IDIAP RESEARCH REPORT

INTEGRATING ARTICULATORY FEATURES USING KULLBACK-LEIBLER DIVERGENCE BASED ACOUSTIC MODEL FOR PHONEME RECOGNITION

Ramya Rasipuram Mathew Magimai.-Doss

Idiap-RR-02-2011

FEBRUARY 2011

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

INTEGRATING ARTICULATORY FEATURES USING KULLBACK-LEIBLER DIVERGENCE BASED ACOUSTIC MODEL FOR PHONEME RECOGNITION

Ramya Rasipuram[1,2] *and Mathew Magimai.-Doss*[1]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Ecole Polytechnique Fédérale, Lausanne (EPFL), Switzerland
{ramya.rasipuram,mathew}@idiap.ch

ABSTRACT

In this paper, we propose a novel framework to integrate articulatory features (AFs) into HMM- based ASR system. This is achieved by using posterior probabilities of different AFs (estimated by multilayer perceptrons) directly as observation features in Kullback-Leibler divergence based HMM (KL-HMM) system. On the TIMIT phoneme recognition task, the proposed framework yields a phoneme recognition accuracy of 72.4% which is comparable to KL-HMM system using posterior probabilities of phonemes as features (72.7%). Furthermore, a best performance of 73.5% phoneme recognition accuracy is achieved by jointly modeling AF probabilities and phoneme probabilities as features. This shows the efficacy and flexibility of the proposed approach.

Index Terms— automatic speech recognition, articulatory features, phonemes, multilayer perceptrons, Kullback-Leibler divergence based hidden Markov model, posterior probabilities

1. INTRODUCTION

State-of-the-art speech recognition systems typically use phonemes as sub-word units. Phonological studies suggest that, each phoneme can be further decomposed into a set of features based on the articulators used to produce the sound, like, manner of articulation, place of articulation, height of vowel etc. In recent years, articulatory features have been used for ASR with the aim of better pronunciation modeling [1], better co-articulation modeling, robustness to noise [2], multi-lingual and cross-lingual portability of systems [3]. Automatic speech recognition using articulatory features poses two main challenges: firstly, estimating articulatory features from the acoustic signal and secondly, integrating them into the conventional hidden Markov model (HMM) based framework.

In literature, pattern recognition techniques like multilayer perceptrons (MLPs) [1, 2, 4], support vector machine classifiers (SVMs) are typically used for the estimation of articulatory features. To integrate articulatory features into HMM framework they are either, transformed suitably for use as features in Tandem based speech recognition systems [1, 2, 5] or converted to phoneme posteriors (by training another MLP) and used as emission probabilities in hybrid HMM/MLP based systems [2].

In a more recent work [6], Kullback-Leibler divergence based hidden Markov model (KL-HMM) is proposed where the posterior probabilities of phonemes (phoneme posteriors) are directly used as features and each HMM state is parameterized using a multinomial posterior distribution. In this work, we use posterior probabilities of articulatory features (articulatory posteriors) *directly* as feature observations in KL-HMM (Section 2). This approach may enable the efficient use of articulatory features in multi-lingual and cross-lingual speech recognition systems, since no transformations or conversions are applied on them.

The phoneme recognition task on the TIMIT database is used to evaluate the system (Section 3). We investigate, two different MLP based approaches to estimate articulatory posteriors. In the first approach, independent MLP classifiers are trained using only spectral features (Section 4). In the second approach, we model the dependencies between different articulatory features using a multi-stage/hierarchical MLP classifier framework (Section 5). Our studies show that, the KL-HMM system using articulatory posteriors estimated from the first approach yields phoneme recognition accuracy worse than the KL-HMM system using phoneme posteriors (67.4% vs. 69.6%). However, the KL-HMM system using articulatory posteriors estimated from the second approach yields phoneme recognition accuracy comparable to the system using phoneme posteriors estimated from hierarchical MLP classifier (72.4% vs. 72.7%). Furthermore, jointly modeling articulatory posteriors and phoneme posteriors by concatenating them yields a phoneme recognition accuracy of 73.5%.

2. KL-HMM ACOUSTIC MODELING

In KL-HMM acoustic modeling [6], posterior probabilities of sub-word units are directly used as features and the state distribution is parameterized by a reference multinomial distribution (as shown in Figure 1). In the original work [6], phonemes are used as sub-word units and the posterior probabilities of phonemes (phoneme posteriors) are estimated using MLP. In such a case the posterior probability feature $\mathbf{z}_t$ estimated at time frame t using MLP is given by,

$$\mathbf{z}_t = [z_t^1, \cdots, z_t^D]^T = [P(/aa/|x_t), \cdots, P(/zh/|x_t)]^T \quad (1)$$

where, D is the number of phoneme classes and x_t is input feature given to the MLP. The KL divergence between the multinomial state distribution $\mathbf{y}_i$ and posterior probability feature $\mathbf{z}_t$ is defined as the local matching score for each state,

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

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). The authors would like to thank Joel Praveen Pinto for the fruitful discussions on the work and Guillermo Aradilla for his help with KL-HMM system.

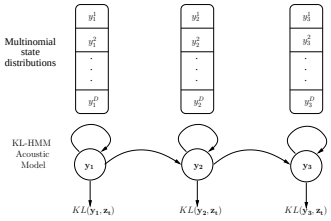

Fig. 1. A three state KL-HMM acoustic model for a phoneme

The multinomial state distributions are estimated using Viterbi expectation maximization algorithm with a cost function based on the KL divergence [6]. Decoding is performed in the usual manner, i.e., Viterbi decoding. By directly using posterior probabilities of sub-word units as features, the outputs of MLP are no more tied to HMM states as in hybrid HMM/MLP, thus providing the flexibility in terms of the choice of posterior feature space without any change in the state representation. In this work, we exploit the flexibility to choose posterior feature space by using posterior probabilities of different articulatory features as feature observations (in place of phoneme posteriors). More specifically, this is done by stacking the posterior estimates of different articulatory features in a single feature observation vector (articulatory posteriors) as shown below,

$$\mathbf{z}_t = [\mathbf{z}_t^{manner}, \cdots, \mathbf{z}_t^{height}]^T, \text{ where,} \quad (3)$$

$$\mathbf{z}_t^{manner} = [P(\text{fric}|x_t), \cdots, P(\text{vowel}|x_t)]^T$$
$$\mathbf{z}_t^{height} = [P(\text{low}|x_t), \cdots, P(\text{high}|x_t)]^T$$

In this case, the reference multinomial state distribution $\mathbf{y}_i$ is also a stack of multinomial distributions i.e.,

$$\mathbf{y}_i = [\mathbf{y}_i^{D_m}, \cdots, \mathbf{y}_i^{D_h}]^T, \text{ where,} \quad (4)$$

$$\mathbf{y}_i^{D_m} = [y_i^1, \cdots, y_i^{D_m}]^T$$
$$\mathbf{y}_i^{D_h} = [y_i^1, \cdots, y_i^{D_h}]^T$$

where, D_m is the cardinality of the manner class and D_h is the cardinality of the height class. The principle advantage of modeling articulatory posteriors using KL-HMM is that, it provides a framework to treat the articulatory posteriors jointly, with out the need to transform them as done in [1, 2].

3. EXPERIMENTAL SETUP

TIMIT acoustic-phonetic corpus is used for all the experiments (excluding the SA sentences). The partitioning of the database as specified in the TIMIT corpus is used. The data consists of 3,000 training utterances from 375 speakers, 696 cross-validation utterances from 87 speakers, and 1,344 test utterances from 168 speakers. The 61 hand labeled phonetic symbols are mapped to set of 39 phonemes with an additional garbage class [7].

The experimental setup is exactly same as the one described in [8]. All the MLPs (for phoneme posterior and articulatory posterior estimation) use the PLP cepstral coefficients with a context window of 9 frames as input. The first 13 PLP coefficients are extracted with a frame size of 25ms and a frame shift of 10ms. These coefficients are mean and variance normalized, and are appended with delta, delta-delta derivatives to obtain a 39 dimensional feature vector.

The output classes of the MLP estimating phoneme posteriors, represent the 40 phonemes. The targets of articulatory features for MLP training are obtained from the phoneme to articulatory feature map given in [9]. The articulatory features consist of manner, place, height, front-back, rounding, glottal state, nasality and vowel, also given in Table 2 along with their cardinality.

The size of the hidden layer of all the MLPs is determined by fixing the total number of parameters to 35% of the training data following the previous work [8]. The articulatory posteriors and phoneme posteriors are estimated from MLP trained using ICSI Quicknet software[1].

In [6], it is shown that hybrid HMM/MLP is a special case of KL-HMM when the state multinomial distributions are delta distributions (i.e., each output unit of MLP is tied to a HMM state). In this work, we build a similar hybrid HMM/MLP system where articulatory posteriors are used as features and the state distributions are replaced with delta distributions obtained using phoneme to articulatory feature map. It is to be noted that this hybrid HMM/MLP system is different from the one used in [2], where articulatory posteriors are converted to phoneme posteriors and are used as emission probabilities in HMM. All the experiments are based on context-independent phoneme sub-word units, where each sub-word unit is represented by a 3 state left-to-right HMM.

4. BASELINE STUDIES

Table 2 shows the articulatory feature classification accuracy of eight articulatory features (first stage classification accuracies in the three-stage MLP classifier) along with their cardinality and chance rates calculated on the cross-validation data. Chance rate is calculated as the accuracy obtained by choosing the most common label value in the reference data [1].

In this section, we present baseline phoneme recognition studies on KL-HMM and hybrid HMM/MLP systems using two different features:

- *base-ph*: Phoneme posteriors estimated using MLP.
- *base-af*: Articulatory posteriors estimated using a set of MLPs.

Table 1 presents the phoneme recognition accuracies of the above systems on the test set of TIMIT database. The KL-HMM system performs slightly better than the hybrid HMM/MLP system using both phoneme posteriors and articulatory posteriors. Also, recognition accuracy of the KL-HMM system using phoneme posteriors is higher than the system using articulatory posteriors (2.2% absolute). In the next section, we propose to estimate the articulatory posteriors by modeling the dependencies between articulatory features and using longer temporal contextual information in a multi-stage MLP framework.

5. MULTI-STAGE MLP ARTICULATORY FEATURE CLASSIFIER

In the previous section, articulatory features are independently modeled by training an MLP for each articulatory feature. Typically, many of the earlier articulatory feature recognition studies have treated them independently [2], [5], i.e., a independent classifier is trained for each articulatory feature. In [4], it has been shown

[1] http://www.icsi.berkeley.edu/Speech/qn.html

Features	System	
	KL-HMM	Hybrid HMM/MLP
base-ph	69.6	69.3
base-af	67.4	66.8

Table 1. Phoneme recognition accuracy expressed in percentage on the TIMIT test set, using phoneme posteriors and articulatory posteriors in KL-HMM and hybrid HMM/MLP systems

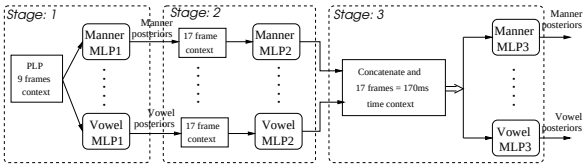

Fig. 2. Multi-stage MLP classifiers for articulatory posterior estimation

that the performance of place feature can be improved by training manner specific place classifiers. In [10], an approach to model inter-feature dependencies was studied using dynamic Bayesian networks (DBNs). This approach showed improvements in articulatory feature classification compared to an equivalent system where they were treated independently. Motivated from the previous studies [4], [10], and the hierarchical MLP framework [8], we investigate a novel multi-stage MLP classifier based approach to model the inter-feature dependencies of articulatory features.

The proposed multi-stage MLP classifier based approach for articulatory feature recognition consist of three stages as shown in the Figure 2 (referred as three-stage MLP classifiers). In the first stage, a set of parallel MLPs are used to estimate articulatory posteriors for the eight articulatory features. Each MLP receives PLP features as input and is trained to classify a specific articulatory feature (Stage 1 in Figure 2). This is the baseline system used to estimate articulatory posteriors in the previous section.

In the second stage, to model the temporal contextual information of articulatory features, a new set of MLPs are trained using articulatory posteriors estimated by the first stage of MLPs as input, with longer temporal context (Stage 2 in Figure 2). The width of the temporal context is fixed at 17 frames, following the results in [8], where it was found that phoneme recognition accuracy saturates at around 170 ms. We can expect that, the second stage of MLPs learn the articulatory feature confusions at the output of first stage of MLPs and model the phonotactics of a language (phonological constraints), both at a individual articulatory feature level [8].

In the third stage, to model the inter-feature dependencies of articulatory features, articulatory posteriors estimated from Stage 2, are used as input to next stage of MLPs, along with the information of other articulatory features (Stage 3 in Figure 2). It is to be noted that, though the number of MLPs used to extract articulatory posteriors have increased, no additional data (apart from TIMIT) is used, also the MLPs at all stages are trained for the same targets.

We also consider a modified case of the multi-stage MLPs where Stage 2 in Figure 2 is omitted, consequently, Stage 1 is followed by Stage 2. This set of MLPs (referred to as two-stage MLP classifiers) are built to ascertain the importance of temporal contextual modeling and inter-feature dependencies of articulatory features.

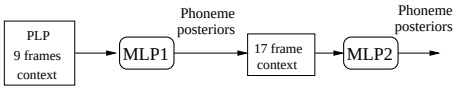

Fig. 3. Hierarchical MLP classifier for phoneme posterior estimation

5.1. Articulatory feature classification

Table 2 compares the articulatory feature classification accuracy of the three-stage MLP classifiers at different stages. From the table, we see that both the contextual information and information of other articulatory features contribute towards improvements. However, classification accuracy benefits more when both the contextual information and inter-feature dependencies are modelled in Stage 3 compared to Stage 2, when only contextual information is modelled. This is further verified in the two-stage MLP classifiers, which achieve the classification accuracy very close to the three-stage MLP classifiers for most of the articulatory features.

Articulatory class	Cardinality	Chance rates	Three-stage MLPs			Two-stage MLPs
			First stage	Second stage	Third stage	
Manner	8	34.1	86.0	86.3	88.1	88.1
Glottal state	5	61.6	92.9	93.4	94.6	94.5
Nasality	4	77.9	96.0	96.2	96.8	96.8
Place	11	34.1	86.3	87.5	88.5	88.5
Height	9	47.7	82.5	83.1	85.5	85.1
Frontedness	8	47.7	84.2	84.6	87.1	86.6
Rounding	4	65.8	89.9	90.2	92.5	91.9
Vowel	22	47.7	81.3	83.0	84.1	84.5

Table 2. Articulatory feature classification accuracy expressed in percentage on the TIMIT development set, using three-stage and two-stage MLP classifiers

5.2. Phoneme recognition studies

In this section, we present the phoneme recognition studies using the articulatory posteriors estimated from multi-stage MLP classifiers. To compare similar systems, phoneme posteriors are also estimated using hierarchical MLP classifier as described in [8], also shown in Figure 3. We compare two systems, KL-HMM and hybrid HMM/MLP using three different features:

1. *hier-ph*: Phoneme posteriors estimated from hierarchical MLP classifier of Figure 3.
2. *3-stage-af*: Articulatory posteriors estimated from three-stage MLP classifiers of Figure 2.
3. *2-stage-af*: Articulatory posteriors estimated from two-stage MLP classifiers.

Table 3 shows that, multi-stage MLP classifiers for posterior estimation help in improving the phoneme recognition accuracy of both the articulatory and phoneme posterior based systems. However, articulatory posteriors achieve an absolute improvement of 5.0%, where as phoneme posteriors achieve an absolute improvement of 3.1% compared to their respective baselines. It is worth noting that the performance gap between systems using hierarchical phoneme posteriors and multi-stage articulatory posteriors is only 0.2% (as opposed to 2.2% on the baselines). The three-stage articulatory posterior based system is slightly better than the two-stage

articulatory posterior based system. Also, it is interesting to note that the hybrid HMM/MLP system using articulatory posteriors performs slightly better than the system using phoneme posteriors (71.9 vs. 71.6).

Features	System	
	KL-HMM	Hybrid HMM/MLP
hier-ph	72.7	71.6
3-stage-af	72.4	71.9
2-stage-af	72.0	71.8

Table 3. Phoneme recognition accuracy expressed in percentage on the TIMIT test set, using phoneme posteriors and articulatory posteriors estimated using multi-stage MLP classifiers in KL-HMM and hybrid HMM/MLP systems

In [10], a DBN framework was proposed to model the inter-feature dependencies of articulatory features. The dependencies between different articulatory features were hierarchically organized and related uni-directionally, i.e., place feature is conditioned on the manner feature but not vice versa. Moreover, the dependencies were determined manually. We used the same set of dependencies for articulatory features specified in [10], but modelled them using MLP classifiers. The resulting articulatory posteriors when integrated into the KL-HMM system resulted in phoneme recognition accuracy of 70.4%. The result shows improvement over the equivalent KL-HMM system where the dependencies are not modelled (67.4%), but, is significantly lower than the proposed multi-stage MLP based approach (72.4%), where the relations between articulatory features are mutually modelled (place feature is conditioned on manner feature and vice versa). This indicates that, it is better to model the dependencies between articulatory features mutually rather than uni-directionally.

The key strength of KL-HMM lies in its ability to incorporate posteriors estimated using different methods. The hierarchical phoneme posteriors (*hier-ph*) and three-stage articulatory posteriors (*3-stage-af*) are concatenated and used as features in KL-HMM. Similarly, the baseline posteriors (*base-ph* and *base-af*) are concatenated and used as features in KL-HMM. Table 4 shows the results of these experiments. The increase in recognition accuracy of the system using combined baseline posteriors is not significant compared to phoneme posterior based system. However, the combined multi-stage posteriors show increase in recognition accuracy over their corresponding single feature systems. This indicates that the information learned through contextual modeling in posterior domain and articulatory domain is complementary. Similar trends were observed in [5], where phoneme and articulatory posteriors are modelled using conditional random fields for phoneme recognition.

Features	Accuracy
base-ph + *base-af*	69.8
hier-ph + *3-stage-af*	73.5

Table 4. Phoneme recognition accuracy expressed in percentage on the TIMIT test set, using phoneme posteriors appended with articulatory posteriors as features in KL-HMM system

6. CONCLUSION

In this paper, we proposed a novel framework using KL-HMM to integrate directly articulatory feature probabilities for ASR. Our studies showed that by modeling the inter-feature dependencies between articulatory features, phoneme recognition accuracy similar to the use of phoneme probabilities in KL-HMM can be achieved. Furthermore, we demonstrated the flexibility of the proposed approach by jointly modeling articulatory feature probabilities and phoneme probabilities which yielded the best phoneme recognition accuracy of 73.5%. Future work includes investigating the proposed framework on continuous speech recognition, and investigating alternate ways to model dependencies between articulatory features such as using multi-tasking learning [11].

7. REFERENCES

[1] K. Livescu et al., "Articulatory Feature-based Methods for Acoustic and Audio-Visual Speech Recognition: 2006 JHU Summer Workshop Final Report," `http://www.clsp.jhu.edu/ws2006/groups/afsr/documents/WS06AFSR_final_report.pdf`, 2008.

[2] K. Kirchhoff, G. A. Fink, and G. Sagerer, "Combining acoustic and articulatory feature information for robust speech recognition ," *Speech Communication*, vol. 37, pp. 303–319, 2002.

[3] S. Stüker, T. Schultz, F. Metze, and A. Waibel, "Multilingual articulatory features," in *Proc. of ICASSP*, 2003, vol. 1, pp. 144–147.

[4] S. Chang, M. Wester, and S. Greenberg, " An elitist approach to automatic articulatory-acoustic feature classification for phonetic characterization of spoken language," *Speech Communication*, vol. 47, pp. 290–311, 2005.

[5] E. Fosler-Lussier and J. Morris, "Crandem systems: Conditional random field acoustic models for hidden Markov models," in *Proc. of ICASSP*, 2008, pp. 4049–4052.

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

[7] K.-Fu Lee and H.-W Hon, "Speaker-Independent Phone Recognition using Hidden Markov Models," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, 1989.

[8] J. Pinto, G. Sivaram, M Magimai.-Doss, H. Hermansky, and H. Bourlard, "Analysis of MLP Based Hierarchical Phoneme Posterior Probability Estimator," *To appear in IEEE Trans. on Audio, Speech, and Language Processing*.

[9] J. Frankel, M. Magimai-Doss, S. King, K. Livescu, and Ö. Çetin, "Articulatory Feature Classifiers Trained on 2000 hours of Telephone Speech," in *Proc. of Interspeech*, 2007.

[10] J. Frankel, M. Wester, and S. King, "Articulatory feature recognition using dynamic Bayesian networks," in *Computer Speech & Language*, 2007, vol. 21(4), pp. 620–640.

[11] Rich Caruana, "Multitask Learning," *Machine Learning*, vol. 28, no. 1, pp. 41–75, 1997.