INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Dysarthric Speech Recognition Using Kullback-Leibler Divergence-based Hidden Markov Model

Myungjong Kim[1], *Jun Wang*[1], *Hoirin Kim*[2]

[1] Speech Disorders & Technology Lab, University of Texas at Dallas, United States
[2] School of Electrical Engineering, Korea Advanced Institute of Science and Technology, Korea

myungjong.kim@utdallas.edu, wangjun@utdallas.edu, hoirkim@kaist.ac.kr

Abstract

Dysarthria is a neuro-motor speech disorder that impedes the physical production of speech. Patients with dysarthria often have trouble in pronouncing certain sounds, resulting in undesirable phonetic variation. Current automatic speech recognition systems designed for the general public are ineffective for dysarthric sufferers due to the phonetic variation. In this paper, we investigate dysarthric speech recognition using Kullback-Leibler divergence-based hidden Markov models. In the model, the emission probability of state is modeled by a categorical distribution using phoneme posterior probabilities from a deep neural network, and therefore, it can effectively capture the phonetic variation of dysarthric speech. Experimental evaluation on a database of several hundred words uttered by 30 speakers consisting of 12 mildly dysarthric, 8 moderately dysarthric, and 10 control speakers showed that our approach provides substantial improvement over the conventional Gaussian mixture model and deep neural network based speech recognition systems.

Index Terms: dysarthria, Kullback-Leibler divergence-based hidden Markov model, speech recognition

1. Introduction

Dysarthria is a neuro-motor speech disorder resulting from neurological injury of the motor speech system [1], [2]; dysarthria damages the physical production of speech, rendering it unintelligible. Dysarthria is often accompanied with a physical disability such as cerebral palsy that limits the speaker's capability to communicate through computers and electronic devices. Although an automatic speech recognition (ASR) system is essential for dysarthria sufferers, current ASR systems for the general public are not well-suited to dysarthric speech because of acoustic mismatch resulting from their articulatory limitation [3]. In other words, dysarthric individuals often fail to pronounce certain sounds, leading to undesirable phonetic variation which is the main cause of performance degradation.

Related works on the recognition of dysarthric speech have been mostly focused on acoustic modeling to capture the acoustic cues of disordered speech. Hasegawa-Johnson *et al*. [4] compared ASR systems based on Gaussian mixture model-hidden Markov models (GMM-HMMs) and support vector machines (SVMs). They reported that HMM-based models may provide robustness against large-scale word-length fluctuations and SVM-based models can handle the deletion or reduction of consonants. Rudzicz [5], [6] compared several acoustic models including GMM-HMM, artificial neural networks (ANNs), conditional random field, and SVMs. Their experimental results show that discriminative models such as ANNs produced better phoneme classification accuracy than GMM-based generative acoustic models. Further, an ANN-HMM hybrid approach in which HMM states are modeled by ANNs was presented to improve the recognition performance of disordered speech [7].

Another research direction is to handle the phonetic variation of dysarthric speech in an explicit or implicit way. Explicit phonetic variation modeling generally creates multiple pronunciations for each word in the lexicon. Mengistu and Rudzicz [8] manually made a pronunciation lexicon for each individual with dysarthria through expert assessment of the individual's pronunciation. Christensen *et al*. [9] automatically generated a speaker-specific pronunciation dictionary using phoneme posterior probabilities of a deep neural network (DNN), which is an ANN with multiple hidden layers, trained on normal speech. Also, weighted finite state transducers (WFSTs) were built using phonetic confusion matrices resulting from a normal ASR system to allow phonetic variation during decoding process [10], [11]. Implicit modeling, on the other hand, depends on the underlying acoustic-phonetic models to account for phonetic variation, such as model parameter tying [28], and therefore it can remove the necessity to explicitly determine and represent phonetic variation in the lexicon. Although implicit phonetic variation modeling is promising, it has rarely been investigated in the field of dysarthric speech recognition.

Recently, Kullback-Leibler divergence-based HMM (KL-HMM) [12], [13] has been emerging since KL-HMM is a very powerful and flexible framework in achieving implicit phonetic variation modeling. KL-HMM is a particular form of HMM in which the emission probability of state is parametrized by a categorical distribution of phoneme classes referred as acoustic units. Since HMM states are generally represented as subword lexical units in the lexicon, KL-HMM can model the phonetic variation against target phonemes. For score computation in training and decoding, KL divergence-based dissimilarity measure between the categorical distribution and phoneme posterior probabilities is used. KL-HMM has been successfully utilized in various speech recognition applications such as non-native speech recognition [14], multilingual speech recognition [15], and grapheme-based speech recognition [16].

In this paper, we investigate the effectiveness of KL-HMM for dysarthric speech recognition. To effectively model the typical phonetic variation of dysarthric speech, the categorical distribution of KL-HMM is trained on speech data from several dysarthric talkers using (context-dependent) phoneme posterior probabilities obtained from a DNN acoustic model. Several DNN-based acoustic models such as normal DNN and DNN

http://dx.doi.org/10.21437/Interspeech.2016-776

adapted on dysarthric speech are compared to explore the effectiveness of an acoustic model in training KL-HMM.

2. Dysarthric speech data

We collected speech data from 78 native Korean speakers of which 68 (40 males and 28 females) were dysarthric and 10 (5 males and 5 females) were non-dysarthric control speakers. All dysarthric speakers were recruited from Seoul National Cerebral Palsy Public Welfare and had been diagnosed with cerebral palsy, which is one of the most prevalent causes of dysarthria [17]. The mean ages of the dysarthric and control participants were 36.6 years (standard deviation of 9.7 years) and 33.1 years (standard deviation of 3.9 years), respectively.

All speakers spoke an average of 628 isolated words, including repetitions of 37 Assessment of Phonology and Articulation for Children (APAC) words, 100 command words, 36 Korean phonetic codes which are used for identifying the Korean alphabet letters in voice communication, a subset from 452 Korean Phonetically Balanced Words (PBW), and a subset from 500 additional command words. Recordings were made in a quiet office with a Shure SM12A head-worn microphone at 16 kHz sampling rate in a mono-channel.

All participants were diagnosed by a speech-language pathologist, who has a top level license for speech therapy and has worked in the field over five years, according to the percentage of consonants correct (PCC) [18] using the APAC words [19]. The APAC words comprised familiar vocabulary words composed of one to four syllables and were phonetically balanced to partially assess the articulation ability on a phonetic basis [20], [21]. Based on this assessment, among the 68 dysarthric subjects, 37 subjects were graded as mildly dysarthric (PCC 85-100%) and 31 subjects were graded as moderately dysarthric (PCC 50-84.9%). All control subjects were graded as PCC 100%.

3. KL-HMM framework

A KL-HMM framework is mainly composed of two models [22]: 1) A neural network-based acoustic model which represents the relationship between acoustic feature observations and acoustic units and 2) a categorical distribution-based lexical model which captures a probabilistic relationship between the subword lexical units in the pronunciation lexicon and the acoustic units. The acoustic units can be chosen as context-independent or clustered context-dependent phonemes.

3.1. DNN-based acoustic model

A DNN has been received great attention since the complex structure of speech sounds can be modeled through multiple layers using powerful optimization techniques such as generative layer-wise pretraining and discriminative fine-tuning, and therefore it has been successfully applied in speech recognition as an acoustic model [23]-[25]. It is expected that the DNN-based acoustic model may also capture the complex acoustic structure of dysarthric speech as well. In this work, we used 40 log mel-filterbank energies with 11 context window $\mathbf{x}_t=\{\mathbf{x}_{t-5},\ldots,\mathbf{x}_t,\ldots,\mathbf{x}_{t+5}\}$ as acoustic feature observations and clustered context-dependent phonemes, i.e., senones, as output units or acoustic units a^d. Given the DNN acoustic model, the probabilities of acoustic units, i.e., D-dimensional acoustic unit posterior probability vectors, can be obtained as

$$\begin{aligned}\mathbf{z}_t &= \left[z_t^1,\ldots,z_t^d,\ldots,z_t^D \right]^T \\ &= \left[P\left(a^1 \mid \mathbf{x}_t\right),\ldots,P\left(a^d \mid \mathbf{x}_t\right),\ldots,P\left(a^D \mid \mathbf{x}_t\right) \right]^T\end{aligned} \quad (1)$$

Then, the acoustic unit posterior probability vectors are used to train categorical distributions in HMM states which correspond to the lexical units.

3.2. Categorical distribution-based lexical model

KL-HMM is a particular type of HMM where the emission probability of state l^i of a lexical unit is parametrized by a categorical distribution $\mathbf{y}_i=[y_i^1,\ldots,y_i^d,\ldots,y_i^D]^T$, where $y_i^d=P(a^d|l^i)$. Therefore, each HMM state can capture a probabilistic relationship between a lexical unit l^i and D acoustic units.

In the KL-HMM framework, the following KL divergence between the acoustic unit posterior vector $\mathbf{z}_t$ and the categorical variable $\mathbf{y}_i$ is used as the local score at each HMM state.

$$d\left(\mathbf{z}_t,\mathbf{y}_i\right)=\sum_{d=1}^{D} z_t^d \log\left(\frac{z_t^d}{y_i^d}\right) \quad (2)$$

Actually, there are a number of ways to obtain the KL divergence such as symmetric variant of the KL divergence. However, recent studies reported that asymmetric KL divergence as in (2) is more robust [15]. Therefore, we used the asymmetric version of the KL divergence as the local score in this work.

Given the acoustic unit probability vectors $Z=[\mathbf{z}_1,\ldots,\mathbf{z}_t,\ldots,\mathbf{z}_T]$ where T represents the number of frames, the categorical variables $Y=[\mathbf{y}_1,\ldots,\mathbf{y}_i,\ldots,\mathbf{y}_L]$ where L denotes the number of lexical units can be trained by minimizing the cost function defined by summing the local scores over time t and state l^i as follows:

$$\min \sum_{t=1}^{T}\sum_{i=1}^{L} d\left(\mathbf{z}_t,\mathbf{y}_i\right)\delta_t^i \qquad s.t.\sum_{d=1}^{D} y_i^d = 1 \quad (3)$$

where $\delta_t^i = 1$ if $\mathbf{x}_t$ is associated with state l^i, otherwise 0. Here, the state association of each $\mathbf{x}_t$ is determined using Viterbi forced alignment. To minimize the cost function in (3), we take the partial derivative with respect to each variable $\mathbf{y}_i$ and set it to zero. Finally, the optimal state distribution is the arithmetic mean of the acoustic unit probability vectors assigned to the state given by

$$y_i^d = \frac{1}{T_i}\sum_{t\in l^i} z_t^d \quad (4)$$

where T_i denotes the number of frames associated with state l^i. Finally, decoding is performed using the standard Viterbi decoder with the KL divergence-based local score defined in (2).

3.3. Application to dysarthric speech recognition

KL-HMM has advantages for dysarthric speech recognition. First, it can effectively represent phonetic variation through categorical distribution-based lexical modeling, which may be particularly useful for dysarthric speech. Therefore, it is expected that KL-HMM is appropriate in recognizing disordered speech. Second, the acoustic model and the lexical model can be trained on an independent set of resources [22]. For example, the acoustic model can be trained on resources from resource-rich domains whereas the lexical model can be trained on a relatively small amount of resources from a target domain. Using this knowledge, the acoustic model is trained on data from a large population with normal speech (or further

Table 1. *Word error rates (%) of DNN_{nor}-KL_{dys}-HMM according to the context dependency of acoustic units and lexical units for dysarthric and control speakers.*

AU \ LU	CI		CD	
	Dys.	Con.	Dys.	Con.
CI	43.9	1.4	38.9	1.2
CD	41.8	1.8	**33.4**	**0.9**

adapted on dysarthric data) whereas the lexical model is trained on a relatively small amount of dysarthric speech data. This strategy is reasonable because the size of an acoustic model is generally much larger than the size of a lexical model.

4. Experimental results

4.1. Experimental setup

The normal training set includes 300k utterances (about 54 hours) of 8k Korean isolated words from several databases (DBs) consisting of the Korean Phonetically Optimized Words (KPOW) DB, Korean Phonetically Balanced Words (KPBW) DB, and Korean Phonetically Rich Words (KPRW) DB, which are widely used for acoustic modeling in Korea. The dysarthric training set includes 20k utterances (about 4 hours) from 48 dysarthric speakers described in Section 2. Also, the evaluation set consists of 23k utterances spoken by 20 dysarthric speakers including 12 mild and 8 moderate subjects, and 10 non-dysarthric control speakers. Specifically, each dysarthric speaker utters 5 repetitions of 100 command words and 36 Korean phonetic codes, and 213 additional command words, i.e., a total of 893 utterances. Each control speaker utters 2 repetitions of 100 command words and 36 Korean phonetic codes, and 213 additional command words, i.e., a total of 485 utterances. The repeated data are obtained in multiple sessions. The speakers in the evaluation set are totally separated from the training set.

We compared three ASR systems: GMM-HMM, DNN-HMM, and KL-HMM systems.

GMM-HMM system: We first train a normal GMM-HMM system (referred as GMM_{nor}-HMM) using 39 dimensional mel-frequency cepstral coefficients, consisting of 12 cepstral coefficients, 1 energy term, and their first and second derivatives with frame size of 25 milliseconds and shift size of 10 milliseconds. The GMM_{nor}-HMM consists of 1480 tied-state (senone) left-to-right triphone HMMs, where each HMM has 3 states and each state is modeled with 16 Gaussian components and is trained on the normal training set. The dysarthric GMM can be obtained by adapting the GMM_{nor} to dysarthric speech using maximum *a posteriori* (MAP) adaptation on the dysarthric training set (referred as GMM_{nor}-MAP_{dys}-HMM).

DNN-HMM system: A normal DNN is trained using 40 dimensional log mel-filterbank energy features with a context window of 11 frames and frame alignment information resulting from the GMM_{nor}-HMM system. The DNN has 3 hidden layers with 1024 hidden units at each layer and the 1480 dimensional softmax output layer, corresponding to the number of senones of the GMM_{nor}-HMM system. The parameter is initialized using layer-by-layer generative pre-training and the network is discriminatively trained using backpropagation [26] (referred as DNN_{nor}-HMM). To further construct dysarthric DNN, linear output network adaptation [27] (DNN_{nor}-LON_{dys}-HMM) and DNN retraining (DNN_{dys}-HMM) using dysarthric training set were considered.

Table 2. *Word error rates (%) of DNN_{nor}-HMM* and *DNN_{nor}-KL_{dys}-HMM with the number of DNN hidden layers for dysarthric and control speakers.*

# of hidden layers	DNN_{nor}-HMM		DNN_{nor}-KL_{dys}-HMM	
	Dys.	Con.	Dys.	Con.
1	47.1	0.7	34.8	1.3
2	45.8	0.7	33.9	1.0
3	**44.8**	**0.4**	**33.4**	**0.9**
4	45.1	0.5	33.6	0.9
5	45.0	0.6	33.6	0.9

KL-HMM system: A KL-HMM is trained using DNN posterior probability vectors obtained from the dysarthric training set and frame alignment information resulting from the DNN_{nor}-HMM system. In this work, DNN_{nor} and DNN_{nor}-LON_{dys} are considered as acoustic models in obtaining posterior probability vectors, and therefore, we can refer to these systems as DNN_{nor}-KL_{dys}-HMM and DNN_{nor}-LON_{dys}-KL_{dys}-HMM, respectively.

4.2. Effectiveness of context dependency of acoustic units and lexical units

We first examine the effectiveness of context dependency of acoustic units (AUs) and lexical units (LUs). Table 1 presents the performances of DNN_{nor}-KL_{dys}-HMM systems according to the types of AUs and LUs, which are context-independent (CI) phonemes or context-dependent (CD) phonemes. The number of CI units is 148 (46 phonemes x 3 states + 2 silences x 5 states) and the number of CD units (senones) is 1480. Interestingly, the CD-LU systems produce better performances than CI-LU systems regardless of AU types for both dysarthric and control speakers. This is the reason why CD-LU systems can utilize more temporal information and more target lexical units. When both context-dependent acoustic and lexical units are used, we can obtain the best results for dysarthric and control speakers. This implies that various phonetic variation can be properly modeled through KL-HMM. In the following experiments, CD units are used as AUs and LUs.

4.3. Effectiveness of DNN-based acoustic model

Table 2 compares the performances of DNN_{nor}-HMM and DNN_{nor}-KL_{dys}-HMM systems by varying the number of DNN hidden layers for dysarthric and control speakers. While the number of hidden layers increases, the speech recognition performances are improved for both dysarthric and control speakers on both systems. In addition, we can observe both DNN_{nor}-HMM and DNN_{nor}-KL_{dys}-HMM show the best performances when the DNN with 3 hidden layers that produces the lowest WER is chosen as an acoustic model. This indicates that choosing a better acoustic model is important in achieving better performance in KL-HMM as well. In addition, DNN_{nor}-KL_{dys}-HMM outperforms DNN_{nor}-HMM for dysarthric speakers whereas its performance is slightly degraded compared with DNN_{nor}-HMM for control speakers. This implies that the general characteristics of phonetic variability of dysarthric speech are reflected to the DNN_{nor}-KL_{dys}-HMM. Since there is a trade-off between dysarthric and control speakers, reducing the gap is important in improving the

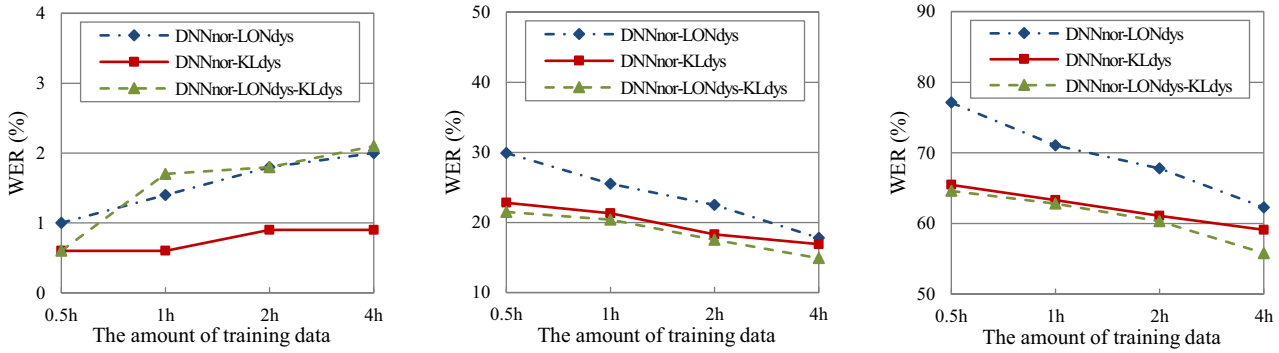

Figure 1: *Performance evaluation with the amount of dysarthric training data from 0.5 hours to 4 hours for (a) control, (b) mildly dysarthric, and (c) moderately dysarthric speakers.*

Table 3. *Performance comparison of GMM-HMM, DNN-HMM, and KL-HMM systems.*

ASR system	WER (%)		
	Dys.	Con.	Avg.
GMM_{nor}-HMM	51.1	0.7	25.9
GMM_{nor}-MAP_{dys}-HMM	42.3	1.5	21.9
DNN_{nor}-HMM	44.8	0.4	22.6
DNN_{nor}-LON_{dys}-HMM	35.3	2.0	18.7
DNN_{dys}-HMM	35.5	2.2	18.9
DNN_{nor}-KL_{dys}-HMM	**33.4**	0.9	17.2
DNN_{nor}-LON_{dys}-KL_{dys}-HMM	**31.0**	2.1	16.6

compatibility of an ASR system. In the following experiments, the DNN with 3 hidden layers is used as the default system for the remainder of this paper.

4.4. Effectiveness of KL-HMM

Table 3 shows the performances of GMM-HMM, DNN-HMM, and KL-HMM systems for both dysarthric and control speakers in terms of the word error rate (WER). Also, we measured unweighted average WERs across dysarthric and control speakers to evaluate the compatibility of each ASR system for universal access. For the comparison of GMM_{nor}-HMM and DNN_{nor}-HMM systems, the performance of the DNN_{nor}-HMM is better than with the GMM_{nor}-HMM for both dysarthric and control speakers. This implies that the DNN acoustic model is more effective in recognizing speech uttered by control speakers as well as patients with dysarthria. It is also observed that the systems trained on dysarthric data such as DNN_{nor}-LON_{dys}-HMM produce better results than with systems trained on only normal data in terms of the unweighted average WER. The performance of DNN_{nor}-LON_{dys}-HMM is slightly better than with DNN_{dys}-HMM. Since the amount of dysarthric training data is quite small in training all hidden layers, it is better to adapt DNN with LON adaptation. For the evaluation of our KL-HMM approach, DNN_{nor}-KL_{dys}-HMM outperforms DNN_{nor}-LON_{dys}-HMM for both dysarthric and control speakers, producing 5.8% relative improvement in the average WER reduction. In DNN_{nor}-LON_{dys}-KL_{dys}-HMM, we can achieve the lowest WER on dysarthric speakers, obtaining 12.2% relative improvement over DNN_{nor}-LON_{dys}-HMM, while for control speakers the performance is comparable. Through these experiments, we found that the KL-HMM approach is very effective for dysarthric speakers while keeping comparable performance for control speakers. Also, a good acoustic model that is better fitted to dysarthric speech is more appropriate in modeling KL-HMM.

4.5. Evaluation with the amount of training data

Next, we perform experiments with varying the amount of dysarthric training data for control, mildly dysarthric, and moderately dysarthric speaker groups in Figure 1. To this end, DNN_{nor}-LON_{dys}-HMM, DNN_{nor}-KL_{dys}-HMM, and DNN_{nor}-LON_{dys}-KL_{dys}-HMM are exploited. As can be seen, the KL-HMM approach consistently outperforms DNN_{nor}-LON_{dys}-HMM regardless of the amount of training data for all speaker groups. Specifically, when the amount of available training data gets small, the performance improvement of KL-HMM gets large over DNN_{nor}-LON_{dys}-HMM for dysarthric speakers. Through this experiment, we also found that KL-HMM is more robust on the data sparseness problem.

5. Conclusions

In this paper, we investigated the effectiveness of KL-HMM to improve the recognition performance of disordered speech. To deal with phonetic variation resulting from the limitation of articulatory movement, the KL-HMM framework composed of DNN acoustic modeling and categorical distribution-based probabilistic lexical modeling was exploited. In order to evaluate the effectiveness of our approach, a series of experiments were performed in terms of the WER on both 20 dysarthric and 10 control speakers. Experimental results showed that the KL-HMM approach provides significant improvement over the conventional ASR systems based on DNN-HMM when even a small amount of dysarthric training data is available. In this work, we tried to develop a speaker-independent speech recognition system for people with dysarthria by modeling the typical phonetic variation of dysarthric speech. Dysarthric speakers often have their own phonetic and articulatory variation patterns. Thus our further work includes applying speaker adaptation techniques and using articulatory information [29] on the KL-HMM framework.

6. Acknowledgements

This work was supported by the National Research Foundation of Korea under a grant number 2014R1A2A2A01007650 and the National Institutes of Health under an award number R03DC013990.

7. References

[1] J. R. Duffy, *Motor Speech Disorders: Substractes, Differential Diagnosis, and Management*, St Louis, MO: Elsevier Mosby, 2005.

[2] H. Kim, K. Mating, M. Hasegawa-Johnson, and A. Perlman, "Frequency of consonant articulation errors in dysarthric speech," *Clinical Linguist. Phonet.*, vol. 24, no. 10, pp. 759-770, Oct. 2010.

[3] 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 Technol.: The Official Journal of RESNA*, vol. 22, no. 2, pp. 99-112, 2010.

[4] M. Hasegawa-Johnson, J. Gunderson, A. Perlman, and T. Huang, "HMM-based and SVM-based recognition of the speech of talkers with spastic dysarthric," in *Proc. IEEE Int. Conf. Acoust. Speech, and Signal Process.*, 2006, pp. 1060-1063.

[5] F. Rudzicz, "Phonological features in discriminative classification of dysarthric speech," in *Proc. IEEE Int. Conf. Acoust. Speech, and Signal Process.*, Taipei, Apr. 2009, pp. 4605-4608.

[6] F. Rudzicz, "Articulatory knowledge in the recognition of dysarthric speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, May 2011, pp. 947-960.

[7] G. Jayaram and K. Abdelhamied, "Experiments in dysarthric speech recognition using artificial neural networks," *J. Rehabil. Res. Develop.*, vol. 32, no. 2, pp. 162-169, May 1995.

[8] K. T. Mengistu and F. Rudzicz, "Adapting acoustic and lexical models to dysarthric speech," in *Proc. IEEE Int. Conf. Acoust. Speech, and Signal Process.*, 2011, pp. 4924-4927.

[9] H. Christensen, P. Green, and T. Hain, "Learning speaker-specific pronunciations of disordered speech," in *Proc. Interspeech*, Lyon, France, Aug. 2013.

[10] S. O. C. Morales and S. J. Cox, "Modelling errors in automatic speech recognition for dysarthric speakers," *EURASIP J. Adv. Signal Process.*, vol. 2009, no. 1 Article ID 308340, 2009.

[11] 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 *Proc. 13th Int. Conf. Comput. Helping People Special Needs*, Linz, Austria, 2012, pp. 475-482.

[12] G. Aradilla, H. Bourlard, and M. Magimai-Doss, "Using KL-based acoustic models in a large vocabulary recognition task," in *Proc. Interspeech*, 2008.

[13] G. Aradilla, J. Vepa, and H. Bourlard, "An acoustic model based on Kullback-Leibler divergence for posterior features," in *Proc. IEEE Int. Conf. Acoust. Speech, and Signal Process*., 2007, pp. IV-657-IV-660.

[14] M. Razavi and M. M. Doss, "On recognition of non-native speech using probabilistic lexical model," in *Proc. Interspeech*, Singapore, Sep. 2014.

[15] D. Imseng, P. Motlicek, H. Bourlard, and P. N. Garner, "Using out-of-language data to improve an under-resourced speech recognizer," *Speech Commun.*, vol. 56, Jan. 2014, pp. 142-151.

[16] M. Magimai.-Doss, R. Rasipuram, G. Aradilla, and H. Bourlard, "Grapheme-based automatic speech recognition using KL-HMM," in *Proc. Interspeech*, Aug. 2011.

[17] B. Maassen, R. Kent, H. Peters, P. V. Lieshout, and W. Hulstijn, *Speech motor control in normal and disordered speech (chap. 12)*, Oxford University Press, 2004.

[18] L. D. Shriberg and J. Kwiatrkowski, "Phonological disorders III: A procedure for assessing severity of involvement," *J. Speech and Hearing Disorders*, vol. 47, no. 3, pp. 256-270, 1982.

[19] M. J. Kim, S. Pae, and C. Park, *Assessment of phonology and articulation for children*, Human Brain Research & Consulting, 2007.

[20] Y. Lee, J. E. Sung, and H. Sim, "Effects of listeners' working memory and noise on speech intelligibility in dysarthria," *Clinical Linguist. Phonet.*, vol. 28, no. 10, pp. 785-795, Oct. 2014.

[21] M. J. Kim, Y. Kim, and H. Kim, "Automatic intelligibility assessment of dysarthric speech using phonologically-structured sparse linear model," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 4, Apr. 2015, pp. 694-704.

[22] R. Rasipuram and M. Magimai.-Doss, "Articulatory feature based continuous speech recognition using probabilistic lexical modeling," *Comput. Speech Lang.*, vol. 36, pp. 233-259, Mar. 2016.

[23] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Ngugen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82-97, Nov. 2012.

[24] A. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 14-22, Jan. 2012.

[25] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. Audio, Speech Lang. Process*., vol. 20, no. 1, Jan. 2012.

[26] G. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput*., vol. 18, pp. 1527-1554, 2006.

[27] D. Yu and L. Deng, *Automatic speech recognition: A deep learning approach*, Springer-Verlag London, 2015.

[28] T. Hain, "Implicit modelling of pronunciation variation in automatic speech recognition," *Speech Commun.*, vol. 46, no. 2, pp. 171-188, 2005.

[29] S. Hahm, D. Heitzman, and J. Wang, "Recognizing dysarthric speech due to amyotrophic lateral sclerosis with across-speaker articulatory normalization," in *Proc. of the ACL/ISCA Workshop on Speech and Language Processing for Assistive Technologies*, 2015, pp. 47-54.