Pronunciation Adaptation For Disordered Speech Recognition Using State-Specific Vectors of Phone-Cluster Adaptive Training

Sriranjani. R ⋆, S. Umesh† and M. Ramasubba Reddy ⋆

⋆Biomedical Engineering Group, Department of Applied Mechanics
†Department of Electrical Engineering
Indian Institute of Technology - Madras
am12s036@smail.iitm.ac.in, umeshs@ee.iitm.ac.in, rsreddy@iitm.ac.in

Abstract

Pronunciation variation is a major problem in disordered speech recognition. This paper focus on handling the pronunciation variations in dysarthric speech by forming speaker-specific lexicons. A novel approach is proposed for identifying mispronunciations made by each dysarthric speaker, using state-specific vector (SSV) of phone-cluster adaptive training (Phone-CAT) acoustic model. SSV is low-dimensional vector estimated for each tied-state where each element in a vector denotes the weight of a particular monophone. The SSV indicates the pronounced phone using its dominant weight. This property of SSV is exploited in adapting the pronunciation of a particular dysarthric speaker using speaker-specific lexicons. Experimental validation on Nemours database showed an average relative improvement of 9% across all the speakers compared to the system built with canonical lexicon.

Index Terms: Dysarthric speech recognition, phone-CAT, lexical modeling, pronunciations, phone confusion matrix

1. Introduction

Clinical applications of speech technology play an important role in aiding communication for people with motor speech disorders. One such motor speech disorder is dysarthria, acquired secondary to stroke, traumatic brain injury, cerebral palsy etc. This affects more than one subsystem of speech production, leading to unintelligible speech. Some of the common characteristics of dysarthria include slurred speech, swallowing difficulty, slow speaking rate with increased effort to speak and muscle fatigue while speaking [1, 2]. All these effects affect the speech intelligibility but also the social interaction ability of people with speech disorders. Clinical applications of speech technology provide way to improve their communication in terms of the alternative and augmentative communication (AAC) devices. Automatic speech recognition (ASR) systems play a major role as an AAC device for aiding communication in terms of command/control in their daily lives. Only handful of databases are available for dysarthric speech, due to the fatigue and discomfort faced by the dysarthric speaker in providing data for longer time. With such constraints, acoustic models are usually built-in speaker adaptation framework [3, 4, 5].

The impairment in phonatory subsystem of a person affected with dysarthria leads to pronunciation errors. The slow rate of speech leads to a single syllable word being misrecognized as two syllable words. Frequent occurrences of non-speech sounds like hesitations false starts occur as part of dysarthric speech. These hesitations also lead to misrecognition of words as explained in [6, 4]. Imprecise consonant production is another characteristic of dysarthric speech. Since consonant production involves complex articulations compared to vowels, the errors are more frequent [7]. Muscle fatigue and lack of breath support increase the pronunciation errors of a dysarthric speaker [8].

All these effects increase the rate of insertions, substitutions, deletions and distortions in the dysarthric ASR systems. Thus the issue of pronunciation errors makes the design of dysarthric ASR system more challenging. The focus of this paper lies in handling these pronunciation errors especially substitutions by improving the lexical models. The lexicon contains the multiple pronunciations for each word expanded in terms of phones. The alternate pronunciations of a word is either formed manually [9] or obtained from the list of phone confusion pairs [10, 11]. This paper introduces a recently developed phone-cluster adaptive training (Phone-CAT) [12] acoustic modeling technique. Phone-CAT method build robust acoustic models using lesser number of parameters and limited amount of data. Thus the method can be used for limited data available domains especially in the case of dysarthric speech recognition. The main contributions of this paper are as follows:

- A novel approach to form speaker-specific phone confusion matrix using the low-dimensional SSV of Phone-CAT
- Using the speaker-specific phone confusion matrix to identify the confusion pairs (substitution phones) to form alternate pronunciations in the speaker-specific lexicon

Our proposed approach helps in forming phone confusion matrix directly from the Phone-CAT acoustic model, compared to the existing methods [10, 11] which align the decoded transcription with canonical transcription to form the phone confusion matrix. Thus we circumvent the usage of expensive decoding process. This preliminary study using Nemours database shows a relative performance improvement of 9% using our proposed approach compared to baseline model built using canonical lexicon.

2. Related work

Multiple pronunciations of a word in the lexicon improves the recognition performance. The lexical models are improved either implicitly or explicitly handling the pronunciation errors [13]. In order to improve the lexical models, the phones mispronounced by each dysarthric speaker need to be identified. Earlier work handled multiple pronunciations using expert knowledge by adapting pronunciations manually [9]. Per-

sonalized speaker articulation patterns were obtained from the speaker-adapted models along with the confusion matrix. These speaker-adapted models were obtained using universal disordered matrix and the posterior probability from the ASR system in an unsupervised fashion [13].

Another approach for identifying the mispronounced phones is by aligning the decoded text with the true transcription. A phone confusion matrix is formed using the decoded transcription and canonical transcription. This phone confusion matrix is used to identify the mispronunciations [10]. The substitution, insertion and deletion errors, were modeled as discrete hidden Markov model (HMM) called metamodels [11]. Another variant of this system is to train the extended metamodels from an integrated confusion matrix using genetic algorithm [14].

The concept of weighted finite state transducer (WFST) improves the performance of speech recognition systems. Composing confusion matrix along with the lexicon and language models in the WFST framework provides complementary information to the system. This concept was used in speech recognition [15] and keyword searching [16]. In dysarthric speech recognition framework, different methods were used to form confusion matrices to be used with WFST. One such method is to use KL distance measure between two context-dependent triphones to form confusion matrix [17, 18]. Deep neural networks (DNN) can also be used to improve pronunciation models. The posterior probabilities from pre-trained DNN were used to identify mispronunciations. They were further analyzed to generate pronunciations to form speaker-specific lexicons [19]. All the above methods, uses confusion matrix obtained by aligning the decoded transcriptions with the canonical transcriptions to improve the lexicons.

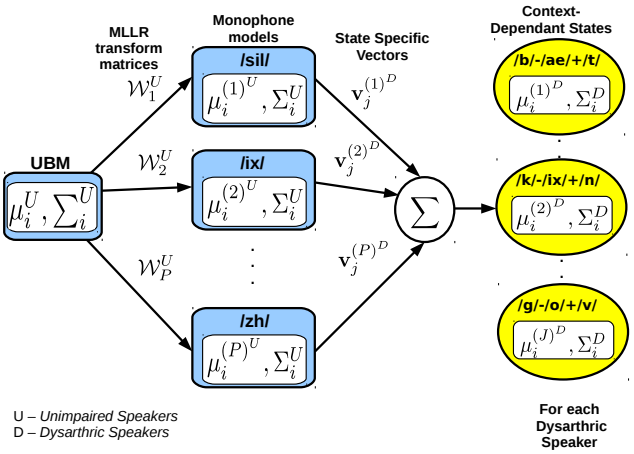

Figure 1: Phone-CAT architecture

In this paper, a novel approach is proposed to form the confusion matrix using the low-dimensional vector from the Phone-CAT acoustic model. Each tied-state in Phone-CAT model is modeled using SSV. The dominant weight of the SSV represents the pronounced phone. The mispronounced phone of each dysarthric speaker obtained using SSV is compared with the canonical phone to form the phone confusion matrix. This matrix is used to improve the lexical models by providing alternate pronunciations of words. Since each speaker has a separate pronunciation pattern, speaker-specific lexicons are formed. Acoustic models rebuilt using these lexicon, improve the performance of the system.

3. Phone-Cluster Adaptive Training Acoustic Models

The acoustic models are usually built using hidden Markov model–Gaussian mixture model (HMM–GMM) framework. The acoustic variations of speech due to age, gender, environmental changes and pronunciation variations are being modeled using GMM. The sequence information involving co-articulation is modeled as HMM. The triphone model represents a phone along with its left and right contexts capturing the co-articulation effects. For example, consider the triphone $/ax/-/b/+/k/$ representing the model for the center phone $/b/$, capturing the effect of its left context $/ax/$ and right context $/k/$. Several triphones with similar acoustic characteristics and same center phone $/b/$ are clustered to form a single tied-state. The GMM parameters are then estimated independently to model each tied-state. This estimation requires huge number of parameters and sufficient amount of data. This issue is handled using the recently proposed phone-CAT acoustic model by robustly modeling the available data with lesser number of parameters.

Phone-CAT is a HMM-GMM system in which the GMM parameters are represented in a compact form. In other words, the GMM for each tied-state is formed by the linear combination of all the monophone GMMs in that language. For example, the tied-state $/ax/-/b/+/k/$ containing triphones $/ax/-/b/+/k/$,$/ch/-/b/+/k/$,$/ae/-/b/+/k/$ is formed from the linear combination of all the monophone GMMs like $/sil/, /ax/, .../k/, ..., /zh/$. The weights of each monophone GMMs are represented by $v_j^{(1)}, v_j^{(2)} \ldots v_j^{(P)}$, where P is the number of monophones. The vector containing the monophone weights is called SSV and is represented as $\mathbf{v}_j = \begin{bmatrix} v_j{}^{(1)} & v_j{}^{(2)} & . & . & v_j{}^{(P)} \end{bmatrix}^T$ with P dimensions. The monophone GMMs are in turn formed by adapting the universal background model (UBM) using maximum likelihood linear regression [20] transformation. The UBM is a GMM built using the available speech data from all the speakers. This UBM is adapted using the transformation matrices $\boldsymbol{\mathcal{W}}_1, \boldsymbol{\mathcal{W}}_2, ..., \boldsymbol{\mathcal{W}}_P$ for each of the P monophones, forming P monophone GMMs. The Phone-CAT architecture is shown in figure 1. The GMM parameters of the tied-state model are: means $\boldsymbol{\mu}_{ji}$, covariances $\boldsymbol{\Sigma}_i$ and Gaussian priors w_{ji}.

The mean parameter for each monophone models $\mu_i^{(p)}$ with Gaussian mixture i is combined to form the mean parameter of the tied-state j using the following equations:

$$\mu_i^{(p)} = \boldsymbol{\mathcal{W}}_P \boldsymbol{\xi}_i = \boldsymbol{\mathcal{W}}_P [\mu_i \;\; 1]^T$$

$$\boldsymbol{\mu}_{ji} = \sum_{p=1}^{P} \mu_i^{(p)} \; v_j^{(p)}$$

Here $\boldsymbol{\xi}_i$ is the extended mean vector $[\mu_i \;\; 1]^T$ with μ_i as the canonical mean of the Gaussian component i of the UBM. Since the mean $\boldsymbol{\mu}_{ji}$ and the Gaussian prior w_{ji} are represented in terms of the vector SSV $\mathbf{v}_j$ as in [12], the parameters are represented in low dimensions. Also the covariances $\boldsymbol{\Sigma}_i$ are estimated in a shared fashion across the tied-states. This reduction in the number of parameters helps in reducing the amount of data needed for estimation. More details of the model training and estimation of each parameters are explained in [12].

4. Importance of state-specific vectors

The SSV is a low-dimensional vector of dimension P representing each tied-state j uniquely. It captures the context information since it represent the weights with which each monophone GMM linearly combine to form a single tied-state. We know that different triphones with the same acoustical characteristics are tied together in order to form tied-state. The SSV plot of the second state of the triphone $/ch/-/ix/+/ng/$ is shown in figure 2.

It is clearly shown that, the dominant weight corresponds to the center phone $/ix/$. Apart from the center phone, the left and right context phones also get some considerable weight. The negative value represents the direction of the vector, but we are interested only in the absolute value of the elements of the SSV.

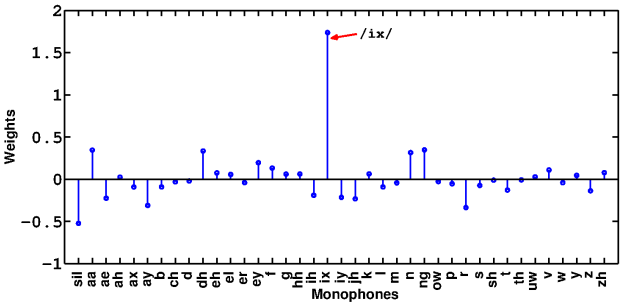

Figure 2: SSV plot of the second state of the triphone $/ch/-/ix/+/ng/$

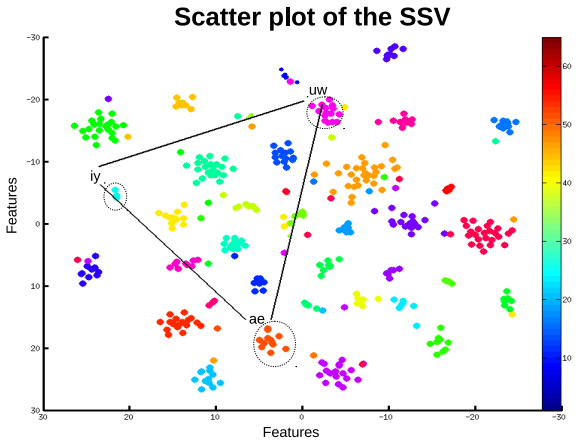

Figure 3: This two-dimensional scatter plot is obtained using the t-SNE toolkit by plotting the SSV of all the tied-states in Nemours database

A statistics of the dominant weight property of SSV was performed for unimpaired (control) speech data from Nemours speech database. The aim of the task was to check the statistics of the SSV picking the center phone of the tied-state correctly. It was found that out of 204 tied-states, the dominant component of SSV correctly picks the center phone 76% of the times and the top three weight values in the SSV picks up center phone 88% of the time. A similar analysis was also performed for the standard Switchboard database (≈300 hours of data), with 2400 tied-states. It was found that 70% of the time, the center phone was correctly picked up by the dominant component of SSV. Also ≈92% of the time, the left/center/right phones are picked up as the dominant component of SSV [21]. This shows that the SSV uniquely represents the enunciated phone (center phone of the tied-state) through its dominant weight most likely. The scatter plot of the P dimensional SSV reduced to two dimension is shown in the figure 3. The SSV related to each cluster represents a particular monophone (each in different color, a total of 39 phones were present in the Nemours database). These clusters are located at articulatory position of the vowel triangle in a well discriminated manner. This shows that SSV has the capacity to capture the phonetic information along with context information. Thus the analysis of SSV in this section leads us to the following conclusions:

- The dominant weight in SSV most likely represents the enunciated phone (center phone) of the tied-state
- Provides discriminable phonetic class information, since each vector is modeled for a particular tied-state
- SSV is hypothesized to capture the pronunciations of each dysarthric speaker when speaker-specific Phone-CAT models are built

This leads us to proceed to the proposed method of building Phone-CAT model specific to each speaker, thereby capturing the pronunciations of each dysarthric speaker.

Table 1: Extract dysarthric enunciated phone from SSV

Tied-states	**Phones**						
(Canl)	***sil***	***aa***	...	***ey***	...	***zh***	**Dysp**
$*-/\mathbf{sil}/+*$	**1.75**	0.21	...	0.38	...	0.12	$/sil/$
$*-/\mathbf{aa}/+*$	0.03	**0.09**	...	0.01	...	0.08	$/aa/$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
$*-/\mathbf{jh}/+*$	0.19	0.11	...	**0.36**	...	0.13	$/ey/$
$ch-/\mathbf{ix}/+ng$	0.48	**0.50**	...	0.01	...	0.22	$/aa/$
$n-/\mathbf{ix}/+k$	0.10	**0.90**	...	0.25	...	0.76	$/aa/$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
$*-/\mathbf{zh}/+*$	2.01	0.02	...	0.10	...	**3.06**	$/zh/$
Canl - canonical pronunciation; Dysp - dysarthric pronunciations							

The numbers inside circle shows the absolute maximum value in each SSV corresponding to dysarthric pronounced phone

5. Proposed Method for Improving Lexical Models

5.1. Phone-CAT model for each dysarthric speaker

The major step of our proposed method is to build speaker-specific Phone-CAT model. Initially, using the unimpaired speaker's data in the dysarthric database, a Phone-CAT model is built. The speaker-specific Phone-CAT model is obtained from the unimpaired speaker model by re-estimating the SSV and providing dysarthric speaker's data in maximum likelihood framework. The SSVs are initialized as (1/number of monophones), to allow the system to learn the weights of the monophone GMM using the available dysarthric speaker's data. At the end of this training process, Phone-CAT speaker-specific models are built. The architecture of speaker-specific Phone-CAT model is shown in figure 1. Finally, we obtain a set of tied-states specific to each dysarthric speaker from the speaker-specific Phone-CAT model.

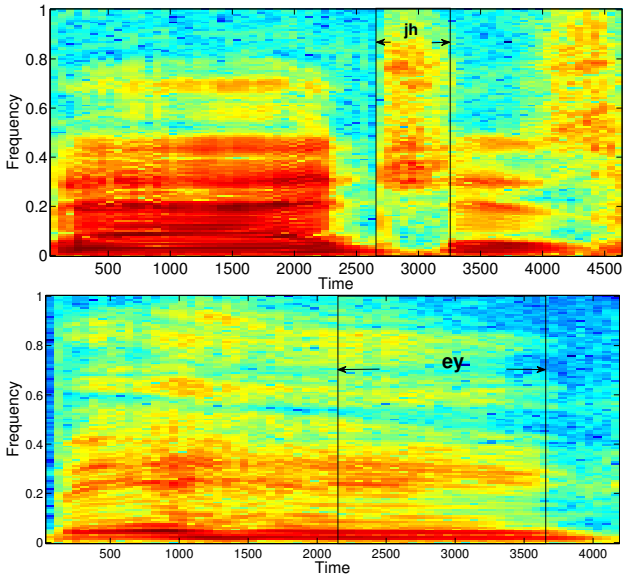

Figure 4: Spectrogram of the word "Badge" spoken by unimpaired speaker (on top) and dysarthric (BV) speaker (bottom). The spectrograms are plotted for a part of the waveform containing "The Badge is lifting the Beige".

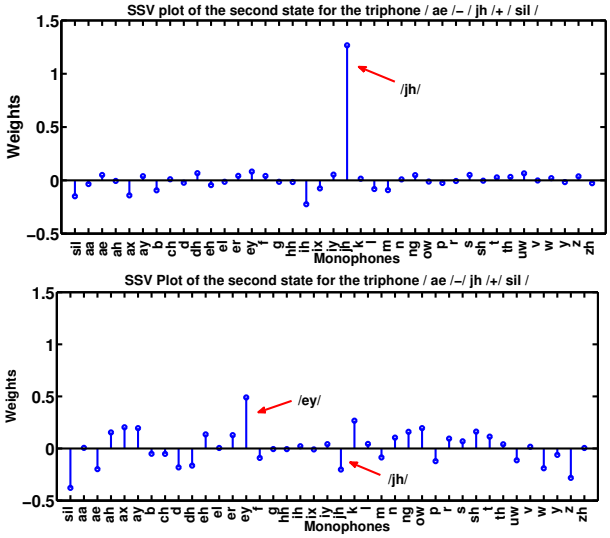

Figure 5: SSV for the second state of the triphone $/ae/-/jh/+/sil/$ for the unimpaired speaker (top) and dysarthric speaker BV (bottom).

5.2. Identification of mispronunciations made by the dysarthric speaker using SSV

Having built the speaker-specific Phone-CAT models, the next step is to extract the unique SSV associated with the set of tied-states. The P-dimensional SSV is extracted from each dysarthric speaker's Phone-CAT model. Using the dominant weight property of SSV discussed in section 4, the absolute maximum weight value of the SSV is obtained for each tied-state from each dysarthric speaker's Phone-CAT model. Since the pronounced phone is captured by the dominant weight of the SSV, the phone corresponding to the absolute maximum weight is hypothesized as the pronunciations made by the dysarthric speaker as shown in table 1. There may be cases where the canonical pronunciation (center phone of the tied-state) does not represent the observed pronunciation (phone associated with the absolute maximum weight of the SSV). In that case, it means that the phone model built for the speaker represents the observed pronounced phone rather than the canonical pronounced phone.

5.2.1. Analysis of the mispronunciations picked up by the SSV

Figure 5 shows the SSV plot for the second state of the triphone $/ae/ - /jh/ + /sil/$ of unimpaired and dysarthric speaker (BV). As discussed in Section 4, the dominant weight indicates the pronounced phone $/jh/$ for the triphone of the unimpaired speaker. But for dysarthric speaker, instead of $/jh/$, the phone $/ey/$ gets the dominant weight. This indicates that the phone $/jh/$ is mispronounced as $/ey/$. In order to analyze our hypothesis of the dysarthric speaker pronunciations captured by the dominant weight of the SSV, perceptual test was conducted. The audio samples of the BV speaker in the context for the word "$Badge - b\,ae\,jh$" (canonical pronunciation) was heard as "$Badge - b\,ae\,ey$" (dysarthric speaker's observed pronunciation). The audio sample was verified by 10 naive listeners and their mean opinion score was taken.

To further support this hypothesis, the spectrogram of the word "$Badge$" pronounced by unimpaired and dysarthric speaker is shown in figure 4. The fricative $/jh/$ is clearly visible in the spectrogram of unimpaired speaker, while for dysarthric speaker the diphthong $/ey/$ occurs instead of actual phone $/jh/$. Hence the second state of the triphone model $/ae/-/jh/+/sil/$ is more acoustically closer and represents the second state of the triphone model $/ae/-/ey/+/sil/$, captured directly by the SSV using its dominant weight. Thus we confirm our hypothesis that the phone captured by the SSV using its dominant weight corresponds to the pronunciations made by the dysarthric speaker.

5.3. Formation of phone confusion matrix using SSV

Using the SSV corresponding to a tied-state for a particular dysarthric speaker's model, a phone confusion matrix is formed. The set of canonical pronunciations (center phone of the tied-state) and the set of observed dysarthric speaker's pronunciations (absolute maximum weight of the SSV for each tied-state) are used to form the phone confusion matrix. From each dysarthric speaker's Phone-CAT model, speaker-specific phone confusion matrix is formed. Each row of the matrix corresponds to canonical pronunciations and each column represents the observed dysarthric speaker's pronunciations. The sum of all elements of the matrix corresponds to the total number of tied-states.

The diagonal elements represent the number of correct pronunciations made by the speaker, where the center phone of the tied-state is correctly picked up by the SSV as its dominant weight. The off-diagonal elements represents the mispronunciations made by that speaker, where the center phone does not correspond to the dominant weight of the SSV. Value in each element of the matrix say aij, represents the frequency of occurrence of the canonical phone i being mispronounced as phone j. This phone confusion matrix also correlates with the intelligibility scores of the different severity levels of the speakers. Since the diagonal elements represent the correct pronunciations, the number of elements across the diagonal varies with respect to the severity level of dysarthria. As the degree of impairment increases, the diagonal pattern disintegrates. Thus phone

confusion matrix helps in the objective assessment of dysarthric speech [22]. Apart from assessment, phone confusion matrix is used for improving the lexical models which is the main focus of this paper.

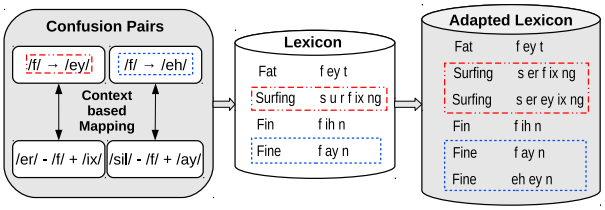

Figure 6: Schematic diagram of the proposed method to build an adapted lexicon from the phone confusion pairs using context dependent mapping. Here the canonical phone /f/ is confused with /ey/ when /f/ occurs in between the context /er/ and /ix/. Hence the word "surfing — s er **f** ix ng" gets the alternate pronunciation as "surfing — s er **ey** ix ng" while the words fat, fin and fine are neglected.

Algorithm 1 Procedure to form phone confusion matrix and modified lexicon from SSV

1. For each dysarthric speaker
   - (a) Build the Phone-CAT acoustic model and extract the P-dimensional SSV from the set of tied-states
   - (b) Absolute maximum weight of each SSV is picked up as the dysarthric speaker's pronounced phones as shown in table 1
   - (c) Using the set of canonical pronunciations (center phones of the tied-states) and the observed dysarthric pronunciations (absolute maximum values of the SSV), a phone confusion matrix is formed
   - (d) The list of substitution phones are obtained from the phone confusion matrix using a threshold
   - (e) Obtained confusion pairs are mapped only for those canonical phones when their context matches with the corresponding tied-state
   - (f) The modified lexicon along with the alternate pronunciations is then used to rebuild the acoustic models

5.4. Improved lexical models using SSV

The phone confusion matrix captures the mispronunciations made by each dysarthric speaker. The list of substitution phones are obtained from the set of mispronunciations in the matrix using a threshold rule. The substitution phones are further used to form alternate pronunciations forming speaker-specific lexicons. For example, if the phone $/f/$ is mispronounced as $/ey/$ in the confusion matrix with high recurrence, then it is taken as substitution phone. For the word "$five$" the alternate pronunciation in the lexicon is given as:

$$[Five] -> /\mathbf{f}\ ay\ v/ \quad \textit{(canonical pronunciation)}$$
$$[Five] -> /\mathbf{ey}\ ay\ v/ \quad \textit{(alternate pronunciation)}$$

It was shown that adding context-dependent pronunciation variation models helps in improving the performance of the system [18]. The triphones corresponding to the phone confusion pairs, are used to substitute the phones in the lexicon, for the words with the corresponding triphone context information as in figure 6. In the figure, $/f/$ is substituted with $/ey/$ only for the word "surfing" which contains the triphone context $/er/-/f/+/ix/$. For other words with phone $/f/$, no alternate pronunciations were given. This helps in reducing the size of the lexicon.

The number of confusion pair to be substituted from the confusion matrix is chosen based on the threshold rule. This helps in reducing the number of confusion pairs avoiding the selection of alternative confusion pair for each canonical phone. This modified lexicon is composed with grammar in WFST framework. In this approach, we mainly focus on modeling the substitution errors using the alternate pronunciations. Further, the modified lexicon is used to rebuild the acoustic model in the HMM-GMM framework.

6. Experimental setup

The experiments were performed in Kaldi [23] open-source speech recognition toolkit. Nemours database [24] was used for our experiments. It contains continuous speech utterances with 16 KHz sampling rate. It has 11 speakers, out of which only 10 speakers were used for our experiments [24]. Each speaker recorded 74 nonsensical sentences of the form "The N1 is Ving the N2" where the N1 and N2 are monosyllabic noun and V is the disyllabic verb. The lexicon is expanded in terms of phones with vocabulary size of 113 words and 39 phones in ARPAbet (advanced research project agency) symbol set is used for experimentation. One unimpaired speaker's data covering all the sentences spoken by each dysarthric speaker was recorded as control subject. The standard Frenchay dysarthric assessment (FDA) scores were also provided for each dysarthric speaker. The train data contains 490 utterances and test data contains 250 utterances, selected using 3-fold cross validation procedure. Trigram language model was used and the performance of the continuous density hidden Markov model (CDHMM) is measured using word error rate (WER). Baseline CDHMM is built with 200 tied-states and 10 Gaussian mixture components. The baseline system uses the canonical lexicon for both training and testing.

Rate	Mdl	FB	MH	BB	LL	JF	RL	RK	BK	BV	SC
Ins	Base	0	0	1	0	2	6	0	8	0	7
	Expt 1	0	0	1	0	2	6	0	7	0	7
	Expt 2	0	0	0	0	0	5	0	6	0	3
Del	Base	0	0	0	0	0	0	4	18	0	0
	Expt 1	0	0	0	0	0	0	2	17	0	0
	Expt 2	0	0	0	0	0	0	0	17	0	0

Table 2: Rate of insertions (Ins) and deletions (Del) for different models (mdl): Baseline (Base), Expt 1 and Expt 2

7. Results and Discussion

7.1. Results with modified lexicons : Proposed method

Two different experiments were performed to compare with the baseline CDHMM (Base) system. First is to train acoustic model using canonical lexicon and decoding the text using the modified lexicon (Expt 1). The second experiment is to use the modified lexicon for both training and testing process (Expt 2). Speaker-wise results for both the experiments are shown in the table 4. Comparing with baseline, all the speakers obtain improved performance for the system rebuilt using the modified lexicons (Expt 2). On an average, the relative improvement is 13.1%. Comparing with baseline system, an relative improve-

ment of 5.4% is obtained across all the speakers for the system only tested using modified lexicon (Expt 1). Severe category speakers shows considerable improvement compared to moderate and mild category speakers. Substitutions form a major portion of the error compared to insertions and deletions in our model. Hence we focused on reducing the number of substitution errors in this paper. Figure 7 shows the reduction in the rate of substitutions for the proposed model compared to baseline system. The rate of insertions and deletions were also reduced which is shown in table 2.

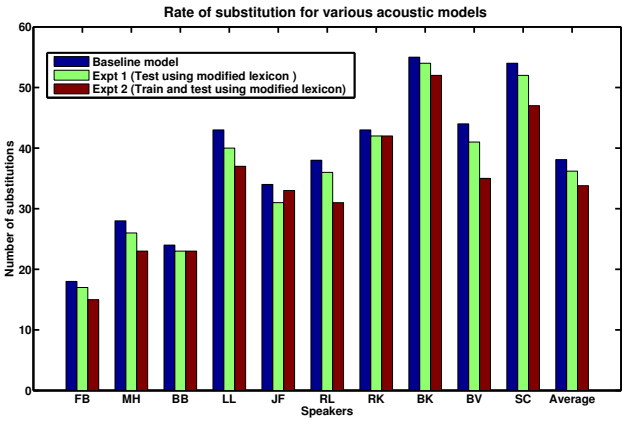

Figure 7: Substitution rate for different models

Table 3: Comparison of proposed method with existing method

Lexicon Usage	**Method/ Model type**	**MH (Mild)**	**RK (Moderate)**	**SC (Severe)**
Training $canonical$ + Testing $canonical$	Baseline	18.7	31.3	29.3
Training $canonical$ + Testing $modified$ (Expt 1)	Existing method	17.3	29.3	27.3
	Proposed method	16.7	29.3	27.3
Training $modified$ + Testing $modified$ (Expt 2)	Existing method	16.0	30.0	26.0
	Proposed method	15.3	29.3	23.3
	% R.I	4.2	2.2	10.3

Here % R.I denotes relative improvement with respect to existing method

7.2. Comparison with Existing Method

Some of the existing methods in literature involve forming phone confusion matrix aligning the recognized phoneme sequence with reference transcriptions [10]. Then using rule-based method, speaker dependent multiple pronunciation lexicons are formed. In [11], the recognized transcription is aligned with the reference transcription to form the phone confusion matrix. The confusion pairs are then used to provide multiple pronunciations in the lexicon. In order to compare our proposed method with the existing method, the confusion pairs from the phone confusion matrix formed by aligning the recognized transcription with the reference transcription on baseline model are used to form the lexicon.

Table 4: Results of lexical modeling for Nemours database in terms of % word error rate (% WER)

Severity	**Speakers**	**Baseline CDHMM**	**Testing using new lexicon (Expt 1)**	**Train + Test using new lexicon (Expt 2)**
Mild	FB	12.0	11.3	10.0
	MH	18.6	17.3	15.3
	BB	16.6	16.0	15.3
	LL	28.6	26.6	24.6
Moderate	JF	24.0	22.0	22.0
	RL	29.3	28.0	24.0
	RK	31.3	29.3	29.3
Severe	BK	54.0	52.0	50.0
	BV	29.3	27.3	23.3
	SC	40.6	39.3	33.3
Average		**28.7**	**26.9**	**24.7**

Similar to section 7.1, two different experiments (Expt 1 and Expt 2) were performed on baseline model using the modified lexicon formed using this phone confusion matrix for three different severity category. As shown in table 3, proposed method using phone confusion matrix formed from SSV shows an relative improvement of 10.3% compared to existing method using phone confusion matrix formed using decoded transcription. In the existing method, a single frame is involved in estimating the likelihood with respect to the corresponding acoustic model. While in case of our proposed approach, a set of frames corresponding to a tied-state label is involved in the estimation of SSV. Thus the estimated SSV are more reliable in identifying the confusion pairs which helps in improving the recognition performance over existing method.

8. Conclusions

This paper focuses on improving the performance of dysarthric speech recognition systems by handling pronunciation errors. A novel approach of forming phone confusion matrix for each dysarthric speaker using SSV from Phone-CAT model is discussed. Phone-CAT model handles the data efficiently by using less number of parameter for estimation. The SSV captures the context and phonetic information. It represent the enunciated phone using the dominant weight. This property is used to identify the mispronunciations made by each dysarthric speaker, by building speaker-specific Phone-CAT model. Using the phone confusion matrix, alternate pronunciations are formed in personalized speaker lexicons. These modified lexicons improves the performance of the dysarthric ASR system. This preliminary study shows that the proposed phone confusion matrix using SSV captures the speaker-specific pronunciation patterns and avoid the usage of decoded transcription. This approach has to be explored in detail to analyze, model the error pattern and handle the insertion and deletion errors which forms our future work.

9. References

[1] J. R. Duffy, “Motor speech disorders: clues to neurologic diagnosis,” in *Parkinsons Disease and Movement Disorders*, pp. 35–53, Springer, 2000.

[2] R. D. Kent, G. Weismer, J. F. Kent, H. K. Vorperian, and J. R. Duffy, “Acoustic studies of dysarthric speech: Meth-

ods, progress, and potential," *Journal of communication disorders*, vol. 32, no. 3, pp. 141–186, 1999.

[3] M. B. Mustafa, S. S. Salim, N. Mohamed, B. Al-Qatab, and C. E. Siong, "Severity-based adaptation with limited data for asr to aid dysarthric speakers," *PloS one*, vol. 9, no. 1, p. e86285, 2014.

[4] P. Raghavendra, E. Rosengren, and S. Hunnicutt, "An investigation of different degrees of dysarthric speech as input to speaker-adaptive and speaker-dependent recognition systems," *Augmentative and Alternative Communication*, vol. 17, no. 4, pp. 265–275, 2001.

[5] F. Rudzicz, "Comparing speaker-dependent and speaker-adaptive acoustic models for recognizing dysarthric speech," in *Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility*, pp. 255–256, ACM, 2007.

[6] E. Rosengren, P. Raghavendra, and S. Hunnicutt, "How does automatic speech recognition handle dysarthric speech?," *Lund Working Papers in Linguistics*, vol. 43, pp. 112–115, 2009.

[7] H. Kim, K. Martin, M. Hasegawa-Johnson, and A. Perlman, "Frequency of consonant articulation errors in dysarthric speech," *Clinical linguistics & phonetics*, vol. 24, no. 10, pp. 759–770, 2010.

[8] E. Sanders, M. B. Ruiter, L. Beijer, and H. Strik, "Automatic recognition of dutch dysarthric speech: a pilot study.," in *INTERSPEECH*, 2002.

[9] K. T. Mengistu and F. Rudzicz, "Adapting acoustic and lexical models to dysarthric speech," in *Proc. ICASSP*, pp. 4924–4927, IEEE, 2011.

[10] W. K. Seong, J. H. Park, and H. K. Kim, "Multiple pronunciation lexical modeling based on phoneme confusion matrix for dysarthric speech recognition," *Advanced Science and Technology Letters*, vol. 14, pp. 57–60, 2012.

[11] S. O. C. Morales and S. J. Cox, "Modelling errors in automatic speech recognition for dysarthric speakers," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, p. 2, 2009.

[12] V. Manohar, C. Srinivas, S. Umesh, *et al.*, "Acoustic modeling using transform-based phone-cluster adaptive training," in *Proc. ASRU*, pp. 49–54, IEEE, 2013.

[13] C.-H. Wu, H.-Y. Su, and H.-P. Shen, "Articulation-disordered speech recognition using speaker-adaptive acoustic models and personalized articulation patterns," *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 10, no. 2, p. 7, 2011.

[14] S.-O. Caballero-Morales and F. Trujillo-Romero, "Evolutionary approach for integration of multiple pronunciation patterns for enhancement of dysarthric speech recognition," *Expert Systems with Applications*, vol. 41, no. 3, pp. 841–852, 2014.

[15] P. Jyothi and E. Fosler-Lussier, "Discriminative language modeling using simulated asr errors.," in *INTERSPEECH*, pp. 1049–1052, 2010.

[16] G. Chen, O. Yilmaz, J. Trmal, D. Povey, and S. Khudanpur, "Using proxies for oov keywords in the keyword search task," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, pp. 416–421, IEEE, 2013.

[17] W. K. Seong, J. H. Park, and H. K. Kim, "Dysarthric speech recognition error correction using weighted finite state transducers based on context–dependent pronunciation variation," in *Computers Helping People with Special Needs*, pp. 475–482, Springer, 2012.

[18] W. K. Seong, J. H. Park, and H. K. Kim, "Performance improvement of dysarthric speech recognition using context-dependent pronunciation variation modeling based on kullback-leibler distance," *Advanced Science and Technology Letters*, vol. 14, no. 1, pp. 53–56, 2012.

[19] H. Christensen, P. D. Green, and T. Hain, "Learning speaker-specific pronunciations of disordered speech.," in *INTERSPEECH*, pp. 1159–1163, 2013.

[20] M. J. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer speech & language*, vol. 12, no. 2, pp. 75–98, 1998.

[21] B. Abraham, N. M. Joy, and N. K. Umesh, "A data-driven phoneme mapping technique using interpolation vectors of phone-cluster adaptive training," in *Spoken Language Technology Workshop (SLT), 2014 IEEE*, pp. 36–41, IEEE, 2014.

[22] R. Sriranjani, S. Umesh, and M. Reddy, "Automatic severity assessment of dysarthria using state-specific vectors.," *Biomedical sciences instrumentation*, vol. 51, pp. 99–106, 2015.

[23] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, *et al.*, "The kaldi speech recognition toolkit," in *Proc. ASRU*, pp. 1–4, 2011.

[24] X. Menendez-Pidal, J. B. Polikoff, S. M. Peters, J. E. Leonzio, and H. T. Bunnell, "The nemours database of dysarthric speech," in *Proc. ICSLP*, vol. 3, pp. 1962–1965, IEEE, 1996.