Published in final edited form as:
IEEE Trans Neural Syst Rehabil Eng. 2017 September ; 25(9): 1581–1591. doi:10.1109/TNSRE.2017.2681691.

Regularized Speaker Adaptation of KL-HMM for Dysarthric Speech Recognition

Myungjong Kim [Member, IEEE],
Department of Bioengineering, University of Texas at Dallas, Richardson, TX 75080, United States

Younggwan Kim,
School of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon 305-701, Korea

Joohong Yoo,
School of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon 305-701, Korea

Jun Wang [Member, IEEE], and
Department of Bioengineering, University of Texas at Dallas, Richardson, TX 75080, United States

Hoirin Kim [Member, IEEE]
School of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon 305-701, Korea

Abstract

This paper addresses the problem of recognizing the speech uttered by patients with dysarthria, which is a motor speech disorder impeding the physical production of speech. Patients with dysarthria have articulatory limitation, and therefore, they often have trouble in pronouncing certain sounds, resulting in undesirable phonetic variation. Modern automatic speech recognition systems designed for regular speakers are ineffective for dysarthric sufferers due to the phonetic variation. To capture the phonetic variation, Kullback-Leibler divergence based hidden Markov model (KL-HMM) is adopted, where the emission probability of state is parametrized by a categorical distribution using phoneme posterior probabilities obtained from a deep neural network-based acoustic model. To further reflect speaker-specific phonetic variation patterns, a speaker adaptation method based on a combination of L2 regularization and confusion-reducing regularization which can enhance discriminability between categorical distributions of KL-HMM states while preserving speaker-specific information is proposed. Evaluation of the proposed speaker adaptation method on a database of several hundred words for 30 speakers consisting of 12 mildly dysarthric, 8 moderately dysarthric, and 10 non-dysarthric control speakers showed that the proposed approach significantly outperformed the conventional deep neural network based speaker adapted system on dysarthric as well as non-dysarthric speech.

This paragraph of the first footnote will contain the date on which you submitted your paper for review.

Index Terms

Dysarthria; speech recognition; speaker adaptation; KL-HMM; regularization

I. Introduction

DYSARTHRIA is a motor speech disorder resulting from neurological injury of the motor speech system. Patients with dysarthria have trouble controlling the motor subsystems including respiration, phonation, resonance, articulation, and prosody [1]. Speech in patients with dysarthria is generally characterized by poor articulation of phonemes, breathy voice, and monotonic intonation [2]; thus, their speech intelligibility is reduced in proportion to the severity of dysarthria.

In general, dysarthria is often accompanied with a physical disability (e.g., cerebral palsy) that limits the speaker's capability to communicate through computers and electronic devices, making keyboard typing about 300 times slower than for regular users [3]. However, dysarthric speech is at most about 15 times slower than regular speech [4]. Consequently, spoken commands become an attractive alternative to normal keyboard and mouse input. In practice, people with dysarthria tend to prefer spoken expression over other physical modes due to its relative naturalness and speed [3], [6]. 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 [5]. That is, dysarthric individuals often fail to pronounce certain sounds, resulting in undesirable phonetic variation which is the main cause of performance degradation. Thus, it is necessary to develop an ASR system specialized for dysarthric speech.

Most studies on the recognition of dysarthric speech have been focused on acoustic modeling based on hidden Markov model (HMM) to capture the acoustic characteristics of disordered speech. Several HMM topologies such as ergodic or left-to-right HMM were investigated [7], [8]. They reported that left-to-right HMMs outperform ergodic HMMs for triphone acoustic modeling. This implies that state transitions not accounted for in left-to-right HMMs capture rather poorly the outlier events that differentiate dysarthric speech from unimpaired speech at the subword level [8].

An HMM state is generally modeled using a Gaussian mixture model (GMM-HMM) [9], which is one of the most widely used generative models. Instead, discriminative acoustic models such as support vector machines (SVMs), conditional random field, and artificial neural networks (ANNs) were applied to dysarthric speech recognition [10], [11], [12]. The works reported that discriminative acoustic models produced better results 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 dysarthric speech [13].

Dysarthric speech deviates from regular speech in various ways. Nonetheless, it can be characterized by highly consistent articulatory errors for each speaker [14]. Therefore, it would be promising to make the ASR system more suitable for an individual. To this end, speaker-adapted (SA) models, which are adjusted to a single user from a speaker-independent (SI) initial model trained on a large population with regular speech, were investigated [15], [16] using conventional adaptation methods such as maximum *a posteriori* (MAP) [17] on GMM-HMM based ASR systems. These studies reported that SA models are more appropriate for dysarthric speakers compared to speaker-dependent (SD) models, which are trained solely to the individual, and SI models, which are trained on regular speech from several regular speakers [18].

In the speaker adaptation method, choosing an appropriate initial model to be adapted directly affects the recognition performance. Sharma and Hasegawa-Johnson [19] proposed an interpolation-based technique to obtain a better initial acoustic model for adaptation. The method computes a speaker-dependent background model to represent the dysarthric talker's general speech characteristics, and the background model is interpolated with regular SI models. Then, MAP-based speaker adaptation is applied to the interpolated SI model. Kim *et al.* [20] also studied the effect of an initial model employing an SI dysarthria-adapted (DA) acoustic model which is adapted from a regular SI model using speech data from several dysarthric talkers. Finally, speaker adaptation is applied to the SI DA acoustic model. The experimental results showed that the DA initial model is better than the regular initial model in terms of word error rates, especially when using a small amount of speaker adaptation data.

Another research direction is to handle the phonetic variation of dysarthric speech in an explicit or implicit way. Explicit modeling generally creates multiple pronunciations for each word in the lexicon [46]. Mengistu and Rudzicz [16] manually made a pronunciation lexicon for each individual with dysarthria through expert assessment of the individual's pronunciation. A speaker-specific pronunciation dictionary was automatically generated using phoneme posterior probabilities of a deep neural network (DNN) trained on regular speech [21] or the state-specific vector of phone-cluster adaptive training-based acoustic models [49]. Weighted finite state transducers (WFSTs) using phonetic confusion matrices resulting from a regular ASR system were built to allow phonetic confusions during decoding process [22], [23], [32].

Implicit modeling, on the other hand, depends on the underlying acoustic models to deal with phonetic variation. Most studies have been focused on model parameter tying in which the acoustic model parameters of a target phoneme are shared with those of its alternative phonemes [46], [47]. Chandrakala and Rajeswari [50] used the log-likelihood scores of generative acoustic models as input to an SVM-based speech recognizer. Therefore, it can remove the necessity to explicitly determine and represent phonetic variation in the lexicon. Although implicit modeling is promising, it has rarely been investigated in the field of dysarthric speech recognition.

This paper addresses the problem of automatic recognition of dysarthric speech, focusing on implicit phonetic variation modeling. The contributions of the paper include the following:

- An effective application of Kullback-Leibler divergence-based HMM (KL-HMM) [24], [25] to dysarthric speech recognition for dealing with phonetic variation. KL-HMM is an emerging method as it offers a powerful and flexible framework for achieving implicit phonetic variation modeling. KL-HMM is a particular form of HMM in which the emission probability of state is parameterized by a categorical distribution of phoneme classes referred to as acoustic units. The categorical distribution is usually trained using phoneme posterior probabilities. In other words, a KL-HMM framework can be regarded as a combination of an acoustic model to obtain phoneme posterior probabilities from acoustic feature observations and a categorical distribution-based lexical model [33]. Since HMM states are generally represented as subword lexical units in the lexicon, KL-HMM can model the phonetic variation against target phonemes. KL-HMM has been successfully utilized in various speech recognition applications such as non-native speech recognition [26] and multilingual speech recognition [27]. Therefore, the KL-HMM is expected to effectively capture the phonetic variation of dysarthric speech.

- Regularized speaker adaptation of KL-HMM to make the system more speaker-specific, since dysarthric individuals generally have their own phonetic variation pattern. To this end, we reformulate the Bayesian adaptation of a categorical distribution as an L2 regularized optimization problem. The L2 regularized adaptation can reflect the speaker-specific phonetic pattern, but confusions between KL-HMM lexical models may arise because dysarthric individuals have a limited phonetic repertoire resulting from the limitation of their articulatory movement [23]. To overcome this problem, we propose lexical confusion-reducing (LCR) regularization which can enhance discriminability between lexical models in addition to the L2 regularization. It can be expected that discriminative power between lexical models increases while keeping the discerning power of phonetic variation within a lexical model. This adaptation method can also be applied to train a DA initial model to make the system better fitted to general dysarthric speech. Therefore, we believe that the proposed adaptation method can effectively represent speaker-specific phonetic variation patterns, which can help in improving recognition performance.

This paper is an extension of [48], including regularized adaptation framework, extensive experiments and analysis. The paper is organized as follows: The background including a database and the motivation for our work is described in Section II. In Section III, we present the proposed KL-HMM based ASR system in detail. Section IV shows experimental results demonstrating the effectiveness of the proposed method. Finally, our conclusions are summarized in Section V.

II. Background

A. Participants & Speech Tasks

Speech data were collected from 78 native Korean speakers of which 68 (40 males and 28 females) were dysarthric and 10 (5 males and 5 females) were regular control speakers. All dysarthric speakers had been diagnosed with cerebral palsy, which is one of the most prevalent causes of dysarthria [28], and were recruited from Seoul National Cerebral Palsy

Public Welfare. The mean ages of the dysarthric and regular participants were 36.6 ± 9.7 years and 33.1 ± 3.9 years, respectively.

All speakers uttered an average of 760 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 assessed by a speech-language pathologist, who has a top level license for speech therapy and has worked in the field over 5 years. The assessment was according to the percentage of consonants correct (PCC) [29], which is defined as the ratio of the number of correctly uttered consonants to the number of total consonants, using the APAC words [1] [30]. Based on this assessment [2], 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%)[3]. All control subjects were graded as PCC 100%.

B. Phonetic Variation

Dysarthric individuals have articulatory limitation, so they often have trouble in pronouncing certain sounds, resulting in phonetic variation which is the main cause of ASR performance degradation. To show the phonetic variation of dysarthric speech, phoneme confusion matrices[4] resulting from phoneme posterior probabilities from a conventional DNN-based ASR system trained on regular speech for a moderately dysarthric speaker and a control speaker are made in Fig. 1. The speakers and speech data are chosen from our database described in Section II-A. As can be seen, there are various phonetic confusions against target phonemes for dysarthric speech including unnatural phonetic substitutions that would not actually be made by humans, e.g., the vowel /e/ sound to the consonant /g/ sound. Also, there is a set of confusable phonemes, which are commonly shared for most target phonemes. Our observations are largely in line with the findings in [22]. For control speech, on the other hand, the confusion matrix shows a clearer pattern of correct recognition and a few confusions.

Conventional ASR systems can model the phonetic variation if the training data contain a variety of phonetic pronunciations or if the system uses clustered triphones via a decision tree. However, the amount of dysarthric training data is not generally enough to train the

[1] 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. This word set is commonly used for assessing articulation disorders in Korea [31], [32].

[2] The intra-rater and inter-rater reliability levels, based on Pearson's correlation,were measured by rechecking the speech data of 24 speakers. It was found that the intra-rater correlation was 0.96 and the inter-rater correlation was 0.90, both of which are highly reliable measures for the use of ground-truth in analyzing the results of speech recognition in terms of dysarthria severity.

[3] In this work, we focus on mildly and moderately dysarthric speakers. It was hard to recruit severely dysarthric speakers and collect speech data from them (We have a few speech data from three patients with severe dysarthria at this time). We think the amount of data is not enough to reliably evaluate the method and analyze the results. Therefore, we excluded severe dysarthria from the current research.

[4] Phonemes are represented based on Korean phonetic symbols. The corresponding International Phonetic Alphabet (IPA) symbols can be found in [44].

acoustic model and the regular training data rarely contain the phonetic variation that dysarthric individuals produce. In the KL-HMM framework, on the other hand, the acoustic model and the lexical model can be trained on an independent set of resources [33]. Specifically, the acoustic model can be trained on data from resource-rich domains such as regular speech whereas the lexical model can be trained on a relatively small amount of resources from a target domain such as dysarthric-specific or speaker-specific domain [48]. Therefore, KL-HMM is expected to effectively deal with the phonetic variation of dysarthric speech.

C. Probabilistic Lexical Modeling Framework

KL-HMM is based on a probabilistic lexical modeling framework and therefore we briefly review the probabilistic lexical modeling framework in this section. In the framework of probabilistic lexical modeling [33], the relationship between the acoustic feature observation $\mathbf{x}_t$ and the HMM state q_t that represents lexical unit l^i , i.e., $q_t \in \mathbb{L} = \{l^1, \ldots, l^i, \ldots, l^L\}$ where L denotes the number of lexical units, is factored using a latent variable a^d which is referred as acoustic unit given by

$$p\left(\mathbf{x}_t|q_t=l^i\right)=\sum_{d=1}^{D}p\left(\mathbf{x}_t|a^d\right)P\left(a^d|q_t=l^i\right) \qquad (1)$$

where $p(\mathbf{x}_t|a^d)$ represents the acoustic unit likelihood at frame t, $P(a^d|q_t=l^i)$ represents the probability of the acoustic unit given the lexical unit, and D is the number of acoustic units. In other words, $p(\mathbf{x}_t|a^d)$ is obtained by the acoustic model which builds the relationship between the acoustic feature observation $\mathbf{x}_t$ and the acoustic unit a^d. Also, $P(a^d|q_t=l^i)$ is obtained by the lexical model which models the relationship between all acoustic units and the lexical unit l^i. In this work, both the acoustic units and lexical units are chosen by clustered context-dependent phonemes (i.e., senones or clustered triphone states) to better represent phonetic patterns. The lexical model can be either deterministic or probabilistic. In the deterministic lexical model, each lexical unit l^i is deterministically mapped to an acoustic unit a^j as follows: $P(a^d \mid q_t=l^i) = 1$ if $d = j$, otherwise 0. In standard HMM-based ASR systems such as GMM-HMM and ANN-HMM, the deterministic lexical model is generally adopted. In this probabilistic lexical model, on the other hand, each lexical unit is probabilistically related to all acoustic units so that $0 \le P(a^d|q_t=l^i) \le 1$ and $\sum_{d=1}^{D} P\left(a^d|l^i\right)=1$. Therefore, the probabilistic lexical model can effectively capture the pronunciation variation of dysarthric speech. To this end, probabilistic classification of HMM states [34], tied posteriors [35], and KL-HMM approaches were proposed. Recently, KL-HMM has been successfully applied in achieving probabilistic lexical modeling [36]. Therefore, we adopt KL-HMM based probabilistic lexical modeling in this work.

III. Proposed Method

In this section, we explain the proposed KL-HMM approach including the DNN acoustic model, the categorical distribution-based lexical model, and speaker adaptation with regularization. The schematic diagram of the KL-HMM framework is depicted in Fig. 2.

A. DNN Acoustic Model

A deep neural network (DNN) is an ANN with multiple hidden layers of units between the input and output layers [37]. Recently, a DNN has been received great attention since the complex structure of speech sounds can be modeled through multiple layers using powerful optimization techniques, and therefore it has been successfully applied in speech recognition as an acoustic model [37]. It is expected that the DNN acoustic model also captures the complex acoustic structures of dysarthric speech as well. Therefore, the DNN-based acoustic model was adopted in this work. We used 40 log mel-filterbank energies with 11 context window $\mathbf{x}_t=\{\mathbf{x}_{t-5},\dots,\mathbf{x}_t,\dots,\mathbf{x}_{t+5}\}$ as acoustic feature observations and senones as output units or acoustic units. Given the DNN acoustic model, the probabilities of acoustic units, i.e., D-dimensional acoustic unit posterior probability vectors can be obtained as

$$\mathbf{z}_t=\left[z_t^1,\dots,z_t^d,\dots,z_t^D\right]^T=\left[P\left(a^1|\mathbf{x}_t\right),\dots,P\left(a^d|\mathbf{x}_t\right),\dots,P\left(a^D|\mathbf{x}_t\right)\right]^T \tag{2}$$

Then, the probabilistic lexical model is trained using the acoustic unit probability vectors. In other words, the acoustic posterior probabilities are used as feature observations to train KL-HMM whose states correspond to the lexical units.

B. KL-HMM Probabilistic Lexical Model

KL-HMM is a type of HMM where the emission probability of lexical state l^i is parametrized by a categorical distribution $\mathbf{y}_i=\left[y_i^1,\dots,y_i^d,\dots,y_i^D\right]^T$, where $y_i^d=P(a^d|l^i)$. A categorical distribution is a multinomial distribution from which only one sample is drawn. Therefore, each state captures a probabilistic relationship between a lexical unit l^i and D acoustic units.

In the KL-HMM framework, the acoustic unit likelihood in (1) is replaced with the acoustic unit posterior probability in (2). Therefore, the local score at each HMM state can be computed using the KL divergence between the acoustic unit posterior feature $\mathbf{z}_t$ and the categorical variable $\mathbf{y}_i$ as

$$d(\mathbf{z}_t,\mathbf{y}_i)=\sum_{d=1}^{D}z_t^d\log\left(\frac{z_t^d}{y_i^d}\right) \tag{3}$$

Recent studies reported that asymmetric KL divergence as in (3) is more robust than other symmetric variants of the KL divergence [27]. Therefore, we used the asymmetric 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 denotes the number of frames, the categorical variables $Y=[\mathbf{y}_1,\ldots,\mathbf{y}_i,\ldots,\mathbf{y}_L]$ can be trained by minimizing the cost function summing the local scores over time t and state l^i as

$$\min \sum_{t=1}^{T}\sum_{i=1}^{L} d(\mathbf{z}_t,\mathbf{y}_i)\delta_t^i \quad s.t. \sum_{d=1}^{D} y_i^d=1 \qquad (4)$$

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 (4), 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 $y_i^d=\sum_{t\in l^i} z_t^d/T_i$, where T_i stands for the number of frames associated with state l^i.

C. KL-HMM Adaptation

In general, dysarthric individuals have their own phonetic variation patterns. Although the KL-HMM is an appropriate framework in modeling such phonetic variations, it still has limitations for each speaker. Therefore, speaker adaptation that modifies speaker-independent (SI) model parameters for a single speaker to make it more speaker-specific is promising to improve ASR performance. One of the widely used speaker adaptation techniques is using the Bayesian estimation since it can properly cope with the overfitting problem resulting from a small amount of speaker-specific adaptation data by using the Bayesian priors. With the Bayesian priors, the categorical distribution can be modeled by the Dirichlet-categorical conjugate distributions [38], [39]. The resulting speaker-adapted (SA) categorical distribution of KL-HMM through the Bayesian estimation [42] is given by

$$\mathbf{y}_i^{\mathrm{SA}}=\eta_i\mathbf{y}_i^{\mathrm{SD}}+(1-\eta_i)\mathbf{y}_i^{\mathrm{SI}} \qquad (5)$$

where $\mathbf{y}_i^{\mathrm{SA}}$ and $\mathbf{y}_i^{\mathrm{SI}}$ stand for the categorical state distributions estimated given a small amount of speaker-dependent (SD) adaptation data and a large amount of speaker-independent (SI) training data, respectively. Also, η_i is the adaptation coefficient in the range of [0, 1] and it is used to determine the balance between the SD and SI model parameters.

1) Adaptation with L2 regularization—As can be seen in (5), it is noticeable that the adapted model parameter is computed by the form of interpolation and the interpolation can be treated as the following L2 regularized optimization problem [41]:

$$\min_{\boldsymbol{\varphi}_i} \|\boldsymbol{\varphi}_i - \left(\mathbf{y}_i^{\mathrm{SD}} - \mathbf{y}_i^{\mathrm{SI}}\right)\|_2^2 + \lambda_i \|\boldsymbol{\varphi}_i\|_2^2 \quad (6)$$

where $\boldsymbol{\varphi}_i$ denotes a vector moving from an initial SI model and $\|\cdot\|_2$ is the L2-norm. Also, $\lambda_i \geq 0$ is a regularization parameter. The solution can be obtained by taking partial derivative with respect to $\boldsymbol{\varphi}_i$ and set it to zero as follows:

$$\boldsymbol{\varphi}_i^{L2} = \frac{\mathbf{y}_i^{\mathrm{SD}} - \mathbf{y}_i^{\mathrm{SI}}}{1 + \lambda_i} \quad (7)$$

Finally, the SA model parameter is given by

$$\mathbf{y}_i^{\mathrm{SA}-L2} = \boldsymbol{\varphi}_i^{L2} + \mathbf{y}_i^{\mathrm{SI}} \quad (8)$$

Here, when λ_i is set to $(1 - \eta_i) / \eta_i$, the solution of the equation (8) equals the solution of the equation (5).

2) Lexical confusion-reducing regularization—Adaptation with L2 regularization can properly represent the phonetic variations of each speaker. However, dysarthric individuals generally have a limited repertoire of pronunciation resulting from their articulatory limitation. That is, there may be commonly shared phonetic variations across target phonemes shown in Section II-B, and therefore, confusions between the categorical distributions of KL-HMM lexical models may also be induced. To reduce the lexical confusions, the lexical confusion-reducing (LCR) regularization is additionally taken into account as second regularization as

$$\min_{\boldsymbol{\varphi}_i} \|\boldsymbol{\varphi}_i - \left(\mathbf{y}_i^{\mathrm{SD}} - \mathbf{y}_i^{\mathrm{SI}}\right)\|_2^2 + \lambda_{i,1} \underbrace{\|\boldsymbol{\varphi}_i\|_2^2}_{\text{L2regul.}} + \lambda_{i,2} \underbrace{\|\boldsymbol{\varphi}_i - \left(\mathbf{y}_i^{\mathrm{SD}} - \overline{\mathbf{y}}\right)\|_2^2}_{\text{LCR regul.}} \quad (9)$$

where $\overline{\mathbf{y}} = \sum_{i=1}^{L} \left(\mathbf{y}_i^{\mathrm{SD}} + \mathbf{y}_i^{\mathrm{SI}}\right) / 2L$ means the common phonetic variations across all lexical models. Here, we consider the SI lexical models as well as the SD lexical models in order to reflect underlying pronunciation patterns induced from a large population for generality. Looking at the LCR regularization, the common phonetic variations are removed from the SD parameters on each lexical state. That is, lexical state-specific characteristics can be captured through the LCR regularization. The solution can also be obtained by taking partial derivative with respect to $\boldsymbol{\varphi}_i$ and set it to zero given by

$$\varphi_i^{\mathrm{LCR}} = \frac{\left(\mathbf{y}_i^{\mathrm{SD}} - \mathbf{y}_i^{\mathrm{SI}}\right) + \lambda_{i,2}\left(\mathbf{y}_i^{\mathrm{SD}} - \overline{\mathbf{y}}\right)}{1 + \lambda_{i,1} + \lambda_{i,2}} \quad (10)$$

where $\lambda_{i,1} \geq 0$ and $\lambda_{i,2} \geq 0$ are regularization parameters for the L2 and LCR regularization terms, respectively. Finally, the SA model parameter can be obtained as

$$\mathbf{y}_i^{\mathrm{SA-LCR}} = \varphi_i^{\mathrm{LCR}} + \mathbf{y}_i^{\mathrm{SI}} = \frac{\mathbf{y}_i^{\mathrm{SD}} + \lambda_{i,1}\mathbf{y}_i^{\mathrm{SI}} + \lambda_{i,2}\left(\mathbf{y}_i^{\mathrm{SD}} + \mathbf{y}_i^{\mathrm{SI}} - \overline{\mathbf{y}}\right)}{1 + \lambda_{i,1} + \lambda_{i,2}} \quad (11)$$

As can be seen in (11), the adapted parameter is represented in the interpolated form of the SD, SI, and lexical confusion-reduced parameters. The $\lambda_{i,1}$ and $\lambda_{i,2}$ are used to control the strength of the SI parameter and the confusion-reduced parameter, respectively. Consequently, it is expected that the LCR regularization may result in enhanced discriminability between categorical distributions of KL-HMM states while preserving speaker-specific information. Therefore, we believe that the ASR performance can be improved by considering the LCR regularization.

In practice, the adaptation performance is affected by an SI initial model. To obtain a better SI initial model, we applied the adaptation technique in constructing an SI dysarthria-adapted (DA) model on speech data from several dysarthric talkers (dysarthric domain adaptation). Through this process, it is expected that the general pronunciation variability of dysarthric speech can be captured and it may be more effective than an SI initial model trained on only regular speech. Finally, decoding is performed using the standard Viterbi decoder where the log-likelihood based score is replaced with the KL divergence-based local score in (3).

IV. Experimental Results

A. Experimental Setup

Our data distribution is summarized in Table I. The SI non-dysarthric regular training set includes 300k utterances of 8k Korean isolated words from several databases (DBs) uttered by regular speakers, 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 SI dysarthria adaptation (dysarthric domain adaptation) set was used to construct an SI DA initial model which is better fitted to general dysarthric speech. It includes 20k utterances from 48 dysarthric speakers including 25 mild and 23 moderate subjects described in Section II-A. 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 uttered 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 uttered 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 were obtained in multiple sessions. For speaker adaptation, the command words and additional command words collected from another session from test speakers were used. 100, 200, 500, and 926 adaptation utterances[5] were used for dysarthric speakers whereas 100 and 313 adaptation utterances were used for control speakers. The speakers in the evaluation set were totally separated from the SI dysarthria adaptation set.

We compared several ASR systems to evaluate the proposed method as follows: GMM-HMM, DNN-HMM, and KL-HMM.

GMM-HMM system—We first trained a regular GMM-HMM system (referred to as GMM_{reg}-HMM) using 39 dimensional mel-frequency cepstral coefficients (MFCCs), 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_{reg}-HMM consists of 1480 tied-state (senone) left-to-right triphone HMMs, where each HMM has 3 states and each state is modeled with 7 Gaussian components on average and is trained on the SI regular training set. The dysarthric GMM can be obtained by adapting the GMM_{reg} to dysarthric domain using MAP adaptation on the SI dysarthria adaptation set (referred to as GMM_{reg}-MAP_{dys}-HMM).

DNN-HMM system—A regular SI DNN was trained using 40 dimensional log mel-filterbank energy features with a context window of 11 frames and frame alignment information resulting from the standard GMM_{reg}-HMM system. The DNN has 5 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_{reg}-HMM system. The parameter was initialized using layer-by-layer generative pre-training and the network was discriminatively trained using backpropagation on the regular training set [42] (referred to as DNN_{reg}-HMM). To further construct SI DA DNN, two kinds of DNN adaptation methods were considered using the SI dysarthria adaptation set: one is L2 regularization [45] (referred to as DNN_{reg}-L2_{dys}-HMM), which adds the L2-norm of all the model parameter difference between the initial model and adapted model to the frame-cross entropy criterion, and the other is linear output network (LON) adaptation [43] (referred to as DNN_{reg}-LON_{dys}-HMM), which adds one extra layer on top of the DNN_{reg} (i.e., 1480 × 1480 hidden units) and perform backpropagation training on this extra layer using the SI dysarthria adaptation set for fair comparison with KL-HMM. In addition, DNN with multicondition training ($\text{DNN}_{\text{multi}}$-HMM), which is trained on both SI regular training and SI dysarthria adaptation sets, was compared. For speaker adaptation, the LON adaptation was exploited on the speaker adaptation set (LON_{spk}).

KL-HMM system—A baseline SI KL-HMM was trained using DNN_{reg} posterior probability vectors obtained from the SI regular training set (DNN_{reg}-KL_{reg}-HMM) or the SI

[5] The small number of speaker adaptation data is reasonable for patients with dysarthria (under 200 utterances) for the practical use. However, the evaluation on the large number of adaptation data may be needed for algorithm testing in case of long-term data acquisition situation. Therefore, we evaluated our proposed method on a variety number of speaker adaptation data.

dysarthria adaptation set (DNN$_{reg}$-KL$_{dys}$-HMM). The SI DA KL-HMM was adapted using the L2 regularization in Section III-C1 (DNN$_{reg}$-KL$_{reg}$-L2$_{dys}$-HMM) or the LCR regularization in Section III-C2 (DNN$_{reg}$-KL$_{reg}$-LCR$_{dys}$-HMM) on the SI dysarthria adaptation set to model the general pronunciation variability of dysarthric speech. Note that the LCR regularization was used together with the L2 regularization as in (9). In addition, we considered SI DA acoustic models such as DNN$_{reg}$-L2$_{dys}$ and DNN$_{reg}$-LON$_{dys}$ in obtaining posterior probability vectors. Also, speaker adaptation was performed using the L2 regularization or the LCR regularization given the trained SI KL-HMM (referred to as L2$_{spk}$ or LCR$_{spk}$, respectively). The regularization parameters $\lambda_{i,1}$ and $\lambda_{i,2}$ were set to fixed constant values for all lexical states and were chosen empirically from the experiments, which mostly ranges between 0.05 and 0.5 for λ_1 and ranges between 0.005 and 0.05 for λ_2. Hereafter, we omit state index i for concise description.

B. Effectiveness of the Proposed SI KL-HMM

We first performed speech recognition experiments on speaker-independent (SI) systems. Table II shows the performances of SI 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 the GMM$_{reg}$-HMM and DNN$_{reg}$-HMM systems, the performance of the DNN$_{reg}$-HMM was better than with the GMM$_{reg}$-HMM for both dysarthric and control speakers. It was also observed that the SI DA systems adapted on SI dysarthria adaptation data such as DNN$_{reg}$-LON$_{dys}$-HMM produce better results than with systems trained on only regular data (DNN$_{reg}$-HMM) and multicondition data (DNN$_{multi}$-HMM) in terms of the unweighted average WER. Specifically, the SI DA system much improved the recognition performance for dysarthric speakers whereas the performance was somewhat degraded for control speakers when compared with the regular system and multicondition system. Since there is a trade-off between dysarthric and control speakers, reducing the gap is important in developing a universally accessible ASR system.

For the KL-HMM approach, DNN$_{reg}$-KL$_{dys}$-HMM outperformed DNN$_{reg}$-L2$_{dys}$-HMM and DNN$_{reg}$-LON$_{dys}$-HMM for both dysarthric and control speakers. The regularized KL-HMM produced a lower WER than with DNN$_{reg}$-KL$_{dys}$-HMM. Specifically, when using LCR regularization (DNN$_{reg}$-KL$_{reg}$-LCR$_{dys}$-HMM), the performance can be improved by obtaining a WER of 31.9% for dysarthric speakers and 0.6% for control speakers which is comparable to DNN$_{reg}$-HMM.

When DNN$_{reg}$-L2$_{dys}$ was used as an acoustic model in the KL-HMM systems, the WERs of dysarthric individuals were further reduced compared with the DNN$_{reg}$ based KL-HMM systems. When we further applied LCR$_{dys}$ regularization (DNN$_{reg}$-L2$_{dys}$-KL$_{reg}$-LCR$_{dys}$-HMM), we were able to achieve the best performance for dysarthric speakers, providing relative improvements of 23.9% over DNN$_{reg}$-L2$_{dys}$-HMM, 12.2% over DNN$_{reg}$-KL$_{reg}$-LCR$_{dys}$-HMM, and 3.1% over DNN$_{reg}$-L2$_{dys}$-KL$_{reg}$-L2$_{dys}$-HMM. In addition, DNN$_{reg}$-L2$_{dys}$-KL$_{reg}$-LCR$_{dys}$-HMM attained better performance than the DNN$_{reg}$-L2$_{dys}$-KL$_{reg}$-L2$_{dys}$-HMM for control speakers as well, obtaining a WER of 1.0% which is somewhat comparable to the performances of other regular systems. DNN_{reg}-LON_{dys}-KL_{reg}-LCR_{dys}-HMM was slightly worse than with DNN_{reg}-L2_{dys} based KL-HMM systems. Therefore, we chose DNN_{reg}-L2_{dys} based KL-HMM systems as a baseline SI model in the following speaker adaptation experiments. Through these experiments, we found that the KL-HMM framework is effective for dysarthric speakers while keeping comparable performance for control speakers. Also, a good acoustic model which is better fitted to general dysarthric speech (DNN_{reg}-L2_{dys}) is more appropriate in modeling KL-HMM for dysarthric speech. Finally, dysarthric domain adaptation with LCR regularization (LCR_{dys}) helped in improving the recognition performance.

C. Effectiveness of Speaker Adaptation

Table III compares the WERs of speaker adaptation to the DNN-HMM and KL-HMM by varying the number of adaptation utterances for dysarthric individuals. For baseline DNN-HMM, LON speaker adaptation (LON_{spk}) was used on top of DNN_{reg}-L2_{dys}-HMM for fair comparison with the KL-HMM approach. Note that the resulting system is adapted to a particular speaker on the number of adaptation utterances using the SA method given the SI ASR system. The performances of the proposed KL-HMM systems were better than with the DNN-HMM system regardless of the amount of adaptation utterances. In DNN_{reg}-L2_{dys}-KL_{reg}-L2_{dys}-HMM, speaker adaptation with LCR regularization (LCR_{spk}) produced a better performance than with L2 regularization (L2_{spk}). When we replaced the initial SI model by DNN_{reg}-L2_{dys}-KL_{reg}-LCR_{dys}-HMM, the recognition performance of LCR_{spk} was further improved. This indicates that choosing a proper initial model is important in representing better speaker-specific phonetic characteristics. In addition, LCR regularized speaker adaptation is helpful in attaining better performances as in SI DA experiments shown in previous section. Finally, when the DNN acoustic model and the KL-HMM lexical model were simultaneously adapted for each speaker (LON_{spk}+LCR_{spk}), we were able to obtain the lowest WER for all adaptation conditions, achieving 28.9%, 31.8%, 38.0%, and 43.6%relative improvements compared with DNN_{reg}-L2_{dys}-LON_{spk},and 11.3%, 14.8%, 20.0%, and 23.5% relative improvements compared with DNN_{reg}-L2_{dys}-KL_{reg}-LCR_{dys}-LCR_{spk} in the WER reduction when using 100, 200, 500, and 926 adaptation utterances, respectively. This result suggests that speaker-specific acoustic and phonetic variation characteristics can be successfully modeled in the KL-HMM framework.

To examine the effectiveness of the proposed method for the test speaker's severity levels, WERs of the SA DNN-HMM and KL-HMM systems with various number of adaptation data depending on their severity levels are indicated in Fig. 3. Note that L2 regularized speaker adaptation (L2_{spk}) is based on DNN_{reg}-L2_{dys}-KL_{reg}-L2_{dys}-HMM while LCR regularized speaker adaptation (LCR_{spk}) is based on DNN_{reg}-L2_{dys}-KL_{reg}-LCR_{dys}-HMM. As can be seen, we observed similar trends for both mildly and moderately dysarthric individuals. For mild speakers in Fig. 3(a), LCR_{spk} produced 24.8% and 6.5% relative WER reductions compared with LON_{spk} and L2_{spk} on average across four adaptation conditions, respectively. When we used both LON_{spk} and LCR_{spk}, a 29.4% relative improvement was obtained on average when compared with LCR_{spk}. For moderate speakers in Fig. 3(b), when LCR_{spk} was applied, we obtained 21.2% and 3.5% relative improvements on average in WER reduction across all adaptation conditions compared to LON_{spk} and L2_{spk} as well, respectively. Also, applying both LON$_{spk}$ and LCR$_{spk}$ provided a 13.4% relative WER reduction on average compared to LCR$_{spk}$.

Fig. 4 represents how the recognition performances were affected by the regularization parameters of the LCR regularized SA KL-HMM, the WERs with varying λ_1 and λ_2 using 926 adaptation utterances. In this experiment, we used DNN$_{reg}$-L2$_{dys}$-KL$_{reg}$-L2$_{dys}$ as an SI initial model to check the effectiveness of the LCR regularization during speaker adaptation process. Note that λ_2=0 corresponds to the L2 regularized SA KL-HMM. As can be seen, when λ_1 was small, it produced lower WERs. In practice, the optimal λ_1 was inversely proportionally related to the amount of adaptation data. That is, as the amount of adaptation data was small, a large value of λ_1 was optimal. It was also observed that considering the LCR regularization parameter λ_2 yields better recognition performances for most cases. However, when λ_2 was relatively larger than λ_1, the performance was worsened. This trend was also observed in other adaptation conditions. Therefore, it is important that λ_1 and λ_2 are set to suitable values. From these observations, we can suggest that λ_2 should be set to a smaller value or a comparable value over λ_1.

We also performed the same experiments for control speakers to examine the universality of the proposed method and the results are summarized in Table IV. As can be seen, the proposed LCR regularization on the KL-HMM outperformed the SA DNN and L2 regularized KL-HMM as in the case of dysarthric speakers. From these results, it ensures that our proposed ASR method is a good framework for dysarthric as well as non-dysarthric speakers.

D. Analysis of the Proposed KL-HMM System

Next, we analyzed the KL-HMM parameters to find out the main cause of the performance improvement. We first checked how the phonetic variations are modeled in the KL-HMM framework. Fig. 5 shows the average percentage of the counted number of categorical variables whose values are above the threshold in KL-HMM states for control, mildly dysarthric, and moderately dysarthric speaker groups. Note that the total number of categorical variables in each state is 1480. That is, it can be interpreted that the dominant phonetic variations increase as the number of counted parameters increases. Here, L2$_{spk}$, LCR$_{spk}$, and LON$_{spk}$+LCR$_{spk}$ regularized KL-HMM systems in Section IV-C were compared for the analysis. For dysarthric speakers, 926 adaptation data were used while 313 adaptation utterances were used for control speakers. As can be seen, the phonetic variation is properly represented in the categorical distribution of KL-HMM in a probabilistic way which is rarely handled in the DNN-HMM system. In Fig. 5(a), as the threshold goes down, the counted number increased. Also, as the dysarthria severity of speakers gets worse, more confusions were observed. For speaker adaptation with LCR regularization in Fig. 5(b), the counted number was reduced across all speaker groups compared with the L2 regularization. This means that dominant phonetic confusions in each lexical state are reduced by using LCR regularization. That is, phonetic confusions are concentrated on a much smaller number of senones. Moreover, when we considered both LON$_{spk}$ and LCR$_{spk}$ in Fig. 5(c), phonetic confusions were further reduced. It was observed that the percentage of phonetic variations for mild speakers is similar to that for control speakers but still moderate speakers have many phonetic variations. Reducing the number of dominant KL-HMM parameters is very important in making the KL-HMM system more stable.

Table V shows the symmetric KL divergence between the categorical distributions of SA KL-HMM states on the same systems in Fig. 5. The KL divergence was computed over all pairs of categorical state distributions and then averaged to measure the average confusability between lexical models. That is, as the KL divergence value is larger, the discriminative power between lexical models is increased. As shown in Table V, the LCR regularized speaker adaptation produced larger KL divergence values than L2 regularized speaker adaptation for all speaker groups consisting of control, mildly dysarthric, and moderately dysarthric speakers. When we adapted DNN and KL-HMM simultaneously for each speaker (LON_{spk}+LCR_{spk}), we obtained the largest KL divergence for all speaker groups. Also, it was observed that as dysarthria becomes severe, smaller KL divergence values are obtained. This result indicates that the performance gain comes from the reduction of confusability between lexical models. Interestingly, mildly dysarthric speakers on LON_{spk}+LCR_{spk} gave a larger KL divergence than control speakers on LCR_{spk} although the actual recognition rate of dysarthric speakers was lower than with control speakers. Through acoustic model adaptation, the confusions between lexical models can be much reduced, but some limitations still remain. One of the factors which give rise to performance degradation is the consistency of speech [5], [20]. There are diverse intra-speaker variations, so it limits the performance improvement although the acoustic and lexical models are well trained through speaker adaptation. We found that the causes which lead to the wide intra-speaker variation are articulatory errors as well as involuntary breathing, stuttering, and accidental pauses between syllables. Therefore, future works include the investigation to deal with the problems.

V. Conclusion

In this paper, a novel and effective method to automatically recognize dysarthric speech was proposed. The method relies on two important parts: 1) To address the 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. 2) To make the system more speaker-specific, the LCR regularized KL-HMM speaker adaptation method was proposed. A series of experiments were performed (measured in WER) on both 20 dysarthric and 10 control speakers to evaluate the effectiveness of the proposed method. The experimental results show the effectiveness in the aspects of 1) the performance through comparison with other ASR systems, achieving significant improvements regardless of the amount of adaptation data for dysarthric speakers, 2) the stability of the KL-HMM system, reducing the number of dominant KL-HMM parameters, 3) the universality of the proposed approach, showing better ASR results for regular speakers as well. Thus, our framework presents a possibility in helping people who suffer from dysarthria to communicate with spoken expression.

Acknowledgments

This work was supported by the National Research Foundation of Korea (No. 2014r1a2a2a01007650) and the National Institutes of Health (No. r03dc013990).

References

1. Duffy, JR. Motor speech disorders: Substrates, differential diagnosis, and management. St Louis, MO: Elsevier Mosby; 2005.
2. Kim H, Marting K, Hasegawa-Johnson M, Perlman A. Frequency of consonant articulation errors in dysarthric speech. Clinical Linguist Phonet. Oct; 2010 24(10):759–770.
3. Hosom JP, Kain AB, Mishra T, van Santen JPH, Fried-Oken M, Staehely J. Intelligibility of modifications to dysarthric speech. Proc IEEE Int Conf Acoust Speech, and Signal Process. Apr. 2003 :924–927.
4. Rudzicz F. Learning mixed acoustic/articulatory models for disabled speech. Proc Workshop on Machine Learning for Assist Technol on Neural Inf Process Syst. Dec.2010 :70–78.
5. Young V, Mihailidis A. Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: a literature review. Assist Technol. 2010; 22(2):99–112. [PubMed: 20698428]
6. Coleman CL, Meyers LS. Computer recognition of the speech of adults with cerebral palsy and dysarthria. Augment Altern Commun. 1991; 7(1):34–42.
7. Polur PD, Miller GE. Experiments with fast fourier transform, linear predictive and cepstral coefficients in dysarthric speech recognition algorithms using hidden Markov model. IEEE Trans Neural Sys Rehabil Eng. Dec; 2005 13(4):558–561.
8. Sharma HV, Hasegawa-Johnson M. State-transition interpolation and MAP adaptation for HMM-based dysarthric speech recognition. Proc NAACL HLT 2010 Workshop on Speech and Lang Process Assistive Tech. Jon;2010 :72–29.
9. Huang, X., Acero, A., Hon, HW. Spoken Language Processing. Englewood Cliffs, JN: Prentice-Hall; 2001.
10. Hasegawa-Johnson M, Gunderson J, Perlman A, Huang T. HMM-based and SVM-based recognition of the speech of talkers with spastic dysarthric. Proc IEEE Int Conf Acoust Speech, and Signal Process. 2006:1060–1063.
11. Rudzicz F. Articulatory knowledge in the recognition of dysarthric speech. IEEE Trans Audio, Speech, Lang Process. May; 2011 19(4):947–960.
12. Jayaram G, Abdelhamied K. Experiments in dysarthric speech recognition using artificial neural networks. J Rehabil Res Develop. May; 1995 32(2):162–169.
13. Polur PD, Miller GE. Investigation of an HMM/ANN hybrid structure in pattern recognition application using cepstral analysis of dysarthric (distorted) speech signals. Med Eng Phys. 2006; 28(8):741–748. [PubMed: 16359906]
14. Yorkston, KM., Beukelman, DR., Bell, KR. Clinical management of dysarthric speakers. San Diego, CA: College-Hill Press; 1988.
15. Christensen, H., Cunningham, S., Fax, C., Green, P., Hain, T. Proc Interspeech. Portland, Oregon: Sep. 2012 A comparative study of adaptive, automatic recognition of disordered speech.
16. Mengistu KT, Rudzicz F. Adapting acoustic and lexical models to dysarthric speech. Proc IEEE Int Conf Acoust Speech, and Signal Process. 2011:4924–4927.
17. Gauvain JL, Lee CH. Maximum a posteriori estimation of multivariate Gaussian mixture observations of Markov chains. IEEE Trans Speech and Audio Process. Apr; 1994 2(2):291–298.
18. Rudzicz, F. Proc ASSETS. Tempe, Arizona, USA: Oct. 2007 Comparing speaker-dependent and speaker-adaptive acoustic models for recognizing dysarthric speech; p. 255-256.
19. Sharma HV, Hasegawa-Johnson M. Acoustic model adaptation using in-domain background models for dysarthric speech recognition. Comput Speech Lang. Sep; 2013 27(6):1147–1162.
20. Kim, MJ., Yoo, J., Kim, H. Proc Interspeech. Lyon, France: Aug. 2013 Dysarthric speech recognition using dysarthria-severity-dependent and speaker-adaptive models; p. 3622-3626.
21. Christensen H, Green P, Hain T. Learning speaker-specific pronunciations of disordered speech. Proc Interspeech. Aug.2013
22. Morales SOC, Cox SJ. Modelling errors in automatic speech recognition for dysarthric speakers. EURASIP J Adv Signal Process. 2009; 2009(1) Article ID 308340.

23. Seong, WK., Park, JH., Kim, HK. Proc 13th Int Conf Comput Helping People Special Needs. Linz, Austria: 2012. Dysarthric speech recognition error correction using weighted finite state transducers based on context-dependent pronunciation variation; p. 475-482.

24. Aradilla G, Bourlard H, Magimai-Doss M. Using KL-based acoustic models in a large vocabulary recognition task. Proc Interspeech. 2008

25. Aradilla G, Vepa J, Bourlard H. An acoustic model based on Kullback-Leibler divergence for posterior features. Proc IEEE Int Conf Acoust Speech, and Signal Process. 2007:IV-657–IV-660.

26. Razavi, M., Doss, MM. Proc Interspeech. Singapore: Sep. 2014 On recognition of non-native speech using probabilistic lexical model.

27. Imseng D, Motlicek P, Bourlard H, Garner PN. Using out-of-language data to improve an under-resourced speech recognizer. Speech Commun. Jan.2014 56:142–151.

28. Maassen, B., Kent, R., Peters, H., Lieshout, PV., Hulstijn, W. Speech motor control in normal and disordered speech. Vol. chap. 12. Oxford University Press; 2004.

29. Shriberg LD, Kwiatrkowski J. Phonological disorders III: A procedure for assessing severity of involvement. J Speech and Hearing Disorders. 1982; 47(3):256–270.

30. Kim MJ, Pae S, Park C. Assessment of phonology and articulation for children. Human Brain Research & Consulting. 2007

31. Lee Y, Sung JE, Sim H. Effects of listeners' working memory and noise on speech intelligibility in dysarthria. Clinical Linguist Phonet. Oct; 2014 28(10):785–795.

32. Kim MJ, Kim Y, Kim H. Automatic intelligibility assessment of dysarthric speech using phonologically-structured sparse linear model. IEEE/ACM Trans Audio, Speech, Lang Process. Apr; 2015 23(4):694–704.

33. Rasipuram R, Magimai-Doss M. Articulatory feature based continuous speech recognition using probabilistic lexical modeling. Comput Speech Lang. Mar.2016 36:233–259.

34. Luo X, Jelinek F. Probabilistic classification of HMM states for large vocabulary continuous speech recognition. Proc IEEE Int Conf Acoust Speech, and Signal Process. 1999:353–356.

35. Rottland J, Rigoll G. Tied posteriors: an approach for effective introduction of context dependency in hybrid NN/HMM LVCSR. Proc IEEE Int Conf Acoust Speech, and Signal Process. 2000:1241–1244.

36. Rasipuram R, Magimai-Doss M. Acoustic and lexical resource constrained ASR using language-independent acoustic model and language-dependent probabilistic lexical model. Speech Commun. 2015; 68:23–40.

37. Hinton G, Deng L, Yu D, Dahl GE, Mohamed A, Jaitly N, Senior A, Vanhoucke V, Ngugen P, Sainath TN, Kingsbury B. Deep neural networks for acoustic modeling in speech recognition. IEEE Signal Process Mag. Nov; 2012 29(6):82–97.

38. Murphy, KP. Binomial and multinomial distributions. 2006. [Online]. Available:http://www.cs.ubc.ca/~murphyk/Teaching/CS340-Fall06/reading/bernoulli.pdf

39. Suh Y, Kim H. Probabilistic class histogram equalization based on posterior mean estimation for robust speech recognition. IEEE Signal Process Lett. Dec; 2015 22(12):2421–2424.

40. Imseng D, Bourlard H. Speaker adaptive Kullback-Leibler divergence based hidden Markov models. Proc IEEE Int Conf Acoust Speech, and Signal Process. May.2013 :7913–7917.

41. Kim, Y., Kim, H. Proc IEEE Int Conf Acoust Speech, and Signal Process. Florence: May. 2014 Constrained MLE-based speaker adaptation with L1 regularization; p. 6369-6373.

42. Hinton G, Osindero S, Teh YW. A fast learning algorithm for deep belief nets. Neural Comput. 2006; 18:1527–1554. [PubMed: 16764513]

43. Yu, D., Deng, L. Automatic speech recognition: A deep learning approach. Springer-Verlag; London: 2015.

44. Kim JM. Computer codes for Korean sounds. J Acoust Soc Korea. Dec; 2001 20(4):3–16.

45. Chen, G., Xu, H., Wu, M., Povey, D., Khudanpur, S. Proc Interspeech. Dresden, Germany: Sep. 2015 Pronunciation and silence probability modeling for asr; p. 533-537.

46. Hain T. Implicit modelling of pronunciation variation in automatic speech recognition. Speech Commun. 2005; 46(2):171–188.

47. Saraclar M, Nock H, Khudanpur S. Pronunciation modeling by sharing Gaussian densities across phonetic models. Comput Speech Lang. 2000; 14(2):137–160.

48. Kim, M., Wang, J., Kim, H. Proc Interspeech. San Francisco, CA: Sep. 2016 Dysarthric speech recognition using kullback-leibler divergence-based hidden Markov model; p. 2671-2675.

49. Sriranjani R, Umesh S, Reddy MR. Pronunciation adaptation for disordered speech recognition using state-specific vectors of phone-cluster adaptive training. Proc ISCA/ACL 6th Workshop on Speech and Lang Process Assist Technol. Sep.2015

50. Chandrakala S, Rajeswari N. Representation learning based speech assistive system for persons with dysarthria. IEEE Trans Neural Syst Rehabil Eng. in press.

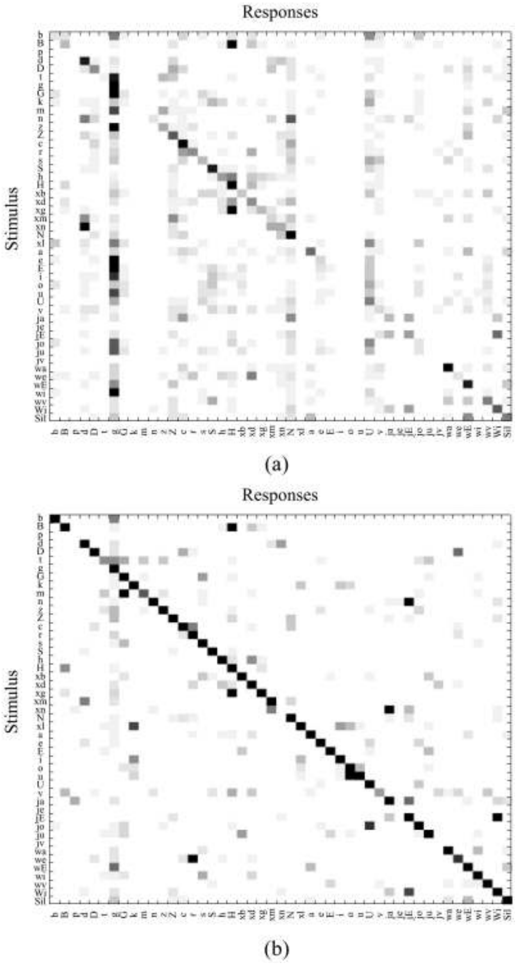


Fig. 1.

Phoneme confusion matrices (a) from a moderately dysarthric speaker and (b) from a control speaker.

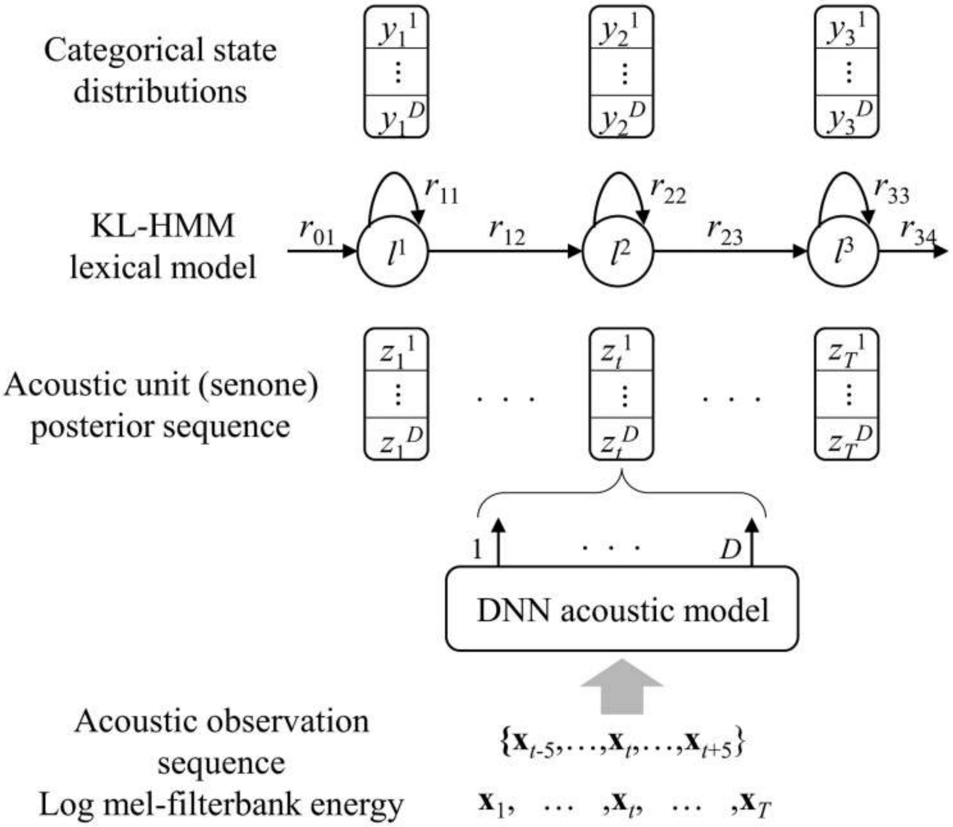
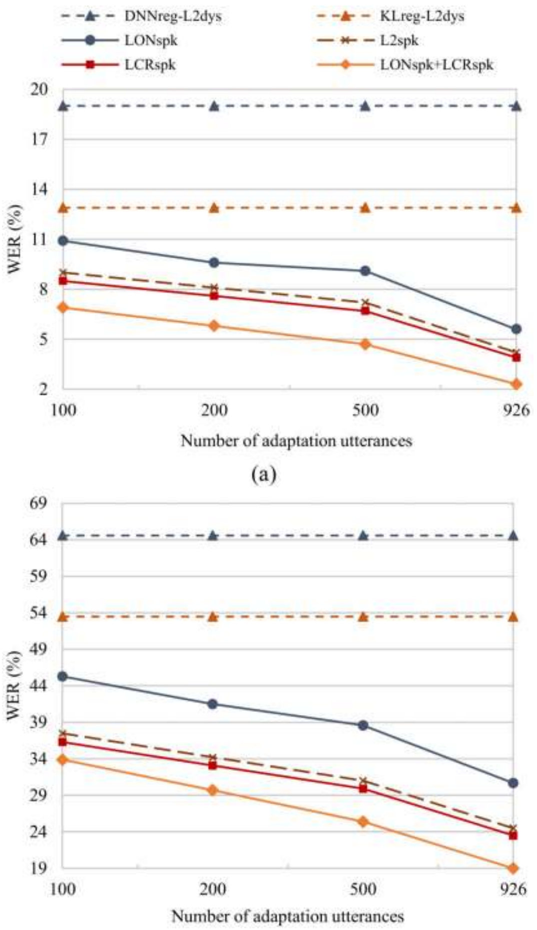

Categorical state distributions
y_1^1 ⋮ y_1^D
y_2^1 ⋮ y_2^D
y_3^1 ⋮ y_3^D
KL-HMM lexical model
r_{01}
l^1
r_{11}
r_{12}
l^2
r_{22}
r_{23}
l^3
r_{33}
r_{34}
Acoustic unit (senone) posterior sequence
z_1^1 ⋮ z_1^D
. . .
z_t^1 ⋮ z_t^D
. . .
z_T^1 ⋮ z_T^D
1
. . .
D
DNN acoustic model
Acoustic observation sequence
Log mel-filterbank energy
$\{\mathbf{x}_{t-5},\ldots,\mathbf{x}_t,\ldots,\mathbf{x}_{t+5}\}$
$\mathbf{x}_1, \ldots, \mathbf{x}_t, \ldots, \mathbf{x}_T$

Fig. 2.

KL-HMM framework used in this work. Here, *r* denotes the state transition probability.



(a)

(b)

Fig. 3.

Performance comparison of the speaker-adapted DNN-HMM and proposed KL-HMM according to the number of adaptation data for (a) mildly and (b) moderately dysarthric speakers.

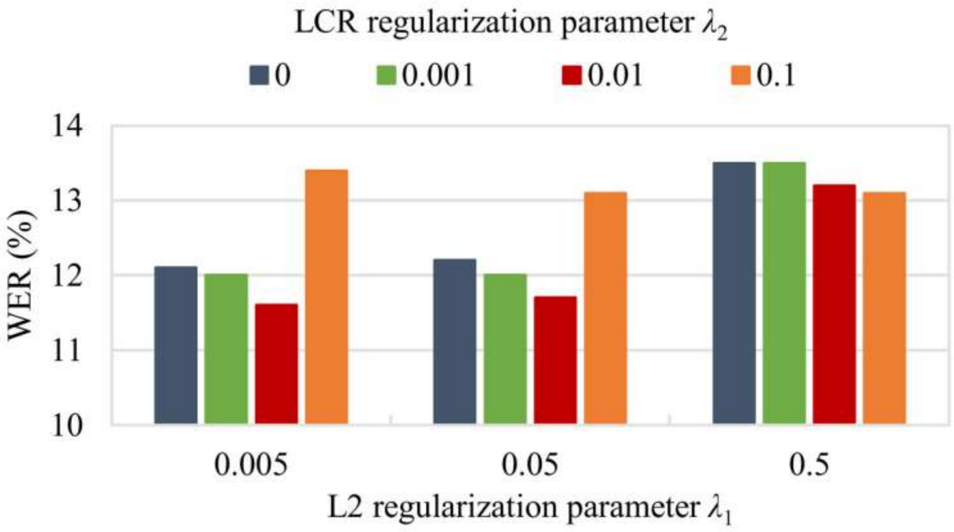


Fig. 4.

WERs with varying the regularization parameters λ_1 and λ_2 for the LCR regularized SA KL-HMM on dysarthric speech.

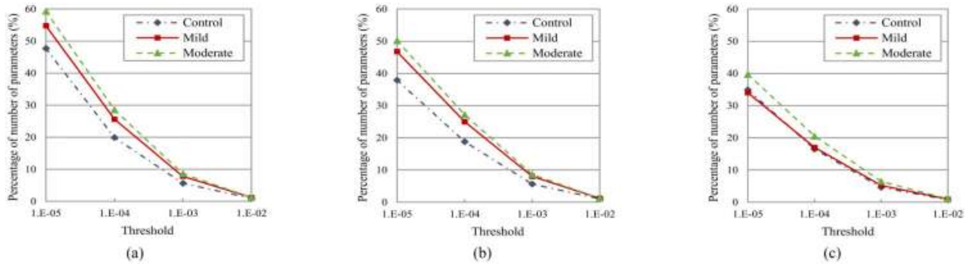


Fig. 5.

Average percentage of the number of KL-HMM parameters whose values are above the threshold for (a) $L2_{spk}$, (b) LCR_{spk}, and (c) $LON_{spk}+LCR_{spk}$.

Table I

Data Distribution

Role	DB	# of speakers	Data
SI regular training	KPOW DB KPRW DB KPBW DB	580 non-dysarthric	300k utterances (about 54 hours)
SI dysarthria adaptation	Section II-A	48 dysarthric (25 mild & 23 moderate)	20k utterances (about 4 hours)
Speaker adaptation	Section II-A	20 dysarthric (12 mild & 8 moderate)	100, 200, 500, and 926 utterances per each speaker
		10 non-dysarthric	100 and 313 utterances per each speaker
Evaluation	Section II-A	20 dysarhtic (12 mild & 8 moderate)	18k utterances (893 utterances per each speaker)
		10 non-dysarthric	5k utterances (485 utterances per each speaker)

Table II

Performances (WERs) of the Proposed KL-HMM, DNN-HMM, and GMM-HMM based ASR Systems

SI ASR system	**WER (%)**		
	Dys.	**Con.**	**Avg.**
GMM_{reg}-HMM	51.1	0.7	25.9
GMM_{reg}-MAP_{dys}-HMM	42.3	1.5	21.9
DNN_{reg}-HMM	45.0	0.6	22.8
DNN_{reg}-L2_{dys}-HMM	36.8	1.4	19.1
DNN_{reg}-LON_{dys}-HMM	35.8	2.2	19.0
$\text{DNN}_{\text{multi}}$-HMM	37.8	0.8	19.3
DNN_{reg}-KL_{reg}-HMM	45.0	0.5	22.8
DNN_{reg}-KL_{dys}-HMM	33.8	0.9	17.4
DNN_{reg}-KL_{reg}-L2_{dys}-HMM	33.6	0.7	17.2
DNN_{reg}-KL_{reg}-LCR_{dys}-HMM	31.9	0.6	16.3
DNN_{reg}-L2_{dys}-KL_{reg}-L2_{dys}-HMM	28.9	1.5	15.2
DNN_{reg}-L2_{dys}-KL_{reg}-LCR_{dys}-HMM	**28.0**	1.0	**14.5**
DNN_{reg}-LON_{dys}-KL_{reg}-LCR_{dys}-HMM	30.1	1.3	15.7

Table III

WERs (%) of the Speaker-Adapted (SA) KL-HMM and DNN-HMM for the Various Number of Adaptation Utterances for Dysarthric Speakers

SI ASR system	SA method	Number of adaptation utterances				
		0	**100**	**200**	**500**	**926**
DNN_{reg}-$\mathrm{L2}_{dys}$	LON_{spk}	36.8	24.2	22.0	20.5	15.6
DNN_{reg}-$\mathrm{L2}_{dys}$-KL_{reg}-$\mathrm{L2}_{dys}$	$\mathrm{L2}_{spk}$	28.9	20.2	18.4	16.7	12.2
	LCR_{spk}	28.9	19.9	17.7	16.0	11.7
DNN_{reg}-$\mathrm{L2}_{dys}$-KL_{reg}-LCR_{dys}	LCR_{spk}	**28.0**	**19.4**	**17.6**	**15.8**	**11.5**
	LON_{spk}+LCR_{spk}	**28.0**	**17.2**	**15.0**	**12.7**	**8.8**

Table IV

WERs (%) of the Speaker-Adapted (SA) KL-HMM and DNN-HMM for the Various Number of Adaptation Utterances For Control Speakers

SI ASR system	SA method	Number of adaptation utterances		
		0	**100**	**313**
DNN_{reg}-L2_{dys}	LON_{spk}	1.4	0.7	0.5
DNN_{reg}-L2_{dys}-KL_{reg}-L2_{dys}	L2_{spk}	1.5	0.7	0.4
	LCR_{spk}	1.5	0.6	0.4
DNN_{reg}-L2_{dys}-KL_{reg}-LCR_{dys}	LCR_{spk}	1.0	0.5	0.3
	LON_{spk}+LCR_{spk}	1.0	0.4	0.3

Table V

Symmetric KL Divergence Between Speaker-Adapted KL-HMM Parameters

SA method	Control	Mild	Moderate
L2$_{\text{spk}}$	6.87	5.26	4.66
LCR$_{\text{spk}}$	6.91	5.37	4.80
LON$_{\text{spk}}$+LCR$_{\text{spk}}$	7.88	7.30	6.08