabilitation and Neural Repair*, vol. 23, no. 3, pp. 275–280, 2009.

[6] L. J. Beijer, A. C. M. Rietveld, M. B. Ruiter, and A. C. H. Geurts, "Preparing an e-learning-based speech therapy (est) efficacy study: Identifying suitable outcome measures to detect within-subject changes of speech intelligibility in dysarthric speakers," *Clinical Linguistics & Phonetics*, vol. 28, no. 12, pp. 927–950, 2014.

[7] H. Strik, "ASR-based systems for language learning and therapy," in *Proceedings of the International Symposium on Automatic Detection of Errors in Pronunciation Training*. Stockholm, Sweden: KTH, Computer Science and Communication, jun 2012, pp. 9–14.

[8] C. Cucchiarini, W. Nejjari, and H. Strik, "My pronunciation coach: Improving english pronunciation with an automatic coach that listens," *Language Learning in Higher Education*, vol. 1, no. 2, pp. 365–376, Nov. 2012.

[9] E. Krasnova and E. Bulgakova, *Proceedings of the 16th International Conference on Speech and Computer: SPECOM 2014*, 2014, ch. The Use of Speech Technology in Computer Assisted Language Learning Systems, pp. 459–466.

[10] B. P. de Vries, C. Cucchiarini, S. Bodnar, H. Strik, and R. van Hout, "Spoken grammar practice and feedback in an asr-based call system," *Computer Assisted Language Learning*, vol. 28, no. 6, pp. 550–576, 2015.

[11] P. M. Kato, S. W. Cole, A. S. Bradlyn, and B. H. Pollock, "A video game improves behavioral outcomes in adolescents and young adults with cancer: A randomized trial," *Pediatrics*, vol. 122, no. 2, pp. e305–e317, 2008.

[12] M. Krause, J. Smeddinck, and R. Meyer, "A digital game to support voice treatment for parkinson's disease," in *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '13. New York, NY, USA: ACM, 2013, pp. 445–450.

[13] W. R. Rodríguez, O. Saz, E. Lleida, C. Vaquero, and A. Escartín, "Comunica - tools for speech and language therapy," in *Proceedings of the 2008 Workshop on Children, Computer and Interaction*, 2008.

[14] H. T. Bunnell, D. Yarrington, and J. B. Polikoff, "Star: articulation training for young children," in *Sixth International Conference on Spoken Language Processing, ICSLP / INTERSPEECH*, Nov 2000, pp. 85–88.

[15] C. Vaquero, O. Saz, E. Lleida, J. M. Marcos, and C. Canalís, "Vocaliza: An application for computer-aided speech therapy in spanish language," in *IV Jornadas en Tecnologia del Habla*. Zaragoza, Spain: Grupo de tecnologias de las communicaciones, Universidad de Zaragoza, Nov. 2006, pp. 321–326.

[16] S.-C. Yin, R. C. Rose, O. Saz, and E. Lleida, "Verifying pronunciation accuracy from speakers with neuromuscular disorders," in *INTERSPEECH*. ISCA, Sept 2008, pp. 2218–2221.

[17] O. Saz, S.-C. Yin, E. Lleida, R. Rose, C. Vaquero, and W. R. Rodríguez, "Tools and technologies for computer-aided speech and language therapy," *Speech Communication*, vol. 51, no. 10, pp. 948–967, 2009.

[18] C. T. Tan, A. Johnston, K. Ballard, S. Ferguson, and D. Perera-Schulz, "speak-man: towards popular gameplay for speech therapy," in *Proceedings of The 9th Australasian Conference on Interactive Entertainment: Matters of Life and Death*, no. 28, 2013.

[19] M. B. Mustafa, F. Rosdi, S. S. Salim, and M. U. Mughal, "Exploring the influence of general and specific factors on the recognition accuracy of an asr system for dysarthric speaker," *Expert Systems with Applications*, vol. 42, no. 8, pp. 3924 – 3932, 2015.

[20] Z. A. Benselama, M. Guerti, and M. A. Bencherif, "Arabic speech pathology therapy computer aided system," *Journal of Computer Science*, vol. 3, no. 9, pp. 685–692, 2007.

[21] T. Pellegrini, L. Fontan, J. Mauclair, J. Farinas, C. Alazard-Guiu, M. Robert, and P. Gatignol, "Automatic assessment of speech capability loss in disordered speech," *ACM Transactions on Accessible Computing*, vol. 6, no. 3, pp. 8:1–8:14, May 2015.

[22] C. Middag, "Automatic analysis of pathological speech," Ph.D. dissertation, Ghent University, Belgium, 2012.

[23] E. Yilmaz, M. Ganzeboom, L. Beijer, C. Cucchiarini, and H. Strik, "A dutch dysarthric speech database for individualized speech therapy research," in *Proc. LREC*, may 2016, pp. 792–795.

[24] N. Oostdijk, "The spoken Dutch corpus: Overview and first evaluation," in *Proc. LREC*. LREC, 2000, pp. 886–894.

[25] C. Cucchiarini, J. Driesen, H. Van hamme, and E. Sanders, "Recording speech of children, non-natives and elderly people for HLT applications: the JASMIN-CGN Corpus," in *Proc. LREC*, May 2008, pp. 1445–1450.

[26] M. De Bodt, C. Guns, and G. Van Nuffelen, "NSVO: handleiding," Vlaamse Vereniging voor Logopedie: Herentals, Tech. Rep., 2006.

[27] J. Van de Weijer and I. Slis, "Nasaliteitsmeting met de nasometer," *Logopedie en Foniatrie*, vol. 63, pp. 97–101, 1991.

[28] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *Proc. ASRU*, Dec. 2011.

[29] D. Povey, L. Burget, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. K. Goel, M. Karafiát, A. Rastrow, R. C. Rose, P. Schwarz, and S. Thomas, "The subspace gaussian mixture model - A structured model for speech recognition," *Computer Speech & Language*, vol. 25, no. 2, pp. 404–439, 2011.

An Impulse Sequence Representation of the Excitation Source Characteristics of Nonverbal Speech Sounds

Vinay Kumar Mittal[1] *and B. Yegnanarayana*[2]

[1]Indian Institute of Information Technology Chittoor, Sri City, India
[2]International Institute of Information Technology, Hyderabad, India
[1]vkmittal@iiits.in, [2]yegna@iiit.ac.in

Abstract

Impulse-sequence representation of the excitation source component of *normal* speech signal has been of considerable interest in speech coding research. If a similar representation can be made for *nonverbal* (i.e., nonnormal or nonneutral) speech sounds, that would immensely help in their acoustic analyses and diverse applications. This paper proposes a representation of the excitation source characteristics of *nonverbal speech sounds* signal, in terms of a time-domain sequence of impulses or impulse-like pulses. The nonverbal speech sounds are examined in three categories, namely, emotional speech, paralinguistic sounds and expressive voices. This categorisation is proposed, based upon the degree of rapid changes in pitch of these sounds. A modified zero-frequency filtering *(modZFF)* method is proposed for obtaining an impulse sequence representation of the excitation source component in the acoustic signal of nonverbal speech sounds. Effectiveness of the proposed representation is validated by analysis-by-synthesis approach and perceptual evaluation for Noh singing voice signals. This representation may also be helpful in significant savings in the terms of signal storage and processing requirement, apart from analysis and speech coding of the nonverbal sounds.

Index Terms: nonverbal speech sounds, impulse sequence representation, modified zero-frequency filtering, speech coding

1. Introduction

Assistive technologies can be developed using acoustic cues, that are produced by the human speech production mechanism. For example, analysis of cough sounds may help medical experts in the diagnosis of the type of ailment, the type of infant cry may indicate to mother the cause of cry, or the Unh/Ahan/Hum/Laugh sounds may indicate to psychologists the attention level or attitude etc. Thus there can be a vast range of clinical or other applications possible, in assistive or augmentative roles, using signal processing methods on the acoustic signals of such sounds. But the methods that work well for normal speech, may not work for such sounds. Hence, there is need to develop appropriate signal processing methods, characterize these sounds and develop the systems for assistive applications.

Human speech sounds can be classified into *verbal* and *nonverbal* sounds. *Verbal speech* is normal speech that consists of phonation and linguistic sounds, and mostly follows syntax rules. An articulatory description exists for these reproducible sounds. *Nonverbal speech sounds* carry nonlinguistic information that may be more effective in communication. For example, accent, native dialect, attitude, gestures, moods (interested or indifferent), emotions (happy, sad, angry etc.), articulation and identity. No clear description of articulation exists for these. Their production is mostly involuntary and spontaneous. Based upon the content (verbal/nonverbal), production (involuntary or controlled) and intelligibility, these can be categorised as: emotional speech, paralinguistic sounds and expressive voices [1].

Emotional speech consists of linguistic content and communicates either emotions (shout, anger, sad, fear, happy etc.) or affective states (boredom, interest, surprise etc.). *Paralinguistic sounds* consist of mostly nonlinguistic content (laughter, cry, cough, sneeze, yawn etc.) and communicate a speaker's emotional state or some acoustic-physiological event. These sounds may occur as interspersed with normal speech. *Expressive voices* (e.g., Opera or Noh singing) consist of mostly nonverbal (singing) sounds, mixed with little linguistic content. These are specially trained artistic voices, whose production is voluntarily controlled and involves rapid changes in their excitation characteristics [1]. *Noh* is a Japanese performance art, that involves high emotional expressivity in singing voice [2]. However, these terminologies proposed by the author, may have overlapping semantics and preferences amongst researchers.

Nonverbal speech sounds have few common characteristics. These are *nonsustainable* (i.e., occur for short bursts of time), *nonnormal* (i.e., deviations from normal), form a *continuum* (are nondiscrete) and indicate *humaneness* (help distinguishing between a human and a humanoid). Analysing their production characteristics is a challenging task, because significant changes occur in their excitation source characteristics. An effective representation of their source characteristics can help in a range of applications, such as spotting these in continuous speech, event detection, classification, speaker identification, man/machine discrimination and speech synthesis etc. [3, 4, 5, 6, 7].

Research challenges unique to nonverbal speech sounds can be related to production, databases and classification. *Production*-specific challenges relate to their spontaneity and production-control. *Databases* issues relate to their continuum nature, quality of emoting and reference. *Classification* issues relate to discriminating between normal-nonverbal, spotting nonverbal sounds (in continuous speech) and identifying its category. The nonverbal and normal speech sounds seem to differ in their production characteristics. For example, nonverbal sounds occur in short-bursts of time, with significant changes in their excitation source characteristics and possibly associated changes in the vocal tract system characteristics. Signal processing methods that work well for analysing the normal speech, have limitations for nonverbal speech sounds [1]. Hence, *how to derive the excitation source characteristics from the acoustic signal for nonverbal speech sounds*, is a challenge.

Impulse-sequence representation of the excitation was attempted in speech coders for achieving *low bit-rates of coding* and natural-sounding *voice quality* of synthesized speech.

SLPAT 2016, 7th Workshop on Speech and Language Processing for Assistive Technologies, pages 69–74,
San Francisco, USA, 13 September, 2016.

Speech coders can be categorised as waveform coders, vocoders and hybrid codecs. *Waveform coders* [8, 9] aimed at mimicking the speech waveform, to the best possible extent. *Vocoders* [10, 11] used *linear prediction (LP) coding* [12, 13] or *residual-excited LP (RELP)* [14] that lead to the development of *code-book excited LP* (CELP) [15] codecs. *Hybrid* or *analysis-by-synthesis* codecs aimed at achieving intelligible speech with bit-rates $\leq$ 4 kbps. Excitation source information was represented using *multi-pulse* [9, 16, 17, 18], *regular-pulse* [19], or *CELP* [15] sequences. These approaches differed in estimating the pulse position, amplitude or phase. Hence, *how to represent that excitation source information in terms of a time-domain sequence of impulses for nonverbal sounds*, is second challenge.

Production of *normal* speech sounds reflects the differences in the *locations* of excitation impulses and their relative *amplitudes* [20]. For example, in fricative sounds the impulses occur at random intervals with amplitudes of low strength, but for the vowel-like regions these impulses occur at nearly regular intervals with smooth changes in their amplitudes [21]. In the production of *nonverbal* speech sounds, these impulses are likely to occur at rapidly changing intervals, with significant changes in impulse amplitudes. For example, expressive voices (e.g., Noh singing) have *aperiodicity* in the excitation component due to unequal intervals between successive impulses and unequal strengths of excitation around these [20]. Production of nonverbal speech sounds also involves the amplitude and frequency modulation related to the *voluntary pitch-control* or other *involuntary changes*, whose effect on the pitch perception could be significant [22]. Hence, the third important question is - *how to determine the locations and amplitudes of the impulses that represent the excitation source information in nonverbal sounds?*

This paper explores answers to these three key questions. The excitation source characteristics is represented in terms of a time-domain sequence of impulses, with their relative strengths. The impulse-sequence representation for *normal speech* can be obtained using the *zero-frequency filtering* (ZFF) [23, 24]. But, when pitch period changes rapidly, the ZFF method needs to be modified, in order to capture the subtle variations in the excitation characteristics. These may be related to irregular intervals between epochs and varying strengths of the impulses, e.g., in laughter [25] or expressive voices [20, 22]. Shorter window lengths ($\leq$ one pitch period) may highlight more information for signals having rapid pitch variations, but it is difficult to interpret few epochs sometimes. In order to eliminate the need for selecting an appropriate window length and also to minimize its effect on the derived impulse sequence for nonverbal speech sounds, a *modified zero-frequency filtering (modZFF)* method is proposed. Analysis-by-synthesis approach is adopted for validating the effectiveness of the proposed representation.

This paper is organized as follows. Section 2 reviews existing methods for representing the excitation source information in normal speech. The proposed *modZFF method* for nonverbal speech sounds is described in Section 3. Representation of the excitation source characteristics of different nonverbal speech sounds is discussed in Section 4. Validation of the proposed method is carried out in Section 5, using analysis-by-synthesis approach. Section 6 gives a summary and scope of further work.

2. Existing methods for normal speech

Excitation source characteristics in *normal* speech signal was extracted using different approaches in speech coding methods. (a) *Waveform coders* used transform coders [13], pulse-code modulation (PCM), differential PCM, delta-modulation [8] or adaptive predictive coding [26], to reproduce the speech with high voice quality and minimum distortion. But speech coding bit-rate was high ($\geq$ 16 kbits/sec). (b) *LPC Vocoders* used LP coders (all-pole filters) [13], voice-excited vocoders with pulse-sequence/noise for voiced/unvoiced excitation [27], or RELP vocoders with LP residual for the excitation [14]. Aim was to reduce coding bit-rate $\leq$ 2.4 kbits/sec with intelligible speech, but it was not natural-sounding. (c) *Hybrid (analysis-by-synthesis) codecs* [9, 28, 29, 30] aimed at high intelligibility of synthesized speech with coding bit-rate $\leq$ 4.8 kbits/sec.

Hybrid codecs have two parts, encoder and decoder [28]. *Encoder* consists of synthesis filter, error-weighting and error-minimisation blocks. It analyses each 20 ms frame of signal $s(n)$ by synthesizing multiple approximations to it, and then transmits to decoder the synthesis filter parameters and the excitation sequence $u(n)$. *Decoder* synthesizes the signal $\tilde{s}(n)$, by passing the excitation $u(n)$ through a *synthesis* (*all-pole*) filter $H(z) = \frac{1}{A(z)}$, with $A(z) = 1 - \sum_{i=1}^{p} a_i z^{-i}$ as prediction error filter [9, 28]. Excitation $u(n)$ can be chosen in 3 ways, to give minimum weighted-error $e(n)$ between the original $s(n)$ and the synthesized speech $\tilde{s}(n)$. *Multi-pulse excited* codecs [9] model the ideal excitation by 8 nonzero pulses for every 10 ms frame, and use suboptimal methods to determine pulse positions and amplitudes. *Regular-pulse excited* (RPE) codecs [19] use nonzero pulses (*regularly spaced* at fixed interval) for excitation, needing to determine only the first pulse position and amplitudes of all pulses. *CELP codecs* [15] use for excitation an entry in a vector quantized *code-book* and a gain term, with low bit-rate. MPE codecs have lesser computational complexity than RPE codecs.

2.1. All-pole model of excitation in LPC vocoders

(i) Generic pole-zero model [31]: For a discrete time-series signal $s[n]$, the system output is *predicted* from past outputs and present inputs, as $s[n] = -\sum_{k=1}^{p} a_k\, s[n-k] + G \sum_{l=0}^{q} b_l\, u[n-l]$, where $b_0 = 1$, a_k are system parameters, G is gain and $u[n]$ the unknown input sequence. Taking its z transform, we get $H(z) = G\, \frac{(1+\sum_{l=1}^{q} b_l\, z^{-l})}{(1+\sum_{k=1}^{p} a_k\, z^{-k})}$, where $H(z)$ $\left(= \frac{S(z)}{U(z)}\right)$ is *transfer function* of the system, i.e., the *general pole-zero model*, $U(z)$ is z transform of $u[n]$ and $S(z)$ is z transform of $s[n]$.

(ii) All-pole model [32, 31]: Signal given by past output values and input $u[n]$ is $s[n] = -\sum_{k=1}^{p} a_k\, s[n-k] + G\, u[n]$, where G is gain. Taking its z transform, we get $H(z) = \frac{G}{(1+\sum_{k=1}^{p} a_k\, z^{-k})}$, where $H(z)$ is an *all-pole transfer function*.

(iii) Method of Least Squares [31]: For unknown input $u[n]$, the output can be *predicted* as $\tilde{s}[n] = -\sum_{k=1}^{p} a_k\, s[n-k]$, and *error (residual)* is given by $e[n] = s[n] - \tilde{s}[n] = s[n] + \sum_{k=1}^{p} a_k\, s[n-k]$. A solution to this excitation representation problem is *multi-pulse excitation (MPE)* model [9].

2.2. MPE model of the excitation

In MPE, an all-pole LPC synthesizer filter $H(z)$ is excited by a sequence of pulses at positions $t_1, t_2, ..., t_n, ...$ with amplitudes $\alpha_1, \alpha_2, ..., \alpha_n, ...$ [9]. This desired impulse-sequence ($d[n]$) excites the filter to produce *synthesized* output $\tilde{s}[n]$. It is passed from a low-pass filter to produce the *reconstructed* speech $\hat{s}(t)$.

(i) Determining the MPE input to the LPC all-pole synthesis filter $H(z)$: The desired MPE ($d[n] = \sum_{k=-\infty}^{\infty} r[k]\, h[n-k]$) is determined by modeling the LPC residual $r[n]$, to minimize the *weighted-mean square error* ϵ computed from the difference

$e[n]$ between original speech $s[n]$ and synthesized speech $\tilde{s}[n]$.

(ii) Transfer function of error-weighting filter [9]: The frequency-weighted error is $\epsilon = \int_0^{f_s} |S(f) - \hat{S}(f)|^2 W(f) df$, where $S(f)$ and $\hat{S}(f)$ are Fourier transforms of $s(t)$ and $\hat{s}(t)$, respectively, and $W(f)$ is a *weighting function*. Transfer function of *error-weighting filter* is $W(z) = \frac{(1-\sum_{k=1}^{p} a_k z^{-k})}{(1-\sum_{k=1}^{p} a_k \gamma^k z^{-k})}$. Parameter γ controls the error weight, i.e., $W(z) = 1 - P(z)$ for $\gamma = 0$, and $W(z) = 1$ for $\gamma = 1$, (typically $\gamma = 0.8$).

(iii) Key objective in MPE-LPC model [33]: To find a sequence $u[n]$ and filter parameters $\{a_k\}$, so as to minimize the perceptually weighted mean-squared error $\bar{e^2}[n]$ w.r.t. the reference $s[n]$. *Synthesized signal* $\tilde{s}[n]$, for predictor order p, is $\tilde{s}[n] = \sum_{k=1}^{p} a_k \tilde{s}[n-k] + u[n]$. To minimize the mean-squared error $\bar{e^2}[n] = \sum_n (s[n] - \tilde{s}[n])$, different approaches determine the amplitudes and positions of impulse-like pulses in $u[n]$.

2.3. Estimating the amplitudes of pulses in MPE

(i) Sequential pulse placement (no re-optimization) [29]: The mean-squared weighted error for N_p excitation pulses is $\bar{e^2} = \sum_n (d_n - A_m h_{n-m})^2$, where h_{n-m} is response of filter $H(\gamma z)$ for the first impulse at position m with amplitude A_m. The desired excitation is $d[n]$. The *optimal pulse amplitude* is $\hat{A}_m = \frac{\sum_n d_n h_{n-m}}{\sum_n h_{n-i} h_{n-j}}$. Denoting the vector of cross-correlation terms in numerator by α_m and matrix of correlation terms in denominator by ϕ_{ij}, the optimal amplitude is $\hat{A}_m = \frac{\alpha_m}{\phi_{mm}}$, where $\alpha_m = \sum_n d_n h_{n-m}$ and $\phi_{ij} = \sum_n h_{n-i} h_{n-j}$. Now error $\bar{e^2} = \sum_n d_n^2 - \frac{\alpha_m^2}{\phi_{mm}}$ depends on only position m of the pulse. Best position for a pulse is for that m, for which $\frac{\alpha_m^2}{\phi_{mm}}$ is maximum. *Optimal position* for *next pulse* is $d'_n = d_n - \hat{A}_m h_{n-m}$, and $\alpha'_m = \alpha_m - \hat{A}_{\hat{m}} \phi_{\hat{m}m}$. Likewise, positions and amplitudes for all pulses can be found *sequentially*.

(ii) Re-optimization after having 'all' pulse positions [29]: Using limits of error $\bar{e^2}$ as $-\infty$ to $+\infty$, the optimal pulse amplitude $\hat{A}_m$ depends on best pulse-position m, for which $|\alpha_m|$ is maximized and ϕ_{mm} is minimized. Mean square error for *all* n_p pulses, after getting positions upto m_i, is $\bar{e^2} = \sum_n (d_n - \sum_{i=1}^{n_p} A_{m_i} h_{n-m_i})^2$. Differentiating it w.r.t. all pulse amplitudes A_{m_i}, we get $\sum_n (\sum_{i=1}^{n_p} h_{n-m_i} . \sum_{i=1}^{n_p} h_{n-m_i} . A_{m_i}) = \sum_n (d_n \sum_{i=1}^{n_p} h_{n-m_i})$. Replacing the cross-correlation terms α_{m_i} and correlation terms $\phi_{m_i m_i}$, we get a *set of simultaneous equations*: $\left[\phi_{m_i m_j}\right] \left[\ \hat{A}_{m_i}\ \right] = \left[\ \alpha_{m_i}\ \right]$, where $i, j = 1, 2, ..., n_p$, $\hat{A}_{m_i}$ is optimal amplitude at position m_i and n_p is number of pulses in N samples block. It can be solved by Cholesky decomposition of the *correlation matrix* of elements ϕ_{ij}. Pulse-amplitude re-optimization can be carried out after having *'all' pulse positions* [29] or *'each' pulse position* [34].

2.4. Estimating the positions of pulses in MPE

(i) Pulse correlation method [29]: Best location for an excitation pulse is m, at which the amplitude $\hat{A}_m$ is optimal and error $\bar{e^2}$ minimum. Impulse response of the synthesis filter $H(\gamma z)$ dies-off quickly due to the factor γ, hence this part can be truncated. In *autocorrelation* analysis the correlation term (ϕ_{ij}) is generated by filtering $\{h_n\}$, using recursive synthesis filter. In *covariance* multi-pulse analysis the correlation $\{\phi_{ij}\}$ is defined recursively as $\phi_{i-1,\,j-i} = \phi_{ij} + h_{N-i}\, h_{N-j}$. Initial cross-correlation ϕ_{ij} can be computed using synthesis filter ($\{d_n\}$).

(ii) Pitch-interpolation method [28]: In this, the pulse-position is obtained by interpolating the pitch-period, to minimize the error $\bar{e^2}[n]$. Synthesis filter parameters $\{a_k\}$ are used with an error weighting filter $H(\gamma z)$, to reduce the perceptual distortion. Use of maximum cross-correlation α_m gives the optimum location m_i of i^{th} pulse, determined by finding maximum absolute amplitude A_m for pulse at location m_i. $A_{m_i} = \frac{\alpha_{hs}(m_i) - \sum_{j=1}^{i-1} A_{m_j} . \phi_{hh}(|m_j - m_i|)}{\phi_{hh}(0)}$, where $1 \le m_i, m_j \le N$, N is number of samples, and $\alpha_h(m_i)$ is cross-correlation between weighted speech $s[n]$ and impulse-response $h[n-m]$. The ϕ_{ij} is autocorrelation of response $h[n-m]$, and A_m are amplitudes of pulses determined upto i^{th} location. The *correlation* terms α_{hs} and *autocorrelation* terms ϕ_{hh} are: $\alpha_{hs}(m_i) = \sum_n s[n] h[n-m_i]$, $\phi_{hh}(ij) = \sum_n h_{n-m_i} h_{n-m_j}$.

(iii) SPE-CELP method [30]: It uses *single-pulse excitation (SPE)* instead of multi-pulse, in a pitch-period. The CELP coding [15] does not provide appropriate periodicity of pulses in synthesized speech for bit-rates $\le$ 4 kbits/sec, because small code-book size and coarse quantization of gain factor cause large fluctuations in the spectral characteristics between two periods. In SPE-CELP [30] a LP coder first classifies speech into periodic and non-periodic intervals, then non-periodic speech is synthesized like in CELP coding [28]. Periodic speech is synthesized using single-pulse excitation, and using an algorithm to determine the *pitch-markers* in short blocks of periodic speech.

Speech coding methods have focused at representing the excitation information in normal speech signal in the terms of a sequence of impulse-like pulses, either to reduce the bit-rate of speech coding or to increase the voice quality of synthesized speech. This impulse-sequence representation of the excitation information for nonverbal speech sounds is not yet attempted, to the best of our knowledge. It is proposed in the next section.

3. Proposed method for nonverbal sounds

Speech coding methods focus at representing the excitation in terms of a sequence of impulse-like pulses, for normal speech. An impulse-sequence representation of the excitation information for nonverbal sounds signals is proposed in this section.

The ZFF method [23, 24] has two limitations when applied for deriving the impulse sequence representation for nonverbal speech sounds: (i) shorter window length would be required for trend removal and (ii) impulse sequence for aperiodic signals may be affected by the choice of shorter window length. Both these limitations are addressed in the recently proposed *modified zero-frequency filtering (modZFF)* method by using gradually reducing window lengths, instead of a fixed window length, for the trend removal operation [22]. Key steps involved in the proposed *modZFF* method are as follows:

1. Preprocess the input signal ($s[n]$) by downsampling it to 8 kHz, smoothen over m sample points and then upsample back to original sampling frequency (f_s) of signal.
2. Get differenced signal ($\tilde{x}[n]$) from the pre-processed signal ($s_p[n]$), to further obtain a zero-mean signal ($\hat{x}[n]$).
3. Pass this $\hat{x}[n]$ through a cascade of two ideal digital resonators at 0 Hz, i.e., $y[n] = \sum_{k=1}^{4} a_k y[n-k] + \hat{x}[n]$, where $a_1 = +4$, $a_2 = -6$, $a_3 = +4$, $a_4 = -1$.
4. Remove the trend in output of the cascaded ZFRs ($y[n]$), using gradually reducing windows of lengths 20 ms, 10 ms, 5 ms, 3 ms, 2 ms and 1 ms in successive stages, by subtracting the local mean, in order to highlight the excitation source information in the signal better. Output of each stage (window size of $2N+1$ sample points) is $\hat{y}[n]) = y[n] - \bar{y}[n]$, where $\bar{y}[n] = \frac{1}{2N+1} \sum_{n=-N}^{N} y[n]$ is the local mean computed over the window. The re-

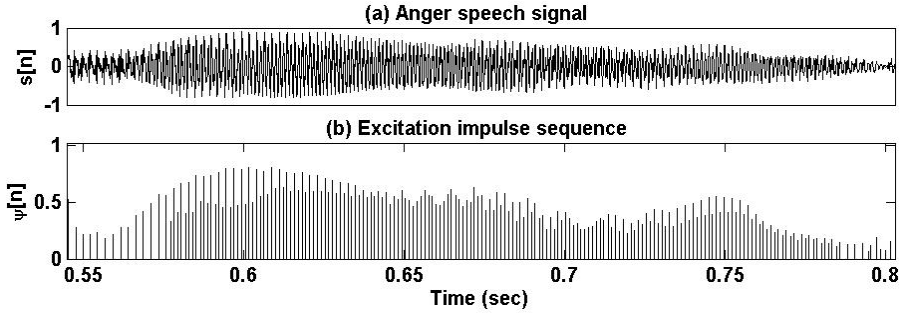


Figure 1: (a) *Emotional (anger) speech* signal (for text "your") and (b) excitation impulse sequence from modZFF output.

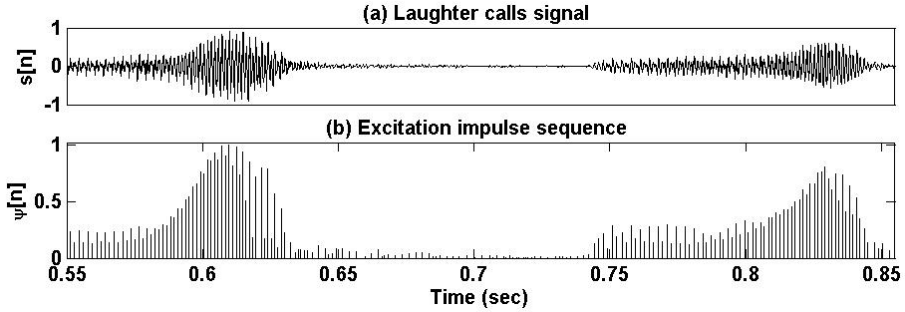


Figure 2: (a) *Paralinguistic sounds* (2 *laugh* calls) signal waveform and (b) excitation impulse sequence from modZFF output.

sultant final trend removed output is called the *modified zero frequency filtered (modZFF)* signal ($z_m[n]$) [22].

5. The positive to negative going zero-crossings of the *modZFF* signal ($z_m[n]$) give locations of impulses (epochs).
6. The slope of the *modZFF* signal ($z_m[n]$) around each of these locations indicates relative *strength of excitation (SoE)* there, and amplitudes of impulses in the sequence.

This sequence represents the excitation source characteristics. The preprocessing step used here for down-sampling to 8 kHz and then upsampling back to the original sampling frequency helps in reducing the number of spurious impulses [22]. The locations and amplitudes of the impulses in SoE based impulse-sequence representation obtained for nonverbal speech sounds, using this modZFF method are not sensitive to the choice of last window length in 1.0 ms to 2.5 ms range [22].

The *modZFF method* helps deriving the impulse sequence to represent the excitation source component of nonverbal speech sound signal, with negligible spurious impulses. But this sparse representation leads to significant savings in terms of storage space and processing requirement.

4. Representing the source characteristics

The *modZFF method* helps deriving the impulse sequence representation of the excitation source component of nonverbal sounds signals. The amplitudes of impulses are the *SoE* at the respective impulse locations. Excitation impulse sequences obtained for *anger (emotional* speech), *laughter* and *cry* (*paralinguistic* sounds), and *Noh* singing (*expressive* voices) are illustrated in figures Fig. 1(b), Fig. 2(b), Fig. 3(b) and Fig. 4(b), respectively. It may be observed from these figures that the impulse sequence representation of the excitation source component in acoustic signals for different nonverbal (nonnormal) sounds seem to have adequate number of impulses and no spurious impulses (i.e., noise-like small magnitude impulses). This indicates efficacy of the *modZFF method* in obtaining the excitation impulse sequence for nonverbal speech sounds. Similar excitation impulse sequences are obtained for the other semi-natural/natural data [35, 25, 2] used in this study.

Table 1: *Average savings in the terms of storage space:* i.e., (%) of sample points saved, for different *nonverbal speech sounds*.

Sl.#	(a) *Acoustic Sound Type*	(b) *Saving (%)*
1.	Emotional speech	97.44
2.	Paralinguistic sounds	98.82
3.	Expressive voices	99.19
	Average	*98.48*

This proposed representation of the excitation source information also results in the savings of storage space, as given in Table 1. Savings in the terms of average number of sample points, is computed for 3-5 files of each of the three types of acoustic signals of nonverbal speech sounds examined in this study. The results appear interesting. The relative space saving (like compression) is less (97.44%) for emotional speech, more (98.82%) for paralinguistic sounds, and further more (99.19%) for expressive voices. It could possibly be related to the relative presence of linguistic speech content and expressivity.

5. Validation by analysis-by-synthesis

Effectiveness of the proposed impulse sequence representation of the excitation source information in nonverbal speech sounds

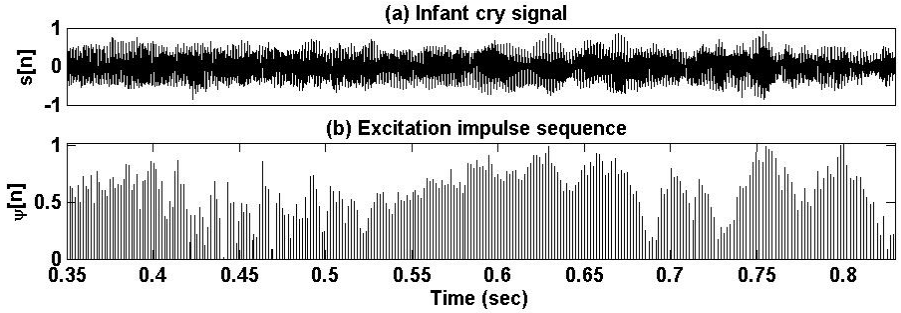


Figure 3: (a) *Paralinguistic sounds* (infant *cry*) signal waveform and (b) excitation impulse sequence from modZFF output.

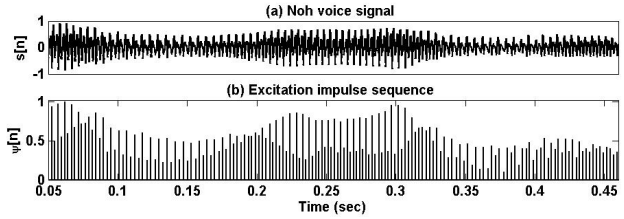


Figure 4: (a) *Expressive voices* (*Noh* singing) signal waveform and (b) excitation impulse sequence from modZFF output.

is validated using analysis parameters' based synthesis and perceptual listening tests. Nonverbal sounds signals for the Noh singing voice are synthesized, by exciting the original vocal tract system characteristics with four different excitation. Three impulse sequences, having impulses at actual intervals, with (i) unit amplitude of impulses (UImps), (ii) amplitudes as per Liljencrants-Fant model between impulses (LF Model), and (iii) respective SoE amplitudes of impulses (SoEImps) are used for the excitation. In the 4^{th} case, LP residual (LPRes) is used for the excitation. The impulse sequences and the SoE are derived using the *modZFF* method. The acoustic signal is synthesized by exciting a 12^{th} order LP model, computed at impulse locations (epochs) for Noh voice signals down-sampled to 8 kHz for each case. Noh voice signal is chosen because of its more rapid changes in pitch, than paralinguistic sounds and emotional speech. Perceptual listening tests are carried out for each case, by 10 subjects (7 male, 3 female). Scores on a scale of 1 to 5 are given by each subject, for perceptual closeness between the original Noh voice and the corresponding synthesized signal. Then the average scores are computed for each case.

In Table 2, the results are given for these 4 cases, in columns (a)-(d), respectively. It may be observed that the synthesized acoustic signal using the SoE impulse sequence for excitation (column (c)) sounds relatively better in comparison to the other two sequences (columns (a) and (b)). The synthesized acoustic signal using the impulse sequences with location information, and amplitude as UImps (column (a)) or LF Model (column (b)), is still intelligible. It indicates that *the impulse location information is relatively more important than the amplitudes information*, and carries more content. The amplitudes of impulses are not very critical. But the perceptual scores are better if the SoE impulse sequence (column (c)) is used for the excitation. It indicates effectiveness of the modZFF method in obtaining the SoE impulse sequence. However, naturalness is lost if the excitation consists of only a sequence of impulses, as it does not have other residual information. This is validated by the synthesized acoustic signal using LP residual for excitation (column (d)). This signal sounds relatively much better and is quite close to the original Noh voice, because the residual information in-between the impulses is also present in this case.

Table 2: *Results of perceptual listening test*: average scores for perceptual closeness between original Noh voice and the speech synthesized using excitation as: impulse-sequences having epoch locations with (a) unit amplitudes, (b) LF Model, (c) SoE amplitudes, and (d) LP residual. The three Noh voice segments considered correspond to Figures 1, 2 and 3 in [2].

Noh voice segment	(a) *UImps*	(b) *LF-Model*	(c) *SoEImps*	(d) *LPRes*
Noh voice segment 1	1.51	2.15	2.42	4.39
Noh voice segment 2	1.61	1.85	2.33	4.69
Noh voice segment 3	1.71	1.95	2.32	4.68
Average	*1.61*	*1.98*	*2.36*	*4.59*

6. Summary and conclusion

Nonverbal speech sounds have subharmonics and aperiodic content in their excitation source component, it was examined earlier. Human perception takes into account all likely values of the changing pitch frequency in these regions. If these relatively important nonuniform intervals and nonuniform amplitudes in the excitation impulse sequence are made uniform, then valuable information is lost. Hence, key challenge lies in estimating the locations and relative amplitudes of these impulse-like pulses in the sequence representing the excitation information. Speech coding methods have focused at obtaining the excitation impulse sequence only for normal speech. This paper proposes an impulse-sequence representation of the excitation source information in acoustic signals of nonverbal speech sounds using a recently proposed *modified zero-frequency filtering* method.

Nonverbal sounds are examined in three categories, namely, emotional speech, paralinguistic sounds and expressive voices. Anger speech, laughter and cry, and Noh singing voices are examined respectively for these three categories. A time-domain impulse-sequence representing the excitation information in the signal, for each case, is obtained using the *modZFF method*. Validation of the proposed representation is carried out by analysis-synthesis and perceptual evaluation.

This representation of excitation information in nonverbal speech sounds signal should be helpful in their analysis, representation and speech-coding. It can also lead to significant savings in-terms of such signals' storage and processing requirement, with minimal loss or intelligibility of the reproduced/synthesized sounds, towards development of assistive technologies for wider applications.

Acknowledgement

The authors are thankful to Prof. Osamu Fujimura and Prof. Hideki Kawahara for providing the data of Noh singing voice.

7. References

[1] V. K. Mittal, "Analysis of Nonverbal Speech Sounds," Ph.D. dissertation, International Institute of Information Technology, Hyderabad, India, Nov. 2014, (No. IIIT/TH/2014/54).

[2] O. Fujimura, K. Honda, H. Kawahara, Y. Konparu, M. Morise, and J. C. Williams, "Noh voice quality," *Logopedics Phoniatrics Vocology*, vol. 34, no. 4, pp. 157–170, 2009.

[3] W. Ruch and P. Ekman, "The Expressive Pattern of Laughter," *Emotion, Qualia, and Consciousness*, pp. 426–443, 2001, edited by A. W. Kaszniak (Word Scientific, Tokyo).

[4] E. Mower, M. J. Mataric, and S. Narayanan, "A framework for automatic human emotion classification using emotion profiles," *Trans. Audio, Speech and Lang. Proc.*, vol. 19, no. 5, pp. 1057–1070, Jul. 2011.

[5] W. Hamza, R. Bakis, E. M. Eide, M. A. Picheny, and J. F. Pitrelli, "The IBM expressive speech synthesis system," in *Proc. of the 8th International Conference on Spoken Language Processing, Jeju, Korea*, 2004, pp. 14–16.

[6] L. S. Kennedy and D. P. W. Ellis, "Laughter detection in meetings," in *Proc. NIST ICASSP 2004 Meeting Recognition Workshop*, Montreal, Mar. 2004, pp. 118–121.

[7] E. Lasarcyk and J. Trouvain, "Imitating conversational laughter with an articulatory speech synthesis," in *Proc. of the Interdisciplinary Workshop on The Phonetics of Laughter*, Aug. 4-5 2007, pp. 43–48.

[8] N. S. Jayant, "Digital coding of speech waveforms: PCM, DPCM and DM Quantization," in *Proc. IEEE*, vol. 62, May 1974, pp. 611–632.

[9] B. S. Atal and J. R. Remde, "A new model of LPC excitation for producing natural-sounding speech at low bit rates," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, May 1982, pp. 614–617.

[10] M. R. Schroeder, "Vocoders: Analysis and Synthesis Speech," in *Proc. IEEE*, ser. 5, vol. 54, 1966, pp. 720–734.

[11] J. L. Flanagan, *Speech Analysis Synthesis and Perception*, 2nd ed. Springer-Verlag, 1972.

[12] J. E. Markel and A. H. Gray, *Linear Prediction of Speech*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 1982.

[13] J. D. Markel and A. H. Gray, "A Linear Prediction Vocoder Simulation Based upon the Autocorrelation Method," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-22, no. 2, pp. 124–134, April 1974.

[14] C. K. Un and D. T. Magill, "The Residual-Excited Linear Prediction Vocoder with Transmission Rate Below 9.6 kbits/s," *IEEE Trans. on Communications*, vol. 23, no. 12, pp. 1466–1474, 1975.

[15] M. Schroeder and B. Atal, "Code-excited linear prediction (CELP): High-quality speech at very low bit rates," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '85*, vol. 10, April 1985, pp. 937–940.

[16] B. S. Atal and B. E. Caspers, "Periodic repetition of multi-pulse excitation," *The Journal of the Acoustical Society of America*, vol. 74, no. S1, pp. S51–S51, 1983.

[17] S. Singhal and B. Atal, "Improving performance of multi-pulse LPC coders at low bit rates," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 9, 1984, pp. 9–12.

[18] B. Caspers and B. Atal, "Role of multi-pulse excitation in synthesis of natural-sounding voiced speech," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '87*, vol. 12, 1987, pp. 2388–2391.

[19] P. Kroon, E. F. Deprettere, and R. Sluyter, "Regular-pulse excitation–a novel approach to effective and efficient multipulse coding of speech," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 34, no. 5, pp. 1054–1063, 1986.

[20] V. K. Mittal and B. Yegnanarayana, "Significance of aperiodicity in the pitch perception of expressive voices," in *INTERSPEECH 2014*, Singapore, Sep. 2014, pp. 504–508.

[21] V. K. Mittal, B. Yegnanarayana, and P. Bhaskararao, "Study of the effects of vocal tract constriction on glottal vibration," *The Jr. of the Acoust. Soc. of Am.*, vol. 136, no. 4, pp. 1932–1941, 2014.

[22] V. K. Mittal and B. Yegnanarayana, "Study of characteristics of aperiodicity in Noh voices," *The Jr. of the Acoust. Soc. of Am.*, vol. 137, no. 6, pp. 3411–3421, 2015.

[23] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1602–1613, 2008.

[24] B. Yegnanarayana and K. S. R. Murty, "Event-based instantaneous fundamental frequency estimation from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 614–624, 2009.

[25] V. K. Mittal and B. Yegnanarayana, "Analysis of production characteristics of laughter," *Computer Speech & Language*, vol. 30, no. 1, pp. 99–115, 2015.

[26] B. Atal and M. R. Schroeder, "Predictive coding of speech signals and subjective error criteria," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 27, no. 3, pp. 247–254, 1979.

[27] M. R. Schroeder, "Recent Progress in Speech Coding at Bell Telephone Laboratories," in *Proc. 3rd Int. Congress on Acoustics*. Elsevier Publishing Co, Amsterdam, 1961, pp. 201–210.

[28] K. Ozawa and T. Araseki, "Low bit rate multi-pulse speech coder with natural speech quality," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '86.*, vol. 11, 1986, pp. 457–460.

[29] M. Berouti, H. Garten, P. Kabal, and P. Mermelstein, "Efficient computation and encoding of the multipulse excitation for LPC," in *Proc. IEEE Int.'l Conf. on Acoustics, Speech and Signal Processing, ICASSP '84*, vol. 9, 1984, pp. 384–387.

[30] W. Granzow, B. Atal, K. Paliwal, and J. Schroeter, "Speech coding at 4 kb/s and lower using single-pulse and stochastic models of LPC excitation," in *Proc. Int.'l Conf. on Acoustics, Speech and Sig. Proc., ICASSP'91*, vol. 1, 1991, pp. 217–220.

[31] J. Makhoul, "Linear prediction: A tutorial review," *IEEE Transaction*, vol. 63, pp. 561–580, Apr. 1975.

[32] B. S. Atal and S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," *The Jr. of the Acoust. Soc. of Am.*, vol. 50, no. 2B, pp. 637–655, 1971.

[33] M. Fratti, G. A. Mian, and G. Riccardi, "An Approach to Parameter Reoptimization in Multipulse-Based Coders," *IEEE Trans. on Speech and Audio Proc.*, vol. 1, no. 4, pp. 463–465, Oct. 1993.

[34] S. Singhal, "Optimizing pulse amplitudes in multipulse excitation," *The Journal of the Acoustical Society of America*, vol. 74, no. S1, pp. S51–S51, 1983.

[35] K. S. Reddy, P. Gangamohan, V. K. Mittal, and B. Yegnanarayana, "Naturalistic Audio-Visual Emotion Database," in *Proc. 11th ICON 2014*, vol. 1, Goa, 2014, pp. 175–182.

Dysarthric Speech Modification Using Parallel Utterance Based on Non-negative Temporal Decomposition

Ryo Aihara, Tetsuya Takiguchi, and Yasuo Ariki

Graduate School of System Informatics, Kobe University, 1-1, Rokkodai, Nada, Kobe, Japan

aihara@me.cs.scitec.kobe-u.ac.jp, {takigu, ariki}@kobe-u.ac.jp

Abstract

We present in this paper a speech modification method for a person with dysarthria resulting from athetoid cerebral palsy. The movements of such speakers are limited by their athetoid symptoms, and their consonants are often unstable or unclear, which makes it difficult for them to communicate. In this paper, duration and spectral modification using Non-negative Temporal Decomposition (NTD) is applied to a dysarthric voice. F0 is also modified by using linear-transformation. In order to confirm the effectiveness of our method, objective and subjective tests were conducted, and we also investigated the relationship between the intelligibility and individuality of dysarthric speech.

Index Terms: speech modification, dysarthria, Non-negative Temporal Decomposition

1. Introduction

Dysarthria refers to a kind of speech disorder resulting from disturbances in the form or function of the speech mechanism. Some nervous system diseases, such as Parkinson's disease or amyotrophic lateral sclerosis (ALS), produce motor paralysis which results in dysarthric speech.

In this paper, we focused on a person with dysarthria resulting from the athetoid type of cerebral palsy. Cerebral palsy is a non-progressive disorder of movement, and most people with cerebral palsy are born with the athetoid type. About two babies in 1,000 are born with cerebral palsy [1]. Cerebral palsy results from damage to the central nervous system, and the damage causes movement disorders. Three general times are given for the onset of the disorder: before birth, at the time of delivery, and after birth. Cerebral palsy is classified into the following types: 1) spastic, 2) athetoid, 3) ataxic, 4) atonic, 5) rigid, and a mixture of these types [2].

Athetoid symptoms develop in about 10-15% of people with cerebral palsy [1]. In the case of a person with this type of dysarthria, his/her movements are sometimes more unstable than usual. That means their utterances (especially their consonants) are often unstable or unclear due to the athetoid symptoms. Athetoid symptoms also restrict the movement of their arms and legs. Most people with athetoid cerebral palsy cannot communicate by sign language or writing, so there is great need for voice systems for them.

In [3], we proposed robust feature extraction based on principal component analysis (PCA), which has more stable utterance data, instead of DCT. In [4], we used multiple acoustic frames (MAF) as an acoustic dynamic feature to improve the recognition rate of a person with dysarthria, particularly for speech recognition using dynamic features only. In spite of these efforts, the recognition rate of dysarthric speech. is still lower than that of non-dysarthric speech. The recognition rate using a speaker-independent model, which is trained by non-dysarthric speech, is 3.5% [3]. This recognition rate suggests that for people who have not communicated with a person with athetoid cerebral palsy, it will be very hard for them to understand what that person is trying to say.

Text-to-speech synthesis (TTS) has been applied to a person with dysarthria in recent years. Veaux *et al.* [5] used HMM-based speech synthesis to reconstruct the voice of individuals with degenerative speech disorders resulting from ALS. Yamagishi *et al.* [6] proposed a project which is named "Voice Banking and reconstruction". In that project, various types of voices are collected, and they proposed TTS for ALS using that database. We also proposed TTS for a person with dysarthria using HMM-based speech synthesis [7]. However, in general, TTS systems need a large amount of training data. In [7], we used more than 500 sentences of dysarthric speech to construct a speaker-dependent model.

Voice conversion (VC) has also been applied to dysarthric speech. The difference between TTS and VC is that TTS needs text input to synthesize speech, whereas VC does not need text input. In [8], we proposed VC system for dysarthric speech and improved the intelligibility of dysarthric words. The amount of the training data for a VC system is less than that for a TTS system; however, more than 200 words, or 50 sentences, are used to construct dysarthric speech model. A large amount of the training data is a high hurdle for practical use of, especially for people with athetoid cerebral palsy.

Speech modification systems for dysarthric speech, that are different from TTS or VC have also been proposed. In this paper, speech modification refers to a kind of voice transformation, which transforms an input labeled speech signal by performing a detailed speech analysis. Kain *et al*. [9] proposed speech modification for the vowel portion of dysarthric speech. Rudzicz [10] proposed a speech modification method for people with dysarthria based on the observations from the database. In general, speech modification needs less training data than TTS. Moreover, with a speech modification system, it is easier to preserve the speaker individuality of dysarthric speech than VC with a system.

This paper proposes a dysarthric speech modification system using parallel utterances in order to improve the intelligibility of dysarthric utterances. Non-negative Temporal Decomposition (NTD) [11] has been proposed in the field of speech coding and it is applied to the rhythm conversion of non-native English. We applied NTD to dysarthric speech. Duration (rhythm) of dysarthric speech is transformed into that of parallel non-dysarthric speech. The consonants of dysarthric speech are also

SLPAT 2016, 7th Workshop on Speech and Language Processing for Assistive Technologies, pages 75–79,
San Francisco, USA, 13 September, 2016.

replaced with the consonants of non-dysarthric speech based on NTD. F0 is also modified by using linear-transformation. The effectiveness of our method is evaluated by using mean opinion score (MOS) [12] test, and we investigated the relationship between the intelligibility and individuality of dysarthric speech.

The rest of this paper is organized as follows: In Section 2, the summary of the NTD algorithm is described. In Section 3, our proposed method is explained. In Section 4, the experimental data are evaluated, and the final section is devoted to our conclusions.

2. Non-negative Temporal Decomposition

In NTD, the i-th dimensional spectrum, $v_i(t)$ of time t is decomposed into spectral event basis $w_{i,l}$ and activity $h_l(t)$. The problem of NTD is defined as follows:

$$\begin{aligned} min \quad & \sum_{l=2}^{L}\sum_{t=t_{l-1}}^{t_l}\sum_{i=1}^{I}(v_i(t)-w_{i,l}h_l(t)-w_{i,l-1}h_{l-1}(t))^2 \\ s.t. \quad & h_l(t)\geq 0, h_{l-1}(t)\geq 0 \\ & w_{i,l}\geq 0, w_{i,l-1}\geq 0 \\ & h_l(t)+h_{l-1}(t)=1 \quad for \quad \forall t,i,l \end{aligned} \tag{1}$$

where l denotes the number of bases and t_l denotes the event timing of l-th basis. By applying the last constraint, activities are restricted to the range $[0, 1]$.

(1) is rewritten into the cost function as follows:

$$\begin{aligned} & d(\mathbf{V}, \mathbf{WH}) \\ = & \sum_{l=2}^{L}\sum_{t=t_{l-1}}^{t_l}\sum_{i=1}^{I}(v_i(t)-w_{i,l}h_l(t)-w_{i,l-1}h_{l-1}(t))^2 \\ & +\alpha\sum_{l=2}^{L}\sum_{t=t_{l-1}}^{t_l}(h_l(t)+h_{l-1}(t)-1)^2 \end{aligned} \tag{2}$$

where $1 = t_1 < t_2 < \cdots < t_l = T$. The second term of (2) is a penalty term to satisfy $h_l(t) + h_{l-1}(t) = 1$ with α as is its weight.

(2) is minimized by iteratively updating (3) - (5), which is shown at the top of the next page. These updating rules are derived in [11]. In [11], line spectral pair (LSP) is used as a spectral feature; however, we use a magnitude spectrum to estimate the event basis and activity more precisely. Moreover in [11], each event basis corresponds to a single phoneme. In order to estimate the event basis and activity more precisely, 3 event bases are extracted from a single phone.

3. Modification of Dysarthric Speech

3.1. Flow of our proposed method

Fig. 1 shows the flow of our speech modification process. First, a dysarthric utterance and a non-dysarthric utterance, which is parallel to the dysarthric utterance, are labeled by using HMM-based forced alignment. Here, parallel means that the utterances are spoken by different speakers, but the text is the same. Then we extract spectral features, F0, and aperiodic features from the parallel utterances by using STRAIGHT analysis [13]. The duration and extracted spectral features are modified by using NTD. The extracted F0 is also modified using linear conversion. The modified spectra and F0, and the aperiodic features of the dysarthric speech are synthesized using STRAIGHT synthesis [13].

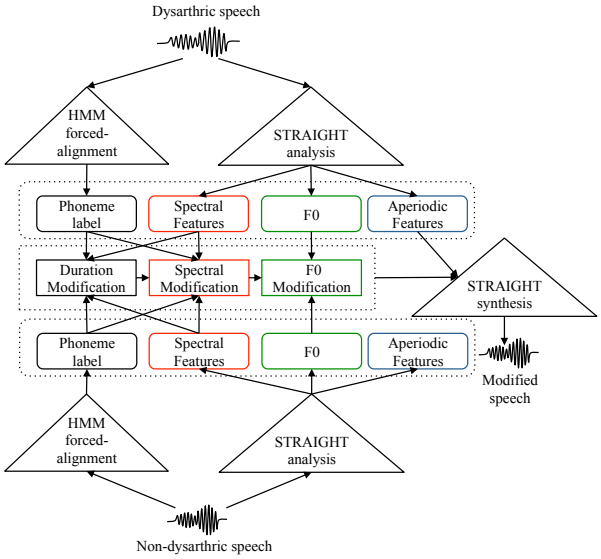


Figure 1: Flow of dysarthric speech modification

3.2. Duration modification

The duration of dysarthric speech tends to be longer than non-dysarthric speech [10]. In [7], we investigated that average duration per mora in 50 dysarthric sentences is 1.3 times slower than that of non-dysarthric speech. We modified the duration of dysarthric speech to that of non-dysarthric speech by using NTD.

First, parallel utterances of dysarthric and non-dysarthric speech are decomposed into the dictionary and activity. We refer to the basis set as the dictionary. Fig. 2 shows the flow of the decomposition. The dysarthric spectrum $\mathbf{V}^s \in \mathbb{R}^{(I\times J)}$ is decomposed into the source dictionary $\mathbf{W}^s \in \mathbb{R}^{(I\times L)}$ and its activity $\mathbf{H}^s \in \mathbb{R}^{(L\times J)}$ using NTD.

$$\mathbf{V}^s \approx \mathbf{W}^s\mathbf{H}^s \tag{6}$$

The non-dysarthric spectrum $\mathbf{V}^t \in \mathbb{R}^{(I\times K)}$ is decomposed into the target dictionary $\mathbf{W}^t \in \mathbb{R}^{(I\times K)}$, and its activity $\mathbf{H}^t \in \mathbb{R}^{(K\times J)}$ is the same way the dysarthric spectra is.

$$\mathbf{V}^t \approx \mathbf{W}^t\mathbf{H}^t \tag{7}$$

In NTD, the l-th event timing t_l is defined with the center frame of l-th phoneme label. Therefore, $\mathbf{W}^s$ and $\mathbf{W}^t$ will be parallel.

The durations of dysarthric spectra is modified as shown in Fig. 3.

$$\mathbf{V}^{s\to t} = \mathbf{W}^s\mathbf{H}^t \tag{8}$$

Because we use the source dictionary for duration modification, only the duration is modified.

3.3. Spectral modification

In general, the vowels voiced by a speaker strongly indicate the speaker's individuality. On the other hand, the consonants of people with dysarthria are often unstable. In [8], in order to improve the intelligibility of dysarthric utterances, we converted dysarthric consonants into non-dysarthric ones. Based on the same idea, we use a "combined-dictionary" that consists of the

$$w_{i,l} \leftarrow \frac{\sum_{t=t_{l-1}}^{t_{l+1}} v_i(t)h_l(t)}{\sum_{t=t_{l-1}}^{t_l-1}(w_{i,l-1}h_{l-1}(t)h_l(t)+w_{i,l}h_l^2(t))+\sum_{t=t_l}^{t_{l+1}}(w_{i,l+1}h_l(t)h_{l+1}(t)+w_{i,l}h_l^2(t))} w_{i,l} \quad (3)$$

$$h_l(t) \leftarrow \frac{\sum_{i=1}^{I} w_{i,l}v_i(t)+\alpha}{\sum_{i=1}^{I}(w_{i,l-1}w_{i,l}h_l(t)+w_{i,l}^2h_l(t))+\alpha(h_{l-1}(t)+h_l(t))} h_l(t) \quad (4)$$

$$h_{l-1}(t) \leftarrow \frac{\sum_{i=1}^{I} w_{i,l-1}v_i(t)+\alpha}{\sum_{i=1}^{I}(w_{i,l-1}w_{i,l}h_l(t)+w_{i,l-1}^2h_{l-1}(t))+\alpha(h_{l-1}(t)+h_l(t))} h_l(t) \quad (5)$$

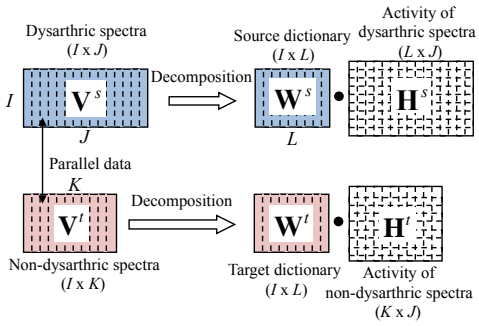


Figure 2: Decomposition using NTD

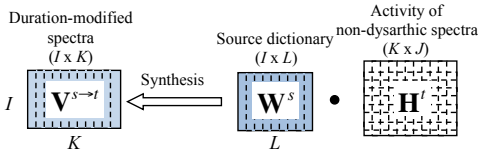


Figure 3: Duration modification

bases of dysarthric vowels from the source dictionary and bases of non-dysarthric consonants from the target dictionary.

The dysarthric spectra $\mathbf{V}^{s\to t}$ are modified as shown in Fig. 4 where $\hat{\mathbf{W}}^{st}$ denotes the combined-dictionary.

$$\hat{\mathbf{V}}^{s\to t} = \hat{\mathbf{W}}^{st}\mathbf{H}^t \quad (9)$$

By using the combined-dictionary, only consonants are modified, and we can preserve the speaker's individuality. Moreover, by using target activity, the duration of dysarthric speech is also modified.

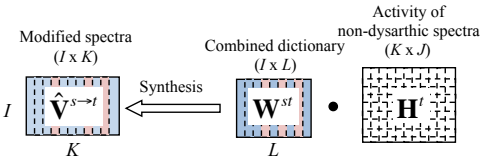


Figure 4: Spectral modification

3.4. F0 modification

Fig. 5 shows an example of non-dysarthric F0. Fig. 6 shows an example of dysarthric F0, which is a parallel utterance of Fig. 5. Although these utterances are parallel, F0 trajectories are different between the two utterances. In a TTS system [7] the F0 model is trained from non-dysarthric speech in order to synthesize an intelligible voice.

In the proposed F0 modification, we use non-dysarthric F0, which is linearly transformed in order to preserve the source speaker's individuality as follows:

$$f0^{conv}(t) = \frac{\sigma^{(s)}}{\sigma^{(t)}}(f0^t(t)-\mu^{(t)})+\mu^{(s)}, \quad (10)$$

where $f0^s(t)$, $f0^t(t)$, and $f0^{conv}(t)$ denote log-scaled F0 of dysarthric speech, non-dysarthric speech, and modified speech at frame t, respectively. $\mu^{(s)}$ and $\sigma^{(s)}$ denote the mean and standard deviation of the log-scaled F0, as calculated from dysarthric speech. $\mu^{(t)}$ and $\sigma^{(t)}$ are the mean and standard deviation of non-dysarthric speech.

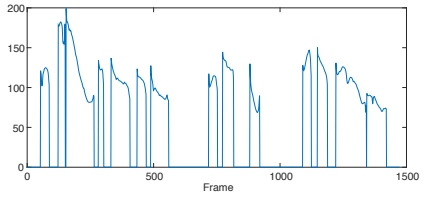


Figure 5: Example of non-dysarthric F0

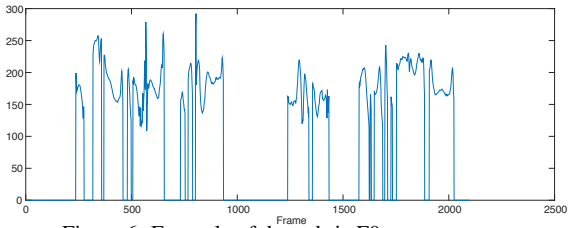


Figure 6: Example of dysarthric F0

4. Experimental Results

4.1. Experimental Conditions

The proposed method was evaluated on sentence-based speech modification for one Japanese male with dysarthric speech. We recorded 50 sentences, which are found in the ATR Japanese database [14]. The speech signals were sampled at 12 kHz, and the frame shift was 5 ms.

Label data were obtained by HMM-based forced alignment using HTK. In the case of dysarthric speech, it is difficult to

obtain precise label data using HMM because some phonemes in dysarthric speech fluctuated due to the speaker's inability to speak clearly. Moreover, dysarthric speech includes the unexpected sound of breath. Therefore, some labels are replaced.

Acoustic and prosodic features were extracted using STRAIGHT. Duration and spectra are modified using NTF. The dictionary is initialized with the spectra of the center frame of each phoneme. The activity of the l-th event area (between t_{l-1} and t_{l+1}) is initialized with positive random values. The number of dimensions of STRAIGHT spectra is 513. α is set at 100.

We conducted the objective evaluation to evaluate the precision of the decomposition of NTF. We used log spectrum distance (LSD) as a measurement.

$$LSD[dB] = \sqrt{\frac{1}{I}\sum_{i}^{I}(20\log_{10}\frac{v_i^s(t)}{v_i^{conv}(t)})^2} \qquad (11)$$

We compared 3 methods: 1) duration modification, 2) duration and spectral modification, and 3) duration, F0, and spectral modification. We conducted subjective evaluations using a 5-scale MOS test. A total of 10 Japanese speakers took part in the listening test using headphones. We evaluated both the aspect of listening intelligibility and the aspect of speaker similarity. For listening intelligibility, dysarthric speech and non-dysarthric speech are presented as reference voices, and the opinion score was set as follows: (5: very intelligible, just like non-dysarthric speech, 4: intelligible, like non-dysarthric speech, 3: fair, 2: not so intelligible, like dysarthric speech, 1: unintelligible, just like dysarthric speech). For speaker similarity, dysarthric speech and non-dysarthric speech are also presented as the references, and the opinion score was set as follows: (5: very similar to a person with dysarthria, 4: similar to a person with dysarthria, 3: fair, 2: similar to a physically unimpaired person, 1: very similar to a physically unimpaired person).

4.2. Results and Discussion

We evaluated log spectrum distance (LSD) using the different number of bases in the dictionary, and the results are shown in Table 1. We obtained a better result when we used three bases for one phoneme than when the number of the bases is the same as that of the phoneme (default setting as [11]). The LSD of dysarthric speech is worse than that of non-dysarthric speech. We assume that this is because dysarthric speech fluctuates.

Table 1: LSD of using different dictinoaries

#basis of phoneme	Dysarthric [dB]	Non-dysarthric [dB]
1	2.33	2.17
3	1.93	1.52

Fig. 7 and Fig. 8 show an example of non-dysarthric spectra and dysarthric spectra, respectively. Comparing Fig. 7 to Fig. 8, the duration of dysarthric speech tends to be long and dysarthric spectra have weak energy. Fig. 9 shows an example of duration-modified spectra. Fig. 10 shows an example of duration and spectrum-modified spectra.

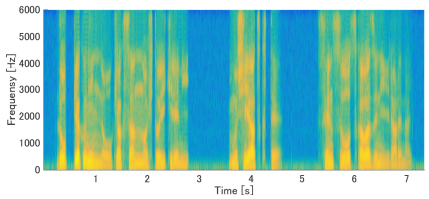


Figure 7: Example of non-dysarthric spectra

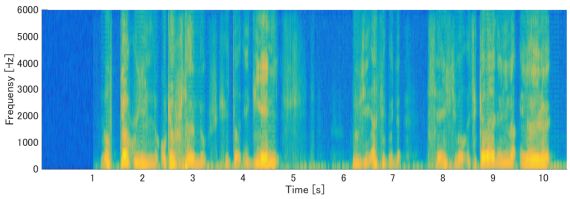


Figure 8: Example of dysarthric spectra

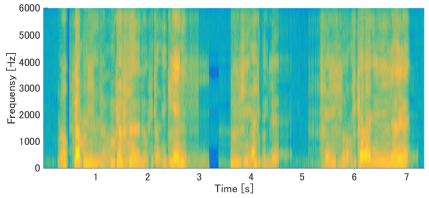


Figure 9: Example of duration-modified spectra

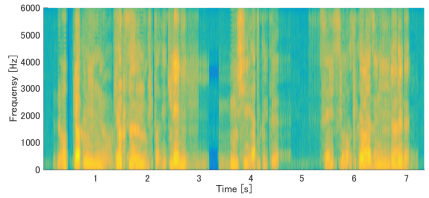


Figure 10: Example of duration and spectrum-modified spectra

Fig. 11 and Fig. 12 show the results of the subjective evaluation on intelligibility and similarity to a person with dysarthria, respectively. Error bars show 95% confidence area, and the results are confirmed with the p-value test result of 0.05. Fig. 11 shows that duration, spectrum, and F0 modification significantly improve the intelligibility of dysarthric speech. Duration- and spectrum-modified speech (without F0 modification) is slightly improved the intelligibility of dysarthric speech. Fig. 12 implies, that because we focus on consonants in spectrum modification, duration and spectrum modification preserve speaker individuality. Considering the results shown in Fig. 11 and Fig. 12, F0 is important for improving intelligibility, and speaker similarity is also impacted greatly by it.

5. Conclusions

We proposed speech modification for dysarthric speech resulting from athetoid cerebral palsy. Input dysarthric speech is labeled by HMM-based forced alignment. Using the label data and parallel non-dysarthric speech, the duration, spectra, and F0

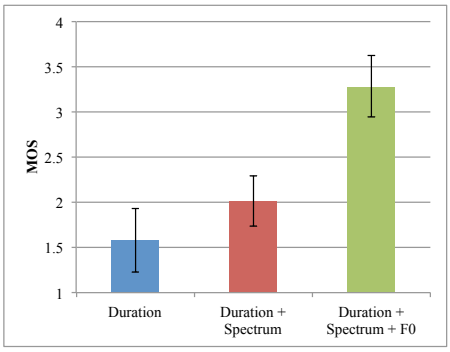


Figure 11: MOS test on intelligibility

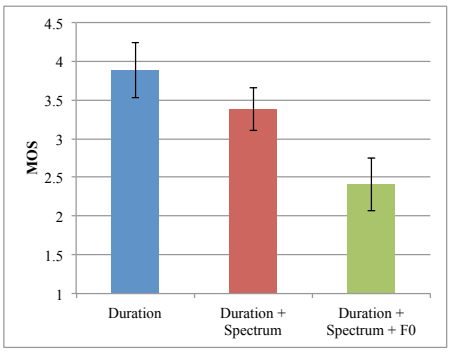


Figure 12: MOS test on similarity

are modified to improve the intelligibility of dysarthric speech. We applied NTF for dysarthric duration and spectrum modification and linear-conversion for dysarthric F0.

Using a subjective testing approach, we investigated the relationship between modified features, intelligibility and similarity to a person with dysarthria. Experimental results show that our speech modification effectively improved the intelligibility of dysarthric speech. However, it was also confirmed that speaker similarity is quite sensitive to F0. Therefore, intelligibility-preserving F0 modification will be the subject of future work. In this paper, there was only one test subject, so in future experiments, we will increase the number of test subjects and further examine the effectiveness of our method. Future work will also include efforts to study the co-articulation effect between phonemes.

6. References

[1] M. V. Hollegaard, K. Skogstrand, P. Thorsen, B. Norgaard-Pedersen, D. M. Hougaard, and J. Grove, "Joint analysis of SNPs and proteins identifies regulatory IL18 gene variations decreasing the chance of spastic cerebral palsy," *Human Mutation, Vol. 34*, pp. 143–148, 2013.

[2] S. T. Canale and W. C. Campbell, "Campbell's operative orthopaedics," Mosby-Year Book, Tech. Rep., 2002.

[3] H. Matsumasa, T. Takiguchi, Y. Ariki, I. Li, and T. Nakabayachi, "Integration of metamodel and acoustic model for dysarthric speech recognition," *Journal of Multimedia*, vol. 4, no. 4, pp. 254–261, 2009.

[4] C. Miyamoto, Y. Komai, T. Takiguchi, Y. Ariki, and I. Li, "Multimodal speech recognition of a person with articulation disorders using AAM and MAF," *in Proc. IEEE International Workshop on Multimedia Signal Processing (MMSP)*, pp. 517–520, 2010.

[5] C. Veaux, J. Yamagishi, and S. King, "Using HMM-based speech synthesis to reconstruct the voice of individuals with degenerative speech disorders," *in Proc. Interspeech*, 2012.

[6] J. Yamagishi, C. Veaux, S. King, and S. Renals, "Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction," *Acoustical Science and Technology*, vol. 33, no. 1, pp. 1–5, 2012.

[7] R. Ueda, R. Aihara, T. Takiguchi, and Y. Ariki, "Individuality-preserving spectrum modification for articulation disorders using phone selective synthesis," *in Proc. Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, 2015.

[8] R. Aihara, T. Takiguchi, and Y. Ariki, "Individuality-preserving voice conversion for articulation disorders using phoneme-categorized exemplars," *ACM Transactions on Accessible Computing (TACCESS)*, vol. 6, no. 4, pp. 13:1–13:17, 2015.

[9] A. B. Kain, J. Hosom, X. Niua, J. Santen, M. Fried-Oken, and J. Staehely, "Improving the intelligibility of dysarthric speech," *Speech Communication*, vol. 49, no. 9, pp. 43–759, 2007.

[10] F. Rudzicz, "Adjusting dysarthric speech signals to be more intelligible," *Computer Speech and Language*, vol. 27, no. 6, pp. 1163–1177, 2014.

[11] S. Hiroya, "Non-negative temporal decomposition of speech parameters by multiplicative update rules," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 2108–2117, 2013.

[12] INTERNATIONAL TELECOMMUNICATION UNION, "Methods for objective and subjective assessment of quality," *ITU-T Recommendation P.800*, 2003.

[13] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequencybased F0 extraction: possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[14] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," *Speech Communication*, vol. 9, pp. 357–363, 1990.

Recognizing Whispered Speech Produced by an Individual with Surgically Reconstructed Larynx Using Articulatory Movement Data

Beiming Cao[1], *Myungjong Kim*[1], *Ted Mau*[3], *Jun Wang*[1,2]

[1]Speech Disorders & Technology Lab, Department of Bioengineering
[2]Callier Center for Communication Disorders
University of Texas at Dallas, Richardson, Texas, United States
[3]Department of Otolaryngology - Head and Neck Surgery
University of Texas Southwestern Medical Center, Dallas, Texas, United States
{beiming.cao, myungjong.kim, wangjun}@utdallas.edu; ted.mau@utsouthwestern.edu

Abstract

Individuals with larynx (vocal folds) impaired have problems in controlling their glottal vibration, producing whispered speech with extreme hoarseness. Standard automatic speech recognition using only acoustic cues is typically ineffective for whispered speech because the corresponding spectral characteristics are distorted. Articulatory cues such as the tongue and lip motion may help in recognizing whispered speech since articulatory motion patterns are generally not affected. In this paper, we investigated whispered speech recognition for patients with reconstructed larynx using articulatory movement data. A data set with both acoustic and articulatory motion data was collected from a patient with surgically reconstructed larynx using an electromagnetic articulograph. Two speech recognition systems, Gaussian mixture model-hidden Markov model (GMM-HMM) and deep neural network-HMM (DNN-HMM), were used in the experiments. Experimental results showed adding either tongue or lip motion data to acoustic features such as mel-frequency cepstral coefficient (MFCC) significantly reduced the phone error rates on both speech recognition systems. Adding both tongue and lip data achieved the best performance.

Index Terms: whispered speech recognition, larynx reconstruction, speech articulation, deep neural network, hidden Markov model

1. Introduction

Larynx is one of the most important articulators for speech and sound production. Vocal fold vibration produces the sounding source for speech. People who have their larynx (vocal fold) impaired due to physical impairment or treatment of laryngeal cancer suffer during their daily life. A surgery can help these patients reconstruct or repair their larynx, but their phonation can hardly be completely recovered [1]. Patients with surgically reconstructed larynx generally have problems in controlling laryngeal functions, thus producing whispered speech with extreme hoarseness [2]. Therefore, assistive automatic speech recognition (ASR) technology is necessary so that they can interact with computers or smart phones in their daily life like normal people do. A standard ASR system that focuses on recognizing normal speech does not work well for these patients, because their speech mostly contains an unvoiced mode of phonation. Thus, ASR systems that are specialized for whispered speech are needed [3].

Whispered speech produced by patients with reconstructed larynx can be treated as a kind of disordered speech, which degrades the performance of conventional speech recognition systems [4,5]. Whispered speech misses glottal excitation, leading to abnormal energy distribution between phone classes, variations of spectral tilt, and formant shifts due to abnormal configurations of the vocal tract [3, 6], which are the main causes of performance degradation of a standard ASR system. To improve the performance of whispered speech recognition, most of the conventional studies used whispered speech data that are collected from normal talkers and focused on reducing the mismatch between normal and whispered speech in acoustic domain through acoustic model adaptation and feature transformation [7–11].

Articulatory information has been proven effective in the applications of normal speech recognition [12–16] and dysarthric speech recognition [17, 18]. Compared to acoustic features, articulatory features are expected to be less affected for these patients who produce whispered speech [19]. There are a few studies applying articulatory or non-audio information in whispered speech recognition [10, 19, 20]. For example in [19], the authors applied articulatory features (also known as phonological attributes) of whispered speech. Most of the existing work using articulatory information focused on descriptive or derived articulatory features in acoustic domain. Articulatory movement data such as tongue and lip motion have rarely been applied in this application.

In this paper, we investigated whispered speech recognition for a larynx reconstructed patient using tongue and lip motion data. To our knowledge, this is the first study for whispered speech recognition with articulatory data. Tongue and lip motion data were collected using an electromagnetic articulograph. Two speech recognizers were used: Gaussian mixture model-hidden Markov model (GMM-HMM) and deep neural network-hidden Markov model (DNN-HMM). In the experiments, we examined several settings on both speech recognition systems to verify the effectiveness of adding articulatory movement data: mel-frequency cepstral coefficient (MFCC)-based acoustic features, lip and tongue movement-based articulatory data, and MFCC with articulatory data. The remaining of the paper is organized as follows: Section 2 describes our acoustic and articulatory data collected from a patient with surgically reconstructed larynx. In Section 3, we present our experimental design including speech recognition systems and experimental setup. Section 4 shows experimental results and discussion. Conclusions are summarized in Section 5.

SLPAT 2016, 7th Workshop on Speech and Language Processing for Assistive Technologies, pages 80–86,
San Francisco, USA, 13 September, 2016.

(a) Wave System (b) Sensor Locations

Figure 1: *Articulatory (tongue and lip) motion data collection setup.*

2. Data Collection

2.1. Participants and stimuli

The patient is a male of 23 years old. He had his larynx damaged in an accident and then took a larynx reconstruction surgery in 2014. His speech showed extreme hoarseness. He is not using assistive device on a daily basis. He participated in the data collection, where he produced a sequence of 132 phrases at his habitual speaking rate. The phrases were selected from the phrases that are frequently spoken by persons who use augmentative and alternative communication (AAC) devices [21, 22].

2.2. Tongue motion tracking device and procedure

An electromagnetic articulograph (Wave system, Northern Digital Inc., Waterloo, Canada) was used for articulatory data collection (Figure 1a). Four small sensors were attached to the surface of patient's articulators, two of them were attached to tongue tip (TT, 5-10mm to tongue apex) and tongue back (TB, 20-30mm back from TT) using dental glue (PeriAcry1 90, GluStitch). The other two were attached to upper lip (UL) and lower lip (LL) using normal double-sided tape. In addition, another sensor was attached to the middle of forehead for head correction. Our prior work indicated that the four-sensor set consisting of tongue tip, tongue back, upper lip, and lower lip are an optimal set for this application [23–25]. The positions of all five sensors are shown in Figure 1b. With this approach, three-dimension movement data of articulators were tracked and recorded. The sampling rate in Wave recording in this project was 100Hz, The spatial precision of movement tracking is about 0.5mm [26].

The patient was seated next to the magnetic field generator, which is the blue box in Figure 1a, and read the 132 phrases. A three-minute training session helped the patient to adapt to speak with tongue sensors before the formal data collection session.

Before data analysis, the translation and rotation of the head sensor were subtracted from the motion data of tongue and lip sensors to obtain head-independent articulatory data. Figure 1b illustrates the derived 3D Cartesian coordinates system, in which x is left-right direction; y is vertical; and z is front-back direction. We assume the tongue and lip motion patterns of the patient remain the same as normal talkers, where the movement in x direction is not significant in speech production. Therefore, only y and z coordinates were used for analysis in this study [27].

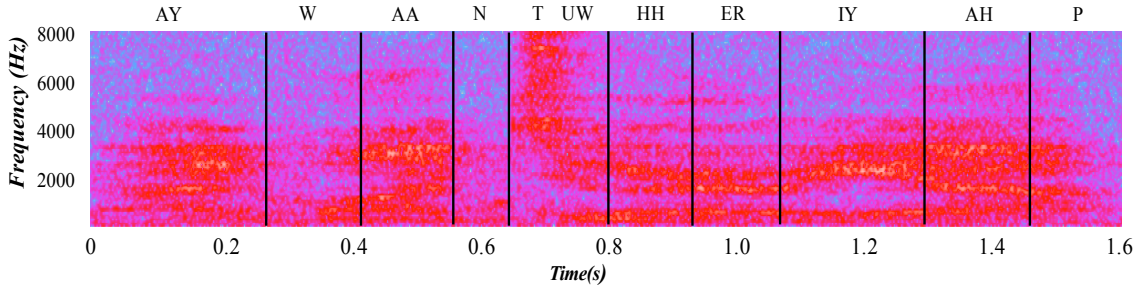
Acoustic data were collected synchronously with the articulatory movement data by built-in microphone in the Wave system. In total, the data set contains 2,292 phone samples of 39 unique phones.

2.3. Acoustic data

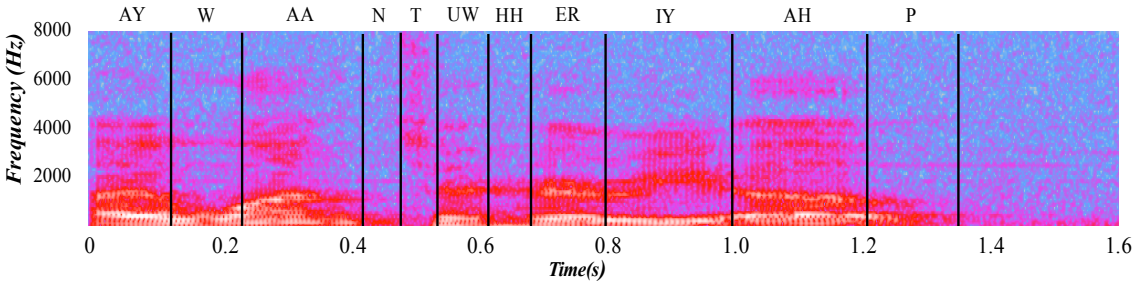
Figure 2 shows the spectrograms of whispered speech and normal (vocalized) speech examples producing the same phrase *I want to hurry up*. Figure 2a is an example spectrogram of whispered speech produced by the patient in this study. Figure 2b is an example of normal speech produced by a healthy speaker. The healthy speaker's data example was just used to illustrate the difference between the spectrograms of whispered and normal speech, and therefore it was not used in analysis. In the figure, brighter color (and reddish) denotes higher energy. As illustrated in Figure 2, for normal speech, the phone boundaries are relatively clear based on spectral energy shape and formant frequencies, and it is easy to distinguish. For whispered speech, however, most of phones have very similar spectral pattern without fundamental frequency, which makes it hard to find the boundary between the phones using acoustic information only. For example, the phone pairs like 'AH' and 'P' in word 'up', can hardly be distinguished in whispered speech, also the 'HH' and 'ER' in word 'hurry' are not easy to classify. On the other hand, those two phone pairs can be clearly distinguished in normal speech, showing that vowels have higher energy and distinct formant frequencies. The ambiguity of phone boundaries contributed to lower performance in whispered speech recognition using standard ASR techniques.

2.4. Articulatory data

Figure 3a and 3b give examples of articulatory movement data, which are obtained from the motion tracking of sensors when uttering same phrase (*I want to hurry up*) in Figure 2, respectively. As mentioned previously, four sensors were attached to articulators (upper lip, lower lip, tongue tip, and tongue back). As illustrated in Figure 3, the articulatory movement pattern of whispered speech somewhat resembles the articulatory movement pattern of normal speech, although the motion range of whispered speech by the patient was larger than that by the healthy talker. Therefore, we expected that articulatory motion data may improve the performance of whispered speech recognition. The larger motion range of tongue and lips of the patient may be because he uses more force than normal talkers during his speech production. For illustration purpose, the two articulatory shapes (tongue and lip sensor motion curves) in Figure



(a) Whispered (hoarse) speech produced by the patient



(b) Normal Speech

Figure 2: *Spectrogram examples of whispered and normal speech producing "I want to hurry up".*

3 were rotated slightly so that the UL and the LL sensors were aligned vertically.

3. Method

Standard speech recognition systems are typically based on hidden Markov models (HMMs), which are an effective framework for modeling time-varying spectral feature vector sequence [28]. A number of techniques for adapting or describing input features are typically used together with HMM. In this study, we used two speech recognizers: the long-standing Gaussian mixture model (GMM)-HMM and the recently available deep neural network (DNN)-HMM. Major parameter configuration of the two recognizers was shown in Table 1.

3.1. GMM-HMM

The Gaussian mixture model (GMM) is a statistical model for describing speech features in a conventional speech recognition system. Given enough Gaussian components, GMMs can model the relationship between acoustic features and phone classes as a mixture of Gaussian probabilistic density functions. More detail explanation of GMM can be found in [29]. GMM-HMM is a model that is "hanging" GMMs to states of HMM, in which GMMs are used for characterizing speech features and HMM is responsible for characterizing temporal properties.

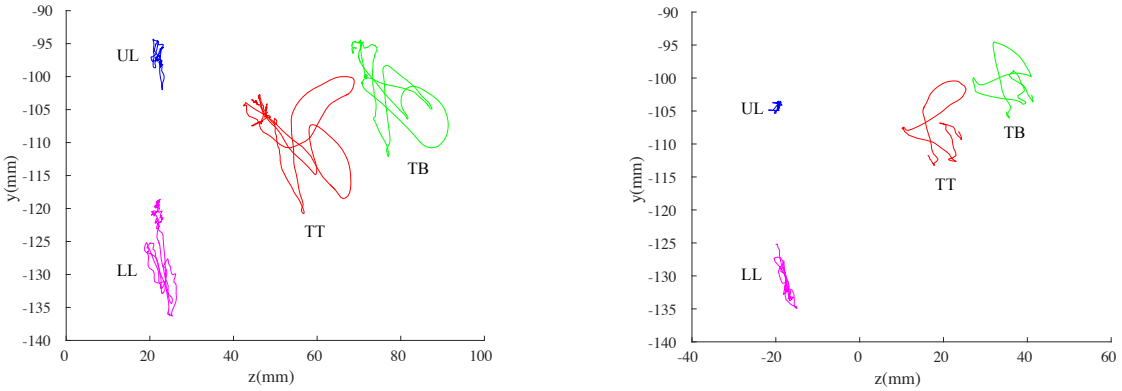
GMM-HMM have been widely used in modeling speech features and as an acoustic model for speech recognition for decades until DNN attracted more interests in the literature recently. However, GMM is still promising when using a small data set. In addition, because of its rapid implementation and execution, we included GMM as a baseline approach. Table 1 gives the major parameters for GMM-HMM.

3.2. DNN-HMM

Deep neural networks (DNNs) with multiple hidden layers have been shown to outperform GMMs on a variety of speech recognition benchmarks [30] including recent works that involved articulatory data [17, 31]. DNN-HMM takes multiple frames of speech features as input and produces posterior probabilities over HMM states as output. The DNN training is based on restricted Boltzmann machines (RBMs). The weights between nodes in neighboring layers at iteration $t+1$ are updated based on iteration t using stochastic gradient descent described by the following equation:

$$w_{ij}(t+1) = w_{ij}(t) + \eta \frac{\partial C}{\partial w_{ij}} \tag{1}$$

in which w_{ij} is the weight between nodes i and j of two layers next to each other, η is the learning rate, and C is the cost function. The output posterior probabilities of DNN are used for decoding. More detailed description of DNN can be found in [30, 32–34]. The similar structure and setup of DNN in [31] were used in this experiment which has 5 hidden layers, each of hidden layers has 512 nodes. We tested all layers from 1 to 6 in each experimental configuration, and the best result was obtained when using 5 hidden layers. The one-subject data set has a relatively small size, thus we used only 512 nodes. The input layer would take 9 frames at a time (4 previous plus current plus 4 succeeding frames), therefore the dimension of input layer changed given different types of data. For example, for the experiments using both MFCC and articulatory data, the dimension of each frame is 13-dimensional MFCC plus 8-dimensional movement plus their delta and delta of delta formed a 63-dimensional vectors that were fed into the DNN. But for the experiments using only MFCC, the frame dimension is 39. The output layer has 122 dimensions (39 phones $\times$ 3 states each phone plus 5 states for silence).



(a) Whispered speech produced by the patient

(b) Normal speech produced by a healthy talker

Figure 3: *Examples of motion path of four articulators TT, TB, UL, and LL of whispered and normal speech for producing "I want to hurry up". In this coordinate system, y is vertical and z is anterior-posterior.*

3.3. Experimental Setup

In this project, frame rate was 10 ms (equivalent to sampling rate of articulatory data recording: 100 Hz). Two dimensional (vertical and anterior-posterior) EMA data of four sensors (tongue tip, tongue body back, upper lip, and lower lip) were used for the experiments. As mentioned previously, for each frame, all either acoustic features, i.e., MFCCs or articulatory movement data plus delta and delta of delta form vectors that were fed into a recognizer. HMM has left-to-right 3-states with a context-independent monophone models. Triphone models were not considered due to the small size of our data set in this work. A bi-gram phone-level language model was used. The training and decoding were performed using the Kaldi speech recognition toolkit [35].

Table 1: *Experimental setup.*

Acoustic Feature	
Feature vector	MFCC (13-dim. vectors) + Δ + $\Delta\Delta$ (39 dim.)
Sampling rate	22050 kHz
Windows length	25 ms
Articulatory Feature (both tongue and lips)	
Feature vector	articulatory movement vector (8 sensors) + Δ + $\Delta\Delta$ (24 dim.)
Concatenated Feature	
Feature vector	MFCC + articulatory movement vector (21-dim vector) + Δ + $\Delta\Delta$ (63 dim.)
Common	
Frame rate	10 ms
GMM-HMM topology	
Monophone	122 states (39 phones $\times$ 3 states, 5 states for silence), total $\approx$ 1000 Gaussians (each state $\approx$ 8 Gaussians) 3-state left to right HMM
Training method	maximum likelihood estimation (MLE)
DNN-HMM topology	
Input	9 frames at a time (4 previous plus current plus 4 succeeding frames)
Input layer dim.	216 (9 $\times$ 24 for articulatory) 351 (9 $\times$ 39 for acoustic) 567 (concatenated)
Output layer dim.	122 (monophone)
No. of nodes	512 nodes for each hidden layer
Depth	5-depth hidden layers
Training method	RBM pre-training, back-propagation
Language model	bi-gram phone language model

Phone error rate (PER) was used as a whispered speech recognition performance measure, which is the summation of deletion, insertion, and substitution phone errors divided by the number of all phones. For each of two recognizers, whispered speech recognition experiments were conducted using different combinations of features, including MFCC only, MFCC concatenated with lip motion data, MFCC with tongue data, and MFCC with both of lip and tongue data.

Three-fold cross validation was used in the experiments. The average performance of the executions was calculated as the overall performance.

4. Results & Discussion

Figure 4 shows the average PERs of speaker-dependent whispered speech recognition for the patient. The baseline results (67.2% for GMM-HMM and 66.1% for DNN-HMM) were obtained using only acoustic (MFCC) features.

The PERs were reduced by adding either tongue motion data (63.6% for GMM-HMM and 63.3% for DNN-HMM) or lip motion data (65.6% for GMM-HMM and 65.6% for DNN-HMM) to MFCC features although the PERs of independent lip motion data or tongue motion data are higher than that obtained with MFCC features only. Particularly, using tongue motion data was more effective than with lip motion data, producing better phone recognition performance. This result is consistent with our previous speech recognition tasks with articulatory data [24], because tongue motion contains more information than lip motion during speech production [25].

The best performance was achieved when both lip and tongue data were applied with acoustic data, 63.0% for GMM-HMM and 62.1% for DNN-HMM. These results indicate that MFCC, lip motion data, and tongue motion data have complementary information in distinguishing phones.

A two-tailed t-test was performed to measure if there were statistical significance between the performances of the configuration with MFCC only and other configurations. As indicated in Figure 4, most data configurations of MFCC+articulatory features showed a statistical significance with the MFCC configuration. The results suggested that adding tongue data or both lip and tongue data to MFCC features significantly im-

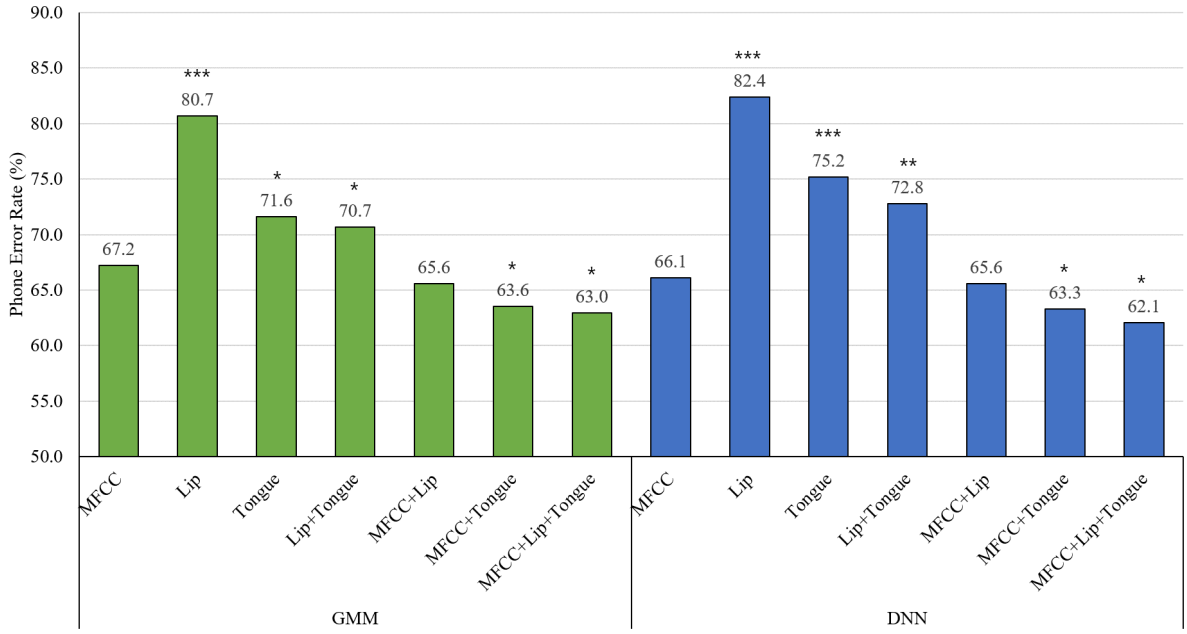


Figure 4: *Phone error rates of whispered speech recognition obtained in the experiments using monophone GMM-HMM and DNN-HMM with various types of data (MFCC, Lip, Tongue, Lip+Tongue, MFCC+Lip, MFCC+Tongue, and MFCC+Lip+Tongue). Statistical significances between the results obtained using MFCC and other data types on each ASR model are marked: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.*

proved the performance. Adding lip movement data, however, did not show a significance, although a slight improvement was observed. The observation may be because of the small data size. A further study with a data set of larger size is needed to verify these findings.

To understand the contribution by adding articulatory movement data in whispered speech recognition, we also tested the recognition performance from articulatory data only (i.e., without using acoustic features, or silent speech recognition). Figure 4 gives the silent speech recognition results of using GMM-HMM and DNN-HMM, respectively. For both GMM-HMM and DNN-HMM, the least performances were obtained when using lip data only; the best performances were obtained when using both tongue and lip data. The results on individual articulatory data configurations (lip, tongue, lip + tongue) were positively correlated by the contribution of adding the data. In other words, using tongue data only obtained a better recognition performance than using lip data only, which explained why adding tongue information better improved the whispered speech recognition than adding lip information. These findings are consistent with our previous work for silent speech recognition using data from combinations of articulators (sensors) [24, 25]. Using only articulatory data always obtained less performance than using acoustic data only, which is also consistent with our prior finding [24].

Table 2: *Numbers of deletion, insertion, and substitution errors in the experiment of whispered speech recognition with articulatory data.*

Model	Feature	Del	Ins	Sub
GMM	MFCC	723	78	740
	MFCC+Lip	640	110	752
	MFCC+Tongue	657	109	691
	MFCC+Lip+Tongue	652	117	674
DNN	MFCC	696	71	752
	MFCC+Lip	782	62	656
	MFCC+Tongue	761	59	632
	MFCC+Lip+Tongue	783	56	610

In addition, Table 2 gives a summary of deletion, insertion, and substitution in the phone recognition errors in these experiments. Table 2 provides more details that different articulatory data decrease the PER in different ways. For GMM-HMM, adding lip data would decrease the number of deletion by 83 but increased the numbers of insertion and substitution. However, adding tongue data decreased the number of substitution and deletion, but increased the number of insertion. For DNN-HMM, adding either tongue or lip data would considerably decrease insertion and substitution errors, although it increased deletion errors. As discussed earlier, we think adding articulatory motion data will help the recognizer to find the boundaries between phones. However, how tongue or lips affect the number of deletion, insertion, and substitution needs to be verified with a larger data set.

DNN typically outperformed GMM in ASR using acoustic data only [30] or using both acoustic and articulatory data [17, 33]. In this work, DNN performance was slightly better than that of GMM as well. Although our data set is small and DNN typically requires a larger data set, DNN still can model the complex structure of whispered speech in this project. This result indicates that DNN will be promising for whispered

speech recognition with articulatory data for a larger, multiple-speaker data set.

In summary, the experimental results demonstrated the effectiveness of applying articulatory movement data to whispered (hoarse) speech recognition. In addition, the results indicated that adding tongue motion data will improve the performance more than that by adding lip motion data in whispered speech recognition. The best performance was obtained when both tongue and lip motion data were used.

Limitation. Although the results are promising, the method (adding articulatory data on top of acoustic data) has been evaluated with only one subject (patient) with whispered speech. A further study with a multiple-speaker data set is needed to verify these findings.

5. Conclusions & Future Work

The effectiveness of articulatory (tongue and lips) movement data in whispered speech recognition has been tested with a data set that was collected from an individual with a surgically reconstructed larynx. The experimental results suggested that adding articulatory movement data decreased the PER of whispered speech recognition for widely used ASR models: GMM-HMM and DNN-HMM. The best performance was obtained when acoustic, tongue, and lip movement data were used together.

Future work includes verifying the findings using a larger data set and using other latest ASR models such as deep recurrent neural networks [36].

6. Acknowledgements

This work was supported by the National Institutes of Health (NIH) under award number R03 DC013990 and by the American Speech-Language-Hearing Foundation through a New Century Scholar Research Grant. We thank Dr. Seongjun Hahm, Joanna Brown, Betsy Ruiz, Janis Deane, Laura Toles, Christina Duran, Amy Hamilton, and the volunteering participants.

7. References

[1] T. Mau, J. Muhlestein, S. Callahan, and R. W. Chan, "Modulating phonation through alteration of vocal fold medial surface contour," *The Laryngoscope*, vol. 122, no. 9, pp. 2005–2014, 2012.

[2] T. Mau, "Diagnostic evaluation and management of hoarseness," *Medical Clinics of North America*, vol. 94, no. 5, pp. 945 – 960, 2010.

[3] T. Ito, K. Takeda, and F. Itakura, "Analysis and recognition of whispered speech," *Speech Communication*, vol. 45, no. 2, pp. 139–152, 2005.

[4] M. Kim, J. Wang, and H. Kim, "Dysarthric speech recognition using kullback-leibler divergence-based hidden markov model," in *Interspeech*, 2016.

[5] H. V. Sharma and M. Hasegawa-Johnson, "Acoustic model adaptation using in-domain background models for dysarthric speech recognition," *Computer Speech & Language*, vol. 27, no. 6, pp. 1147–1162, 2013.

[6] S. Ghaffarzadegan, H. Bořil, and J. H. L. Hansen, "Model and feature based compensation for whispered speech recognition," in *Interspeech 2014*, Singapore, Sept 2014, pp. 2420–2424.

[7] ——, "UT-VOCAL EFFORT II: Analysis and constrained-lexicon recognition of whispered speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing 2014*, Florence, Italy, May 2014, pp. 2563–2567.

[8] B. P. Lim, "Computational differences between whispered and non-whispered speech," Ph.D. dissertation, University of Illinois at Urbana-Champaign, 2011.

[9] A. Mathur, S. M. Reddy, and R. M. Hegde, "Significance of parametric spectral ratio methods in detection and recognition of whispered speech," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, no. 1, pp. 1–20, 2012.

[10] C.-Y. Yang, G. Brown, L. Lu, J. Yamagishi, and S. King, "Noise-robust whispered speech recognition using a non-audible-murmur microphone with vts compensation," in *Chinese Spoken Language Processing (ISCSLP), 2012 8th International Symposium on*. IEEE, 2012, pp. 220–223.

[11] C. Zhang and J. H. Hansen, "Analysis and classification of speech mode: whispered through shouted." in *Proc. INTERSPEECH*, 2007, pp. 2289–2292.

[12] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, "Speech production knowledge in automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 121, no. 2, pp. 723–742, 2007.

[13] A. Wrench and K. Richmond, "Continuous speech recognition using articulatory data," in *Proc. International Conference on Spoken Language Processing*, Beijing China, 2000, pp. 145–148.

[14] P. K. Ghosh and S. S. Narayanan, "Automatic speech recognition using articulatory features from subject-independent acoustic-to-articulatory inversion," *The Journal of the Acoustical Society of America*, vol. 130, no. 4, pp. EL251–El257, Sep. 2011.

[15] K. Kirchhoff, G. A. Fink, and G. Sagerer, "Combining acoustic and articulatory feature information for robust speech recognition," *Speech Communication*, vol. 37, no. 3, pp. 303–319, 2002.

[16] S.-C. S. Jou, T. Schultz, M. Walliczek, F. Kraft, and A. Waibel, "Towards continuous speech recognition using surface electromyography," in *Interspeech*, 2006.

[17] S. Hahm, H. Daragh, and J. Wang, "Recognizing dysarthric speech due to amyotrophic lateral sclerosis with across-speaker articulatory normalization," in *Proc. the ACL/ISCA Workshop on Speech and Language Processing for Assistive Technologies*, 2015, pp. 47–54.

[18] F. Rudzicz, "Articulatory knowledge in the recognition of dysarthric speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 947–960, May 2011.

[19] S.-C. S. Jou, T. Schultz, and A. Waibel, "Whispery speech recognition using adapted articulatory features," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2005, pp. 1009–1012.

[20] ——, "Adaptation for soft whisper recognition using a throat microphone," in *Interspeech*, 2004.

[21] D. R. Beukelman, K. M., Yorkston, M. Poblete, and C. Naranjo, "Analysis of communication samples produced by adult communication aid users," *Journal of Speech and Hearing Disorders*, vol. 49, pp. 360–367, 1984.

[22] J. Wang, A. Samal, and J. Green, "Preliminary test of a real-time, interactive silent speech interface based on electromagnetic articulograph," in *Proc. ACL/ISCA Workshop on Speech and Language Processing for Assistive Technologies*, Baltimore, USA, 2014, pp. 38–45.

[23] J. Wang, J. Green, and A. Samal, "Individual articulator's contribution to phoneme production," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vancouver, Canada, 2013, pp. 7785–7789.

[24] J. Wang, S. Hahm, and T. Mau, "Determining an optimal set of flesh points on tongue, lips, and jaw for continuous silent speech recognition," in *Proc. ACL/ISCA Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, 2015, pp. 79–85.

[25] J. Wang, A. Samal, P. Rong, and J. R. Green, "An optimal set of flesh points on tongue and lips for speech-movement classification," *Journal of Speech, Language, and Hearing Research*, vol. 59, pp. 15–26, 2016.

[26] J. Berry, "Accuracy of the ndi wave speech research sysetm," *Journal of Speech, Language, and Hearing Research*, vol. 54, no. 5, pp. 295–301, 2011.

[27] J. Wang, J. Green, A. Samal, and Y. Yunusova, "Articulatory distinctiveness of vowels and consonants: A data-driven approach," *Journal of Speech, Language, and Hearing Research*, vol. 56, no. 5, pp. 1539–1551, 2013.

[28] M. Gales and S. Young, "The application of hidden markov models in speech recognition," *Foundations and trends in signal processing*, vol. 1, no. 3, pp. 195–304, 2008.

[29] L. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. PTR Prentice Hall Englewood Cliffs, 1993, vol. 14.

[30] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[31] S. Hahm and J. Wang, "Silent speech recognition from articulatory movements using deep neural network," in *Proc. the 18th Intl. Congress of Phonetic Sciences*, 2015.

[32] G. Hinton, "A practical guide to training restricted Boltzmann machines," *Momentum*, vol. 9, no. 1, p. 926, 2010.

[33] C. Canevari, L. Badino, L. Fadiga, and G. Metta, "Cross-corpus and cross-linguistic evaluation of a speaker-dependent DNN-HMM ASR system using EMA data," in *Proc. Workshop on Speech Production in Automatic Speech Recognition*, Lyon, France, 2013.

[34] ——, "Relevance-weighted-reconstruction of articulatory features in deep-neural-network-based acoustic-to-articulatory mapping." in *Interspeech*, Lyon, France, 2013, pp. 1297–1301.

[35] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, and V. K., "The Kaldi speech recognition toolkit," in *Proc. IEEE 2011 workshop on automatic speech recognition and understanding*, Waikoloa, USA, 2011, pp. 1–4.

[36] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002, pp. 6645–6649.

`flexdiam` – flexible dialogue management for problem-aware, incremental spoken interaction for all user groups (Demo paper)

Ramin Yaghoubzadeh, Stefan Kopp

Social Cognitive Systems Group, CITEC, Bielefeld University, Germany
ryaghoubzadeh@uni-bielefeld.de, skopp@uni-bielefeld.de

Abstract

The dialogue management framework `flexdiam` was designed to afford people across a wide spectrum of cognitive capabilities access to a spoken-dialogue controlled assistive system, aiming for a conversational speech style combined with incremental feedback and information update. The architecture is able to incorporate uncertainty and natural repair mechanisms in order to fix problems quickly in an interactive process – with flexibility with respect to individual users' capabilities. It was designed and evaluated in a user-centered approach in cooperation with a large health care provider. We present the architecture and showcase the resulting autonomous prototype for schedule management and accessible communication.

Index Terms: human-computer interaction, conversational spoken dialogue, user models, incremental processing, flexible grounding, assistive systems

1. Introduction and outline

Making spoken human-machine interaction both easy and effortless, and also robust in presence of contradictory pieces of information, is one of the central challenges in providing universal accessibility over this modality. Two of the user groups that would benefit most from this are, on the one hand, older adults, who may be reluctant or lack the capacities to interact with technology using more widely supported modalities, but also people with cognitive impairments, for whom accessing even well-designed classical interfaces can be a challenging task. Spoken interaction is overall reported as the preferred modality by older adults with little technological experience [1]. While speech recognition for these user groups can present specific difficulties [2], the available technology for word recognition has improved in the last few years to a degree that it is now feasible. Given robust – and engaging – spoken interaction, these user groups could benefit from easily accessible and understandable interfaces to technological solutions that help them to maintain an autonomous lifestyle.

In our cooperation with the large health and social care provider v. Bodelschwinghsche Stiftungen Bethel, we have explored the paradigm of a spoken-language controlled virtual assistant for schedule management, to aid in maintaining a client's day structure. Initially, in Wizard-of-Oz explorations, we established that both user groups are, in general, capable of conducting such interactions in a brief and effortless conversational style. We also found that the approach was subjectively judged as pleasant, effective and appropriate.

Building on our existing architecture for incremental dialogue processing, we created a dialogue management framework that aims to address several issues critical to making autonomous interactions with these user groups work robustly, the central requirements being:

- being aware, and addressing interactively, ambiguities in user input,
- being able to react rapidly and give feedback before problems can cascade,
- presenting and negotiating information in a way that supports individual capabilities, and
- allowing the user to feel in charge and being served well.

The resulting architecture was used to build a dialogue system that is able to provide basic schedule management and access to video communication with a conversational, incremental spoken interface represented by an embodied assistant, which we are presenting here. A subset of this functionality, namely completing a weekly schedule if events, was evaluated with older adults and people with cognitive impairments, leading to comparable performance and subjective ratings as the earlier WOz system.

2. Architecture overview

We present the architecture in an abridged account here, please refer to our previous work [3] for more details on the internal mechanics. `flexdiam` builds on our general architecture for incremental processing, IPAACA [4]. This this architecture, based on an abstract model by Schlangen et al. [5], information is represented as so-called 'Incremental Units' (IUs), which are globally exchanged information packages that can form functional networks. It is designed to be used to represent data in both the input (and interpretation) channels and processing, and also in output planning and realization (cf. Fig. 1, left).

The temporal structure of dialogue is represented in the `TimeBoard`, which stores all past, ongoing, and projected future events in thematically grouped tiers (Fig. 2). It serves as the interface between input processing, dialogue management proper, and behavior planning and realization. Events are most often either a single IU or a specific sequence of IUs. A set of interval relations on sets of tiers is used to determine higher-level events.

Data other than events with temporal extent, i.e. knowledge and propositional information, are represented via a structure termed `VariableContext` (Fig. 1, right), a blackboard satisfying two requirements: firstly, all information may reside there in the form of distributions. Moreover, all changes are stored as time-stamped deltas, enabling both rollbacks and for analysis between two points in time. Task and discourse states are represented a forest of structures called `Issues`, terminology adapted from Larsson [6], that represent (attributed) common current topics or current questions that have to be resolved cooperatively. In `flexdiam`, they are independent agents that

SLPAT 2016, 7th Workshop on Speech and Language Processing for Assistive Technologies, pages 87–90,
San Francisco, USA, 13 September, 2016.

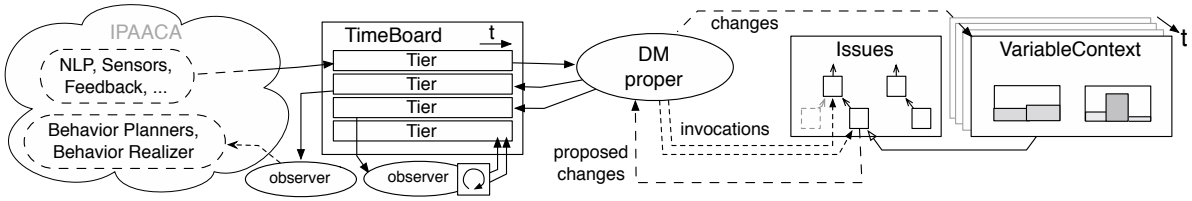


Figure 1: *Architecture overview. The cloud on the left represents external input/output modules that are not part of flexdiam proper, but connected via the common middleware IPAACA. Data structures and processing are described in the text.*

encapsulate the structure of the task addressed so far, localized planning, as well as situated interpretative capability and situated capability for abstract actions (multimodal dialogue contributions and side effects). The *dialogue manager proper* relays information hierarchically though the Issue forest (see Fig. 3 for an example of this in the interpretation process).

In line with the general notion of temporal variability and uncertainty, all operations that do not have immediate effect are treated as asynchronously performed operations that can fail.

3. Input and output

As mentioned above, all input and output components are connected to flexdiam using the IPAACA middleware.

Speech input can be delivered by several components, alternatively or concurrently (there are bridges for Windows ASR, Dragon NaturallySpeaking Client SDK and an experimental one for Google's ASR). A parser component is used to preprocess all ASR hypotheses, identifying the points of deviation in hypotheses, performing an early classification of portions of an utterance using pattern matching, and offering an interface for triggering external NLU accessories, such as POS taggers. Other input modalities accessible over IPAACA include two types of eye tracker, touch screens, keyboard and mouse input.

Output is realized by emitting request IUs that realizer components can listen for and handle. The virtual agent is controlled by the ASAPrealizer [7], which accepts action descriptions in the Behavior Markup Language. Speech generation is realized using a CereVoice [8] TTS component, which is driven by ASAPrealizer. There is a separate controller for GUI elements that can either be addressed directly or in a speech-synchronized manner by ASAPrealizer. Language output is not generated directly in flexdiam, but relayed to an associated dedicated NLG component that can offer multiple alternative realizations for an abstract request (though currently, flexdiam always chooses the first one to appear).

Fig. 4 depicts the typical interface setup, in an interaction scene between an older subject and the virtual agent "Billie". Subjects interacted using the table microphone and touchscreen (red 'panic button' in the corner).

4. Experiments

A basic dialogue system constructed with flexdiam has been subjected to small-scale evaluations with both older adults (n=6) [3] and people with cognitive impairments (n=5). The task for participants was to enter a freely chosen set of appointments into their fictional calendar, the same domain as an earlier Wizard-of-Oz experiment [9], in which we showed that people with cognitive impairments in particular benefit from a much more explicit information grounding strategy compared to con-

Figure 4: *flexdiam driving a virtual agent, "Billie", in an autonomous interaction study with an older adult (anonymized).*

trols when their ability to detect system errors is observed. We also found inter-group differences in preferred verbalizations (e.g. more frequent first-person requests in older adults vs. more frequent neutral dictation in people with cognitive impairments) [10].

For the interactions with the autonomous prototype, we provided some ideas for events on a paper sheet with textual and iconic representations. Subjects were instructed to stick to the task and be to the point, but not primed as to how to phrase their requests or replies. In general, participants were able to enter appointments successfully. Some leeway was given by participants if the agent paraphrased only (a relevant) part of their event descriptions – a simple heuristic approach was used to extract candidate topics from the free-form utterances.

The system in that state was configured to always yield the floor and let the user talk at their leisure. One subject from each group used very verbose interaction styles and attempted to provide a lot of tangential information, despite a clarifying instructive intervention that could be inserted after an initial free practice phase. The current focus of development is hence on subtle and acceptable approaches to pre-emptive floor management.

Subjective ratings of the autonomous system in terms of effectiveness and usability did not differ significantly from the earlier WOz experiment that targeted the same interface and task domain [3].

5. Demo system

The demo system showcases flexdiam in a schedule planning scenario controlled by spoken language, enabling the user to go through their (fictional) week, modifying events, decid-

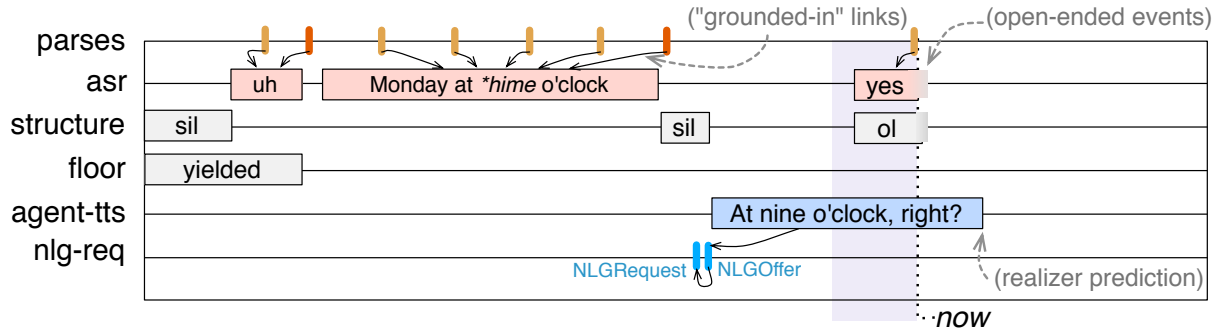


Figure 2: *Temporal representation of dialogue events on the TimeBoard. In a situation where the user (red) wanted to enter a new appointment, they produced an utterance that was mispronounced, leading to ambiguities. The DM posted a clarification question (blue), its predicted end time is shown extending beyond the time marked 'now'. In the current situation, the reply by the user has already started (producing an overlap). In the default configuration, the system would yield the floor to the user immediately if an overlap over a threshold length is encountered.*

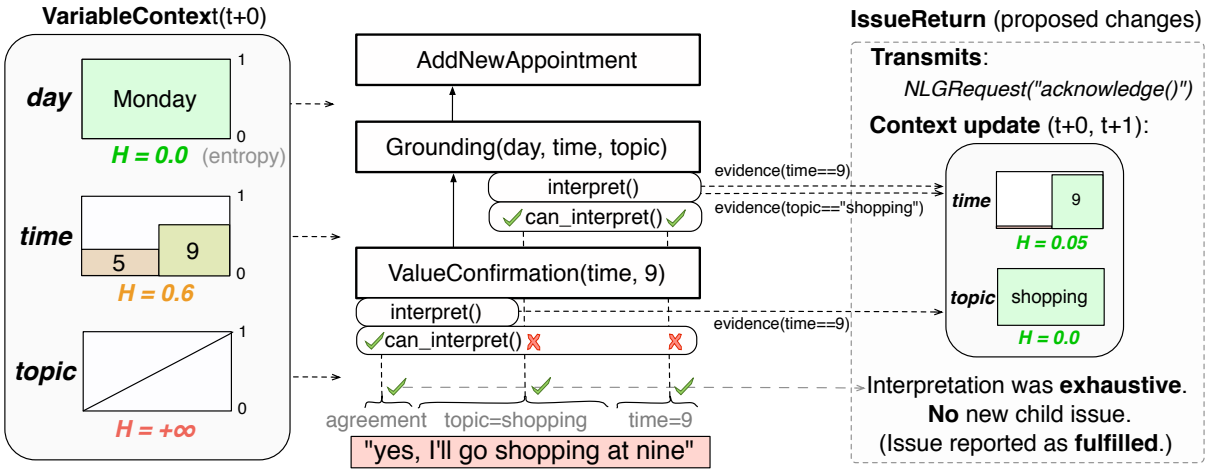


Figure 3: *Current context (left) and relevant subset of the Issue forest (center) for a situation just past the one in Fig. 2: the user has completed their utterance. Interpretation is performed locally first (in the ValueConfirmation issue), then deferred to its parent (Grounding). Both contribute to the evidence that leads to a proposed Context update (right). Since the requirements for a confirmation question were met, the bottommost Issue reports itself as fulfilled. The mechanism during incremental interpretation is identical.*

ing about events offerd by third parties, as well as an interface to an encrypted video telephony application that can be triggered from inside the dialogue situation. Different modes of information grounding can be selected (e.g. concise summaries vs. fine-grained confirmation requests), these correspond to user models for different cognitive abilities; and a live view into the information update process is possible. The demo setup includes a computer, a desk microphone, and an eye tracker. A recording of a previous version is accessible online at purl.org/net/ramin/slpat2016/.

6. Outlook

In the ongoing project with our health care partner, larger-scale evaluations of various aspects of the system are underway, later extending to prolonged experimental deployment in private home environments of interested participants. Current and future steps of development include proactive floor management and improved reference resolution mechanisms in the framework. We are planning to release documentation and a source code package for flexdiam in the future under a permissive license.

7. Acknowledgements

This research was partially supported by the German Federal Ministry of Education and Research (BMBF) in the project 'KOMPASS' (FKZ 16SV7271K) and by the Deutsche Forschungsgemeinschaft (DFG) in the Cluster of Excellence 'Cognitive Interaction Technology' (CITEC).

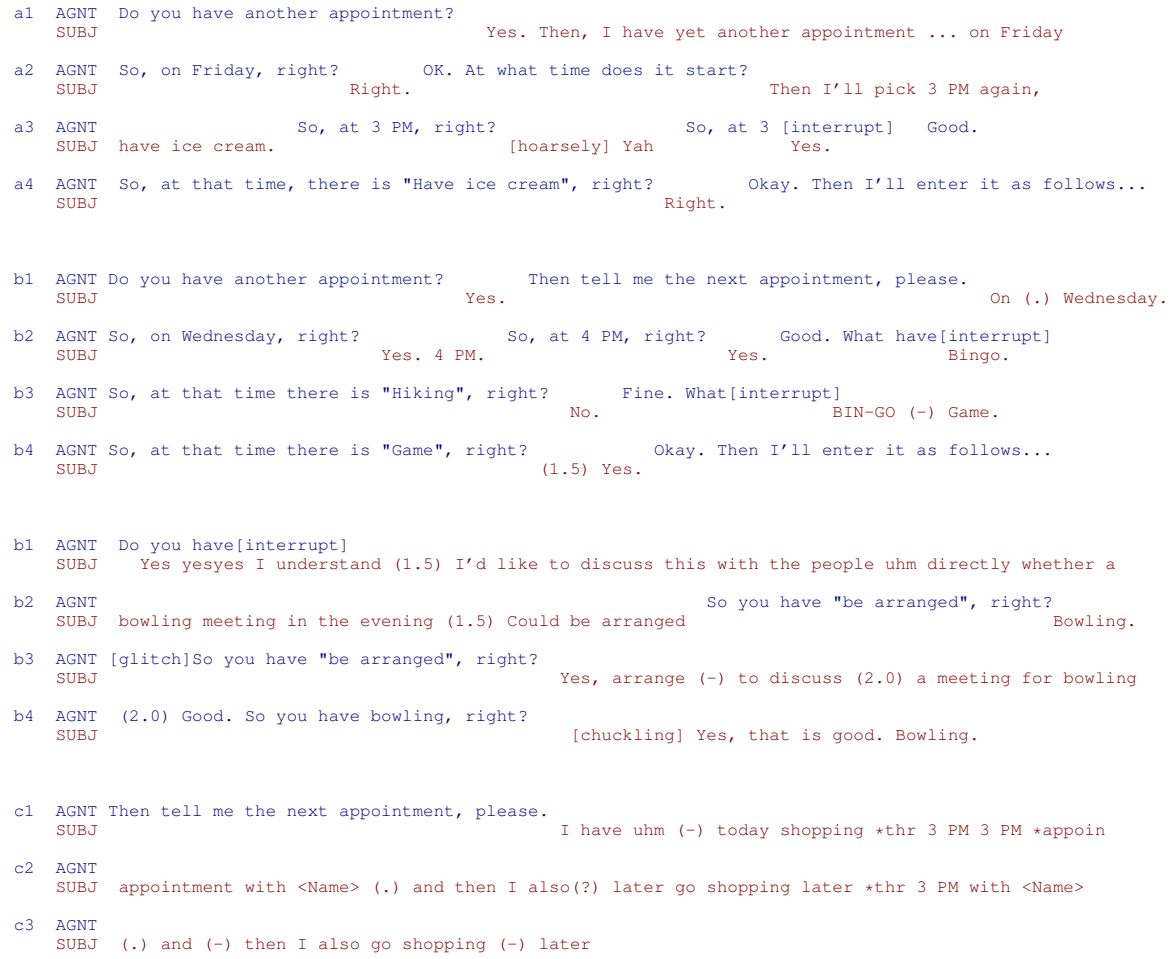


Figure 5: Examples of observed interaction styles (autonomous study): **Top:** older adult, brief but casual style; **second from top:** older adult, brief style; settling on alternative / partial event description; **second from bottom:** older adult, more verbose style; renegotiation; **bottom:** person with noticeable cognitive impairment, verbose turns, exacerbated by dysfluent and unclear articulation.

8. References

[1] GUIDE Consortium, *User Interaction & Application Requirements - Deliverable D2.1*, 2011.

[2] V. Young and A. Mihailidis, "Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review," *Assistive Technology*, vol. 22, no. 2, pp. 99—-112, 2010.

[3] R. Yaghoubzadeh, K. Pitsch, and S. Kopp, "Adaptive grounding and dialogue management for autonomous conversational assistants for elderly users," in *Proceedings of the 15th International Conference on Intelligent Virtual Agents*, ser. LNCS (LNAI), vol. 9238, 2015, pp. 28–38.

[4] D. Schlangen, T. Baumann, H. Buschmeier, O. Buß, S. Kopp, G. Skantze, and R. Yaghoubzadeh, "Middleware for incremental processing in conversational agents," in *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2010, pp. 51–54.

[5] D. Schlangen and G. Skantze, "A general, abstract model of incremental dialogue processing," in *EACL '09 Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, 2009, pp. 710–718.

[6] S. Larsson, "Issue-based dialogue management," Ph.D. dissertation, University of Gothenburg, Sweden, 2002.

[7] H. van Welbergen, D. Reidsma, and S. Kopp, "An incremental multimodal realizer for behavior co-articulation and coordination," in *Proceedings of the 12th International Conference on Intelligent Virtual Agents*, ser. LNCS (LNAI), vol. 7502, 2012, pp. 175–188.

[8] M. P. Aylett and C. J. Pidcock, *The CereVoice Characterful Speech Synthesiser SDK*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 413–414.

[9] R. Yaghoubzadeh, M. Kramer, K. Pitsch, and S. Kopp, "Virtual agents as daily assistants for elderly or cognitively impaired people," in *Proceedings of the 13th International Conference on Intelligent Virtual Agents*, ser. LNCS (LNAI), vol. 8108, 2013, pp. 79–91.

[10] I. Grishkova, R. Yaghoubzadeh, S. Kopp, and C. Vorwerg, "How do human interlocutors talk to virtual assistants? A speech act analysis of dialogues of cognitively impaired people and elderly people with a virtual assistant," *Cognitive Processing*, vol. 15, p. 40, 2014.

Predicting Intelligible Speaking Rate in Individuals with Amyotrophic Lateral Sclerosis from a Small Number of Speech Acoustic and Articulatory Samples

Jun Wang[1,2]*, Prasanna V. Kothalkar*[1]*, Myungjong Kim*[1]*, Yana Yunusova*[3]
Thomas F. Campbell[2]*, Daragh Heitzman*[4]*, Jordan R. Green*[5]

[1]Speech Disorders & Technology Lab, Department of Bioengineering
[2]Callier Center for Communication Disorders
University of Texas at Dallas, Richardson, Texas, United States
[3]Department of Speech-Language Pathology
University of Toronto, Toronto, Canada
[4]MDA/ALS Center, Texas Neurology, Dallas, Texas, United States
[5]Department of Communication Sciences and Disorders
MGH Institute of Health Professions, Boston, MA, United States
wangjun@utdallas.edu

Abstract

Amyotrophic lateral sclerosis (ALS) is a rapidly progressive neurological disease that affects the speech motor functions, resulting in dysarthria, a motor speech disorder. Speech and articulation deterioration is an indicator of the disease progression of ALS; timely monitoring of the disease progression is critical for clinical management of these patients. This paper investigated machine prediction of intelligible speaking rate of nine individuals with ALS based on a small number of speech acoustic and articulatory samples. Two feature selection techniques - decision tree and gradient boosting - were used with support vector regression for predicting the intelligible speaking rate. Experimental results demonstrated the feasibility of predicting intelligible speaking rate from only a small number of speech samples. Furthermore, adding articulatory features to acoustic features improved prediction performance, when decision tree was used as the feature selection technique.

Index Terms: amyotrophic lateral sclerosis, intelligible speaking rate, support vector regression

1. Introduction

Amyotrophic lateral sclerosis (ALS), also referred to as Lou Gehrig's disease, is a fast progressive neurological disease that causes degeneration of both upper and lower motor neurons and affects various motor functions, including speech production [1, 2]. The typical survival time is 2-5 years from the onset time [2]. ALS affects between 1.2 and 1.8 /100,000 individuals and the incidence is increasing at a rate that cannot be accounted for by population aging alone [3]. Approximately 30% of patients present with significant speech abnormalities at disease onset; of the remaining patients, nearly all will develop speech deterioration as the disease progresses [4, 5]. Technology for objective, accurate monitoring of speech decline is critical for providing timely management of speech deterioration in ALS and for extending their functional speech communication. Currently, ALS Functional Rating Scale-Revised (ALSFRS- R) - a self-report evaluation - is used for monitoring the progression of changes across motor function [6]. ALS-FRS-R includes 3 questions pertaining to speech, swallowing, and salivation. Commonly used clinical measures for communication efficiency include speech intelligibility (percentage of words that are understood by listeners) and speaking rate (number of spoken words per minute, WPM), which are not closely correlated. Intelligible speaking rate (also called the communication efficiency index) combines intelligibility and rate in a form of speech intelligibility $\times$ speaking rate, providing an index of intelligible spoken words per minute (WPM) [7, 8, 9].

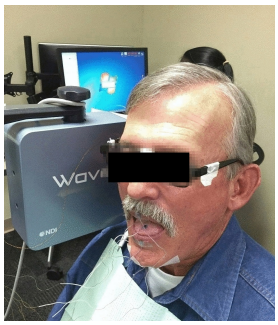
Recent studies have tried to predict the rate of speech intelligibility decline of ALS using an interpretable model based on a comprehensive data set with measures from articulatory, respiratory, resonatory, and phonatory subsystems [10, 11, 12]. Although this approach is promising for understanding the mechanisms of speech decline in ALS, it may not be suitable for clinical environment, given the skill level and the significant time demands required for the data collection and analysis. Novel turnkey and automated speech assessment approaches are, therefore, needed to facilitate clinical diagnosis and management.

Speech signals can be collected using any audio collection devices such as a smart phone and thus can be a great source of information for dysarthria severity estimation. The feasibility of using speech signals revealed promising results in a number of recent studies for disease detection and severity estimation in depression [13, 14], traumatic brain injury [15], and Parkinson's disease detection or severity estimation [16, 17, 18, 19, 20, 21]. Our recent work also showed the feasibility of detection of ALS from speech samples [22]. Estimating the progression of ALS from speech samples using data-driven approaches, however, has rarely been attempted.

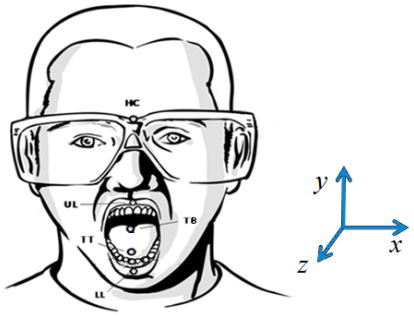
Automatic speech recognition (ASR) systems are a promising but relatively unexplored solution [23, 24]. One significant limitation of ASR for this application is, however, that most approaches require a prohibitively large number of speech samples, since the approach is based on counting the percentage of correctly recognized words. This might be impractical for persons with motor speech disorders due to patient fatigue or variable responses. Yet another challenge of ASR approach is the potential performance variability caused by different speech recognition systems, which is currently understudied.

This project investigated the estimation of speech deteriora-

SLPAT 2016, 7th Workshop on Speech and Language Processing for Assistive Technologies, pages 91–97,
San Francisco, USA, 13 September, 2016.



(a) *Wave System*



(b) *Sensor Locations. Labels are described in text.*

Figure 1: *Data collection setup.*

tion due to ALS from a number of short speech samples. Data-driven approaches were used to predict intelligible speaking rates of individuals with ALS. As ALS is a motor neuron disease, it affects the articulatory movements including tongue and lip motion patterns [9]. Thus we also tested if the inclusion of articulatory movement data on top of acoustic data can benefit the prediction. Previous studies by Hahm and colleagues used quasi-articulatory features that were inversely mapped from acoustic data, which resulted in improvement for detections of Parkinson's condition estimation [25]. We hypothesized that adding articulatory information to the acoustic data might also benefit the speech performance prediction in ALS.

To our knowledge, this project is the first that aims to predict communication efficiency (intelligible speaking rate or intellgible rate) in ALS from a small number of speech (acoustic and articulatory) samples using data-driven approaches. Speech samples are short phrases that are spoken in daily life (e.g., *How are you doing?*). A pre-defined set of speech features was extracted from acoustic and articulatory samples to represent various characteristics of the speech. Two feature selection techniques were used together with support vector regression (SVR) to predict the intelligible speaking rate. We chose to predict intelligible speaking rate (rather than speech intelligibility and speaking rate) at this stage, because intelligible speaking rate is the measure that better represents the communication efficiency level of ALS patients [8]. To understand if articulatory movement data can improve the prediction, three combinations of features (acoustic, acoustic + lip data, acoustic + lip + tongue data) were tested.

Table 1: Speech intelligibility, speaking rate, and intelligible speaking rate in each recorded session.

Session ID	Speech Intelligibility (%)	Speaking Rate (WPM)	Intelligible Rate (WPM)
S01	95.45	136.60	130.38
S02	80.00	147.98	118.38
S03	100.00	182.33	182.33
S04	98.18	172.54	169.40
S05	79.09	121.10	95.78
S06	99.00	164.189	162.54
S07	98.18	110.47	108.46
S08	0.00	41.05	0.00
S09	94.55	111.11	105.05
S10	80.91	108.20	87.54
S11	23.64	80.29	18.98
S12	99.00	108.73	107.64
S13	96.36	33.33	32.12
S14	79.09	71.88	56.85
Average	80.25	113.56	92.59
SD	29.37	40.04	53.51

2. Data Collection

2.1. Participants

Nine patients (five females) with ALS participated in 14 sessions of data collection. The average age at their first visit was 61 years (SD = 11). Table 1 gives the speech intelligibility, speaking rate, and intelligible speaking rate values for each recorded session. Three of the participants contributed data more than once. S04-S05 were from the same participant but with a year gap. S06-08 were from another patient, with five months and nine months intervals between each two consecutive visits. S09-11 were from another patient with four months and eight months gaps between each two consecutive visits.

2.2. Setup and Procedure

An electromagnetic articulograph (Wave system, NDI Inc., Waterloo, Canada) was used for collecting speech acoustic and articulatory data synchronously. Wave is one of the two commonly used electromagnetic motion tracking technologies by tracking small wired sensors that are attached to the subject's tongue, lips, and head [26]. Figure 1a pictures the device and the patient setup. The spatial accuracy of motion tracking using Wave is 0.5 mm when sensors are in the central space of the magnetic field [27].

After a participant was seated next to the Wave magnetic field generator, sensors were attached to the participant's forehead, tongue, and lips. The head sensor was used to track head movement for head-correction of other sensor's data. The four-sensor set - tongue tip (TT, 5-10 mm to tongue apex), tongue back (TB, 20-30 mm back from TT), upper lip (UL), and lower lip (LL) - was used for our experiments as previous studies indicated that the set is optimal for this application [28, 29, 30]. The positions of five sensors attached to a participant's head, tongue and lips were shown in Figure 1b.

All participants were asked to repeat a list of pre-defined phrases multiple times. The phrases were selected based on lists of phrases that are commonly spoken by AAC (alternative and augmentative communication) users in their daily life [31, 32]. The acoustic and articulatory data were recorded synchronously.

Speech intelligibility and speaking rate were obtained by a certified speech-language pathologist with the assistance of SIT software [33]. Intelligible rate was the multiplication of speech intelligibility and speaking rate. The range of intelligible rate in this data set was between 0-182 words per minute (WPM).

2.3. Data Processing

While raw acoustic data (sampling rate 16Khz, 16 Bit resolution) were used directly for feature extraction, a processing procedure was performed on the articulatory data prior to analysis. The two steps of articulatory data processing included head correction and low pass filtering. The head translations and rotations were subtracted from the tongue and lip data to obtain head-independent tongue and lip movements. The orientation of the derived 3D Cartesian coordinates system is displayed in Figure 1b, in which x is left-right, y is vertical, and z is front-back directions. A low pass filter (i.e., 20 Hz) was applied to remove noise [26].

Invalid samples were rare and were excluded from the analysis. A valid sample contained both valid acoustic and articulatory data. A total of 944 valid samples were recorded. The range of number of samples from individual patients was from 39 to 80.

3. Method

The method of intelligible speaking rate prediction in this project involved two major steps: feature preparation and regression, where feature preparation included feature extraction and selection. The goal of feature extraction was to obtain content-independent acoustic and articulatory features from the data samples. Feature selection was to reduce the data size by choosing the best features for regression. Regression aimed to predict a target score (intelligible speaking rate) from features that are extracted from a data sample.

3.1. Feature Extraction

The script provided in [22] was used for extracting acoustic and articulatory features from acoustic and articulatory motion data, respectively. The script was modified based on that provided in [34]. The window size was 70 ms and the frame shift was 35 ms. The script extracted up to 6,373 pre-defined acoustic features that were categorized in groups such as jitter, shimmer, MFCC, and spectral features. However, low frequency articulatory data do not contain these information. The following feature groups were disabled for articulatory feature extraction [22] :

Jitter, Shimmer, logHNR, Rfilt, Rasta, MFCC, Harmonicity, and Spectral Rolloff.

For each feature group, the following features were calculated and fed into the final feature set before fed into a feature selection technique: mean, flatness, posamean (position of the algorithmic mean), range, maxPos, minPos, centroid, stddev, skewness (a measure of the asymmetry of the spectral distribution around its centroid), kurtosis (an indicator for the peakedness of the spectrum), etc. Please refer to [34, 35] for details of these features.

Therefore, for each dimension (x, y, or z) of a sensor, 1,200 features were extracted. In total, 20,733 features (6,373 acoustic feature + 3,600 articulatory features $\times$ 4 sensors (Tongue Tip, Tongue Body Back, Upper Lip, and Lower Lip) were used in the regression test.

3.2. Feature Selection

Feature selection [36] was performed to reduce the data to the most significant features. We used decision tree regression and gradient boosting as the feature selection procedures.

3.2.1. Decision Tree

Decision trees are rule-based, non-linear classification/regression models that perform recursive partitioning on the data by separating the data into disjoint branches (thus forming a tree structure) for classification or regression [37]. There are a number of ways to measure the quality of a split or branching. We used MSE (mean squared error) as the measure in this project, which is equal to variance reduction as feature selection criterion.

Decision tree-based regression fits the best least squared error criterion to the data. The expected value at each leaf node that minimizes this least squared error is the average of the target values within each leaf l.

$$v_l = \frac{1}{|D_l|} \sum_{D_l} y_i \tag{1}$$

where D_l is the set of samples that are partitioned to leaf l and y_i is the target value of sample i in set D_l.

The splitting criterion is to minimize the fitting error of the resultant tree. The fitting error was defined as the average of the squared differences between the target values Y_l at a leaf node l and the mean value v_l. Error of a tree was defined as the weighted average of the error in its leaves and the error of a split is the weighted average of the error of its resulting sub-nodes.

3.2.2. Gradient Boosting

Gradient boosting [38] applies boosting to regression models by selecting simpler base learners to current pseudo residuals by minimizing least squares loss at each iteration. The pseudo residuals are the gradient of the loss functional that is to be minimized, with respect to model values at each training data point, evaluated at the current step. Given training samples $x_i \in R^d, i = 1, \ldots, n$, and a regression vector $\mathbf{y} \in R^n$ such that $y_i \in R$ we want to find a function $F^{(*)}(\mathbf{x})$ that maps $\mathbf{x}$ to y, to minimize the expected value of some specified loss function $\Delta(y, F(\mathbf{x}))$ over the joint distribution of all ($\mathbf{x}$,y) values. Boosting approximates $F^{(*)}(\mathbf{x})$ by a stage-wise summation of the form

$$F(\mathbf{x}) = \sum_{\mathbf{i=0}}^{\mathbf{N}} \gamma_\mathbf{i} \mathbf{g_i}(\mathbf{x}; \mathbf{a_i}) \tag{2}$$

where the functions $g_i(\mathbf{x}; \mathbf{a_i})$ are chosen as base classifiers of $\mathbf{x}$ in stage i where $\mathbf{a_i}$ is set of parameters. γ_i is the expansion coefficient for stage i.

Gradient boosting solves for arbitrary loss functions for each stage in two steps. First, it fits the function $g_i(\mathbf{x}; \mathbf{a_i})$ to current pseudo residuals by minimizing the least squares loss. Second, the optimal value of the expansion coefficient γ_i was found by single parameter optimization based on a general loss criterion. We selected gradient boosting in this experiment because the model generally works well with small datasets [38].

3.3. Selected Features

The partial lists of features that were selected by decision tree and gradient boosting are given below. Decision tree selected 57 features in total; while gradient boosting selected 517 features. The features selected from articulatory data are indicated in parenthesis; otherwise, the features are selected from acoustic data. Below are the top 10 selected features by decision tree:

1. audspec_lengthL1norm_sma_lpgain
2. shimmerLocal_sma_de_iqr1-3
3. logHNR_sma_percentile99.0
4. F0final_sma_quartile3
5. mfcc_sma[7]_quartile1
6. pcm_fftMag_spectralFlux_sma_quartile1
7. audSpec_Rfilt_sma_de[2]_quartile1
8. pcm_fftMag_fband3-8_sma_de_stddevRisingSlope (TTz)
9. F0final_sma_stddev
10. pcm_fftMag_spectralKurtosis_sma_peakMeanAbs (TTy)

where mfcc stands for mel-frequency cepstral coefficients 1-12; fft denotes fast Fourier transform; pcm means pulse-code modulation, the standard digital representation of analog signals; quartile1 denotes the first quartile (the 25% percentile); quartile 2 denotes the second quartile (the 50% percentile); quartile 3 denotes the third quartile (the 75% percentile); iqr1-3 means the inter-quartile range: quartile3-quartile1; Mag means magnitude; Rfilt means Relative Spectral Transform (RASTA)-style filtered; F0final means the smoothed fundamental frequency (pitch) contour; stddev denotes the standard deviation of the values of the contour; kurtosis is an indicator for the peakedness of the spectrum; sma means smoothing by moving average; de means delta; stddevRisingSlope is the standard deviation of rising slopes, i.e. the slopes connecting a valley with the following peak. The suffix sma appended to the names of the low-level descriptors indicates that they were smoothed by a moving average filter with window length 3 [35]. Spectral flux (F_S^t) for N FFT bins at time frame t is computed as

$$F_S^t = \sqrt{\frac{1}{n}\sum_{f=1}^{N}(\frac{X^t(f)}{E^t} - \frac{X^{t-1}(f)}{E^{t-1}})^2} \tag{3}$$

where E^t is energy at time frame t; $X^t(f)$ is the FFT bin f based on data X at time t. Further, audspec stands for auditory spectrum; shimmerLocal is the local (frame-to-frame) Shimmer (amplitude deviations between pitch periods); lpgain implies the linear predictive coding gain; lengthL1norm is the magnitude of the L1 norm; percentile99.0 is the outlier-robust maximum value of the contour, represented by the 99% percentile and logHNR is the log of the ratio of the energy of harmonic signal components to the energy of noise like signal components. A more descriptive explanation, for example for mfcc_sma[7]_quartitle1, is the 25% percentile of the 7^{th} MFCC that was smoothed using an averaging filter with window length 3.

Below are the top 10 selected features by gradient boosting:

1. audspec_lengthL1norm_sma_lpgain
2. pcm_fftMag_fband1000-4000_sma_percentile1.0
3. F0final_sma_linregc2
4. logHNR_sma_percentile99.0
5. mfcc_sma[6]_quartile2
6. pcm_fftMag_fband3-8_sma_de_lpgain(TBx)
7. audspecRasta_lengthL1norm_sma_peakDistStddev
8. pcm_fftMag_spectralFlux_sma_stddevRisingSlope
9. F0final_sma_percentile99.0
10. audSpec_Rfilt_sma[19]_iqr1-3

where percentile1.0 is the outlier-robust minimum value of the contour, represented by the 1% percentile; linregc2 is the offset (c from y = mx+c) of a linear approximation of the contour; fband denotes frequency band; audspecRasta is the Relative Spectral Transform applied to Auditory Spectrum.

The features were selected based on a feature importance score, which is based on the (normalized) total reduction of the variance brought by that feature [37]. These features with highest importance scores were selected.

3.4. Support Vector Regression

Support vector regression is a regression technique that is based on support vector machine [39], was used as the regression model in this project. SVR is a soft-margin regression technique that depends only on a subset of the training data, because the cost function for building the model does not care about training points that are beyond the margin [40], which is similar with SVM. Details on the introduction of SVR can be found in [41]. We used LIBSVM to implement the experiment [42]. After a preliminary test, ν-SVR [43] outperformed or was comparable to others, thus was selected for regression in this experiment. ν-SVR is a variation of standard SVR, which uses ν to control the ϵ. Given training vector $x_i \in R^d, i = 1, \ldots, n$, and a regression vector $\mathbf{y} \in R^n$ such that $y_i \in R$, the SVR optimization problem is

$$\begin{aligned} \min_{\mathbf{w},\mathbf{b},\xi,\epsilon\in\mathcal{R}^\mathbf{n}} \quad & \frac{1}{n}\sum_{i=1}^{n}(\xi_i) - \nu\epsilon + \frac{1}{2}\|w\|^2 \\ \text{subject to} \quad & y_i(\mathbf{w}^\intercal\phi(\mathbf{x_i}) + \mathbf{b}) \geq \epsilon - \xi_\mathbf{i}, \\ & \xi_i \geq 0, i = 1, \ldots, n, \qquad \epsilon \geq 0 \end{aligned} \tag{4}$$

A kernel function is used to describe the distance between two samples (i.e., r and s in Equation 5). The following radial basis function (RBF) was used as the kernel function K_{RBF} in this study, where γ is an empirical parameter ($\gamma = 1/n$, by default, where n is the number of features) [26]:

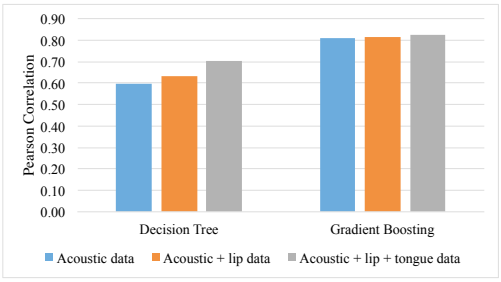
$$K_{RBF}(r, s) = exp(1 - \gamma||r - s||). \tag{5}$$

Please refer to [42] for more details about the implementation of the SVR. All feature values were normalized using z-score before they were fed into SVR.

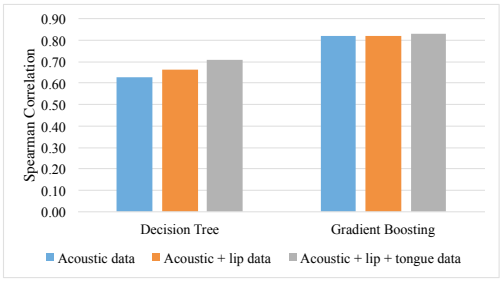
3.5. Experimental Design

As mentioned previously, we tested the prediction on three configurations of data to understand the performance using acoustic signals only and if adding articulatory information is beneficial for the regression. The three configurations of data were acoustic data only, acoustic + lip data, and acoustic + lip data + tongue data.

Three-fold cross validation strategy was used, where all 14 sessions of data were divided into three groups with a balanced distribution of intelligible rates. Initially all the 14 data collections were arranged in ascending order by intelligible rate



(a) *Results measured by Pearson correlation.*



(b) *Results measured by Spearman correlation.*

Figure 2: *Experimental results based on acoustic data only, acoustic + lip data, and acoustic + lip + tongue data using support vector regression and two feature selection techniques.*

(labelled from 1 to 14). Then a jack-knife strategy was used to choose the groups as testing data and the rest as training data. The three folds are sessions (1, 4, 7, 10, 13), (2, 5, 8, 11, 14), (3, 6, 9, 12) (in Table 1). The last validation had four sessions for testing. The data size for testing was about 120 - 360 samples (and the rest for training) in each validation.

Two correlations, Pearson and Spearman, were used to evaluate the performance of the regression. We used both correlations just in case they provide complementay information, becuase of their different characteristics. Pearson correlation is more sensitive than Spearman correlation for outliers [34]; Pearson is typically applied for normally distributed data. The data size is relatively small and the distribution was unknown in this project. Thus, using both correlations (rather than just one of them) may provide more detailed information for interpreting the experimental results. A higher correlation between the estimated rate and the actual rate indicates a better performance.

4. Results and Discussion

Figure 2 gives the results of the regression experiments using SVR and two feature selection techniques, decision tree and gradient boosting, based on acoustic data only, acoustic + lip data, and acoustic + lip + tongue data. The results were measured by Pearson correlation (Figure 2a) and Spearman correlation (Figure 2b). As shown in Figure 2a, the three data configurations obtained Pearson correlations, 0.60, 0.63, and 0.70, respectively when using decision tree, and 0.81, 0.82, and 0.83 when using gradient boosting. The three data configurations obtained Spearman correlations, 0.62, 0.66, and 0.71 respectively when using decision tree, and 0.82, 0.82, and 0.83 when using gradient boosting. There was no difference among the values measured by Pearson or Spearman correlation.

The experimental results indicated the feasibility of predicting intelligible speaking rate from a small number of speech acoustic (and articulatory) samples.

In addition, the results demonstrated that adding articulatory data could improve the performance when using decision tree as the feature selection but not when using gradient boosting. When lip data were added to the acoustic data, the prediction performance was improved when decision tree was used for feature selection. Adding both lip and tongue data obtained the best performance. These findings are consistent with the literature that speech motor function decline (particularly in the articulatory subsystem) are early indicators of the bulbar deterioration in ALS [7]. The added benefit of articulatory data was not obtained when using gradient boosting possibly because this approach was more effective in selecting acoustic features than using the decision tree approach, which required the added articulatory features.

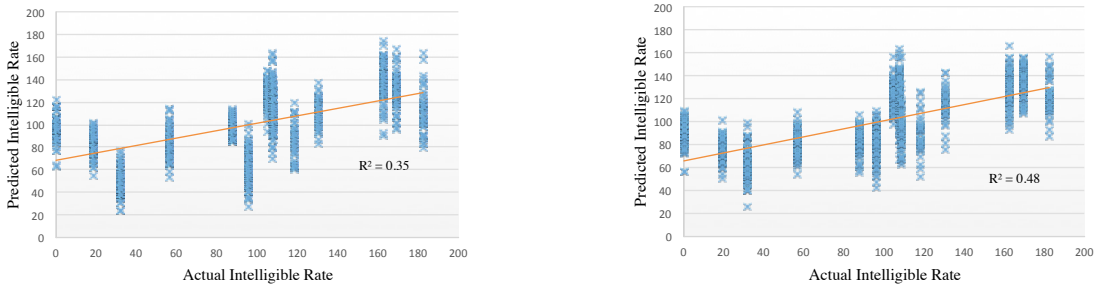
These finding suggested the possibility, in the future, of developing mobile technologies that can collect speech acoustic and lip (via a webcam) as a practical tool for monitoring the ALS speech performance decline as an indicator of disease progression. There are currently logistical obstacle for acquiring tongue data [26] (compared with acoustic data). However, with the availability of portable devices such as portable ultrasound, we anticipated that tongue data will be more accessible in the near future. An alternative solution for tongue data collection is acoustic-to-articulatory inverse mapping [25].

Although comparison of the feature selection techniques was not a focus in this paper, the experimental results indicated that gradient boosting outperformed decision tree. Gradient boosting was so powerful such that adding articulatory information did not show benefit. This finding suggested that feature selection is critical. More feature selection techniques will be explored in the next step of this study.

Figure 3 gives the scatter plots of the measured intelligible rate and predicted intelligible rates using SVR + decision tree on acoustic features only (Figure 3a) and using both acoustic and articulatory features (Figure 3b). Each marker (cross) in the figure represents the measured and predicted intelligible rates on one data sample (a short phrase produced by a patient). As described earlier, each patient produced multiple samples in one session.

A linear regression was applied on both Figure 3a and 3b. The R-squared values illustrated how close the data are to the fitted regression line. A larger value is better. As illustrated in Figure 3, adding articulatory data on top of acoustic data obtained a larger R^2 value, which indicated articulatory features (tongue + lips) improved the prediction on top of acoustic features (using decision tree as the feature selection technique). Specifically, adding articulatory data significantly improved the prediction for some sessions, for example, S13 (with intelligible rate 32.12 WPM) and S05 (with intelligible rate 95.78 WPM). A further analysis is needed to discover how articulatory data affect the prediction performance for these sessions (or patients).

Limitation. The current approach was purely data-driven and used a large number of low-level acoustic and articulatory features. Inclusion of high-level, interpretable features would help the understanding of how these individual features could contribute to the speech decline. Examples of interpretable features include formant centralization ratio [19], intonation [20], and prosody [44], which have already been used for other diseases (e.g., Parkinson's disease).



(a) *Acoustic data only.*

(b) *Acoustic + lip + tongue data.*

Figure 3: *Scatter plots of actual intelligible speaking rate (words per minute) and the predicted values using SVR + decision tree for two data configurations: (a) acoustic data only, and (b) acoustic + lip + tongue data.*

5. Conclusions and Future Work

This paper investigated the automatic assessment of speech performance in ALS from a relatively small number of speech acoustic and articulatory samples. Support vector regression with two feature selection techniques (decision tree and gradient boosting) were used to predict intelligible speaking rate from speech acoustic and articulatory samples. Experimental results showed the feasibility of intelligible speaking rate prediction from acoustic samples only. Adding articulatory data further improved the performance when decision tree was used as the feature selection technique. Particularly, even only lip information was added, the prediction performance was significantly improved. The best results were obtained when both lip and tongue data were added.

The next step of this research would further verify this finding using a larger data set and other feature selection and regression techniques (e.g., deep neural network [25]).

6. Acknowledgements

This work was in parted supported by the National Institutes of Health through grants R01 DC013547 and R03 DC013990, and the American Speech-Language-Hearing Foundation through a New Century Scholar grant. We would like to thank Dr. Panying Rong, Dr. Anusha Thomas, Jennifer McGlothlin, Jana Mueller, Victoria Juarez, Saara Raja, Soujanya Koduri, Kumail Haider, Beiming Cao and the volunteering participants.

7. References

[1] M. C. Kiernan, S. Vucic, B. C. Cheah, M. R. Turner, A. Eisen, O. Hardiman, J. R. Burrell, and M. C. Zoing, "Amyotrophic lateral sclerosis," *The Lancet*, vol. 377, pp. 942–955, 2011.

[2] M. Strong and J. Rosenfeld, "Amyotrophic lateral sclerosis: A review of current concepts," *Amyotrophic Lateral Sclerosis and Other Motor Neuron Disorders*, vol. 4, pp. 136–143, 2003.

[3] E. Beghi, G. Logroscino, A. Chi, O. Hardiman, D. Mitchell, R. Swingler, and B. J. Traynor, "The epidemiology of ALS and the role of population-based registries," *Biochimica et Biophysica Acta*, vol. 1762, pp. 1150–1157, 2011.

[4] R. D. Kent, R. L. Sufit, J. C. Rosenbek, J. F. Kent, G. Weismer, R. E. Martin, and B. Brooks, "Speech deterioration in amyotrophic lateral sclerosis: a case study," *Journal of Speech, Language and Hearing Research*, vol. 34, pp. 1269–1275, 1991.

[5] S. E. Langmore and M. Lehman, "The orofacial deficit and dysarthria in ALS," *Journal of Speech and Hearing Research*, vol. 37, pp. 28–37, 1994.

[6] J. M. Cedarbaum, N. Stambler, E. Malta, C. Fuller, D. Hilt, B. Thurmond, A. Nakanishi, and BDNF_ALS_Study_Group, " The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function," *Journal of the Neurological Sciences*, vol. 169, pp. 13–21, 1999.

[7] J. R. Green, Y. Yunusova, M. S. Kuruvilla, J. Wang, G. L. Pattee, L. Synhorst, L. Zinman, and J. D. Berry, "Bulbar and speech motor assessment in ALS: Challenges and future directions," *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, vol. 14, pp. 494–500, 2013.

[8] K. M. Yorkston and D. R. Beukelman, " Communication efficiency of dysarthric speakers as measured by sentence intelligibility and speaking rate," *Journal of Speech and Hearing Disorders*, vol. 46, pp. 296–301, 1981.

[9] Y. Yunusova, J. R. Green, L. Greenwoode, J. Wang, G. Pattee, and L. Zinman, "Tongue movements and their acoustic consequences in ALS," *Folia Phoniatrica et Logopaedica*, vol. 64, pp. 94–102, 2012.

[10] Y. Yunusova, J. S. Rosenthal, J. R. Green, P. Rong, J. Wang, and L. Zinman, " Detection of bulbar ALS using a comprehensive speech assessment battery," in *Proc. of the International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, 2013, pp. 217–220.

[11] P. Rong, Y. Yunusova, J. Wang, and J. R. Green, " Predicting early bulbar decline in amyotrophic lateral sclerosis: A speech subsystem approach," *Behavioral Neurology*, no. 183027, pp. 1–11, 2015.

[12] P. Rong, Y. Yunusova, J. Wang, L. Zinman, G. L. Pattee, J. D. Berry, B. Perry, and J. R. Green, "Predicting speech intelligibility decline in amyotrophic lateral sclerosis based on the deterioration of individual speech subsystems," *PLoS ONE*, vol. 11, no. 5, p. e0154971, 2016.

[13] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, 2015.

[14] T. F. Quatieri and N. Malyska, " Vocal-source biomarkers for depression: A link to psychomotor activity," in *Proc. of INTERSPEECH*, 2012, pp. 1059–1062.

[15] M. Falcone, N. Yadav, C. Poellabauer, and P. Flynn, " Using isolated vowel sounds for classification of mild traumatic brain injury," in *Proc. of ICASSP*, 2012, pp. 7577–7581.

[16] A. Tsanas, M. Little, P. McSharry, and L. Ramig, "Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity," *Journal of the Royal Society Interface*, vol. 8, pp. 842–855, 2011.

[17] M. Little, P. E. McSharry, E. J. Hunter, J. Spielman, and L. O. Ramig, "Suitability of dysphonia measurements for telemonitoring of Parkinsons disease," *IEEE Transactions on Biomedical Engineering*, vol. 56, pp. 1015–1022, 2009.

[18] A. Tsanas, M. Little, P. McSharry, J. Spielman, and L. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of Parkinsons disease," *IEEE Transactions on Biomedical Engineering*, vol. 59, pp. 1264–1271, 2012.

[19] S. Sapir, L. O. Ramig, J. L. Spielman, and C. Fox, "Formant Centralization Ratio (FCR): A proposal for a new acoustic measure of dysarthric speech," *Journal of Speech, Language, and Hearing Research*, vol. 53, pp. 114–125, 2010.

[20] S. Skodda, W. Grnheit, and U. Schlegel, "Intonation and speech rate in parkinson's disease: General and dynamic aspects and responsiveness to levodopa admission," *Journal of Voice*, vol. 25, no. 4, pp. e199 – e205, 2011.

[21] J. C. Vasquez-Correa, J. R. Orozco-Arroyave, J. D. Arias-Londono, J. F. Vargas-Bonilla, and E. Noth, "New computer aided device for real time analysis of speech of people with parkinson's disease," *Revista Facultad de Ingenieria Universidad de Antioquia*, no. 72, pp. 87–103, 2014.

[22] J. Wang, P. V. Kothalkar, B. Cao, and D. Heitzman, "Towards automatic detection of amyotrophic lateral sclerosis from speech acoustic and articulatory samples," in *Proc. of INTERSPEECH*, 2016.

[23] T. Haderlein, C. Moers, B. Mobius, F. Rosanowski, and E. Noth, "Intelligibility rating with automatic speech recognition, prosodic, and cepstral evaluation," *Proceedings of Text, Speech and Dialogue (TSD), ser. Lecture Notes in Artificial Intelligence*, vol. 6836, pp. 195–202, 2011.

[24] R. Vich, J. Nouza, and M. Vondra, "Automatic speech recognition used for intelligibility assessment of text-to-speech systems," *In Verbal and Nonverbal Features of Human-Human and Human-Machine Interactions, Lecture Notes in Computer Science*, vol. 5042, pp. 136–148, 2008.

[25] S. Hahm and J. Wang, "Parkinsons condition estimation using speech acoustic and inversely mapped articulatory data," in *Proc. of INTERSPEECH*, 2015, pp. 513–517.

[26] J. Wang, J. Green, A. Samal, and Y. Yunusova, "Articulatory distinctiveness of vowels and consonants: A data-driven approach," *Journal of Speech, Language, and Hearing Research*, vol. 56, no. 5, pp. 1539–1551, 2013.

[27] J. Berry, "Accuracy of the NDI wave speech research system." *Journal of Speech, Language, and Hearing Research*, vol. 54, no. 5, pp. 1295–301, 2011.

[28] J. Wang, J. Green, and A. Samal, "Individual articulator's contribution to phoneme production," in *Proc. of ICASSP*, Vancouver, Canada, 2013, pp. 7785–7789.

[29] J. Wang, S. Hahm, and T. Mau, "Determining an optimal set of flesh points on tongue, lips, and jaw for continuous silent speech recognition," in *Proc. of ACL/ISCA Workshop on Speech and Language Processing for Assistive Technologies*, 2015, pp. 79–85.

[30] J. Wang, A. Samal, P. Rong, and J. R. Green, "An optimal set of flesh points on tongue and lips for speech-movement classification," *Journal of Speech, Language, and Hearing Research*, vol. 59, pp. 15–26, 2016.

[31] D. R. Beukelman, K. M., Yorkston, M. Poblete, and C. Naranjo, "Analysis of communication samples produced by adult communication aid users," *Journal of Speech and Hearing Disorders*, vol. 49, pp. 360–367, 1984.

[32] J. Wang, A. Samal, J. Green, and F. Rudzicz, "Sentence recognition from articulatory movements for silent speech interfaces," in *Proc. of ICASSP*, Kyoto, Japan, 2012, pp. 4985–4988.

[33] D. R. Beukelman, K. M. Yorkston, M. Hakel, and M. Dorsey, "Speech Intelligibility Test (SIT) [Computer Software]," 2007.

[34] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. H. nig, J. R. Orozco-Arroyave, E. Noth, Y. Zhang, and F. Weninger, "The INTERSPEECH 2015 Computational Paralinguistics Challenge: Nativeness, Parkinsons & Eating Condition," in *Proc. of INTERSPEECH*, 2015, pp. 478–482.

[35] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, 2010, pp. 1459–1462.

[36] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Resarch*, vol. 3, pp. 1157–1182, 2003.

[37] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.

[38] J. H. Friedman, "Stochastic gradient boosting," *Computational Statistics & Data Analysis*, vol. 38, no. 4, pp. 367–378, 2002.

[39] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[40] H. Drucker, C. J. C. Burges, L. Kaufman, A. J. Smola, and V. N. Vapnik, *Support Vector Regression Machines*. MIT Press, 1997, vol. 9.

[41] A. J. Smola and B. Schlkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, 2004.

[42] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[43] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett, "New support vector algorithms," *Neural computation*, vol. 12, no. 5, pp. 1207–1245, 2000.

[44] S. Skodda, H. Rinsche, and U. Schlegel, "Progression of dysprosody in Parkinson's disease over time - A longitudinal study," *Movement Disorders*, vol. 24, no. 5, pp. 716–722, 2009.

Combining word prediction and r-ary Huffman coding for text entry

Seung Wook Kim[1], *Frank Rudzicz*[2,1]

[1]Department of Computer Science, University of Toronto;
[2]Toronto Rehabilitation Institute, University Health Network;
seungwook.kim@mail.utoronto.ca, frank@cs.toronto.edu

Abstract

Two approaches to reducing effort in switch-based text entry for augmentative and alternative communication devices are word prediction and efficient coding schemes, such as Huffman. However, character distributions that inform the latter have never accounted for the use of the former. In this paper, we provide the first combination of Huffman codes and word prediction, using both trigram and long short term memory (LSTM) language models. Results show a significant effect of the length of word prediction lists, and up to 41.46% switch-stroke savings using a trigram model.

1. Introduction

There are approximately 270,000 people in North America with spinal cord injuries, approximately 47% of whom develop quadriplegia (also called tetraplegia), which is a partial or total paralysis of the limbs and torso [1, 2]. In addition to these traumatic losses of motor function, millions more are affected by neuromotor disorders, collectively called *dysarthria*, that impair the production of speech secondary to various congenital or traumatic conditions, including cerebral palsy, stroke, Parkinson's disease, and multiple sclerosis.

Individuals with communication disorders often use augmentative and alternative communication (AAC) technologies to express themselves, specifically to synthesize speech from typed text or symbol sequences. These systems can employ a wide range of inputs, including hand gestures, typing, or eye and head movements [3] that are designed to minimize muscle movement, given global motor deficits. These modalities can interact with either screen-based or screen-free paradigms in which input is transduced to a cursor position [4]. Typically, symbols are selected when the user either dwells on them or performs a specific action, such as blinking or activating a switch, as shown in Figure 1.

Figure 1: Example of a two-button head-switch mounted on a wheelchair. Image used by permission of the Tetra Society of North America.

Through a series of local interviews with AAC users, we have found that screen-based approaches can interfere with certain social aspects of conversation. In particular, users emphasized that screens often form a barrier to eye contact between conversants and that, given a shared screen, conversation partners will often "read-as-they-go," and interrupt the speaker, resulting in editorialization. For these reasons, we are optimizing a screen-free system using eye and head movements.

In this paper, we assign codes to alphanumeric English characters using r-ary Huffman coding, as is typical. However, since AAC devices are also likely to benefit from *word prediction*, the distribution of those characters in training data will not necessarily resemble actual use. For example, although the letter '*e*' is quite frequent, if it tends to occur to-

SLPAT 2016, 7th Workshop on Speech and Language Processing for Assistive Technologies, pages 98–103,
San Francisco, USA, 13 September, 2016.

wards the ends of words, it is less likely to be typed if those words can be accurately predicted from context. We therefore provide the first work that adjusts Huffman codes given distributions subsequent to word prediction, using both trigram and long short term memory (LSTM) language models. The result is up to 41.46% switch-stroke savings using a trigram model.

1.1. Previous work

Previous AAC systems for gestural text entry have sought to minimize selection complexity by limiting the number of possible inputs. The H4, EdgeWrite, and 'Left, Up, Right, Down' Writer systems all relied on codes that used four discrete inputs [5, 6, 7], typically target regions placed at the edges or corners of a screen. The MDITIM (Minimal Device Independent Text Input Method) system uses a similar convention, with four inputs dedicated to the coding of characters and one input reserved as a modifier, for example, to achieve capitalization [8]. In order to further simplify the input process, the H4 and MDITIM systems, unlike EdgeWrite, have used prefix-free codes to avoid the need for a unique termination event, such as a finger-up or blink, to designate the end of each character [5, 8].

Expert users, with about 2.5 hours of experience using the EyeS eye gesture communication system, had text communication rates of 6.8 words per minute (wpm), as compared with typical speech rates of 130-200 wpm, and typing rates of 30-40 wpm for unimpaired typists [9, 10, 11]. Similarly, users with 5.0 hours of practice using the MDITIM had an average text entry speed of less than 10 wpm [8]. One approach to improving communication rate is to reduce the number of inputs needed to enter each character. The H4 system uses Huffman codes to form a prefix-free code, and resulted in an average text entry rate of 20 wpm after 6.5 hours of experience [12]. Roark *et al.* [13] also uses Huffman coding to select the symbols to highlight during character scanning process, minimizing the expected bits per symbol.

Word prediction is another strategy for optimizing text entry. Trnka *et al.* [14] showed that word prediction, using a recency-of-use model, increased communication rates in an AAC-like onscreen keyboard system and that more advanced methods based on statistical language modelling proved more effective, increasing communication rates by 56.8%. The number of options presented is an important factor – longer lists increases the chances that the desired word will be found, but this also increases the visual or auditory scan time to evaluate the list. Mackenzie [15] suggested that a list size of $N = 5$ is optimal.

2. Data

We use three data sets:

Wall Street Journal (WSJ) Selected 2,499 stories from a three-year WSJ collection consisting of 1,098,785 word tokens (43,283 word forms). This dataset contains the most formal language of the three databases.

Essays A collection of essays, poems, and short stories from Grade 11 students in high-schools across Ontario recorded as part of their regular curricula. This consists of 5,831,405 word tokens (114,113 word forms) across 5,448 documents. The formality of the writing is appropriate for teenage writers.

NUS Short Message Service (SMS) Corpus [16] A collection of 55,835 SMS messages collected by the NLP group of the National University of Singapore. This dataset consists of 548,210 word tokens (33,694 word forms) and represents the least formal language of the three databases here.

For our purposes, alphabets are reduced to lowercase alphanumerics and 'space'. All capital letters are changed to lowercase equivalents and extraneous characters are deleted. Additional datasets were considered, including some artificial simulations of AAC text, but these were either too small for our purposes, or provided no additional benefit to the data sets described above.

3. Methods

We train language models to build our word prediction system that produces the list of N most probable next words given the history of characters typed. Each alphanumeric English characters and the indices of the prediction list is assigned a code using r-ary Huffman coding based on the information we get from the word prediction system, assuming that the

user types with an AAC system that has r switches. Input savings by using the word prediction system is calculated for varying N and r values.

We describe the two language models used in word prediction in section 3.1, and our implementation of r-ary Huffman coding in section 3.2.

3.1. Language models

We train two types of language model for each data set. Each produces an N-best prediction list for each word w_i given the previous words $w_{i-n+1}, ..., w_{i-1}$. That is, we choose the top N probabilities from the list L such that

$$L = \{P(w^j | w_{i-n+1} ... w_{i-1}) : 0 \leq j < |V|\} \quad (1)$$

where $|V|$ is the size of the vocabulary V, and w^j is the j-th word in V.

3.1.1. Trigram model

We compute the probability of corpus C:

$$P(C) = \prod_{i=1}^{||C||} P(w_i | w_1 ... w_{i-1}) \quad (2)$$

To address sparseness, we apply Witten-Bell smoothing [17] which linearly interpolates the trigram probability and lower-order smoothed probabilities recursively. In general, the nth-order Witten-Bell probability is:

$$\begin{aligned} &P_{wb}(w_i | w_{i-n+1} ... w_{i-1}) = \\ &\lambda_{w_{i-n+1}^{i-1}} P(w_i | w_{i-n+1} ... w_{i-1}) + \\ &(1 - \lambda_{w_{i-n+1}^{i-1}}) P_{wb}(w_i | w_{i-n+2} ... w_{i-1}) \end{aligned} \quad (3)$$

The parameters $\lambda_{w_{i-n+1}^{i-1}}$ are computed by

$$\lambda_{w_{i-n+1}^{i-1}} = 1 - \frac{N(w_{i-n+1}^{i-1})}{N(w_{i-n+1}^{i-1}) + SC(w_{i-n+1}^{i-1})} \quad (4)$$

where

$$N(w_{i-n+1}^{i-1}) = |\{w_i : count(w_{i-n+1} ... w_{i-1} > 0\}| \quad (5)$$

$$SC(w_{i-n+1}^{i-1}) = \sum_{w_j} count(w_{i-n+1} ... w_{i-1} w_j) \quad (6)$$

The intuition is to give more weight to trigrams in the training set, and to back off to the lower-order probabilities for those that are not.

Trigram t	Count	$logP(t)$	$logP_{wb}(t)$
come up with	20	-0.1140	-0.1760
come up to	0	$-\infty$	-1.8059
come up sing	0	$-\infty$	-6.1763

Table 1: Example trigram probabilites and smoothed probabilites from the WSJ dataset.

3.1.2. Long Short-Term Memory model

Long short-term memory (LSTM) [18] units are a special type of unit in recurrent neural networks (RNNs) designed to solve the vanishing-exploding gradient problem.

Let $s_1, ..., s_N$ be sentences in corpus C, which has N sentences. Suppose $w_1^i, ..., w_J^i$ are words in sentence s_i with J words. We define x_k^i to be the vector word *embedding* of w_k^i. We also define $h_t^l \in \Re^n$ to be the hidden state of layer l at timestep t. Then, for a sequence $w_1^i, ..., w_j^i$, we have $x_k^i = h_k^1$ for $1 \leq k \leq j$. We apply dropout regularization only to non-recurrent connections:

$$\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \rho \end{pmatrix} T_{2n,4n} \begin{pmatrix} D(h_t^{l-1}) \\ h_{t-1}^l \end{pmatrix} \quad (7)$$

$$c_t^l = f \cdot c_{t-1}^l + i \cdot g \quad (8)$$

$$h_t^l = o \cdot \rho(c_t^l) \quad (9)$$

The vectors $i, f, o, c \in \Re^n$ represent input, forget, output, and cell vectors respectively. $T_{2n,4n}$ represents a linear transformation from $\Re^{2n}$ to $\Re^{4n}$, D represents the dropout operator which sets a percentage of its parameter to zero, σ represents the element-wise sigmoid activation function and ρ represents the element-wise tanh activation.

The hidden states h_t^L of the top layer L are used to infer y_t^i given a sequence $x_1^i, ..., x_t^i$. The model is trained to maxmize the probability
$\prod_{i=1}^N \prod_{t=1}^J P(w_t^i | x_1^i ... x_{t-1}^i)$.

We train a 2-layer LSTM language model with 1,500 hidden units in each layer. Our vocabulary V contains the 50,000 most frequent words in the given corpus, and replaces all other words with $< unk >$. We use a dropout rate of 65% to the non-recurrent connections as described above.

3.2. r-ary Huffman coding

The r-ary Huffman coding method constructs trees in which each leaf node is a unique character from the alphabet. This results in a prefix-free coding in which the coded string for each character cannot be a prefix of the coded string of some other character. This is in contrast to Morse code in which, e.g., the letter '*e*' is encoded as '$\cdot$', which is the prefix of 17 other alphanumeric characters, including '*s*' ('$\cdot\cdot\cdot$'), which can lead to ambiguities.

Huffman coding depends on the prior probability of each character, c_i, in the alphabet, which is simply the frequency of that character in the training corpus (i.e, $P_H(c_i) = Count(c_i)/\sum_j Count(c_j)$ [19].

The key to the present work is that the corpora upon which these frequencies are based are first processed by the word prediction software. A corpus to be studied is divided into training, development, and test sets. The development and test sets are processed using language models trained on the training set, so that as soon as a word appears in the prediction list, all remaining characters are replaced with a single special character corresponding to their index in the prediction list as exemplified in in Table 2. This processed *development* set is used to compute r-ary Huffman codes, and the input savings are calculated between the original and processed *test* sets as follows:

$$IS(org, proc) = \frac{len(org) - len(proc)}{len(org)} \cdot 100 \quad (10)$$

where org and $proc$ represent the original and processed $test$ sets respectively, and len calculates the number of switch-strokes needed to type characters in the sets which would be equal to the number of characters if Huffman coding is not used. For example, if 'p' has code length 3, 'o' has code length 2, and '$\#$' has code length 1, then len('*pop*') is 8 where len('$p\#$') is 4.

4. Experiments

We partition each dataset into five chunks; each is iteratively used for development and testing, and the others are used for training. As described in the previous section, all reported input savings count the proportion of *code symbols* saved – not characters; *this is an important distinction* – counting the latter, as is typical in AAC research, would not take our application of the Huffman code into account.

Original sentence
the results met estimates of analysts who had already slashed their projections after the company said in late august that its 1989 earnings could
Processed sentence
t* re* met es@ **$** an**%** w@ **#** a* sla@ th**$** proj* af@ **@** **$** **%** i**%** l a@ t* i@ 1**%** **$** c**#**

Table 2: Example of pre-processing, with $N = 5$ from WSJ dataset. Each special character (*,@, $, %, #) in **bold** represents different indices in the prediction list.

WSJ	$r=3$	$r=4$	$r=5$	$r=6$	$r=\infty$
$N=3$	37.85	37.82	37.29	37.72	35.52
$N=4$	39.37	39.25	39.18	38.80	37.69
$N=5$	40.60	39.93	40.32	39.83	39.22
$N=6$	41.46	40.39	41.16	40.62	40.51
Essay	$r=3$	$r=4$	$r=5$	$r=6$	$r=\infty$
$N=3$	30.64	30.42	30.03	30.07	29.06
$N=4$	32.22	31.70	32.07	31.29	31.35
$N=5$	33.52	32.30	33.34	32.44	33.02
$N=6$	34.45	32.77	34.17	33.39	34.39
SMS	$r=3$	$r=4$	$r=5$	$r=6$	$r=\infty$
$N=3$	20.70	20.02	20.42	19.45	18.99
$N=4$	22.10	20.54	21.73	20.81	20.81
$N=5$	22.70	20.61	22.40	21.50	22.19
$N=6$	22.96	20.72	22.76	21.90	23.27

Table 3: Input savings (in %) on each dataset for different N and r values using the trigram model.

We run 5-fold cross validation for each number of coding symbols $r = \{3, 4, 5, 6, \infty\}$, where $r = \infty$ is the baseline character code length of 1, which mimics the case where Huffman coding is *not* used, and the length of the prediction list $N = \{3, 4, 5, 6\}$.

Table 3 shows input savings for each dataset using the trigram model for word prediction. As N increases, we get more input savings because the probability of the target word being in the prediction list goes up. However, a two-way F-test on N and r (Table 4) shows that the value of r does not affect the savings and that N and r do not interact.

Table 5 shows results obtained from the LSTM model for word prediction. The trigram model per-

	$SumSq$	$MeanSq$	F	$Pr(>F)$
N	0.0274	0.027399	5.554	0.0193
r	0.0012	0.001227	0.249	0.6184
$N:r$	0.0000	0.00010	0.002	0.9645

Table 4: Two-way F-test on N and r.

WSJ	$r=3$	$r=4$	$r=5$	$r=6$	$r=\infty$
$N=3$	25.57	25.24	25.81	25.78	25.47
$N=4$	27.11	26.96	27.64	27.81	27.70
$N=5$	28.39	28.15	28.71	29.20	29.38
$N=6$	29.35	29.22	29.66	30.06	30.70
Essay	$r=3$	$r=4$	$r=5$	$r=6$	$r=\infty$
$N=3$	18.56	18.28	18.81	17.82	19.15
$N=4$	20.02	19.74	20.43	19.78	21.23
$N=5$	21.21	20.76	21.52	21.10	22.87
$N=6$	22.08	21.87	22.17	22.08	24.16
SMS	$r=3$	$r=4$	$r=5$	$r=6$	$r=\infty$
$N=3$	14.43	13.51	14.53	14.01	14.26
$N=4$	15.61	14.20	15.66	15.24	16.06
$N=5$	16.25	14.94	16.16	15.96	17.42
$N=6$	16.83	15.55	16.52	16.36	18.52

Table 5: Input savings (in %) on each dataset for different N and r values using the LSTM model.

forms much better than the LSTM model; validating this finding on other sets of data should be the subject of future work.

The most input savings are obtained from the WSJ, and the least from the SMS dataset. This may be due to a more consistent grammatical structure in the former. The results clearly show that the input savings vary a lot depending on the corpus used. Trnka *et al.* [14], who showed that word prediction increased communication rates, also reported higher input savings when experimented on the Switchboard corpus, which has a different topic domain and vocabulary size. Analyzing the relation between the characteristics of a corpus (e.g., vocabulary size, level of formality, and grammatical structure) and the input saving rate is ongoing.

5. Conclusion and Future Work

In this paper, we examined input savings by combining word prediction models and r-ary Huffman coding on datasets with different levels of formality. Future work should evaluate performance 'online' with human participants, which may affect the optimal value of N, given a possible interaction effect with scanning time. Moreover, even though the value of r is not significant in terms of input savings, people might have different levels of difficulty in memorizing different code lengths[1]. Piloting the combination of word prediction and Huffman coding with real users is the next step, but the theoretical basis established in this paper is a requisite *first* step, since recruiting and training a sufficient number of participants will depend on constraining N and r, in order to obtain the appropriate statistical power. Alternatives to N-gram and LSTM models, initialized with pre-trained word embedding vectors, should also be applied to increasingly large datasets.

6. References

[1] R. Walls, J. J. Ratey, and R. I. Simon, *Rosen's Emergency Medicine: Expert Consult (Premium ed.)*, premium ed. St. Louis Missouri: Mosby, 2009.

[2] R. A. Spears and A. Holtz, *Spinal Cord Injury*. Oxford UK: Oxford University Press, 2010.

[3] S. L. Glennen and D. C. DeCoste, *The Handbook of Augmentative and Alternative Communication*. San Diego, CA: Singular, 1996.

[4] E. Dymond and R. Potter, "Controlling assistive technology with head movements a review," *Clinical Rehabilitation*, vol. 10, no. 2, pp. 93–103, 1996.

[5] S. J. Castellucci and I. S. Mackenzie, "Gestural text entry using Huffman codes," in *Proceedings of the International Conference on Multimedia and Human-Computer Interaction*, vol. 119, 2013, pp. 1–8.

[6] J. O. Wobbrock, B. A. Myers, and J. A. Kembel, "EdgeWrite: A Stylus-Based Text Entry Method Designed for High Accuracy and Stability of Motion," in *Proceedings of the 16th Annual ACM Conference on User Interface Software and Technology (UIST 03)*, 2003, pp. 61–70.

[1] We note that the fact that characters in the Morse code can be encoded in up to five symbols.

[7] T. Felzer and R. Nordmann, "Alternative text entry using different input methods," in *Proceedings of the 8th ACM Conference on Computers and Accessibility (ASSETS 06)*, 2006.

[8] P. Isokoski and R. Raisamo, "Device independent text input: A rationale and an example," in *Proceedings of the ACM Working Conference on Advanced Visual Interfaces (AVI 00)*, 2000, pp. 76–83.

[9] M. Porta and M. Turina, "Eye-S: a Full-Screen Input Modality for Pure Eye-Based Communication," in *Proceedings of the ACM 2008 Symposium on Eye Tracking Research and Applications (ETRA 08)*, 2008, pp. 27–34.

[10] B. Arons, "Techniques, perception, and applications of time-compressed speech," in *Proceedings of the 1992 Conference of the American Voice I/O Society*, 1992, pp. 169–177.

[11] A. Newell, S. Langer, and M. Hickey, "The role of natural language processing in alternative and augmentative communication," *Natural Language Engineering*, vol. 4, no. 1, pp. 1–16, 1998.

[12] I. S. Mackenzie, R. W. Soukoreff, and J. Helga, "1 Thumb, 4 Buttons, 20 Words Per Minute: Design and Evaluation of H4-Writer," in *Proceedings of the 24th Annual ACM Conference on User Interface Software and Technology (UIST 11)*, 2011, pp. 471–480.

[13] B. Roark, R. Beckley, C. Gibbons, and M. Fried-Oken, "Huffman scanning: using language models within fixed-grid keyboard emulation," *Comput Speech Lang*, vol. 27, no. 6, Sep 2013.

[14] K. Trnka, J. McCaw, D. Yarrington, K. F. McCoy, and C. Pennington, "Word Prediction and Communication Rate in AAC," in *Proceedings of the Fourth IASTED Conference on Telehealth and Assistive Technologies (Telehealth/AT 2008)*, 2008.

[15] I. S. Mackenzie, "KSPC as a characteristic of text entry techniques," in *Proceedings of the 4th ACM International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI 02)*, 2002, pp. 195–210.

[16] T. Chen and M.-Y. Kan, "Creating a live, public short message service corpus: the nus sms corpus," *Language Resources and Evaluation*, vol. 47, no. 2, pp. 299–335, 2012.

[17] I. H. Witten and T. C. Bell, "The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression," *IEEE Transactions on Information Theory, July*, vol. 37, no. 4, pp. 1085–1094, 1991.

[18] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[19] J. van Leeuwen, "On the construction of Huffman trees," in *Proceedings of ICALP*, 1976, pp. 382–410.