On using Articulatory Features in Posterior-based ASR System: Representation, Estimation and Integration

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

[1] Idiap Research Institute, CH-1920 Martigny, Switzerland
[2] Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland
{mathew,ramya.rasipuram}@idiap.ch

Abstract

Articulatory features describe the properties of speech production, i.e., each sound unit of a language (phone) can be decomposed into a set of features based on the articulators used to produce the sound. Articulatory features (AFs) have been used for automatic speech recognition (ASR) with the aim of better pronunciation modeling [1], robustness to noise [2], multilingual and cross-lingual portability of the systems [3] etc. In case of text-to-speech (TTS) conversion systems AFs are explored to achieve emotional speech synthesis. ASR or TTS using AFs poses three main challenges:

- firstly, the type of AF representation;
- secondly, the estimation of AFs from the acoustic signal;
- and finally, the integration into conventional hidden Markov model (HMM) based ASR framework.

The goal of this presentation is to present recent advances made at Idiap Research Institute in addressing the above mentioned challenges in the context of ASR, and open the discussion on natural extension of the work presented here such as, grapheme-based ASR, generation of AF based dictionaries, template-based ASR using AFs, use of deep learning approaches to improve AF estimation [4].

The remainder of the extended abstract briefly describes the recent advances, presents some related experimental results, and potential future research directions.

a) Type of AF Representation: 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 are binary valued features, for example +voice and -voice, +sonorant and -sonorant [5]. However, according to Ladefoged, it is more natural to allow the features to take multiple values [6]. In government phonological feature system, speech sounds are destructed into a set of primes and can be represented by fusing these primes structurally [7].

In this work, our interest lies in multi-valued AFs. We present our investigations on different phoneme to AF maps as given in John Hopkin's workshop (referred to as JHU map) [1] and Hosom's work (referred to as Hosom map) [8]. The main difference between the two mappings being, Hosom map is more compact, i.e., it has less number of features (4) and the cardinality of each of the features is high, while the JHU map has more features (7) but the cardinality of the features is low compared to Hosom map.

b) Estimation of AFs: In the literature, typically, each AF is learned by training a classifier like multilayer perceptron (MLP) [1, 2], or support vector machine. In our recent work, we showed that AF classification accuracy and thereby the phoneme recognition accuracy can be improved by modeling the inter-feature dependencies using a hierarchy of MLP classifiers and/or multitask learning [9, 10, 11] .

Briefly, as illustrated in Figure 1, the hierarchical MLP classifier based approach for AF recognition consists of two stages. In the first stage, a set of parallel MLPs are used to estimate articulatory posteriors for 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 longer temporal context as input.

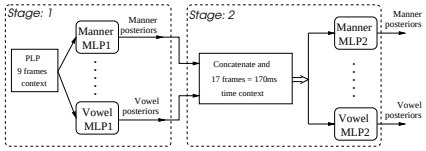

Figure 1: Multi-stage MLP classifiers for articulatory posterior estimation

c) Integration of AFs in HMM-based ASR: To integrate AFs into the standard HMM-based ASR systems, the posterior probabilities of AFs are typically converted to phoneme posterior probabilities by training another classifier. The AFs converted to phoneme posterior probabilities are either used as state emission probabilities in hybrid HMM/MLP systems or transformed suitably for use as features in Tandem systems [1, 2]. In [9], we proposed an approach where the estimated a posteriori probabilities of AFs are directly integrated into HMM-based ASR by using them (without any kind of transformation) as feature observations in Kullback-Leibler divergence based hidden Markov model (KL-HMM).

Briefly, KL-HMM is an approach where the a posteriori probabilities of acoustic classes are directly used as feature observation and each HMM state is parameterized using a categorical distribution [12]. The categorical state distributions are estimated using Viterbi expectation maximization algorithm by minimizing a cost function based on the KL divergence. Figure 2 shows the proposed approach to integrate AFs in HMM-based ASR framework using KL-HMM. We present studies showing how KL-HMM approach can be exploited to perform phoneme recognition and continuous speech recognition by learning the probabilistic relationship between subword units and AFs.

d) Brief Overview of Experiments and Results: We present

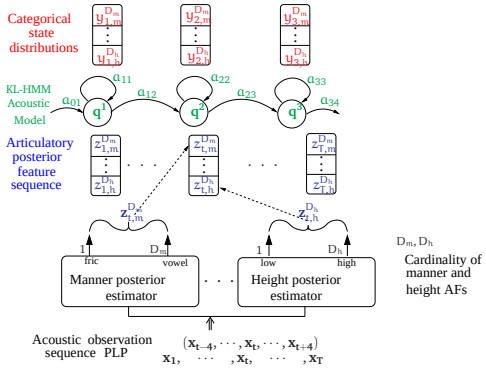

Figure 2: Three state KL-HMM acoustic model for phoneme where posterior probabilities of AFs are used as feature observations

phoneme recognition and ASR studies related to the above described advances. The reader can find phoneme recognition studies in our previous publications [9, 10, 11]. For both phoneme recognition tasks and ASR tasks we have observed that the Hosom map yields better system compared to the JHU map. Here, we briefly present part of the unpublished ASR work using the Hosom map.

ASR studies are conducted on the DARPA Resource Management task. The training set consists of 2,880 utterances spoken by 109 speakers. The test set contains 1,200 utterances amounting to 1.1 hours of speech data (formed by combining Feb'89, Oct'89, Feb'91 and Sep'92 test sets). The test set is completely covered by a word pair grammar (perplexity 60) included in the task specification. The phoneme lexicon was obtained from UNISYN dictionary.

The MLPs used for phoneme posterior and articulatory posterior estimation are trained on auxiliary Wall Street Journal (WSJ) corpus. In case of hierarchical MLP, first set of MLPs (referred to as WSJ-MLP) are trained on the WSJ corpus, where as the second set of MLPs (referred to as WSJ-RM-MLP) are trained on the RM corpus. The size of the hidden layer for all the MLPs is determined by fixing the total number of parameters to 35% of the training data. Table 1 presents the ASR results. It can be observed that stand-alone AF posterior probability based system (Af) yields inferior performance compared to phoneme posterior probability based system (Ph). However, the system using both phoneme posterior and AF posterior based system ($Ph + Af$) yields the best system. Thus, indicating the complementarity between the two posterior features.

Features	Dimension	MLP	
		WSJ-MLP	WSJ-RM-MLP
Ph	45	4.7	4.4
Af	54	5.7	5.2
$Ph + Af$	99	4.3	3.9

Table 1: Word error rate (WER) expressed in terms of percentage. On this setup, standard context-dependent phoneme-based HMM/GMM system with state tying achieves 4.9% WER.

One of the reasons for System Af yielding low performance could be that the state distribution for AF or a single categorical distribution may not be sufficient to capture the dynamic nature of AFs. We are presently investigating approaches where, a) multiple categorical distributions are learned per state, and b) temporal context of AFs posterior probabilities is modeled.

e) Future Directions: The posterior-based ASR using AFs can be naturally extended to develop

- an ASR system that models the relationship between grapheme subword units and AFs similar to our recent work on grapheme-based ASR using KL-HMM [13, 14]. This is interesting from both multilingual speech processing and under-resourced ASR perspectives.
- AF based pronunciation dictionary development (which is particularly interesting for AF-based TTS) following the recently proposed acoustic data-driven grapheme-to-phoneme conversion approach using KL-HMM [15]. More specifically, the probabilistic relationship between graphemes or phonemes and AFs learned by KL-HMM in the posterior-based ASR system could be exploited to generate AF based pronunciation dictionary.
- template-based ASR system using AF posterior probabilities similar to the approach of using phoneme posterior probabilities [16].

We will present a few outcomes based on our on-going work and discuss the potential implications of these future directions.

1. References

[1] K. Livescu *et al.*, "Articulatory Feature-based Methods for Acoustic and Audio-Visual Speech Recognition: 2006 JHU Summer Workshop Final Report," 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*, vol. 1, 2003, pp. 144–147.

[4] G. Hinton *et al.*, "Deep Neural Networks for Acoustic Modeling in Speech Recognition ," *IEEE Signal Processing Magazine*, 2012.

[5] N. Chomsky and M. Halle, *The Sound Pattern of English*. MIT Press, 1968.

[6] P. Ladefoged, *A Course in Phonetics*. Harcourt Brace College Publishers, 1993.

[7] J. Harris, *English Sound Structure*. Blackwell, 1994.

[8] J. Hosom, "Speaker-independent phoneme alignment using transition-dependent states," *Speech Communication*, vol. 51, pp. 352–368, 2009.

[9] R. Rasipuram and M. Magimai.-Doss, "Integrating articulatory features using kullback-leibler divergence based acoustic model for phoneme recognition," in *Proc. of ICASSP*, 2011.

[10] ——, "Multitask learning to improve articulatory feature estimation and phoneme recognition," Idiap Research Report, Idiap-RR-21-2011, 2011.

[11] ——, "Improving articulatory feature and phoneme recognition using multitask learning," in *Artificial Neural Networks and Machine Learning - ICANN 2011*, 2011.

[12] 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.

[13] R. Rasipuram and M. Magimai.-Doss, "Improving grapheme-based asr by probabilistic lexical modeling approach," in *Proceedings of Interspeech*, 2013.

[14] ——, "Probabilistic lexical modeling and grapheme-based automatic speech recognition," Idiap Research Report, Idiap-RR-15-2013, 2013.

[15] ——, "Acoustic Data-driven Grapheme-to-Phoneme Conversion using KL-HMM," in *Proc. of ICASSP*, 2012.

[16] S. Soldo, M. Magimai.-Doss, J. P. Pinto, and H. Bourlard, "Posterior Features for Template-based ASR," in *Proc. of ICASSP*, 2011.