Individuality-Preserving Spectrum Modification for Articulation Disorders Using Phone Selective Synthesis

Reina Ueda, Ryo Aihara, Tetsuya Takiguchi, Yasuo Ariki

Graduate School of System Informatics, Kobe University
1-1, Rokkodai, Nada, Kobe, 6578501, Japan
{reina_1102, aihara}@me.cs.scitec.kobe-u.ac.jp, {takigu, ariki}@kobe-u.ac.jp

Abstract

This paper presents a speech synthesis method for people with articulation disorders resulting from athetoid cerebral palsy. For people with articulation disorders, there are duration, pitch and spectral problems that cause their speech to be less intelligible and make communication difficult. In order to deal with these problems, this paper describes a Hidden Markov Model (HMM)-based text-to-speech synthesis approach that preserves the voice individuality of those with articulation disorders and aids them in their communication. For the unstable pitch problem, we use the F0 patterns of a physically unimpaired person, with the average F0 being converted to the target F0 in advance. Because the spectrum of people with articulation disorders is often unstable and unclear, we modify generated spectral parameters from the HMM synthesis system by using a physically unimpaired person's spectral model while preserving the individuality of the person with an articulation disorder. Through experimental evaluations, we have confirmed that the proposed method successfully synthesizes intelligible speech while maintaining the target speaker's individuality.

Index Terms: Articulation disorders, Speech synthesis system, Hidden Markov Model, Assistive Technologies

1. Introduction

In this study, we focus on a person with an articulation disorder resulting from the athetoid type of cerebral palsy. 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. It 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 cerebral palsy sufferers [1]. In the case of persons with articulation disorders resulting from the athetoid type of cerebral palsy, his/her movements are sometimes more unstable than usual. That means their utterances (especially their consonants) are often unstable or unclear due to their athetoid symptoms, and there is a great need for voice systems that can assist them in their communication.

An HMM-based speech synthesis system [3] is a text-to-speech (TTS) system that can generate signals from input text data. A TTS system may be useful for those with articulation disorders because they have difficulty moving their lips. In an HMM-based speech synthesis system, the spectrum, F0 and duration are modeled simultaneously in a unified framework. Mel-cepstral coefficients are used as spectral features, which are modeled by continuous density HMMs. F0 patterns are modeled by a hidden Markov model based on multi-space probability distribution (MSD-HMM [4]), and state duration densities are modeled by single Gaussian distributions [5].

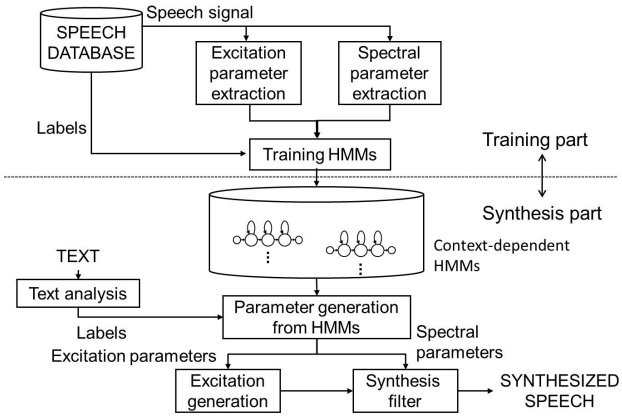


Figure 1: HMM-based sound synthesis system

In the field of assistive technology, Veaux *et al.* [6] used HMM-based speech synthesis to reconstruct the voice of individuals with degenerative speech disorders resulting from Amyotrophic Lateral Sclerosis (ALS). They have proposed a reconstruction method for degenerative speech disorders using an HMM sound synthesis system. In this method, the subject's utterances are used to adapt an average voice model pre-trained on many speakers. Creer *et al.* [7] also adapt the average voice model of multiple speakers to the severe dysarthria data. And Khan *et al.* [8] uses such adaption method to the laryngectomy patient's data. Yamagishi *et al.* [9] proposed a project called "Voice Banking and Reconstruction". In that project, various types of voices were collected, and they proposed TTS for ALS using that database. Also, Rudzicz [10] proposed a speech adjustment method for people with articulation disorders based on observations from the database.

In this paper, we propose an HMM-based speech synthesis method for articulation disorders because there are several problems in the recorded voice of persons with articulation disorders, and this causes the output synthesized signals to be unintelligible. To deal with these problems, it is necessary to develop a speech synthesis system in which the output signals become more intelligible and include the subject's individuality.

To generate an intelligible voice while preserving the speaker's individuality, we train the speech synthesis system using training data from both a person with an articulation disorder and a physically unimpaired person. Because the utterance rate of persons with articulation disorders differs from that of a

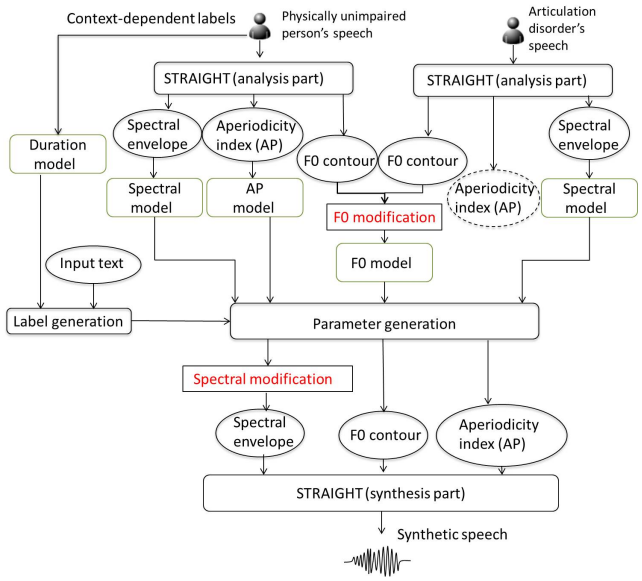


Figure 2: Diagram of HMM-based sound synthesis method for articulation disorders

physically unimpaired person, we utilize the duration model of a physically unimpaired person only in our method. In addition to the utterance rate problem, the F0 patterns of persons with articulation disorders are often unstable compared to those of physically unimpaired persons. In our method, the F0 model is trained from a physically unimpaired person's F0 patterns, and the average F0 is used as the F0 pattern for the person with an articulation disorder.

As for the spectral problem associated with persons with articulation disorders, the consonant parts of their speech are often unstable or unclear, which causes their voice to be unintelligible. To resolve this consonant problem, we conduct different operations on the consonant and vowel parts. For the consonants parts, we basically generate the output spectrum from the spectral model of a physically unimpaired person. For the vowel parts, we generate the output spectrum from the spectral model of a person with an articulation disorder in order to preserve the person's individuality.

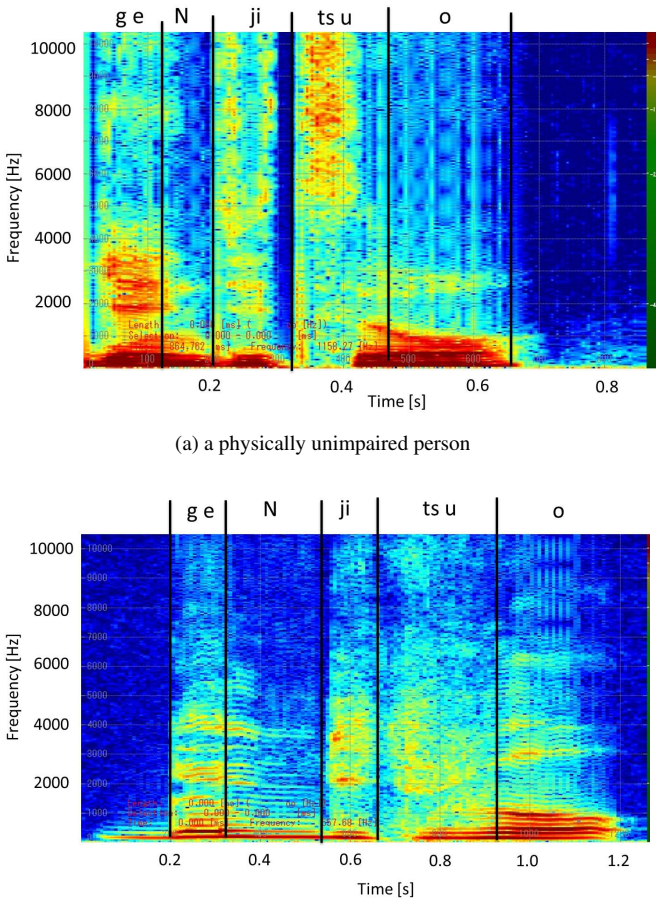
2. HMM-based sound synthesis

2.1. Basic approach

Fig. 1 shows the overview of the basic approach to text-to-speech synthesis (TTS) based on HMMs. This figure shows the training and synthesis parts of the HMM-based TTS system. In the training part, parameters (spectral, F0, and aperiodicity) are extracted as feature vectors. These features are modeled by context-dependent HMMs. Also, by installing the duration model, it is able to model each parameter, as well as the duration in the unified framework.

In the synthesis part, a context-dependent label sequence is obtained from an input text by text analysis. A sentence HMM is constructed by concatenating context-dependent HMMs according to the context-dependent label sequence. Then, HMM state sequences $q = [q_1, \cdots, q_T]$ are decided from the duration model as follows:

$$\hat{q} = \arg\max_{q} P(q|\lambda) \tag{1}$$



(a) a physically unimpaired person

(b) a person with an articulation disorder

Figure 3: Examples of spectrogram uttered for // g e N j i ts u o

where T, q_t, and λ represent the number of frames, index of the HMM-state of the t-th input frame, and the parameter sets of HMM, respectively. The explicit constraint between static and dynamic features, and signal parameter sets are generated with maximizing HMM likelihood [11].

$$c = \arg\max_{c} P(\mathbf{W}c|\hat{q}, \lambda) \tag{2}$$

In Eq. (2), $c = [{c_1}^{\mathsf{T}}, \cdots, c_t^{\mathsf{T}}, \cdots, {c_T^{\mathsf{T}}}]^{\mathsf{T}}$ represents signal parameter sequences, $c_t = [c(1), \cdots, c(D)]^{\mathsf{T}}$ represents a signal parameter vector of the t-th frame, and $\mathbf{W}$ represents the matrix constructed from weights which are used for calculating dynamic features [12].

Finally, by using an MLSA (Mel-Log Spectrum Approximation) filter [13], speech is synthesized from the generated parameters.

2.2. HMM-based sound synthesis for articulation disorders

If each feature parameter is trained using the acoustic features obtained from a person with an articulation disorder, the synthesized sound becomes unintelligible. Therefore, we created a more intelligible synthesized sound while preserving the speaker's individuality by mixing the voices of a person with an articulation disorder and a physically unimpaired person.

Fig. 2 shows the overview of our method. In this method, we train the speech synthesis system using training data from both a person with an articulation disorder and a physically unimpaired person. First, we extract three acoustic parameters (F0 contour, spectral envelope, and aperiodicity index (AP)) from these two person's speaking voices by using STRAIGHT analysis [14]. After extracting the features, the F0 patterns of a physically unimpaired person are modified as explained in Section 2.3.

Because the duration of persons with articulation disorders is slower than that of physically unimpaired people, the duration model is generated using only the context-dependent label sequences of a physically unimpaired person. With the input text and the duration model, context-dependent label sequences are generated. Then, spectral, F0 and AP parameters are generated based on the label sequences and trained HMMs, where F0 parameters are generated from the modified F0 model and AP parameter sequences are generated from the AP model of a person with an articulation disorder.

Each spectral parameter is generated from each person's spectral model. After parameter generation, the spectral parameters of a person with an articulation disorder are modified as explained in Section 2.4. Finally, the output signal is synthesized from the features (spectral envelope, F0 contour, and aperiodicity index) by using the synthesis part of the STRAIGHT. In the following section, we explain the details of the operations related to spectral and F0 parameters.

2.3. F0 modification

In this method, the F0 patterns of a physically unimpaired person are used for training the F0 model in HMM synthesis because the F0 patterns of a person with an articulation disorder are often unstable. To make the F0 feature's characteristics close to those of a person with an articulation disorder, the F0 features of a physically unimpaired person are modified to those of a person with an articulation disorder. The F0 model is trained from the converted F0 sequences, which means that the F0 model includes the individuality of a person with articulation disorder.

The F0 features of a physically unimpaired person are modified by using the following linear transformation:

$$\hat{x}_t = \frac{\sigma_y}{\sigma_x}(x_t - \mu_x) + \mu_y \tag{3}$$

where x_t represents the log-scaled F0 of the physically unimpaired person at the frame t, μ_x and σ_x represent the mean and standard deviation of x_t, respectively. μ_y and σ_y represents the mean and standard deviation of the log-scaled F0 of a person with an articulation disorder, respectively.

2.4. Spectral modification

Fig. 3 shows the original spectrograms for the word "genjitsuo" ("real" in English) of a physically unimpaired person and a person with an articulation disorder. As shown in Fig. 3, the high-frequency spectral power of a person with an articulation disorder is weaker compared to that of a physically unimpaired person. This fact implies that the synthesized spectrum of the consonant components for a person with an articulation disorder becomes weak, which makes the person's speech difficult to understand.

For the spectral vowel components, the spectral parameters of a person with an articulation disorder are needed in order to preserve the target individuality. As shown in Fig. 2, after being given the input text, we generate spectral parameter sequences from each person's spectral model. Then, we create the combined spectral parameter sequences, which include the parameters of a physically unimpaired person at the high-frequency part and the parameters of a person with an articulation disorder at the low-frequency part. This combination of spectral parameters is given by

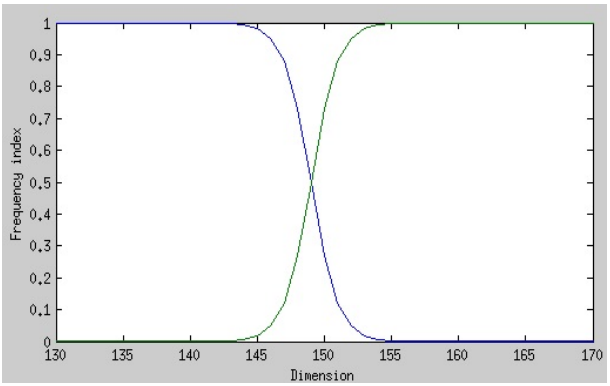


Figure 4: Plot of the function f_m and f_g (green: f_m blue: f_g)

$$\hat{S}^{(ij)} = f_m^{(j)} S_m^{(ij)} + f_g^{(j)} S_g^{(ij)} \tag{4}$$

where S_m, S_g, $\hat{S}$, i and j represent the spectrum of a physically unimpaired person, the spectrum of a person with an articulation disorder, the modified spectrum, the index of spectral frames, and the frequency index, respectively. The weight functions are given by

$$f_m^{(j)} = \frac{1}{1 + e^{(-j+S)}} \tag{5}$$

$$f_g^{(j)} = \frac{1}{1 + e^{(j-S)}} \tag{6}$$

where f_m represents the weight function for a physically unimpaired person's spectrum, f_g represents that of a person with an articulation disorder, and S represents the control parameter, respectively.

Fig. 4 shows an example of the functions f_m and f_g. The function, f_m, emphasizes the high-frequency components and weakens the low-frequency components of spectral parameters. The function, f_g, emphasizes the low frequency components and weakens the high-frequency components of spectral parameters.

By using Eq. (4), at the high-frequency part, the spectrum is complemented by that of a physically unimpaired person in order to make the consonants clear. At the low-frequency part, we need to preserve the spectrum of a person with an articulation disorder in order to preserve the individuality. The spectral modification is calculated at each frame using Eq. (4), and the frequency thresholds are determined for the vowel part and consonant part. In our study, the total number of spectral dimensions (indexes) is 513, S is set to 150 for the vowel part, and S is set to 80 for the consonant part.

3. Experiments

3.1. Experimental conditions

We prepared the training data for two men. One is a physically unimpaired person, and the other is a person with an articula-

Table 1: Voices compared in the evaluation tests

Type	Duration Model	F0 Model	AP Model	Spectral Model
ADM	AD	AD	AD	AD
Ref1	PU	AD	AD	AD
Prop	PU	convPU	AD	MIX
Ref2	PU	convPU	AD	AD
PUM	PU	PU	PU	PU

Note
ADM: Articulation disorder person's model
Prop: Proposed method
PUM: Physically unimpaired person's model
AD: Articulation Disordered
PU: Physically Unimpaired
convPU: Creating the model from a physically unimpaired person's pa
which are converted to those of the person with an articulati
MIX: mixing ADM and PUM spectra using Eq. (4)

tion disorder. We used 513 sentences from the ATR Japanese database for a physically unimpaired person, and recorded 429 sentences in the same database uttered by a person with an articulation disorder. The speech signals were sampled at 48 kHz and the frame shift was 5 ms. Acoustic and prosodic features were extracted by using STRAIGHT. As spectral parameters, mel-cepstrum coefficients, their dynamic, acceleration coefficients were used. As excitation parameters, log-F0 and 5 band-filtered aperiodicity measures [15] were used and their dynamic and acceleration coefficients were also used. Context-dependent phoneme HMMs with five states were used in the speech synthesis system [3].

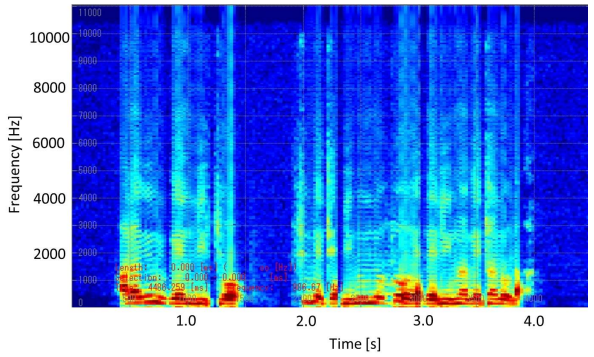
In order to confirm the effectiveness of our method, we evaluated both the aspect of listening intelligibility and the aspect of speaker similarity by listening to voices synthesized under the five conditions shown in Table 1. Ten sentences included in the ATR Japanese database were synthesized under those five conditions. A total of 8 Japanese speakers took part in the listening test using headphones. For speaker similarity, we performed a MOS (Mean Opinion Score) test [16]. In the MOS test, the opinion score was set to a 5-point scale (5: Identical, 4: Very Similar, 3: Quite Similar, 2: Dissimilar, 1: Very Dissimilar). For the listening intelligibility, a paired comparison test was carried out, where each subject listened to pairs of speech converted by the two methods, and then selected which sample was more intelligible.

3.2. Results and discussion

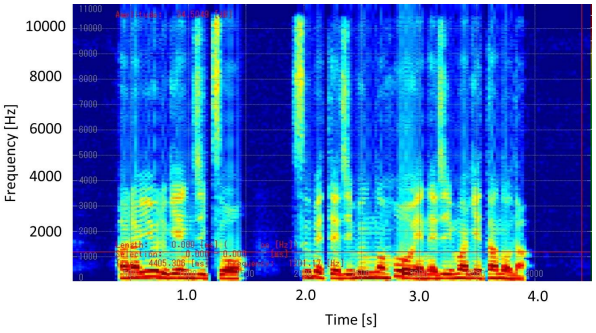
Table 2: Average duration per mora in 50 sentences

	Average time [ms/mora]
ADM	219.768
PUM	179.69

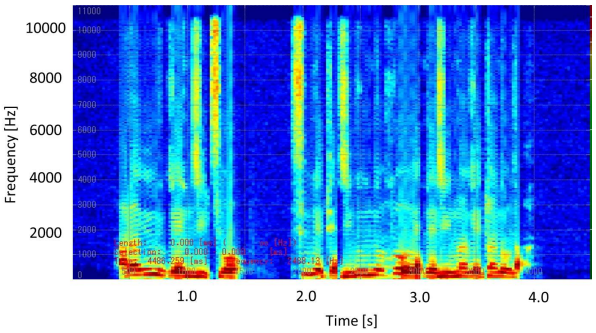
We calculated the average synthesized signal's duration per mora in 50 sentences. As shown in Table 2, the average duration of ADM (Articulation disorder person's model) is 219.768 [ms/mora] and that of PUM (Physically unimpaired person's model) is 179.69 [ms/mora]. As compared to the duration of



(a) ADM spectrogram



(b) PUM spectrogram



(c) Modified spectrogram

Figure 5: Examples of synthesized spectrograms

PUM, that of ADM is quite slower, which causes the unintelligibility of the synthesized sound.

In the proposed method, we generated the modified spectral parameters by mixing both ADM and PUM spectral parameters. Fig. 5a shows the generated spectrum from the ADM spectral model and Fig. 5b shows the generated spectrum from the PUM spectral model. Both spectral parameters are generated from the same text and the same PUM duration model so that they have the same number of frames and dimensions. As shown in Fig. 5a, the high-frequency component is weaker compared to Fig. 5b, which means that the consonant parts of ADM spectral parameters are weak. This causes the output synthesized signals to be less intelligible. Fig. 5c shows the modified spectrum created from both ADM and PUM spectral parameters by using Eq. (4). As shown in Fig. 5c, the consonant parts are complemented by the high-frequency parameters of PUM while preserving ADM's low-frequency components.

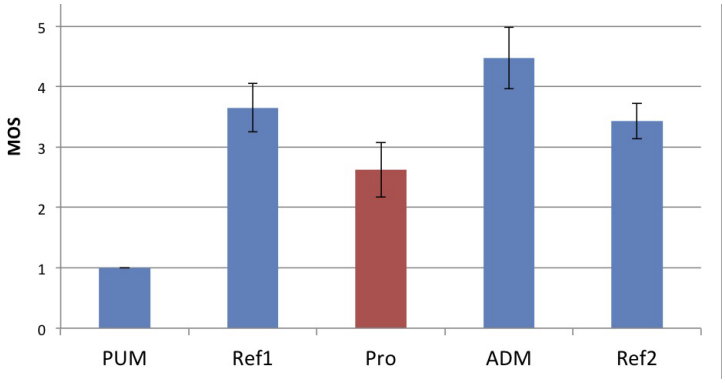


Figure 6: Speaker similarity to the articulation disorder person's speech

Fig. 6 shows the results of the MOS test on speaker similarity, where the error bar shows a 95% confidence score. As shown in Fig. 6, the ADM score was the highest score of all. This is because the signal from ADM is synthesized only from the feature parameters of a person with an articulation disorder. The Prop score is slightly less than those of Ref1 and Ref2 because of the modification of the spectral parameters.

Fig. 7 shows the preference score for the listening intelligibility, where the error bar shows a 95% confidence score. As shown in Fig. 7, our method obtained a higher score than Ref1 and ADM. These results show that the proposed method is effective. By replacing the physically unimpaired person's duration model and converting his F0 patterns to those of the person with an articulation disorder improves intelligibility. Our method also obtained a higher score than Ref2. This result shows that modifying the output spectral parameters is quite effective in improving intelligibility. Therefore, considering from Figs. 6 and 7, it is confirmed that our proposed method implements the synthesized signals which is intelligible and includes individuality of a person with an articulation disorder.

4. Conclusion

We have proposed a text-to-speech synthesis method based on HMMs for a person with an articulation disorder. In our method, to generate synthesized sounds that are more intelligible, the duration model of a physically unimpaired person is used, and the F0 model is trained using the F0 features of a physically unimpaired person, where the average F0 is converted to the articulation disorder person's F0 using a linear transformation. In order to complement the consonant parts of the spectrum of a person with an articulation disorder, we replaced the high-frequency parts with those of a physically unimpaired person. The experimental results showed that our method is highly effective in improving the listening intelligibility of speech spoken by a person with an articulation disorder. In future research, we will complement the consonant parts of the spectral parameters at the training part.

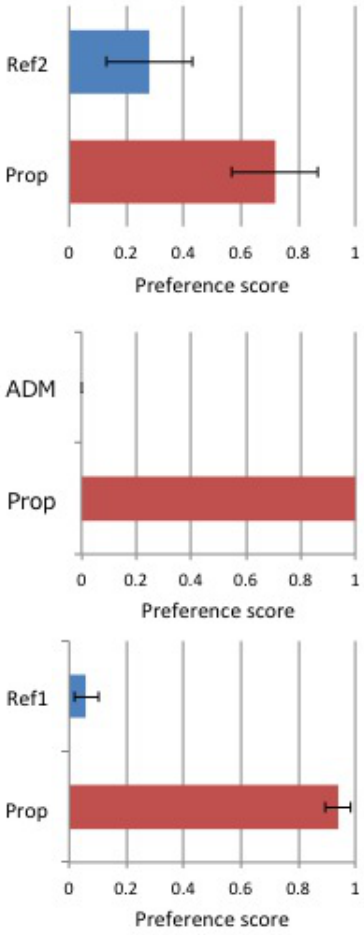


Figure 7: Preference scores for listening intelligibility

5. 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, January 2013.

[2] T. Canale and W. C. Campbell, *Campbell's operative orthopaedics*. Technical report, Mosby Year Book, June 2002, vol. 12.

[3] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. of Eurospeech*, 1999, pp. 2347–2350.

[4] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling," in *Proc. of ICASSP*, 1999, pp. 229–232.

[5] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Duration Modeling in HMM-based Speech Synthesis System," in *Proc. of ICSLP*, 1998, pp. 29–32.

[6] C. Veaux, J. Yamagishi, and S. King, "Using HMM-based speech synthesis to reconstruct the voice of individuals with degenerative speech disorders," in *Proc. of Interspeech*, 2012.

[7] S. Creer, S. Cunningham, P. Green, and J. Yamagishi, "Building personalised synthetic voices for individuals with severe speech impairment," *Computer Speech & Language*, vol. 27, no. 6, pp. 1178–1193, 2013.

[8] Z. Ahmad Khan, P. Green, S. Creer, and S. Cunningham, "Reconstructing the voice of an individual following laryngectomy," *Augmentative and Alternative Communication*, vol. 27, no. 1, pp. 61–66, 2011.

[9] 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.

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

[11] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. of ICASSP*, 2000, pp. 1315–1318.

[12] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, pp. 1039–1064, 2009.

[13] S. Imai, K. Sumita, and C. Furuichi, "Mel log spectrum approximation (MLSA) filter for speech synthesis," *Electronics and Communications in Japan (Part I: Communications)*, vol. 66, pp. 10–18, 1983.

[14] H. Kawahara, I. Masuda-Katsuse, and A. D. Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech communication*, vol. 27, pp. 187–207, 1999.

[15] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT." in *Proc. of MAVEBA*, 2001, pp. 59–64.

[16] I. T. Union, "ITU-T Recommendation P.800.1: Mean Opinion Score (MOS) terminology," International Telecommunication Union, Tech. Rep., July 2006.