IDIAP RESEARCH REPORT

ARTICULATORY FEATURE BASED CONTINUOUS SPEECH RECOGNITION USING PROBABILISTIC LEXICAL MODELING

Ramya Rasipuram Mathew Magimai.-Doss

Idiap-RR-19-2014

NOVEMBER 2014

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Articulatory Feature based Continuous Speech Recognition using Probabilistic Lexical Modeling

Ramya Rasipuram[a,b], Mathew Magimai.-Doss[a]

{*ramya.rasipuram,mathew*}*@idiap.ch*

[a] *Idiap Research Institute, Martigny, Switzerland*
[b] *Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland*

Abstract

Phonological studies suggest that the typical subword units such as phones or phonemes used in automatic speech recognition systems can be decomposed into a set of features based on the articulators used to produce the sound. Most of the current approaches to integrate articulatory feature (AF) representations into an automatic speech recognition (ASR) system are based on deterministic knowledge-based phoneme-to-AF relationship. In this paper, we propose a novel two stage approach in the framework of probabilistic lexical modeling to integrate AF representations into an ASR system. In the first stage, the relationship between acoustic feature observations and various AFs is modeled. In the second stage, a probabilistic relationship between subword units and AFs is learned using transcribed speech data. Our studies on a continuous speech recognition task show that the proposed approach effectively integrates AFs into an ASR system. Furthermore, the studies show that either phonemes or graphemes can be used as subword units. Analysis of the probabilistic relationship captured by the parameters has shown that the approach is capable of adapting the knowledge-based phoneme-to-AF representations using speech data; and allows different AFs to evolve asynchronously.

Keywords: Automatic speech recognition; articulatory features; probabilistic lexical modeling, Kullback-Leibler divergence based hidden Markov model; phoneme subword units; grapheme subword units.

1. Introduction

Articulatory features describe the properties of speech production, i.e., each sound unit of a language, a phone or a phoneme, can be decomposed into a set of features based on the articulators used to produce it. The use of articulatory feature (AF) representations in an automatic speech recognition (ASR) system is motivated by their abilities such as:

- Better pronunciation modeling: AFs are hypothesized to capture acoustic variation at finer level than the phoneme-based representation (Deng et al., 1997; Richardson et al., 2003; Livescu et al., 2008)
- Robustness to noise: Different AFs may have variable noise sensitivity. The "divide and conquer" approach provides a framework to exploit the variable noise sensitivity of AFs (Kirchhoff et al., 2002).
- Multilingual and crosslingual portability: AFs can provide better sharing capabilities than phonemes across languages (Stüker et al., 2003; Lal and King, 2013; Siniscalchi et al., 2012).

To incorporate the articulatory knowledge in an ASR system, the following main concerns have to be addressed:

1. AF representations: There exist different types of articulatory representations of speech like: binary features, multi-valued features, and government phonological features. AFs defined by Chomsky and Halle (1968) are binary valued features, for example +voice and -voice, +sonorant and -sonorant. However, according to Ladefoged (1993), it is more natural to allow each AF to take multiple values. In government phonological feature system, speech sounds are destructed into a set of primes and can be represented by fusing them structurally (Harris, 1994). In this paper, we are interested in multi-valued AFs.
2. Estimation of AFs from acoustic speech signal: In the literature, many approaches have been explored to extract articulatory features from the acoustic speech signal. For example, techniques based on acoustic-to-articulatory feature codebooks (Hogden et al., 1996; Suzuki et al., 1998), artificial neural networks (Livescu et al., 2008; Kirchhoff et al., 2002; Chang, 2002; Rasipuram and Magimai.-Doss, 2011b; Siniscalchi et al., 2012), support vector machines (Juneja and Espy-Wilson, 2004; Scharenborg et al., 2007), Gaussian mixture models (Metze and Waibel, 2002; Stüker et al., 2003), hidden Markov models (Hiroya and Honda, 2004), conditional random fields (Prabhavalkar et al., 2011) nearest neighbour (Næss et al., 2011), dynamic Bayesian networks (Frankel and King, 2005; Frankel et al., 2007) are used.
3. Integration: Integrating AFs into the conventional hidden Markov model (HMM) based ASR framework is a challenging task mainly because of the multiple AF estimators. The dynamic Bayesian network (DBN) based approaches for AF integration preserve the articulatory representation in DBN state space (Livescu and Glass, 2004; Livescu et al., 2008; King et al., 2007). These approaches have shown promising results in lexical access experiments. Posterior probabilities of AFs can be transformed for use as features in tandem speech recognition systems (Cetin et al., 2007a,b; Lal and King, 2013). Posterior probabilities of AFs are also used to enhance phoneme-based acoustic models (Cetin et al., 2007a,b; Lal and King, 2013). These approaches however lose other benefits of articulatory representation such as finer granularity and asynchronous evolution.

In this paper, we propose an approach in the framework of probabilistic lexical modeling to integrate multi-valued AFs. In a probabilistic lexical model based ASR system, the relationship between subword units in the lexicon and acoustic feature observations is factored into two models using latent variables: An acoustic model which models the relationship between acoustic feature observations and latent variables; and a lexical model which models a probabilistic relationship between subword units in the lexicon and latent variables. In this paper, we show that by choosing the latent variables as multiple multi-valued AFs, the approach effectively integrates AFs into the HMM-based ASR framework. The lexical model parameters in the proposed approach capture a probabilistic relationship between subword units and AFs learned through transcribed speech data.

The potential of the proposed approach for AF integration is demonstrated on a continuous speech recognition task through experiments and comparisons with the tandem approach. In the proposed framework we explore the use of domain-independent data for acoustic model training; and phonemes and graphemes as subword units. Furthermore, through the analysis of the lexical model parameters we show that the approach adapts the knowledge-based phoneme-to-AF or grapheme-to-AF relationship and allows different AFs to evolve asynchronously.

The rest of the paper is organized as follows: Section 2 gives an overview of the HMM-based ASR and the framework of probabilistic lexical modeling. Section 3 presents the literature review of approaches that integrate multi-valued AFs for ASR in the light of the background information given in Section 2. In Section 4, the approach for AF integration is presented. Sections 5 and 6 present the experimental setup and results, respectively. Section 7 presents an analysis on the subword-unit-to-AF relationship captured by the lexical model parameters. Finally, in Section 8 we provide a discussion and conclusion.

2. Background

The goal of ASR is to find the most likely word sequence $W^* = \{\mathbf{w}_1, \ldots, \mathbf{w}_m, \ldots, \mathbf{w}_M\}$ given the acoustic observation sequence $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_t, \ldots, \mathbf{x}_T\}$ where M is the number of words in the utterance and T represents the number of frames in the speech signal. The most likely word sequence W^* given the acoustic observation sequence is obtained as follows:

$$W^* = \arg\max_{W \in \mathcal{W}} P(W|X) \tag{1}$$

$$= \arg\max_{W \in \mathcal{W}} p(X|W)P(W) \tag{2}$$

where $\mathcal{W}$ denotes the set of all possible word sequences and W denotes a word sequence. The first term on the right hand side of Eqn (2) is the likelihood of acoustic observation sequences given the word sequence and is referred to as the acoustic likelihood. The second term on the right hand side of Eqn (2) is the prior probability of the word sequence or the language model probability.

In general, speech recognition systems model words as a sequence of subword units, which are further modeled as a sequence of HMM states. The sequence of subword units for a word is given by its pronunciation model as specified in the pronunciation lexicon. The acoustic likelihood in an HMM-based ASR system is computed as follows:

$$
\begin{aligned}
p(X|W, \Theta_A) &= \sum_{Q \in \mathcal{Q}} p(X|Q, W, \Theta_A) P(Q|W, \Theta_A) && (3)\\
&= \sum_{Q \in \mathcal{Q}} p(X|Q, \Theta_A) P(Q|W, \Theta_A) && (4)\\
&\approx \max_{Q \in \mathcal{Q}} p(X|Q, \Theta_A) P(Q|W, \Theta_A) && (5)\\
&\approx \max_{Q \in \mathcal{Q}} \Big[\prod_{t=1}^{T} p(\mathbf{x}_t|q_t, \Theta_A) P(q_t|q_{t-1}, \Theta_A) \Big] && (6)
\end{aligned}
$$

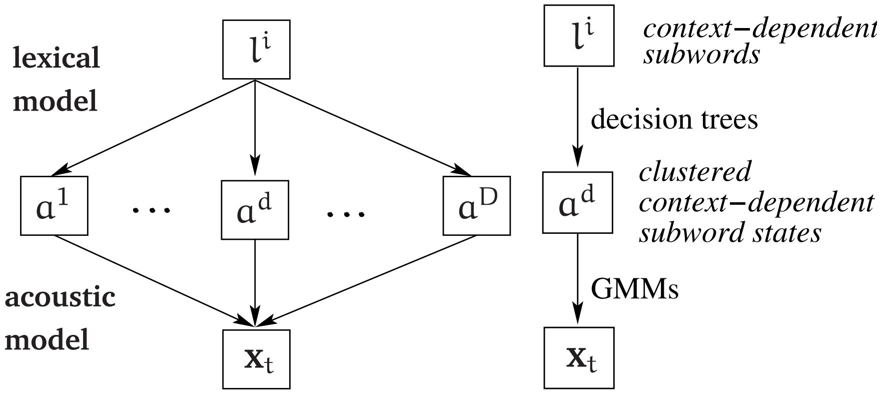
In Eqn (3), the acoustic likelihood is obtained by summing over all possible state sequences $\mathcal{Q}$ where each $Q = \{q_1, \ldots, q_t, \ldots, q_T\}$ denotes a sequence of HMM states corresponding to a word sequence hypothesis. Eqn (4) assumes that acoustic likelihood is independent of words given the state sequence. In Eqn (5), a Viterbi approximation is employed where the sum over all possible state sequences is replaced with the most probable state sequence. Eqn (6) arises from the two HMM assumptions i.e., acoustic feature observations are conditionally independent of each other and the HMM state at time t depends only on the HMM state at time $t - 1$.

In subword unit based ASR systems, HMM states represent lexical units i.e., $q_t \in \mathcal{L} = \{l^1, \ldots l^i \ldots l^I\}$ and I is the number lexical units. If context-independent phonemes are used as subword units then the number of lexical units $I = M \times K$ where K is the number of context-independent subword units in the lexicon and M is the number of HMM states for each context-independent phoneme. If context-dependent phonemes are used as subword units then the number of lexical units $I = M \cdot K^{c_r + c_l + 1}$ where c_l is the preceding context length, c_r is the following context length. Typically, each context-independent or context-dependent phoneme is modeled with three HMM states i.e., $M = 3$.

2.1. Framework of Probabilistic Lexical Modeling

In the framework of probabilistic lexical modeling (Rasipuram and Magimai.-Doss, 2014), the relationship between acoustic feature observation $\mathbf{x}_t$ and lexical

Figure 1: Graphical model representations



(a) Graphical model representation of probabilistic lexical model based ASR system at time frame t

(b) Graphical model representation of deterministic lexical model based ASR system at time frame t

unit l^i is factored through a *latent* variable a^d as follows:

$$p(\mathbf{x}_t|q_t = l^i, \Theta_A) = \sum_{d=1}^{D} p(\mathbf{x}_t, a^d|q_t = l^i, \Theta_A) \tag{7}$$

$$= \sum_{d=1}^{D} p(\mathbf{x}_t|a^d, q_t = l^i, \theta_a, \theta_l) \cdot P(a^d|q_t = l^i, \theta_l) \tag{8}$$

$$= \sum_{d=1}^{D} \underbrace{p(\mathbf{x}_t|a^d, \theta_a)}_{\textbf{acoustic model}} \cdot \underbrace{P(a^d|q_t = l^i, \theta_l)}_{\textbf{lexical model}} \tag{9}$$

The parameters of the acoustic likelihood estimator Θ_A encompass the *acoustic model* (θ_a), the *pronunciation lexicon* (θ_{pr}) and the *lexical model* (θ_l) parameters, therefore, $\Theta_A = \{\theta_a, \theta_{pr}, \theta_l\}$. The relationship in Eqn (9) is as a result of the assumption that given a^d, $p(\mathbf{x}_t|a^d, q_t = l^i, \theta_a, \theta_l)$ is independent of l^i. In Eqn (9), $p(\mathbf{x}_t|a^d, \theta_a)$ is the acoustic unit likelihood and $P(a^d|l^i, \theta_l)$ is the probability of the latent variable given the lexical unit. We refer to $p(\mathbf{x}_t|a^d, \theta_a)$ as the acoustic model, $P(a^d|l^i, \theta_l)$ as the lexical model, the latent variable a^d as the acoustic unit, the set of acoustic units $\mathcal{A} = \{a^1, \dots a^d, \dots a^D\}$ and D as the number of acoustic units. Figure 1(a) shows the Bayesian network of an ASR system that uses the factorization of Eqn (9). The lexical unit is given deterministically by the current word and its subword units. The lexical unit is mapped to all the acoustic units probabilistically and the acoustic feature observation is conditioned on all the acoustic units.

2.2. Lexical and Acoustic Units

In the case of context-independent ASR systems, the lexical unit set $\mathcal{L}$ and the acoustic unit set $\mathcal{A}$ are knowledge driven and defined based on the subword units in the pronunciation lexicon. The number of lexical units or acoustic units $I = D = M \times K$, typically, $M = 3$.

In the case of context-dependent ASR systems, the number of lexical units $I = M \cdot K^{c_r + c_l + 1}$. Generally, not all context-dependent subword units will appear sufficiently often in the training data. Hence a sharing approach is used to enable multiple lexical units to share an acoustic model. This is done using a decision-tree based state clustering and tying technique that uses a pronunciation lexicon, linguistic knowledge to prepare a phonetic question set and acoustic data (Young et al., 1994). The number of acoustic units D varies depending on hyper parameters such as the state occupancy count and the log-likelihood threshold that are used during decision-tree based state clustering. However, the number of acoustic units D is well below the number of lexical units I. The resulting acoustic units are typically referred as clustered context-dependent states or tied-HMM states.

Other possibilities for the choice of the acoustic units include fenones (Bahl et al., 1988), senones (Hwang and Huang, 1992), automatically derived units from the acoustic data (T.Holter and T.Svendsen, 1997) etc. In this paper, we show that HMM-based ASR systems can be built using multi-valued articulatory features as the acoustic units.

2.3. Acoustic Model

In the literature, there are many approaches for acoustic modeling, depending on the way acoustic units are modeled and the acoustic score $p(\mathbf{x}_t|a_t^d, \theta_a)$ is estimated.

- **Gaussian mixture models (GMMs) (Rabiner, 1989):** Each acoustic unit is modeled by a mixture of Gaussians. The acoustic model score is estimated given a mixture of C^d Gaussians that model an acoustic unit a^d, i.e.,

$$p(\mathbf{x}_t|a^d, \theta_a) = \sum_{c=1}^{C_d} w_c^d \mathcal{N}(\mathbf{x}_t, \mu_c^d, \Sigma_c^d) \tag{10}$$

  where w_c^d, μ_c^d and Σ_c^d are the weight, means and covariances of the mixture c of the acoustic unit a^d.

- **Semi-continuous GMMs (SCGMMs) (Huang and Jack, 1989; Bellegarda and Nahamoo, 1990):** In this case, the Gaussian means and variances are shared among all the acoustic units and only the mixture weights for each acoustic unit differ. Given a set of large number of Gaussians $\{\mu_c, \Sigma_c\}_{c=1}^{C}$, the acoustic score is estimated as:

$$p(\mathbf{x}_t|a^d, \theta_a) = \sum_{c=1}^{C} w_c^d \mathcal{N}(\mathbf{x}_t, \mu_c, \Sigma_c) \tag{11}$$

  where w_c^d is the weight of the mixture c of the acoustic unit a^d.

- **Subspace GMMs (SGMMs) (Povey et al., 2011a):** In SGMMs, similar to SCGMMs, all the acoustic units share the same C Gaussians. However, for each acoustic unit, the Gaussians are selected from a subspace spanned by all the 'C' Gaussians and the acoustic unit specific parameters. The acoustic model score in SGMM is computed as:

$$p(\mathbf{x}_t|a^d, \theta_a) = \sum_{c=1}^{C} w_c^d \mathcal{N}(\mathbf{x}_t, \mu_c^d, \Sigma_c) \tag{12}$$

  The mixture weights w_c^d and means μ_c^d are derived from the state specific parameter $\mathbf{v}_d$ via globally shared parameters $\mathbf{M}_c$ and $\mathbf{w}_c^T$. Covariances Σ_c are shared between states.

$$\mu_c^d = \mathbf{M}_c \mathbf{v}_d \tag{13}$$

$$w_c^d = \frac{\exp \mathbf{w}_c^T \mathbf{v}_d}{\sum_{c=1}^{C} \exp \mathbf{w}_c^T \mathbf{v}_d} \tag{14}$$

- **Artificial neural networks (ANN) (Morgan and Bourlard, 1995):** The ANN computes the probability of acoustic units given the acoustic feature observations $p(a^d|\mathbf{x}_t, \theta_a)$ which is then converted to scaled-likelihood, i.e.,

$$p_{sl}(\mathbf{x}_t|a^d, \theta_a) = \frac{p(\mathbf{x}_t|a^d, \theta_a)}{p(\mathbf{x}_t)} = \frac{P(a^d|\mathbf{x}_t, \theta_a)}{P(a^d)} \tag{15}$$

2.4. Lexical Model

The lexical model in an ASR system can be deterministic or probabilistic.

2.4.1. Deterministic Lexical Model based ASR Approaches

When the lexical model is deterministic, each lexical unit l^i is deterministically mapped to an acoustic unit a^j ($l^i \mapsto a^j$) i.e.,

$$P(a^d|q_t = l^i, \theta_l) = \begin{cases} 1, & \text{if } d = j\ ; \\ 0, & \text{otherwise.} \end{cases} \tag{16}$$

As a result of the deterministic mapping, the only term contributing to the summation in Eqn (9) is the acoustic unit that is mapped to a lexical unit. Figure 1(b) shows the Bayesian network of an ASR system in which the lexical model is deterministic. The lexical unit is given deterministically by the current word and its subword units. A lexical unit is mapped to an acoustic unit and the acoustic feature observation is conditioned on an acoustic unit.

In standard HMM-based ASR approaches such as semi-continuous HMMs (Huang and Jack, 1989; Bellegarda and Nahamoo, 1990), HMM/GMM (Rabiner, 1989), hybrid HMM/ANN (Morgan and Bourlard, 1995), SGMM (Povey et al., 2011a), the lexical model is deterministic. In the case of context-independent ASR systems, it is a *knowledge-based look-up table* that maps each lexical unit to an acoustic unit. In the case of context-dependent

ASR systems, typically the lexical model is the *decision trees* trained using the pronunciation lexicon, acoustic data and the linguistic knowledge. The decision trees map each context-dependent subword unit to a tied HMM state (or an acoustic unit). The following ASR approaches differ primarily in the way the acoustic units are modeled and the acoustic score is computed:

- In semi-continuous HMMs, the acoustic model is semi-continuous GMMs (see Eqn (11)).
- In HMM/GMM, the acoustic model is GMMs (see Eqn (10)).
- In hybrid HMM/ANN, the acoustic model is an ANN (see Eqn (15)).
- In SGMMs, the acoustic model is subspace Gaussians (see Eqns(12)).

Figure 2 illustrates various steps in an HMM-based ASR system where the lexical model is deterministic. As shown in the figure, a sequence of words constrained by language model are represented by a sequence of subword units /ih/ /z/ /ih/ /t/ as given in the pronunciation lexicon. The sequence of subword units are converted to a sequence of context-dependent subword units and later to a sequence of lexical units. In this illustration, each context-dependent subword unit is composed of one lexical unit. However, normally each context-dependent subword unit is represented with three lexical units. Each of the lexical units is deterministically mapped to an acoustic unit i.e., a tied HMM state using decision trees. Finally, the acoustic feature observations are conditioned on acoustic units. Acoustic units can be modeled using any of the methods given in Section 2.3.

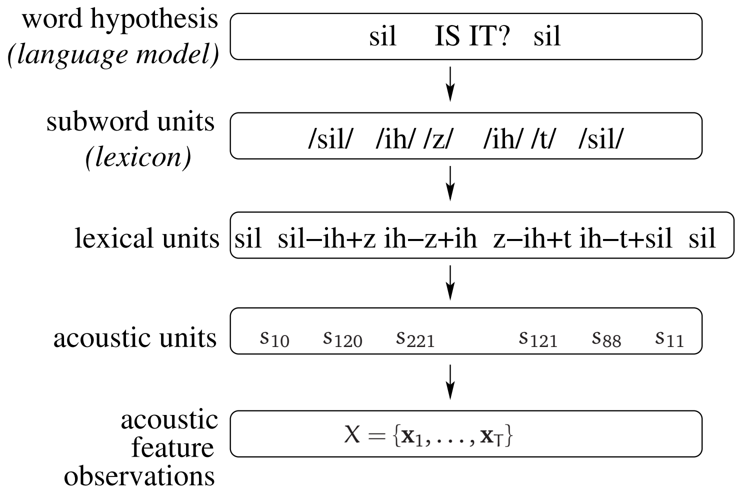


Figure 2: Sequence of steps taken in an context-dependent HMM-based ASR system where the lexical model is deterministic

2.4.2. Probabilistic Lexical Model based ASR Approaches

The Eqn (9) with the two conditions, namely, $0 < P(a^d|l^i, \theta_l) < 1$ and $\sum_{d=1}^{D} P(a^d|l^i, \theta_l) = 1$ characterizes an ASR approach where each lexical unit is probabilistically related to all acoustic units. In the following approaches, lexical units are based on context-dependent subword units and the lexical model is probabilistic. The approaches [1] differ in two main aspects: the way lexical model parameters are computed; and the way acoustic units are modeled.

- **Probabilistic classification of HMM states (PC-HMM) (Luo and Jelinek, 1999):** In the PC-HMM approach, lexical units are based on context-dependent phonemes and the lexical model is probabilistic. Acoustic units are clustered context-dependent phoneme states and each acoustic unit is modeled with a GMM. The acoustic and lexical model parameters are learned together through the expectation maximization (EM) algorithm.
- **Tied posteriors (Rottland and Rigoll, 2000):** In the tied posterior approach, the acoustic units are modeled using an ANN. The lexical model parameters are learned through the scaled-likelihood estimates $p_{sl}(\mathbf{x}_t|a^d, \theta_a)$ using the EM algorithm or the Viterbi EM algorithm. The tied-posterior approach is equivalent to the PC-HMM approach when the acoustic units are clustered context-dependent phoneme states and the acoustic units are modeled with GMMs instead of ANNs.
- **Kullback-Leibler divergence based hidden Markov model (KL-HMM) (Aradilla et al., 2007, 2008; Aradilla, 2008):** In the KL-HMM approach, lexical model parameters are learned through acoustic unit posterior probability estimates $P(a^d|\mathbf{x}_t, \theta_a)$ using the Viterbi EM algorithm that minimizes a function based on the Kullback-Leibler (KL) divergence. It has been shown that acoustic units can be modeled using an ANN (Aradilla et al., 2007, 2008; Rasipuram and Magimai.-Doss, 2014) or using GMMs (Rasipuram and Magimai.-Doss, 2013a).

In similar vein, in the hidden model sequences HMM (HMS-HMM) approach, the deterministic mapping between phoneme-to-HMM or phoneme-to HMM-state is replaced with a stochastic model (Hain and Woodland, 1999; Hain, 2005). More specifically, each phoneme is represented by a mixture of HMM state sequences corresponding to different variants.

Despite the similarities in the formulation of the KL-HMM, PC-HMM and tied posterior approaches, KL-HMM has additional advantages (Rasipuram and Magimai.-Doss, 2014, Section 4.4). The approaches differ in terms of the local score at each HMM state (KL-divergence vs scalar product), in the way lexical model parameters are estimated (minimizing the KL-divergence vs maximizing the likelihood) and the way decoding is performed (capability of the KL-HMM approach to reverse the roles of acoustic and lexical models by exploiting the

1. It is important to note that the notion of acoustic units and lexical units was not explicitly defined in these previous works. These approaches can be viewed from the point of view of probabilistic lexical modeling (Rasipuram and Magimai.-Doss, 2014).

asymmetric property of KL-divergence). Furthermore, the experimental studies indicated that the KL-HMM approach performs better than that of the tied posterior approach on various ASR tasks (Rasipuram and Magimai.-Doss, 2014). Therefore, in this paper, we use the KL-HMM approach for probabilistic lexical modeling.

2.5. Advantages of Probabilistic Lexical Modeling

Pronunciation variability modeling: Standard HMM-based ASR systems like HMM/GMM and hybrid HMM/ANN deterministically model the relationship between lexical units and acoustic feature observations. As a result of the deterministic relationship, these systems rely on a well developed phoneme lexicon to handle the variability in the acoustic training data. However, when the pronunciations in the lexicon do not reflect the underlying speech data then such a model may poorly represent the training data. For example, this can happen in the case of non-native speakers (where pronunciations normally reflect native speakers) or in the case of spontaneous and conversational speech (where spoken words are pronounced differently from lexicon pronunciations) or in the case of a grapheme lexicon (where pronunciations are based on orthography of the word). To account for such a variation, typically, phoneme-based ASR systems add pronunciation variants to the lexicon (Strik and Cucchiarini, 1999). However, the manual addition of pronunciation variants may require explicit human knowledge.

In the context of pronunciation variability modeling, it has been shown that the limitation of the standard HMM/GMM system imposed by deterministic mapping can be handled by modeling a probabilistic relationship between lexical and acoustic units (Saraclar et al., 2000; Hain, 2005; Rasipuram and Magimai.-Doss, 2013b, 2014). ASR studies suggest that implicit modelling of pronunciation variation through probabilistic lexical modeling can perform equally or better than the use of explicit knowledge-based addition of pronunciation variants. This is observed for phoneme-based ASR systems when multiple pronunciations of a word were collapsed to a single pronunciation (Hain, 2005) and for grapheme-based ASR systems where there is a single pronunciation for each word (Rasipuram and Magimai.-Doss, 2013a).

Resource optimization: In probabilistic lexical model based approaches such as KL-HMM and tied posteriors, acoustic model and lexical model are trained one after another and can be trained on independent set of resources. For example, the acoustic model can be trained on resources from resource-rich languages and domains whereas the lexical model can be trained on a relatively small amount of target language data (Imseng et al., 2012; Rasipuram and Magimai.-Doss, 2014).

Flexibility in the choice of acoustic and lexical units: In probabilistic lexical model based ASR approaches, it is not necessary that the subword unit set used for defining the acoustic units should be the same as the subword unit set used for defining the lexical units. The lexical model can capture the relationship between the distinct subword unit sets through acoustics. This flexibility has been exploited to build ASR systems where the acoustic unit set is based on phonemes and the lexical unit set is based on graphemes (Magimai.-Doss et al., 2011; Rasipuram and Magimai.-Doss, 2013a). Furthermore, lexical and acoustic units can model different contextual units. For instance, lexical units can be based on context-dependent subword units while the acoustic units can be based on context-independent subword units (Rottland and Rigoll, 2000; Magimai.-Doss et al., 2011; Imseng et al., 2012).

In Section 4, we propose a novel approach to integrate articulatory features into HMM-based ASR in the framework of probabilistic lexical modeling that can exploit all the above advantages.

3. Literature Survey

There has been a sustained interest in incorporating speech production knowledge into an ASR system for reasons already stated in the Introduction (Section 1). AFs have been incorporated at various levels of an ASR system. Here we provide a brief overview of ASR systems that used multi-valued AFs according to the background of the previous section.

3.1. Lexical units are phonemes and acoustic units are AFs

In these works, the acoustic units are based on AFs or both AFs and phonemes (Metze and Waibel, 2002; Stüker et al., 2003; Juneja and Espy-Wilson, 2004; Livescu et al., 2008). Each acoustic unit is modeled with a GMM or with discriminative classifiers like ANNs or support vector machines. The lexical model is deterministic, i.e., each phoneme-based lexical unit is deterministically mapped to its AF attributes. The scores from different AF-based acoustic models are combined to arrive at the local emission score $p(x_t|q_t = l^i)$. On continuous speech recognition and cross-lingual adaptation tasks, the use of AF-based acoustic models in combination with phoneme-based acoustic models has resulted in a relative reduction in word error rate (WER) of about 5-10% compared to the use of phoneme-based acoustic models alone (Metze and Waibel, 2002; Stüker et al., 2003).

3.2. Lexical units are AFs and acoustic units are AFs

These are the systems analogous to standard HMM-based ASR systems where both lexical and acoustic units are either based on context-independent or context-dependent subword units (Deng et al., 1997; Richardson et al., 2003; Kirchhoff, 1996; Wester et al., 2004; Livescu et al., 2008). The subword units are now AFs determined from the AF-based pronunciation lexicon. The AF-based pronunciation lexicon transcribes each word in terms of the positions of the articulators. Each AF is associated with its own hidden state variable. The multiple hidden state variables can follow an independent path to a certain extent and can allow certain amount of asynchrony. In the initial works, hidden state variables of various AFs were required to re-synchronize at phoneme level (Deng et al., 1997; Richardson et al., 2003) or syllable level (Kirchhoff, 1996). In more recent works, the flexible DBN framework allows synchronization to happen at word level or even across word boundaries (Livescu and Glass, 2004; Livescu et al., 2008).

Similar to standard context-dependent subword unit-based ASR systems, in these systems, the lexical units are based on context-dependent subword units and the acoustic units are clustered context-dependent subword states obtained using decision tree-based state tying methods, however, unlike standard ASR systems, the subword units are AFs. These systems have obtained improvements in lexical access [2] experiments (Livescu and Glass, 2004; Livescu et al., 2008; Jyothi et al., 2011).

An approach was proposed by Jyothi et al. (2013) to convert a DBN-based pronunciation model into an equivalent set of factored WFSTs. The utility of this approach was demonstrated using phoneme-based pronunciation models on isolated word and continuous word speech recognition tasks; and using AF-based pronunciation models on lexical access tasks. Along the similar lines, an approach was outlined to convert an AF-based DBN pronunciation models into equivalent WFSTs for ASR (Jyothi, 2013, Chapter 8).

3.3. Lexical units are phonemes and acoustic units are phonemes

In these systems, similar to standard HMM-based ASR systems, both lexical and acoustic units are based on phonemes. However, AF representations are used as auxiliary information to enhance the performance of the acoustic model (Kirchhoff et al., 2002; Siniscalchi et al., 2012). For example, the acoustic model can be seen as two stage classifier. In the first stage, a set of AF-based ANNs model the relationship between acoustic features and AFs. In the second stage, a phoneme-based ANN models the relationship between all AFs and phonemes. The resulting phoneme-based ANN is used as an acoustic model in hybrid HMM/ANN systems. These systems have achieved a relative reduction in WER of about 5-6% on noise robust ASR tasks and cross-lingual ASR tasks compared to the systems where acoustic-to-phoneme information is directly modeled (Kirchhoff et al., 2002; Siniscalchi et al., 2012).

Alternatively, AF-based neural networks have also been used in tandem speech recognition systems (Cetin et al., 2007a,b; Livescu et al., 2008; Lal and King, 2013). In the tandem approach, the posterior probabilities of AFs and/or phonemes replace conventional cepstral features in HMM-based ASR systems (Hermansky et al., 2000). In order to model the output of ANN that is typically non-Gaussian with GMMs, posterior probabilities of the acoustic units are Gaussianized using the log function and then decorrelated using the Karhunen-Loeve transform (KLT).

For tandem systems, the use of AF-based ANNs trained on language-independent data was investigated and compared with the use of AF-based ANNs trained on language-dependent data (Cetin et al., 2007b). Cetin et al. (2007b) observed that the AF-based ANNs trained on language-independent

2. The task of lexical access involves predicting a word given its phonetic or broad phonetic transcription (Huttenlocher and Zue, 1984).

data reduced the performance (or the WER has relatively increased by about 2%) of the tandem system compared to the phoneme-based ANNs trained on the same language-independent data. Work by Lal and King (2013) compared the use of AF-based ANNs trained on data from multiple languages (also including the target-language) with AF-based ANNs trained on data from target-language data for tandem systems. It was observed that irrespective of whether the ANNs are trained on data from multiple languages or on target-language, the AF-based ANNs performed better (relative improvement in WER of 1-9%) than the phoneme-based ANNs. However, the AF-based (or phoneme-based) ANNs trained on data from target-language performed better than the AF-based (or phoneme-based) ANNs trained on data from multiple languages (Lal and King, 2013).

The reason for the difference in conclusion by Cetin et al. (2007b) and Lal and King (2013) could be due to the differences in the number of languages used to train the MLPs and their relationship with the target-language. Cetin et al. (2007b) used AF-based ANNs trained on English to generate tandem features for Mandarin ASR task; and English and Mandarin belong to different language families. Whereas Lal and King (2013) used ANNs trained on data from multiple languages that also included target-language.

To summarize, most of the approaches to integrate AFs into an ASR system are based on the deterministic knowledge-based phoneme-to-AF relationship. The approaches summarized in Section 3.1 use the knowledge-based phoneme-to-AF relationship to define the deterministic lexical model parameters. The approaches given in Section 3.2 allow asynchronous evolution of various AFs using AF-based pronunciation lexicon. However, AF-based pronunciation lexicon is prepared using the knowledge-based phoneme-to-AF relationship.

In the next section, we present an ASR approach that integrates a model for lexical access using articulatory features into the HMM-based ASR framework. The approach adapts the knowledge-based phoneme-to-AF relationship using transcribed speech data and incorporates a probabilistic phoneme-to-AF relationship in the model parameters.

4. Proposed Approach

In probabilistic lexical model based ASR systems, a lexical unit is probabilistically related to all acoustic units (Section 2.4.2). For each lexical unit l^i, let $\mathbf{y}_i$ be the D-dimensional probability vector or the categorical variable that captures a probabilistic relationship between the lexical unit l^i and D acoustic units, i.e.,

$$\mathbf{y}_i = [y_i^1, \ldots, y_i^d, \ldots y_i^D]^T \text{ where } y_i^d = P(a^d|l^i) \tag{17}$$

Therefore, the lexical model parameter set $\theta_l = \{\mathbf{y}_i\}_{i=1}^{I}$. In this paper we use the KL-HMM approach to estimate the parameters of the lexical model (Aradilla et al., 2007, 2008; Aradilla, 2008).

4.1. KL-HMM Approach

The KL-HMM approach for lexical model parameter estimation is summarized below:

- The approach assumes that the acoustic unit set $\mathcal{A}$ is defined and a trained acoustic model is available. As described in Section 2, the GMMs or the ANN modeling the acoustic units can be used as the acoustic model. In this paper, we use an ANN as the acoustic model.
- Given the acoustic model, the probabilities of acoustic units or the acoustic unit posterior probability vectors for the training data are estimated. At time t, the acoustic unit posterior probability vector $\mathbf{z}_t$ is a D dimensional probability vector.

$$\begin{aligned}\mathbf{z}_t &= [z_t^1, \ldots, z_t^d, \ldots, z_t^D]^T \\ &= [P(a^1|\mathbf{x}_t), \ldots, P(a^d|\mathbf{x}_t), \ldots, P(a^D|\mathbf{x}_t)]^T \end{aligned} \tag{18}$$

- The acoustic unit probability vector sequences are used along with the pronunciation lexicon and word level transcriptions to train the parameters of the probabilistic lexical model. More precisely, the acoustic unit probability vector sequences are used as feature observations to train a HMM where the states represent the lexical units. Each state l^i is parameterized by a categorical variable $\mathbf{y}_i$ that captures a probabilistic relationship between a lexical unit l^i and D acoustic units.
- As both feature observations and state distributions are probability vectors, the local score at each HMM state can be the KL-divergence between the feature observation $\mathbf{z}_t$ and the categorical variable $\mathbf{y}_i$. KL-divergence being an asymmetric measure, there are the following three possible ways to estimate the KL-divergence:
    1. KL-divergence (S_{KL}): In this case, the state distribution $\mathbf{y}_i$ is the reference distribution.
    $$S_{KL}(\mathbf{y}_i, \mathbf{z}_t) = \sum_{d=1}^{D} y_i^d \log\left(\frac{y_i^d}{z_t^d}\right) \tag{19}$$
    2. Reverse KL-divergence (S_{RKL}): In this case, the acoustic unit probability vector $\mathbf{z}_t$ is the reference distribution.
    $$S_{RKL}(\mathbf{z}_t, \mathbf{y}_i) = \sum_{d=1}^{D} z_t^d \log\left(\frac{z_t^d}{y_i^d}\right) \tag{20}$$
    3. Symmetric KL-divergence (S_{SKL}): The local score S_{SKL} is the average of the local scores S_{KL} and S_{RKL}.
    $$S_{SKL}(\mathbf{y}_i, \mathbf{z}_t) = \frac{1}{2} \cdot [S_{KL}(\mathbf{y}_i, \mathbf{z}_t) + S_{RKL}(\mathbf{z}_t, \mathbf{y}_i)] \tag{21}$$
- The parameters $\{\mathbf{y}_i\}_{i=1}^{I}$ are trained using the Viterbi EM algorithm optimizing a function based on one of the KL-divergence based local scores.

- Decoding is performed using the standard Viterbi decoder where the log-likelihood based score is replaced with the KL-divergence based local score.

The details of the parameter estimation are elaborated in Appendix A. The details on the role of different local scores in estimating the lexical model parameters can found in the work by Rasipuram and Magimai.-Doss (2013c) whereas the role of local scores during decoding can be found in the work by Rasipuram and Magimai.-Doss (2014).

4.2. Integrating AFs using KL-HMM

In the proposed approach, the relationship between lexical units and acoustic features is factored into two parts through the use of AFs as latent variables or acoustic units:

1. *The acoustic model* where the relationship between AFs and acoustic features is modeled.
2. *The lexical model* where a probabilistic relationship between AFs and lexical units is modeled.

The proposed approach exploits the advantage of probabilistic lexical modeling that the subword unit set used for defining the acoustic units need not be the same as the subword unit set used for defining the lexical units (Section 2.5). The lexical units can be based on context-independent subword units or context-dependent subword units. The proposed approach for AF integration can be summarized in the following steps:

- The acoustic unit set includes AFs such as manner and place of articulation. Therefore, the acoustic unit set can be seen as a superset of the individual AF sets, i.e.,

$$\mathcal{A} = \{\{\mathcal{A}_1\}, \ldots, \{\mathcal{A}_\mathsf{F}\}\} \tag{22}$$

  where $\{\mathcal{A}_1\}, \ldots, \{\mathcal{A}_\mathsf{F}\}$ denote the individual AF sets and F the total number of AFs. For example, the set $\{\mathcal{A}_1\}$ may include all the classes specifying the manner of articulation such as vowel, stop, fricative, and so on; the set $\{\mathcal{A}_2\}$ may include all the classes specifying the place of articulation such as alveolar, back, dental, dorsal, front and so on. The total number of acoustic units

$$\mathsf{D} = \mathsf{D}_1 + \cdots + \mathsf{D}_\mathsf{F} \tag{23}$$

  where $\mathsf{D}_1, \ldots, \mathsf{D}_\mathsf{F}$ denote the cardinality of the individual AFs.
- Each AF is associated with an acoustic model, in our case, with an ANN.
- Given the AF-based acoustic models, posterior probabilities of AFs are estimated. The posterior probability estimates of various AFs are concatenated to produce a D dimensional acoustic unit probability vector $\mathbf{z}_\mathsf{t}$, i.e.,

$$\begin{aligned} \mathbf{z}_\mathsf{t} &= [\mathbf{z}_{\mathsf{t},1}, \ldots, \mathbf{z}_{\mathsf{t},\mathsf{F}}]^\mathsf{T} \text{ where} \\ \mathbf{z}_{\mathsf{t},1} &= [z^1_{\mathsf{t},1}, \ldots, z^{\mathsf{D}_1}_{\mathsf{t},1}] \\ \mathbf{z}_{\mathsf{t},\mathsf{F}} &= [z^1_{\mathsf{t},\mathsf{F}}, \ldots, z^{\mathsf{D}_\mathsf{F}}_{\mathsf{t},\mathsf{F}}] \end{aligned} \tag{24}$$

- The lexical model parameters $\theta_l = \{\{\mathbf{y}_{i,f}\}_{f=1}^{F}\}_{i=1}^{I}$ where $\mathbf{y}_{i,f}$ captures a probabilistic relationship between the lexical unit l^i and the AF classes $\{\mathcal{A}_f\}$. That is,

$$\mathbf{y}_{i,1} = [y_{i,1}^1, \ldots, y_{i,1}^{D_1}], \ \sum_{d=1}^{D_1} y_{i,1}^d = 1$$

$$\mathbf{y}_{i,F} = [y_{i,F}^1, \ldots, y_{i,F}^{D_F}], \ \sum_{d=1}^{D_F} y_{i,F}^d = 1$$

- The parameters of the lexical model are trained using the KL-HMM approach. For a lexical unit l^i, the state distribution $\mathbf{y}_i$ can be seen as a stack of F categorical variables, i.e.,

$$\mathbf{y}_i = [\mathbf{y}_{i,1}, \ldots, \mathbf{y}_{i,F}]^T \tag{25}$$

- The local score at each HMM state is the KL-divergence between the feature observation and the state distribution. If the local score is S_{RKL}, then the KL-divergence is computed as:

$$S_{RKL}(\mathbf{z}_t, \mathbf{y}_i) = \sum_{d=1}^{D_1} z_{t,1}^d \log\left(\frac{z_{t,1}^d}{y_{i,1}^d}\right) + \cdots + \sum_{d=1}^{D_F} z_{t,F}^d \log\left(\frac{z_{t,F}^d}{y_{i,F}^d}\right) \tag{26}$$

Figure 3 illustrates the proposed AF-based ASR approach. The ANNs are trained to classify various AFs. The speech data is forward passed through the AF-based ANNs to obtain posterior probabilities of AFs. The posterior probabilities of the individual AFs are stacked and used as feature observations to train a HMM. As mentioned in Section 2.2, each context-independent or context-dependent subword unit is modeled with three-HMM states.

4.3. Previous findings

The proposed approach for AF integration was studied for phoneme recognition using the TIMIT corpus (Rasipuram and Magimai.-Doss, 2011b,a,c). However, in the previous work the notion of probabilistic lexical modeling was not introduced and the studies were presented from the perspective of acoustic modeling (which is not the case). Given the formulation of Section 4.2, the findings from the previous studies are refined and re-summarized below:

- It was demonstrated that using AF-based acoustic models in the KL-HMM approach results in better system than using the same acoustic models in the hybrid HMM/ANN approach (Rasipuram and Magimai.-Doss, 2011b). The study illustrated that it is beneficial if the lexical model is probabilistic.
- Performance of the KL-HMM system using both phoneme-based and AF-based acoustic models was better than the KL-HMM system using only the phoneme-based acoustic model. The study indicates that the KL-HMM approach has the potential to reduce the error rates by incorporating articulatory knowledge into an ASR system.

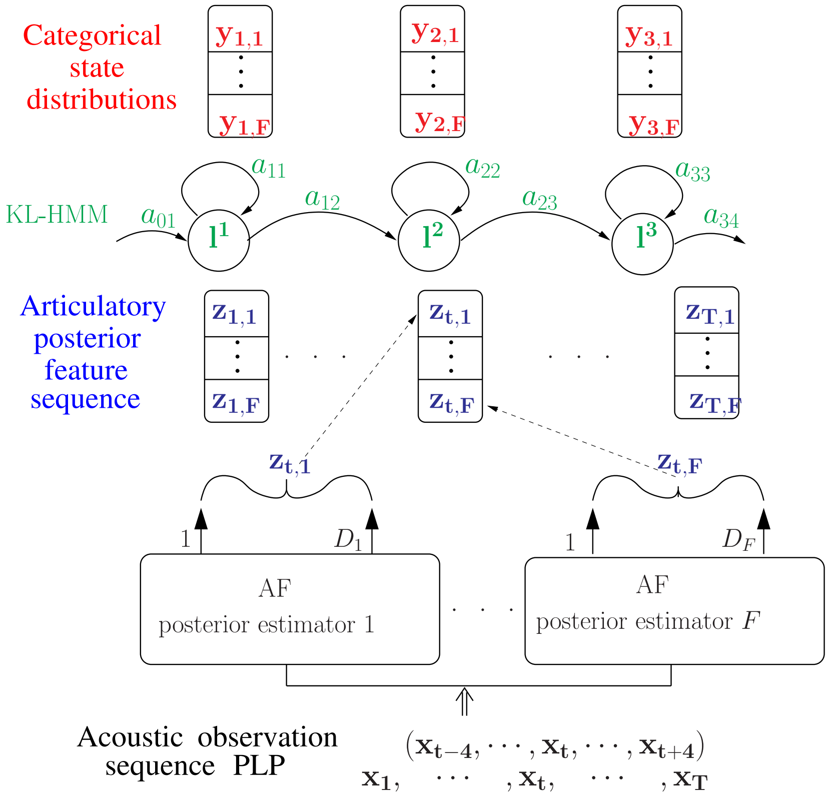


Figure 3: Illustration of the KL-HMM approach for AF integration. Posterior probabilities of AFs estimated using ANNs are combined and used as feature observations. HMM states represent the lexical units.

- The phoneme recognition performance of a system can be improved by improving the AF acoustic models. It was observed that the AF classification accuracy can be improved by modeling the inter-feature dependencies using multistage MLP classifiers and/or multitask learning. In doing so, the performance gap between systems using phonemes as acoustic units and AFs as acoustic units was greatly reduced.

4.4. Contributions of the Present Paper

The contributions of this paper include:

- The proposed approach for AF integration is evaluated on a continuous speech recognition task (see Section 5 for the experimental setup). In the evaluation, acoustic models are ANNs estimating various multi-valued AFs. The lexical units are based on context-dependent phonemes and the lexical model is probabilistic. The proposed approach is compared with the tandem approach for AF integration (Section 6.1). The tandem systems also use the same ANNs as the KL-HMM systems, but as feature extractors.

- In the framework of probabilistic lexical modeling, parameters of the acoustic model θ_a and the lexical model θ_l can be trained on independent set of resources, since they are trained one after another. Exploiting this advantage of probabilistic lexical modeling, we study the case where AF-based acoustic models are trained on domain-independent resources and the lexical model parameters are trained on domain-dependent resources (see Section 6.2). Furthermore, the use of multistage MLP classifiers is studied for a continuous speech recognition task (Section 6.3).
- In the framework of probabilistic lexical modeling it is possible to build grapheme-based ASR systems where the lexical units are based on graphemes and the acoustic units are based on phonemes (Magimai.-Doss et al., 2011; Rasipuram and Magimai.-Doss, 2013a, 2014). These grapheme-based ASR systems, where the lexical model parameters capture a probabilistic graphemes-to-phoneme relationship, performed better than the grapheme-based ASR approaches where the lexical model is deterministic. Motivated by this, in this paper we hypothesize that it is possible to build grapheme-based ASR systems where the lexical units are based on graphemes and the acoustic units are AFs. In this case, the lexical model parameters capture a probabilistic grapheme-to-AF relationship. The resulting grapheme-based ASR approach, in addition to exploiting the advantages of AFs, can also address pronunciation lexical resource constraints in ASR system development.
- The lexical model parameters capture a probabilistic relationship between phonemes and AFs. We analyze the parameters of the lexical model to understand if the captured phoneme-to-AF relationship associates well with the knowledge-based phoneme-to-AF relationship (see Section 7).

5. Experimental Studies

In this paper, we evaluate the proposed approach for AF integration on a speaker-independent continuous speech recognition task using the DARPA Resource Management (RM) corpus.

5.1. Database

The RM corpus consists of read queries on the status of Naval resources (Price et al., 1988). The task is artificial in aspects such as speech type, range of vocabulary, and grammatical constraint. The training set includes 2880 utterances and the development set consists of 1110 utterances. The training and development set together consist of 3,990 utterances spoken by 109 speakers corresponding to approximately 3.8 hours of speech data.

There are four test sets provided by DARPA, namely, feb89, oct89, feb91, and sep92. Each test set contains 300 utterances spoken by 10 speakers. The test set used in this work is obtained by combining the four test sets and thus contains

1,200 utterances amounting to 1.1 hours in total. The test set is completely covered by a word pair grammar included in the task specification.

The phoneme-lexicon consists of 991 words. Similar to the setup reported by Dines and Magimai-Doss (2007), the phoneme-lexicon is obtained from the UNISYN[3] lexicon with general American accent. There are 42 context-independent phonemes including silence. About 35 words in the phoneme lexicon have more than one pronunciation.

The grapheme-lexicon for the RM task is transcribed using 79 graphemes. The first grapheme and the last grapheme of a word are treated as separate graphemes. Therefore, the grapheme set included 26 English graphemes ({[A],[B],...[Z]}), 26 English graphemes occurring at the begin of word ({[b_A],[b_B],...[b_Z]}), 26 English graphemes occurring at the end of word ({[e_A],[e_B],...[e_Z]}) and silence.

5.2. MLPs

KL-HMM systems use a multilayer perceptrons (MLPs) as the acoustic models whereas tandem systems use the same MLPs as feature extractors. The input to all the MLPs is 39-dimensional perceptual linear prediction (PLP) features with a nine frame temporal context (i.e., four frame preceding and four frame following context). We use three-layer (one input, one hidden and one output layer) MLPs that are trained with the frame level cross entropy error criteria using the Quicknet software[4]. The number of hidden units of the MLPs are selected based on the optimal frame accuracy on the development set.

The target labels for the MLPs with phonemes as output units were obtained from the HMM/GMM system. The target labels for the MLPs with AFs as output units are obtained from the phoneme-to-AF map given in Table B.1. The AFs consist of manner, place, height, and vowel. The phoneme-to-AF map is adapted from the mapping defined by Hosom (2009). In order to distinguish all phonemes of the RM task some changes were made to the mapping defined by Hosom (2009). The place class is expanded by adding features like mid-front and mid-back. The height class is expanded by adding features like mid, mid-low, mid-high. Also, the vowel AF is added.

5.2.1. Domain-dependent MLPs

We use the following MLPs trained on the RM corpus:-

1. *MLP-RM-PH*: An *off-the-shelf* MLP (Dines and Magimai-Doss, 2007) trained on the RM corpus to classify 45 context-independent phonemes.
2. *MLP-RM-AF*: Set of MLPs trained on the RM corpus to classify AFs.
3. *MLP-RM-PH+AF*: The phoneme and the AF MLPs together are referred as *MLP-RM-PH+AF*.

---

3. http://www.cstr.ed.ac.uk/projects/unisyn/
4. http://www.icsi.berkeley.edu/Speech/qn.html

5.2.2. Domain-independent MLPs

Exploiting the resource optimization advantage of probabilistic lexical modeling (Section 2.5) we study the case where the acoustic model is trained on domain-independent data whereas the lexical model is trained on target-domain data. Similarly, a tandem system can use the MLP trained on domain-independent data for feature extraction. In this paper, we use the Wall street journal corpus (Paul and Baker, 1992; Woodland et al., 1994) as domain-independent data to train the MLPs whereas the RM corpus is used as domain-dependent data for which we are interested to build an ASR system. The WSJ corpus has two parts - WSJ0 with 14 hours of speech (7,193 utterances from 84 speakers) and WSJ1 with 66 hours of speech (29322 utterances from 200 speakers). In this paper, we use only the WSJ1 corpus as the domain-independent data. Though RM and WSJ are similar domains, among 1000 words present in the RM task, WSJ corpus includes only 568 words. The phoneme-lexicon for the WSJ corpus was also obtained from the UNISYN lexicon. Therefore, the phoneme sets and the AF sets for the RM and WSJ corpora are identical.

In this paper, we use the following MLPs trained on the WSJ corpus:-

1. *MLP-WSJ-PH*: An *off-the-shelf* MLP (Aradilla et al., 2008) trained on the WSJ corpus to classify 45 context-independent phonemes.
2. *MLP-WSJ-AF*: Set of MLPs trained on the WSJ corpus to classify AFs.
3. *MLP-WSJ-PH+AF*: The phoneme and articulatory MLPs together are refereed as *MLP-WSJ-PH+AF*.

5.2.3. Multistage MLPs

In our recent work, we showed that the AF classification accuracy and thereby the phoneme recognition accuracy can be improved by modeling the inter-feature dependencies using multistage MLP classifiers and/or multitask learning (Rasipuram and Magimai.-Doss, 2011b,a,c). In this paper, we use the multistage MLP classifier illustrated in Figure 4. In the first stage, a set of parallel MLPs are used to estimate posteriors of the AFs. Each MLP receives PLP features as input and is trained to classify a specific AF. In the second stage, to model the temporal contextual information and inter-feature dependencies of AFs, a new set of MLPs are trained using articulatory posteriors estimated by the first stage of MLPs (along with other AFs) with a longer temporal context (eight frame preceding and eight frame following context) as input.

In this paper, the first set of AF MLPs are trained on the WSJ corpus, whereas the second set of AF MLPs are trained on the RM corpus.

1. *MULTI-WSJRM-PH*: The multistage MLP trained on both the WSJ and RM corpora to classify context-independent phonemes.
2. *MULTI-WSJRM-AF*: Set of multistage MLPs trained on both WSJ and RM corpora to classify AFs.
3. *MULTI-WSJRM-PH+AF*: The phoneme and articulatory multistage MLPs together are referred as *MULTI-WSJRM-PH+AF*.

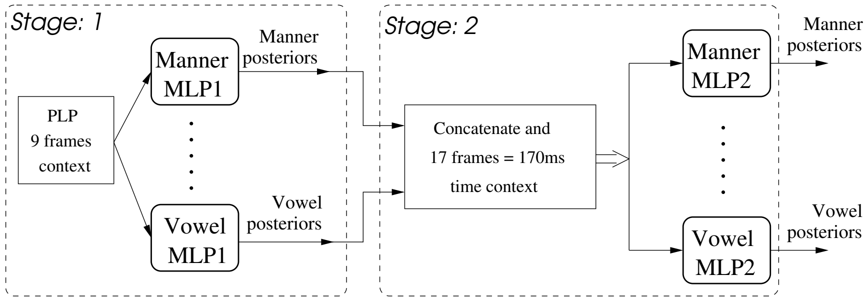


Figure 4: Multistage MLP classifiers as acoustic models

The input, output and hidden layer sizes of all the MLPs used in the paper along with the total number of parameters and frame accuracies are given in Table B.2 of Appendix.

5.3. Systems

We compare the KL-HMM, tandem and HMM/GMM systems. All the systems model crossword context-dependent subword units (either phonemes or graphemes) and each subword unit is modeled as a 3-state HMM. Table 1 summarizes the different systems and their capabilities.

- KL-HMM systems: The acoustic units in the KL-HMM system can be context-independent phonemes, or AFs or both context-independent phonemes and AFs. KL-HMM systems use an MLP as the acoustic model. The lexical units are based on context-dependent phonemes and the lexical model is probabilistic. The lexical model is trained using the local score (S_{SKL} or S_{RKL} or S_{KL}) that results in minimum KL-divergence on the training data compared to other local scores. It was observed that for phoneme-based KL-HMM systems, the local score that resulted in minimum KL-divergence was S_{SKL} whereas for grapheme-based KL-HMM systems it was either S_{SKL} or S_{RKL}.
- Tandem systems: The tandem systems use the MLPs as feature extractors. The MLPs used for feature extraction are the same MLPs that are used as acoustic model in the KL-HMM systems. The output of the MLPs is Gaussianized with log transformation followed by KLT. The dimensionality of the features is reduced by retaining the feature components that contribute to 99% of the variance. The resulting features are used to train a HMM/GMM system where the acoustic units are the clustered context-dependent subword states and the lexical units are based on the context-dependent subword units. The lexical and acoustic units are deterministically related. Each acoustic unit is modeled with an eight mixture Gaussian as it resulted in an optimal ASR performance.
- HMM/GMM systems: The 39-dimensional PLP features used to train the MLP are also used to train the HMM/GMM systems where the acoustic units are clustered context-dependent subword states and the lexical units

are based on context-dependent subword units. The lexical and acoustic units are deterministically related. Each acoustic unit is modeled with an eight mixture Gaussian. The state tying resulted in 1611 tied states for the phoneme-based system and 1536 tied states for the grapheme-based system. Performance in terms of word accuracy on the test set of the RM corpus of the standard crossword context-dependent phoneme subword based HMM/GMM system is 95.9% word accuracy and the crossword context-dependent grapheme subword based HMM/GMM system is 94.8% word accuracy. The performances of the baseline phoneme-based HMM/GMM systems of this paper and the phoneme-based HMM/GMM systems in the literature (Hain and Woodland, 1999; Povey et al., 2011b) on the RM corpus are the same despite of the difference in the phoneme-lexica used. In this work we used a phoneme-lexicon based on the UNISYN lexicon whereas Hain and Woodland (1999) used a phoneme-lexicon based on the CMUDict.

The KL-HMM, tandem and HMM/GMM systems are trained on both the training and development sets (3,990 utterances). The details of various systems such as number of lexical units, acoustic units, dimensionality of tandem features, tied states in tandem systems, total number of parameters in different systems are given in Table B.3. The total number of parameters in the phoneme-based and grapheme-based HMM/GMM systems are about 1.0M.

Table 1: Overview of different systems. *ph* denotes that phonemes are acoustic units, *af* denotes that AFs are acoustic units and *af+ph* denotes that both phonemes and AFs are acoustic units; cCD denotes clustered context-dependent subword states. In the tandem approach, the ANN trained to classify the acoustic units (*ph*, *af* or *af+ph*) is used to extract features for the HMM/GMM system. This is indicated through (ANN) notation. *Det* denotes that the lexical model is deterministic and *Prob* denotes that the lexical model is probabilistic.

System	Acoustic units	Acoustic model	Lexical units	Lexical model
KL-HMM	{ph}, or {af} or {af+ph}	ANN	CD	Prob
Tandem	cCD	(ANN) + GMM	CD	Det
HMM/GMM	cCD	GMM	CD	Det

6. Results

In this section, we compare the KL-HMM and tandem approaches for AF integration.

6.1. Baseline Study

In the baseline study, only the RM corpus is used. Performance of the various systems in terms of word accuracy on the test set of the RM corpus is given in Table 2. The results show that:

- The KL-HMM system using the MLP *MLP-RM-PH* performs better [5] than the KL-HMM system using the MLPs *MLP-RM-AF* as the acoustic model. The KL-HMM system using the MLP *MLP-RM-PH-AF* performs better than the KL-HMM systems using either the MLP *MLP-RM-PH* or the MLP *MLP-RM-AF*.
- Performance of the tandem systems compared to the KL-HMM systems is similar [6] when the MLPs *MLP-RM-PH* and *MLP-RM-AF* are used. However, the performance of the KL-HMM system using the MLP *MLP-RM-PH+AF* is better than the tandem system using the same MLP.
- Performance of the KL-HMM system using the MLP *MLP-RM-PH+AF* is better than all other systems. However, the tandem system using the MLP *MLP-RM-PH+AF* does not perform better than other tandem systems.
- Performances of the KL-HMM and tandem systems using the MLP *MLP-RM-PH* and the baseline HMM/GMM system are similar.

Table 2: Performance in terms of word accuracy on the test set of the RM corpus for phoneme-based KL-HMM and tandem systems. The MLPs *MLP-RM-PH*, *MLP-RM-AF* and *MLP-RM-PH+AF* are trained on the RM corpus.

MLP	System	
	KL-HMM	Tandem
MLP-RM-PH	95.6	95.7
MLP-RM-AF	94.9	95.3
MLP-RM-PH+AF	96.2	95.3

The results show that the performance of the KL-HMM and tandem systems using AF-based MLPs is worse than the respective systems using phoneme-based MLPs. This difference in performance between the systems could be due to the difficulty in estimating articulatory positions directly from speech. Estimating AFs from speech or (the acoustic-to-AF mapping) is challenging because multiple articulatory configurations can produce an identical acoustic realization. In this paper, acoustic-to-AF map is performed using MLPs. MLPs are data-driven estimators and therefore the classification accuracy of an MLP may depend on the amount of training data used. In this section, the AF-based MLPs are trained on a limited amount of domain-dependent data. In the next section, we will verify if by improving the AF-based MLPs (or the acoustic models) using a larger domain-independent data set could benefit the KL-HMM and tandem systems.

5. In the paper, “better” implies that the difference in performance between the compared systems is statistically significant with confidence greater than 95.0%. Statistical significance tests for the systems are performed using the approach proposed by Bisani and Ney (2004).

6. In the paper, “similar” implies that the difference in performance between the compared systems is not statistically significant.

6.2. Domain-independent MLPs

In this case, the KL-HMM and tandem systems use the MLPs trained on the domain-independent WSJ corpus. Performance of the various systems in terms of word accuracy on the test set of the RM corpus is given in Table 3. The results show that:

- The KL-HMM system using the MLPs *MLP-WSJ-AF* performs better than the KL-HMM system using the MLPs *MLP-RM-AF*. Whereas the KL-HMM systems using the MLPs *MLP-WSJ-PH* or *MLP-WSJ-PH+AF* perform similarly to the KL-HMM systems using the MLPs *MLP-RM-PH* or *MLP-RM-PH+AF*, respectively.
- The performance of the KL-HMM systems using the MLPs *MLP-WSJ-PH* and *MLP-WSJ-AF* is similar.
- Performance of the tandem systems using the MLPs *MLP-WSJ-PH* or *MLP-WSJ-AF* or *MLP-WSJ-PH-AF* is similar to the tandem systems using the MLPs *MLP-RM-PH* or *MLP-RM-AF* or *MLP-RM-PH-AF* respectively.
- Similar to the baseline study, performance of the KL-HMM system using the MLPs *MLP-WSJ-PH+AF* is better than all other systems, however, the tandem system using the MLPs *MLP-WSJ-PH+AF* does not perform better than the other KL-HMM and tandem systems.

Results show that the KL-HMM system using AF-based MLPs trained on a large set of domain-independent data performs better than the KL-HMM system using the AF-based MLPs trained on a small set of domain-dependent data. Furthermore, the performance gap between the KL-HMM system using phonemes as acoustic units and the KL-HMM system using AFs as acoustic units has reduced and the two systems perform similarly after the MLPs are trained on a large domain-independent data. However, none of the tandem systems could benefit from the MLPs trained on the domain-independent data.

Table 3: Performance in terms of word accuracy on the test set of RM corpus for phoneme-based KL-HMM and tandem systems. The MLPs *MLP-WSJ-PH*, *MLP-WSJ-AF* and *MLP-WSJ-PH+AF* are trained on the domain-independent WSJ corpus.

MLP	System	
	KL-HMM	Tandem
MLP-WSJ-PH	95.9	95.8
MLP-WSJ-AF	95.5	95.4
MLP-WSJ-PH+AF	96.4	94.9

6.3. Multistage MLPs

Performance of the KL-HMM and tandem systems using multistage MLPs in terms of word accuracy on the test set of the RM corpus is given in Table 4. Results show that the performance of the KL-HMM and tandem systems using the multistage MLPs *MULTI-WSJRM-PH*, *MULTI-WSJRM-AF*, and *MULTI-WSJRM-PH+AF* is similar to the respective systems using the MLPs *MLP-WSJ-PH*, *MLP-WSJ-AF* and *MLP-WSJ-PH+AF*.

Table 4: Performance in terms of word accuracy on the test set of the RM corpus for phoneme-based KL-HMM and tandem systems. The multistage MLPs *MULTI-WSJRM-PH*, *MULTI-WSJRM-AF* and *MULTI-WSJRM-PH+AF* are trained on both RM and WSJ corpora.

MLP	System	
	KL-HMM	Tandem
MULTI-WSJRM-PH	96.1	96.0
MULTI-WSJRM-AF	95.4	95.2
MULTI-WSJRM-PH+AF	96.6	95.1

6.4. Grapheme Subword Units

In this study, the MLPs used in previous sections are used to train the crossword context-dependent grapheme-based KL-HMM and tandem systems. Performance of the grapheme-based KL-HMM and tandem systems using the various MLPs in terms of word accuracy on the test set of the RM corpus is given in Table 5. The results show that:

- Performance of the grapheme-based KL-HMM systems using the MLPs *MLP-WSJ-PH*, *MLP-WSJ-AF* and *MLP-WSJ-PH-AF* is similar to the systems using the MLPs *MLP-RM-PH*, *MLP-RM-AF* and *MLP-RM-PH-AF*, respectively.
- Unlike phoneme-based KL-HMM systems of Section 6.3, performance of the grapheme-based KL-HMM systems using the multistage MLPs *MULTI-WSJRM-AF* and *MULTI-WSJRM-PH-AF* is better than the systems using the MLPs *MLP-WSJ-AF* and *MLP-WSJ-PH-AF*, respectively.
- The grapheme-based KL-HMM systems using the MLPs *MULTI-WSJRM-PH* and *MULTI-WSJRM-AF* perform similarly.
- The performance gap between grapheme-based and phoneme-based KL-HMM systems is greatly reduced when multistage MLPs are used.
- Similar to the phoneme-based KL-HMM systems, performance of the KL-HMM systems that use both phoneme and AF acoustic units together is consistently better than the KL-HMM systems using either of them as acoustic units.
- The KL-HMM and tandem systems perform similarly when the MLPs *MLP-RM-PH*, *MLP-RM-AF*, *MLP-WSJ-AF*, *MULTI-WSJRM-PH*, and *MULTI-WSJRM-AF* are used. Furthermore, the KL-HMM systems using the MLPs *MLP-RM-PH-AF*, *MLP-WSJ-PH-AF* and *MULTI-WSJRM-PH-AF* perform better than the tandem systems using the same MLPs. These two observations are consistent with the phoneme-based systems.

6.5. Summary of the Experimental Results

To summarize, the following conclusions can be drawn from the experimental study:

Table 5: Performance in terms of word accuracy on the test set of RM corpus for grapheme-based KL-HMM and tandem systems.

MLP	System	
	KL-HMM	Tandem
MLP-RM-PH	94.9	94.6
MLP-RM-AF	94.4	94.1
MLP-RM-PH+AF	95.8	94.3
MLP-WSJ-PH	95.5	94.5
MLP-WSJ-AF	94.9	94.8
MLP-WSJ-PH+AF	95.9	93.9
MULTI-WSJRM-PH	95.8	95.6
MULTI-WSJRM-AF	95.8	95.2
MULTI-WSJRM-PH+AF	96.4	95.0

1. The proposed approach for AF integration resulted in an ASR system that performs similar to the tandem approach for AF integration.
   Though both approaches perform similarly, the primary advantage of the proposed approach compared to the tandem approach comes from the fact that the articulatory representations are kept intact in the model parameters in the form of probabilistic phoneme-to-AF or grapheme-to-AF relationship learned from the transcribed speech data. Furthermore, as we will see in the next section, the approach also adapts the knowledge-based phoneme-to-AF and grapheme-to-AF relationship on the transcribed speech data, and allows different AFs to evolve asynchronously. However, the tandem approach tends to lose the two primary benefits of articulatory representation i.e., finer granularity and asynchronous evolution.
2. The performance of the grapheme- or phoneme-based KL-HMM systems using both AFs and phonemes as acoustic units is always better (about 10 to 12% relative reduction in WER) than the KL-HMM system using either of them as acoustic units. Furthermore, the proposed approach performs better than the tandem approach if both AFs and phonemes are used as acoustic units. We speculate the following two reasons:
   (a) When both AFs and phonemes are used as acoustic units, not only the acoustic model is improved but also the lexical model, as both probabilistic phoneme-to-phoneme and phoneme-to-AF (or grapheme-to-phoneme and grapheme-to-AF) relationships are modeled.
   (b) The AF-based and phoneme-based MLPs are trained independently. However, the probabilistic phoneme-to-phoneme and phoneme-to-AF (or grapheme-to-phoneme and grapheme-to-AF) relationships are learned together during lexical model training. Therefore, the approach can learn the inter-feature dependencies among various AFs and phonemes or graphemes.

The ASR studies also indicate that the information captured by the lexical model with phonemes as acoustic units and with AFs as acoustic units is complementary, hence, the combined use of phonemes and AFs as acoustic units significantly improves the performance compared to the use of either of them.

3. The results indicate that the use of a larger domain-independent data set helps both phoneme-based and grapheme-based KL-HMM systems using AFs as acoustic units.
4. In our previous phoneme recognition studies it was observed that the multistage articulatory MLPs improved the phoneme recognition accuracy (Rasipuram and Magimai.-Doss, 2011b). In this paper, multistage AF-based MLPs did not improve the performance of the phoneme-based KL-HMM systems but improved the performance of the grapheme-based KL-HMM systems. We conjecture the following two reasons for the observed trends. Firstly, the multistage AF-based MLP was motivated from the work by Pinto et al. (2011). In that work it was shown that the second MLP in the multistage MLP classifier learns the phonetic temporal patterns (i.e., the phonetic confusions at the output of the first MLP) and the phonotactics of the language observed in the training data. In our case, the second set of MLPs in the multistage AF-based MLP classifier could model phonotactic constraints at the articulatory feature level.
   - In the present work, we model context-dependent phonemes as lexical units which incorporates phonotactic constraints at the lexical model level. The results indicate that it may be redundant to model the phonotactic constraints twice, once at the acoustic model level and again at the lexical model level.
   - However, in the case of context-dependent grapheme-based KL-HMM systems, the lexical model is modeling graphemic constraints and the acoustic model is modeling phonotactic constraints, which could be complementary to each other especially given the fact that the grapheme-to-phoneme relationship in English is irregular.

   Furthermore, though the frame accuracy and phoneme recognition accuracy are considered as important factors for word recognition, they may not be the only indicators of word level performance (Greenberg et al., 2000). The relationship between frame accuracy and word accuracy depend on more than one factor: the pronunciation lexicon, the acoustic model, the lexical model and the language model components of an ASR system. In other words, lexical constraints and syntactic constraints incorporated while decoding can handle the shortcomings of the acoustic model or may render some of the gains obtained through the acoustic model redundant.

7. Analysis

In the proposed approach, with phonemes or graphemes as lexical units and AFs as acoustic units, the parameters of the lexical model capture a probabilistic phoneme-to-AF or grapheme-to-AF relationship, respectively. In this section, we analyze the parameters of the lexical model to understand the following:

- Is the phoneme-to-AF or grapheme-to-AF relationship captured by the lexical model parameters close to the knowledge-based phoneme-to-AF or grapheme-to-AF relationship?
- Does the model allow different AFs to evolve asynchronously?

The target labels for the MLPs with AFs as output units are obtained from the knowledge-based phoneme-to-AF map. Hence, during training, the MLPs do not account for the AFs changing asynchronously. However, as shown by King and Taylor (2000), it is typical that at the output of the MLP, different AFs change at different times (especially at the phoneme boundaries) and exhibit asynchronous behaviour. Since the KL-HMM system models the output of a set of AF-based MLPs, we hypothesize that the lexical model parameters can capture asynchronous AF behaviour especially at the HMM states modeling subword unit boundaries (i.e., the first and third HMM states of a subword unit).

7.1. Subword Level Analysis

Table B.4 shows the manner and place of articulation for context-independent phonemes in three HMM states captured by the lexical model parameters. The analysis is presented only on manner and place of articulation for the sake of simplicity. However, similar trends are observed even when height of articulation is included. The denoted AF values correspond to the dimension with maximum probability captured by the lexical model parameters of context-independent subword units. The first part of the table presents a few context-independent phonemes where the manner and place of articulation do not change (i.e., are synchronous) between three HMM states, and the second part of the table presents all the context-independent phonemes where the manner and place of articulation evolved asynchronously. It can be observed that the phoneme-to-AF relationship of Table B.4 relates well with the knowledge-based relationship given in Table B.1

Table B.4 indicates that for context-independent phonemes that are diphthongs such as /aw/, /ay/, /ow/, and /oy/ and for phonemes that are vowels such as /ah/, /aw/, and /uh/ the captured place of articulation changed between the HMM states whereas the captured manner of articulation is always a “vowel”. For phonemes that are stops i.e., /b/, and /g/, the captured place of articulation is the same across the three HMM states whereas the captured manner of articulation changed between the HMM states. More specifically, the initial states of /b/ and /g/ captured a “stop” and the third state captured a “vowel”. For phonemes /ch/ and /jh/, the captured manner of articulation changed at the second HMM state from “stop” to “fricative” whereas the captured place of articulation changed at the first HMM state from “alveolar” to “front”.

We have computed the percentage of context-dependent phonemes where the lexical model parameters exhibited asynchrony between manner and place of articulation at the HMM state level. A context-dependent phoneme model is said to be synchronous if manner and place of articulation change at the same state transition or remain the same across three HMM states. A context-dependent phoneme model is said to be asynchronous if manner and place of articulation change at different HMM states. For example, in Table 6, the context-dependent phonemes /aa-n+t/ and /ah-z+sh/ exhibit synchronous changes in manner and place of articulation whereas the phonemes /ao-d+ey/ and /ao-hh+aa/ exhibit asynchronous changes at the HMM state level.

Table 6: Examples of the context-dependent phonemes that exhibit synchronous and asynchronous manner and place of articulation at the HMM state level.

Synchronous Examples					**Asynchronous Examples**				
Phoneme	**AF**	**st1**	**st2**	**st3**	**Phoneme**	**AF**	**st1**	**st2**	**st3**
aa-n+t	Manner	nas			ao-d+ey	Manner	stp	vow	
	Place	alv				Place	alv		
ah-z+sh	Manner	vow	frc		ao-hh+aa	Manner	vow	asp	vow
	Place	midf	frt			Place	bck		

In Table 7, the first column indicates the set of MLPs used to train the KL-HMM system, and the second and the third columns indicate the percentage of context-dependent phonemes where the changes in manner and place of articulations at the HMM state level are synchronous and asynchronous, respectively. It is important to note that the classification of context-dependent phoneme models in terms of synchronous and asynchronous does not take into account the errors in the phoneme-to-AF map captured by the lexical model parameters. For example, in Table 9, the place and height of articulation for the grapheme model [W] are asynchronous. However, the captured place of articulation in the first state of [W] as “lateral” could be considered as an error. Therefore, the percentage of context-dependent phoneme models exhibiting synchronous or asynchronous behaviour are only an indicative of the asynchronous nature of the context-dependent phoneme models.

Table 7 indicates that asynchronous articulatory movements among manner and place of articulation are relatively lower when multistage AF-based MLP classifiers are used. We argue the following two reasons for this. Firstly, as discussed in Section 6.5, the second stage of AF-based MLPs in the multistage MLP classifiers model the phonotactics of the language, therefore various AFs may be more synchronous. Secondly, the frame accuracies of the multistage AF-based MLP classifiers are better than the frame accuracies of other MLPs. As a result the number of context-dependent phonemes exhibiting asynchronous changes due to the errors in the captured relationship could be relatively less.

The analysis of the parameters indicated that the model is able to capture asynchronous AF configurations at the subword unit level. In the next section we will see that asynchronous articulatory configurations are more meaningful at the word level as the context-dependent subword models also capture information of the neighbouring phonemes.

Table 7: For various KL-HMM systems, the percentage of context-dependent phonemes where the changes in manner and place of articulations between the three HMM states are synchronous and asynchronous.

MLP	Synchronous	Asynchronous
MLP-RM-AF	57.64%	42.35%
MLP-WSJ-AF	59.30%	40.69%
MULTI-WSJRM-AF	63.57%	36.42%

7.2. Word Level Analysis

The phoneme-to-AF and grapheme-to-AF relationships captured by the lexical model parameters of phoneme-based and grapheme-based KL-HMM systems for the word "BELOW" are given in Tables 8 and 9, respectively. The tables indicate the manner, place and height of articulation.

Table 8: The phoneme-to-AF relationship captured by the lexical model parameters of context-dependent phonemes for the word "BELOW". Pronunciation of the word BELOW in the phoneme lexicon is /b/ /ih/ /l/ /ow/. AF values for manner, place and height of articulation are given.

Phoneme	b			ih			l			ow		
State	st1	st2	st3	st1	st2	st3	st1	st2	st2	st1	st2	st2
Manner	sil	stp	vow				app			vow		
Place	sil	lab		midf		lat				bck		midb
Height	max		high			vhi				mid		vhi

Table 9: The grapheme-to-AF relationship captured by the lexical model parameters of context-dependent graphemes for the word "BELOW". Pronunciation of the word BELOW in the grapheme lexicon is [B] [E] [L] [O] [W]. AF values for manner, place and height of articulation are given.

Grapheme	[B]			[E]			[L]			[O]			[W]		
State	st1	st2	st3	st1	st2	st3	st1	st2	st2	st1	st2	st2	st1	st2	st3
Manner	stp		vow				app		vow						
Place	sil	lab	midf			lat				bck			*lat*	midb	sil
Height	max		high			vhi				mid		*midl*	vhi		sil

It can be observed that the phoneme-to-AF relationship of Table 8 relates well with the knowledge-based relationship given in Table B.1. Table 9 shows that even if subword units are graphemes, articulatory patterns similar to the system using phoneme subword units are captured. The two differences between grapheme-to-AF and phoneme-to-AF are indicated in red italic font in the Table 9. The number of subword units in the pronunciation of the word "BELOW" are five in the case of graphemes and four in the case of phonemes. It can be observed from the table that this irregularity in the grapheme pronunciation has been taken care, as the sequence of graphemes [O] and [W] together capture the information of phoneme /ow/.

Furthermore, the tables also indicate that various AFs evolve asynchronously. For example, in Table 8, the captured manner of articulation in the second state of /b/ is "stop" and in the third state of /b/ is "vowel", whereas the place of articulation in both second and third states of /b/ is "labial".

It was observed by Magimai.-Doss et al. (2011); Rasipuram (2014) that the lexical model parameters of the context-dependent graphemes tend to model the transition information to the next context-dependent grapheme in the sequence. Similar observations can also be made from Tables 8 and 9, but at the finer articulatory feature level. For example, in Table 8, the captured manner of articulation in the third state of /b/ is "vowel" which corresponds to the next phoneme in the sequence i.e., /ih/. Similarly in Table 9, the captured manner, place and height of articulation in the third state of grapheme [B] correspond to the next grapheme in the sequence i.e., [E]. The analysis shows that the lexical model parameters of the context-dependent phonemes and graphemes are capable of capturing some information about preceding and following articulatory configurations.

8. Discussion and Conclusion

In this paper, we proposed an approach to integrate articulatory feature representations into HMM-based ASR in the framework of probabilistic lexical modeling. The proposed approach involves two stages: acoustic model and lexical model. The acoustic model is a posterior probability estimator that models the relationship between acoustic feature observations and articulatory features. The lexical model, models a probabilistic relationship between lexical units and articulatory features. The approach has the following potential advantages:

- Lexical access: As opposed to knowledge-based approaches, the parameters of the lexical model in the proposed approach are learned using transcribed speech data by training a HMM whose states represent lexical units and the parameters of the state capture a probabilistic relationship between a lexical unit and articulatory features. Consequently, the approach integrates a model for lexical access using articulatory features into the HMM-based ASR framework.
- Asynchrony of AFs: As observed in Section 7, the model also allows different AFs to evolve asynchronously. Thus, overcoming some of the limitations of knowledge-based approaches.
- Combination of various AFs: A challenge often faced in using articulatory features for ASR is the combination of evidences from different AFs. In that regard, the proposed approach can be seen as a multi-channel approach where each AF serves as a separate channel and various AFs are combined at the local score computation level. Also, as seen in this paper, the multi-channel approach can be trivially extended to combine other relevant information such as the phoneme information.

Our investigations on a continuous speech recognition task have shown that the proposed approach effectively integrates articulatory features into the HMM-based ASR framework; improves ASR performance if combined with phoneme-based acoustic models; and offers flexibility to use either phonemes or graphemes as subword units.

The probabilistic grapheme-to-AF relationship captured in the lexical model parameters of the KL-HMM system with acoustic units as AFs and lexical units based on context-dependent graphemes can be exploited to generate AF-based pronunciation lexica using the acoustic data-driven grapheme-to-phoneme conversion approach proposed by Rasipuram and Magimai.-Doss (2012). The AF-based pronunciation lexica can be used in DBN-based approaches for AF integration that require such lexica (Livescu and Glass, 2004; Livescu et al., 2008) or in AF-based text-to-speech synthesis systems.

In this paper, we focussed mainly on the integration of articulatory features into an ASR system. More precisely, we focussed on the lexical model aspect of the proposed approach. The three-layer or multistage MLPs classifying context-independent articulatory features (or phonemes) were used as acoustic models. The approach can be potentially improved by improving the acoustic model or articulatory feature estimators along the following directions:

1. Context-dependent articulatory features: The articulatory feature acoustic model could be improved by considering context-dependent articulatory features that take into account the neighbouring articulatory context.
2. Deep architectures for AF estimation: More recently, ANNs with deep architectures have gained lot of attention (Dahl et al., 2012; Hinton et al., 2012). In similar vein, the articulatory feature model can be based on deep ANN architectures (Siniscalchi et al., 2012).

In this paper, we have shown that the AF-based acoustic models can be trained on domain-independent data where as the lexical model can be trained on domain-dependent data. In our recent work we found that in the framework of probabilistic lexical modeling, the acoustic model can be trained on language-independent resources and the lexical model on a relatively small amount of language-dependent data (Rasipuram and Magimai.-Doss, 2014). Articulatory features are considered to be more language-independent and effective for cross-linguistic adaptation (Lal and King, 2013; Siniscalchi et al., 2012). Therefore, we hypothesize that the use of articulatory feature based language-independent acoustic model in the proposed approach can offer potential advantages in building ASR systems for under-resourced and minority languages [7]. Our future work will focus on extending the proposed approach along this direction.

Acknowledgment

This work was supported by the Swiss NSF through the grants "Flexible Grapheme-Based Automatic Speech Recognition (FlexASR)" and the National Center of Competence in Research (NCCR) on "Interactive Multimodal Information Management" (www.im2.ch).

7. A language that lacks one or more resources required to build an ASR system is referred to as under-resourced language (Besacier et al., 2014). Minority languages are languages spoken by minority of population. A minority language need not be under-resourced and an under-resourced language may or not be a minority language

References

Aradilla, G., 2008. Acoustic Models for Posterior Features in Speech Recognition. Ph.D. thesis. EPFL. Switzerland.

Aradilla, G., Bourlard, H., Magimai.-Doss, M., 2008. Using KL-Based Acoustic Models in a Large Vocabulary Recognition Task , in: Proc. of Interspeech, pp. 928–931.

Aradilla, G., Vepa, J., Bourlard, H., 2007. An Acoustic Model Based on Kullback-Leibler Divergence for Posterior Features, in: Proc. of ICASSP, pp. 657–660.

Bahl, L., Brown, P., de Souza, P., Picheny, M., 1988. Acoustic Markov models used in the Tangora speech recognition system, in: Proc. of ICASSP, pp. 497–500 vol.1.

Bellegarda, J., Nahamoo, D., 1990. Tied mixture continuous parameter modeling for speech recognition. IEEE Trans. on Acoustics, Speech, and Signal Processing 38, 2033–2045.

Besacier, L., Barnard, E., Karpov, A., Schultz, T., 2014. Automatic Speech Recognition for Under-resourced Languages: A Survey. Speech Communication 56, 85–100.

Bisani, M., Ney, H., 2004. Bootstrap estimates for confidence intervals in ASR performance evaluation, in: Proc. of ICASSP, pp. I–409–12.

Cetin, O., Kantor, A., King, S., Bartels, C., Magimai-Doss, M., Frankel, J., Livescu, K., 2007a. An Articulatory Feature-Based Tandem Approach and Factored Observation Modeling, in: Proc. of ICASSP, pp. 645–648.

Cetin, O., Magimai-Doss, Livescu, K., Kantor, A., King, S., Bartels, C., Frankel, J., 2007b. Monolingual and crosslingual comparison of tandem features derived from articulatory and phone MLPS, in: Proc. of ASRU, pp. 36–41.

Chang, S., 2002. A Syllable, Articulatory-Feature, and Stress-Accent Model of Speech Recognition. Ph.D. thesis. University of California, Berkeley.

Chomsky, N., Halle, M., 1968. The Sound Pattern of English. MIT Press.

Dahl, G., Yu, D., Deng, L., Acero, A., 2012. Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition . IEEE Trans. on Audio, Speech, and Language Processing 20, 30–42.

Deng, L., Ramsay, G., Sun, D., 1997. Production Models As a Structural Basis for Automatic Speech Recognition. Speech Communication 22, 93–111.

Dines, J., Magimai-Doss, M., 2007. A Study of Phoneme and Grapheme based Context-Dependent ASR Systems, in: Proc. of Machine Learning for Multimodal Interaction (MLMI), pp. 215–226.

Frankel, J., King, S., 2005. A Hybrid ANN/DBN Approach to Articulatory Feature Recognition, in: Proc. of Interspeech, pp. 3045–3048.

Frankel, J., Wester, M., King, S., 2007. Articulatory feature recognition using dynamic Bayesian networks, in: Computer Speech & Language, pp. 620–640.

Greenberg, S., Chang, S., Hollenback, J., 2000. An introduction to the diagnostic evaluation of Switchboard-corpus automatic speech recognition systems, in: Proceedings of the NIST Speech Transcription Workshop.

Hain, T., 2005. Implicit Modelling of Pronunciation Variation in Automatic Speech Recognition. Speech Communication 46, 171–188.

Hain, T., Woodland, P.C., 1999. Dynamic HMM Selection for Continuous Speech Recognition, pp. 1327–1330.

Harris, J., 1994. English Sound Structure. Blackwell.

Hermansky, H., Ellis, D., Sharma, S., 2000. Tandem connectionist feature extraction for conventional HMM systems, in: Proc. of ICASSP, pp. 1635–1638.

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A.R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., Kingsbury, B., 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine 29, 82–97.

Hiroya, S., Honda, M., 2004. Estimation of articulatory movements from speech acoustics using an HMM-based speech production model. IEEE Trans. on Speech and Audio Processing 12, 175–185.

Hogden, J., Lofqvist, A., Gracco, V., Zlokarnik, I., Rubin, P., Saltzman, E., 1996. Accurate recovery of articulator positions from acoustics: New conclusions based on human data. The Journal of The Acoustical Society of America 100, 1819–1834.

Hosom, J., 2009. Speaker-independent phoneme alignment using transition-dependent states. Speech Communication 51, 352–368.

Huang, X., Jack, M., 1989. Semi-continuous hidden Markov models for speech signal. Computer Speech and Language 3, 239–251.

Huttenlocher, D., Zue, V., 1984. A model of lexical access from partial phonetic information, in: Proc. of ICASSP, pp. 391–394.

Hwang, M.Y., Huang, X., 1992. Subphonetic modeling with Markov states-Senone, in: Proc. of ICASSP, pp. 33–36 vol.1.

Imseng, D., Dines, J., Motlicek, P., Garner, P., , Bourlard, H., 2012. Comparing different acoustic modeling techniques for multilingual boosting, in: Proc. of Interspeech.

Juneja, A., Espy-Wilson, C., 2004. Significance of invariant acoustic cues in a probabilistic framework for landmark-based speech recognition, in: From Sound to Sense: Fifty+ Years of Discoveries in Speech Communication, MIT, Cambridge, Massachusetts.

Jyothi, P., 2013. Discriminative And Articulatory Feature-Based Pronunciation Models For Conversational Speech Recognition. Ph.D. thesis. The Ohio State University, Ohio.

Jyothi, P., Fosler-lussier, E., Livescu, K., 2013. Discriminative training of wfst factors with application to pronunciation modeling, in: Proc. of Interspeech, pp. 1961–1965.

Jyothi, P., Livescu, K., Fosler-Lussier, E., 2011. Lexical access experiments with context-dependent articulatory feature-based models, in: Proc. of ICASSP, pp. 4900–4903.

King, S., Frankel, J., Livescu, K., McDermott, E., Richmond, K., Wester, M., 2007. Speech production knowledge in automatic speech recognition . The Journal of The Acoustical Society of America 121, 723–742.

King, S., Taylor, P., 2000. Detection of Phonological Features in Continuous Speech using Neural Networks. Computer Speech and Language 14, 333–353.

Kirchhoff, K., 1996. Syllable-level desynchronisation of phonetic features for speech recognition, in: Proceedings of Int. Conf. Spoken Language Processing.

Kirchhoff, K., Fink, G.A., Sagerer, G., 2002. Combining acoustic and articulatory feature information for robust speech recognition . Speech Communication 37, 303–319.

Ladefoged, P., 1993. A Course in Phonetics. Harcourt Brace College Publishers.

Lal, P., King, S., 2013. Cross-Lingual Automatic Speech Recognition Using Tandem Features. IEEE Trans. on Audio, Speech, and Language Processing 21, 2506–2515.

Livescu, K., Cetin, O., Hasegawa-Johnson, M., King, S., Bartels, C., Borges, N., Kantor, A., Lal, P., Yung, L., Bezman, A., Dawson-Haggerty, S., Woods, B., 2008. Articulatory Feature-based Methods for Acoustic and Audio-Visual Speech Recognition: 2006 JHU Summer Workshop Final Report. Technical Report. John Hopkins University, Center for Language and Speech Processing.

Livescu, K., Glass, J.R., 2004. Feature-based pronunciation modeling with trainable asynchrony probabilities, in: Proceedings of Int. Conf. Spoken Language Processing.

Luo, X., Jelinek, F., 1999. Probabilistic Classification of HMM States for Large Vocabulary Continuous Speech Recognition, in: Proc. of ICASSP, pp. 353–356.

Magimai.-Doss, M., Rasipuram, R., Aradilla, G., Bourlard, H., 2011. Grapheme-based Automatic Speech Recognition using KL-HMM, in: Proc. of Interspeech, pp. 2693–2696.

Metze, F., Waibel, A., 2002. A flexible stream architecture for ASR using articulatory features, in: Proceedings of Int. Conf. Spoken Language Processing.

Morgan, N., Bourlard, H., 1995. Continuous speech recognition: An introduction to the hybrid HMM/connectionist approach. IEEE Signal Processing Magazine , 25–42.

Næss, A.B., Livescu, K., Prabhavalkar, R., 2011. Articulatory Feature Classification Using Nearest Neighbors, in: Proc. of Interspeech, pp. 2301–2304.

Paul, D., Baker, J., 1992. The Design for the Wall Street Journal-based CSR Corpus, in: DARPA Speech and Language Workshop, Morgan Kaufmann Publishers.

Pinto, J., Sivaram, G.S.V.S., Magimai.-Doss, M., Hermansky, H., Bourlard, H., 2011. Analysis of MLP Based Hierarchical Phoneme Posterior Probability Estimator. IEEE Trans. on Audio, Speech, and Language Processing 19, 225–241.

Povey, D., Burget, L., Agarwal, M., Akyazi, P., Kai, F., Ghoshal, A., Glembek, O., Goel, N., Karafiát, M., Rastrow, A., Rose, R., Schwarz, P., Thomas, S., 2011a. The subspace Gaussian mixture model - A structured model for speech recognition. Computer Speech and Language .

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K., 2011b. The Kaldi Speech Recognition Toolkit, in: Proc. of ASRU.

Prabhavalkar, R., Fosler-Lussier, E., Livescu, K., 2011. A factored conditional random field model for articulatory feature forced transcription, in: Proc. of ASRU, pp. 77–82.

Price, P.J., Fisher, W., Bernstein, J., 1988. The DARPA 1000-word resource management database for continuous speech recognition, in: Proc. of ICASSP, pp. 651–654.

Rabiner, L., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. of IEEE 77, 257–286.

Rasipuram, R., 2014. Grapheme-based Automatic Speech Recognition using Probabilistic Lexical Modeling. Ph.D. thesis. EPFL. Switzerland.

Rasipuram, R., Magimai.-Doss, M., 2011a. Improving articulatory feature and phoneme recognition using multitask learning, in: Artificial Neural Networks and Machine Learning - ICANN 2011.

Rasipuram, R., Magimai.-Doss, M., 2011b. Integrating articulatory features using kullback-leibler divergence based acoustic model for phoneme recognition, in: Proc. of ICASSP.

Rasipuram, R., Magimai.-Doss, M., 2011c. Multitask Learning to Improve Articulatory Feature Estimation and Phoneme Recognition. Idiap Research Report, Idiap-RR-21-2011.

Rasipuram, R., Magimai.-Doss, M., 2012. Acoustic Data-driven Grapheme-to-Phoneme Conversion using KL-HMM, in: Proc. of ICASSP.

Rasipuram, R., Magimai.-Doss, M., 2013a. Improving grapheme-based asr by probabilistic lexical modeling approach, in: Proceedings of Interspeech.

Rasipuram, R., Magimai.-Doss, M., 2013b. Probabilistic Lexical Modeling and Grapheme-based Automatic Speech Recognition. `http://publications.idiap.ch/downloads/reports/2013/Rasipuram_Idiap-RR-15-2013.pdf`. Idiap Research Report.

Rasipuram, R., Magimai.-Doss, M., 2013c. Probabilistic Lexical Modeling and Grapheme-based Automatic Speech Recognition. Idiap Research Report, Idiap-RR-15-2013.

Rasipuram, R., Magimai.-Doss, M., 2014. Acoustic and Lexical Resource Constrained ASR using Language-Independent Acoustic Model and Language-Dependent Probabilistic Lexical Model. `http://publications.idiap.ch/downloads/reports/2014/Rasipuram_Idiap-RR-02-2014.pdf`. Idiap Research Report.

Richardson, M., Bilmes, J., Diorio, C., 2003. Hidden-Articulatory Markov Models for Speech Recognition. Speech Communication 41, 713–716.

Rottland, J., Rigoll, G., 2000. Tied Posteriors: An Approach for Effective Introduction of Context Dependency in Hybrid NN/HMM LVCSR, in: Proc. of ICASSP, pp. 1241–1244.

Saraclar, M., Nock, H., Khudanpur, S., 2000. Pronunciation Modeling By Sharing Gaussian Densities Across Phonetic Models, in: Computer Speech and Language, pp. 137–160.

Scharenborg, O., Wan, V., Moore, R.K., 2007. Towards capturing fine phonetic variation in speech using articulatory features. Speech Communication 49, 811–826.

Siniscalchi, S., Lyu, D.C., Svendsen, T., Lee, C.H., 2012. Experiments on Cross-Language Attribute Detection and Phone Recognition With Minimal Target-Specific Training Data. IEEE Trans. on Audio, Speech, and Language Processing 20, 875–887.

Strik, H., Cucchiarini, C., 1999. Modeling pronunciation variation for ASR: A survey of the literature. Speech Communication 29, 225–246.

Stüker, S., Metze, F., Schultz, T., Waibel, A., 2003. Integrating Multilingual Articulatory Features into Speech Recognition, in: Proc. of Eurospeech, pp. 1033–1036.

Suzuki, S., Okadome, T., Honda, M., 1998. Determination of articulatory positions from speech acoustics by applying dynamic articulatory constraints, in: Proceedings of Int. Conf. Spoken Language Processing.

T.Holter, T.Svendsen, 1997. Incorporating Linguistic Knowledge And Automatic Baseform Generation In Acoustic Subword Unit Based Speech Recognition.

Wester, M., Frankel, J., King, S., 2004. Asynchronous articulatory feature recognition using dynamic Bayesian networks, in: Proc. of IEICI Beyond HMM Workshop.

Woodland, P.C., Odell, J.J., Valtchev, V., Young, S.J., 1994. Large Vocabulary Continuous Speech Recognition using HTK , in: Proc. of ICASSP, pp. 125–128.

Young, S.J., Odell, J.J., Woodland, P.C., 1994. Tree-based State Tying for High Accuracy Acoustic Modelling, in: Proceedings of the Workshop on Human Language Technology (HLT), pp. 307–312.

Appendix A. Parameter Estimation and Decoding in the KL-HMM Approach

Given a trained ANN and a training set of N utterances $\{X(n), W(n)\}_{n=1}^{N}$, the set of acoustic unit probability vectors $\{Z(n), W(n)\}_{n=1}^{N}$ are estimated. For each training utterance n, $X(n)$ represents the sequence of cepstral features of length $T(n)$, $W(n)$ represents the sequence of underlying words and $Z(n)$ represents a sequence of acoustic unit probability vectors of length $T(n)$.

The KL-HMM system is parameterized by $\Theta_{kull} = \{\{\mathbf{y}_i\}_{i=1}^{I}, \{a_{ij}\}_{i,j=1}^{I}\}$. The lexical model parameters $\{\mathbf{y}_i\}_{i=1}^{I}$ are initialized uniformly i.e., initially $y_i^d = \frac{1}{D}$

$\forall i, d$. The training data $\{Z(n), W(n)\}_{n=1}^{N}$ and the current parameter set Θ_{kull}, are used to estimate the new set of parameters $\hat{\Theta}_{kull}$ by the Viterbi expectation maximization algorithm. In the case of the local score S_{RKL} the cost function minimized is,

$$\hat{\Theta}_{kull} = \arg\min_{\Theta_{kull}} \Big[\sum_{n=1}^{N} \min_{Q \in \mathcal{Q}} \sum_{t=1}^{T(n)} \big[S_{RKL}(\mathbf{y}_{q_t}, \mathbf{z}_t(n)) - \log a_{q_{t-1}q_t} \big] \Big] \tag{A.1}$$

where $Q = \{q_1, \cdots q_t, \cdots, q_{T(n)}\}$, $q_t \in \{1, \cdots, I\}$ and $\mathcal{Q}$ denotes set of all possible HMM state sequences.

The training process involves iteration over the segmentation and the optimization steps until convergence. Given the current set of parameters, the segmentation step yields an optimal state sequence for each training utterance using the Viterbi algorithm. Given optimal state sequences and acoustic unit posterior vectors belonging to the states, the optimization step then estimates new set of model parameters by minimizing Eqn. (A.1) subject to the constraint that $\sum_{d=1}^{D} y_i^d = 1$.

The optimal state distribution for the local score S_{RKL} (Equation (20)), is the arithmetic mean of the training acoustic unit probability vectors assigned to the state, i.e.,

$$y_i^d = \frac{1}{M(i)} \sum_{\mathbf{z}_t(n) \in Z(i)} z_t^d(n) \quad \forall d \tag{A.2}$$

where $Z(i)$ denotes the set of acoustic unit probability vectors assigned to state l^i and $M(i)$ is the cardinality of $Z(i)$. More details about the parameter estimation for the KL-HMM approach can be found in the thesis by Aradilla (2008).

Appendix B. Details of the Experimental Study

Appendix B.1. Phoneme-to-Articulatory Feature Map

Table B.1: Phoneme-to-articulatory feature map used in this paper

Phoneme	Manner	Place	Height	Vowel
sil	sil	sil	sil	sil
aa	vowel	back	low	aa
ae	vowel	mid-front	low	ae
ah	vowel	mid	mid	ah
ao	vowel	back	mid-low	ao
aw1	vowel	mid-front	low	aw1
aw2	vowel	mid-back	high	aw2
ax	vowel	mid	mid	ax
axr	approximant	retroflex	mid	consonant
ay1	vowel	back	low	ay1
ay2	vowel	mid-front	high	ay2

Table B.1 – continued from previous page

Phoneme	Manner	Place	Height	Vowel
b	voiced-stop	labial	max	consonant
ch	stop	front	max	consonant
d	voiced-stop	alveolar	max	consonant
dh	voiced-fricative	dental	max	consonant
eh	vowel	mid-front	mid	eh
el	approximant	lateral	very-high	consonant
em	nasal	labial	max	consonant
en	nasal	alveolar	max	consonant
er	vowel	mid	mid	er
ey1	vowel	front	mid-high	ey1
ey2	vowel	mid-front	high	ey2
f	fricative	labial	max	consonant
g	voiced-stop	dorsal	max	consonant
hh	aspirated	unknown	max	consonant
ih	vowel	mid-front	high	ih
iy	vowel	front	very-high	iy
jh	voiced-stop	front	max	consonant
k	stop	dorsal	max	consonant
l	approximant	lateral	very-high	consonant
m	nasal	labial	max	consonant
n	nasal	alveolar	max	consonant
ng	nasal	dorsal	max	consonant
ow1	vowel	back	mid	ow1
ow2	vowel	mid-back	high	ow2
oy1	vowel	back	mid-low	oy1
oy2	vowel	mid-front	high	oy2
p	stop	labial	max	consonant
r	approximant	retroflex	mid-low	consonant
s	fricative	alveolar	max	consonant
sh	fricative	front	max	consonant
t	stop	alveolar	max	consonant
th	fricative	dental	max	consonant
uh	vowel	mid-back	high	uh
uw	vowel	back	very-high	uw
v	voiced-fricative	labial	max	consonant
w	approximant	back	very-high	consonant
y	approximant	front	very-high	consonant
z	voiced-fricative	alveolar	max	consonant
zh	voiced-fricative	front	max	consonant

Appendix B.2. Details of the MLPs

The number of hidden units of the MLPs are selected based on the optimal frame accuracy on the development set of the RM or WSJ corpora. Hence, the

Table B.2: Overview of the various MLPs used in the paper. The input, output and hidden layer sizes of MLPs are denoted as $\mathtt{i}$, $\mathtt{h}$ and $\mathtt{o}$. The size of the MLP is equal to $\mathtt{h} * (\mathtt{i} + \mathtt{o} + 1)$. The frame accuracies of MLPs *MLP-RM-PH*, *MLP-RM-AF*, *MULTI-WSJRM-PH* and *MULTI-WSJRM-AF* are computed on the development set of the RM corpus. The frame accuracies of MLPs *MLP-WSJ-PH*, *MLP-WSJ-AF* are computed on the development set of the WSJ corpus.

MLP	acoustic units	input size ($\mathtt{i}$)	hidden size ($\mathtt{h}$)	output size ($\mathtt{o}$)	MLP size	frame acc.
MLP-RM-PH	phonemes	39*9=351	1260	45	0.5M	73.77
MLP-WSJ-PH	phonemes	39*9=351	3652	45	1.5M	69.34
MULTI-WSJRM-PH	phonemes	17*45=765	730	45	2.1M	80.16
MLP-RM-AF	manner	39*9=351	682	10	0.2M	82.48
	place	39*9=351	676	13	0.2M	76.20
	height	39*9=351	685	8	0.2M	79.26
	vowel	39*9=351	658	23	0.2M	79.38
MLP-WSJ-AF	manner	39*9=351	4005	10	1.4M	89.18
	place	39*9=351	3972	13	1.4M	86.83
	height	39*9=351	4027	8	1.4M	88.10
	vowel	39*9=351	3866	23	1.4M	88.29
MULTI-WSJRM-AF	manner	17*54=918	1063	10	6.6M	86.47
	place	17*54=918	1059	13	6.6M	82.42
	height	17*54=918	1065	8	6.6M	84.25
	vowel	17*54=918	1048	23	6.6M	84.36

total number of parameters are different for different MLPs. For the AF-based MLPs trained on the RM corpus, optimal frame accuracy was obtained when the total number of MLP parameters were about 15% of the training frames. For the phoneme-based and AF-based MLPs trained on the WSJ corpus, it was observed that the optimal frame accuracy on the development set was obtained when the total number of parameters were about 5% of the training frames (28M). Therefore, the total number of parameters of the MLPs trained on WSJ corpus is about 28M * 0.05 $\approx$ 1.5M. The total number of parameters in the multistage MLPs include the parameters of the first stage MLP(s) and the second stage MLP.

Table B.3: The details of the various KL-HMM systems namely the total number of MLP parameters (N_{θ_a}), lexical units (I), acoustic units (D), lexical model parameters (N_{θ_l}), KL-HMM system parameters ($N_{\theta_a}+N_{\theta_l}$). The details of the various tandem systems namely the number tied states (ts), Gaussian components in each tied state (ts), dimensionality of the tandem features that contribute to 99% of the variance (ν), tandem system parameters excluding the MLP parameters ($N_{\theta_{tan}}$) and tandem system parameters with the MLP parameters ($N_{\theta_a} + N_{\theta_{tan}}$).

MLP	N_{θ_a}	KL-HMM system				Tandem system			
		#lexical units (I)	#acoustic units (D)	N_{θ_l} $I * D$	Total $N_{\theta_a}+N_{\theta_l}$	#tied states (ts), #Gaussians (g)	dim (ν)	Tandem ($N_{\theta_{tan}}$) $ts * (g * (2\nu + 1))$	Total $N_{\theta_a} + N_{\theta_{tan}}$
MLP-RM-PH	0.5M	10988	45	0.5M	**1.0M**	1726,8	39	1.1M	**1.6M**
MLP-RM-AF	0.8M	8127	54	0.4M	**1.2M**	1737,8	42	1.2M	**2.0M**
MLP-RM-PH+AF	1.3M	11678	99	1.1M	**2.4M**	1723,8	75	2.1M	**3.4M**
MLP-WSJ-PH	1.5M	11723	45	0.5M	**2.0M**	2367,8	37	1.4M	**2.9M**
MLP-WSJ-AF	5.6M	8059	54	0.4M	**6.0M**	1751,8	42	1.2M	**6.8M**
MLP-WSJ-PH+AF	7.1M	9753	99	1.0M	**8.1M**	1718,8	75	2.1M	**9.2M**
MULTI-WSJRM-PH	2.1M	10857	45	0.5M	**2.6M**	2546,8	36	1.5M	**3.6M**
MULTI-WSJRM-AF	9.6M	7753	54	0.4M	**10.0M**	1807,8	40	1.2M	**10.8M**
MULTI-WSJRM-PH+AF	11.7M	9975	99	1.0M	**12.7M**	1784,8	73	2.1M	**13.8M**

Table B.4: The phoneme-to-manner of articulation and phoneme-to-place of articulation relationship captured by the lexical model parameters of context-independent phonemes.

Synchronous Examples				
Phoneme	**AF**	**state1**	**state2**	**state3**
aa	Manner	vow		
	Place	bck		
d	Manner	stp		
	Place	alv		
l	Manner	app		
	Place	lat		
m	Place	lab		
	Manner	nas		
Asynchronous Examples				
Phoneme	**AF**	**state1**	**state2**	**state3**
ah	Manner	vow		
	Place	bck	mid	
aw	Manner	vow		
	Place	midf	bck	midb
ay	Manner	vow		
	Place	bck	midf	
b	Manner	sil	stp	vow
	Place	lab		
ch	Manner	stp		frc
	Place	alv	fnt	
er	Manner	app		vow
	Place	ret		
g	Manner	stp		vow
	Place	dor		
hh	Manner	asp		vow
	Place	lab	midf	
jh	Manner	stp		vow
	Place	alv	fnt	
ng	Manner	nas		
	Place	dor		alv
ow	Manner	vow		
	Place	bck		midb
oy	Manner	vow		
	Place	bck	lat	midf
th	Manner	frc	stp	frc
	Place	alv		den
uh	Manner	vow		
	Place	bck	mid	midb
zh	Manner	vow	vfr	
	Place	midf	alv	fnt