AN ARTICULATORY FEATURE-BASED TANDEM APPROACH AND FACTORED OBSERVATION MODELING

Özgür Çetin[1] *Arthur Kantor*[2] *Simon King*[3] *Chris Bartels*[4]
Mathew Magimai-Doss[1] *Joe Frankel*[1,3] *Karen Livescu*[5]

[1]International Computer Science Institute, Berkeley, CA, USA
[2]University of Illinois, Urbana-Champaign, USA
[3]University of Edinburgh, Edinburgh, UK
[4]University of Washington, Seattle, WA, USA
[5]Massachusetts Institute of Technology, Cambridge, MA, USA

ABSTRACT

The so-called *tandem* approach, where the posteriors of a multilayer perceptron (MLP) classifier are used as features in an automatic speech recognition (ASR) system has proven to be a very effective method. Most tandem approaches up to date have relied on MLPs trained for phone classification, and appended the posterior features to some standard feature hidden Markov model (HMM). In this paper, we develop an alternative tandem approach based on MLPs trained for articulatory feature (AF) classification. We also develop a factored observation model for characterizing the posterior and standard features at the HMM outputs, allowing for separate hidden mixture and state-tying structures for each factor. In experiments on a subset of Switchboard, we show that the AF-based tandem approach is as effective as the phone-based approach, and that the factored observation model significantly outperforms the simple feature concatenation approach while using fewer parameters.

Index Terms—
Speech Recognition, Multilayer Perceptrons

1. INTRODUCTION

The *tandem* approach refers a data-driven signal processing method for extracting features for acoustic modeling of speech [6, 4, 15]. The tandem approach involves first training a MLP to perform phone classification at the frame level, and then using the post-processed frame-level phone posterior estimates of the MLP as the acoustic observations in HMMs. Commonly, the tandem features are appended to some standard feature vector such as the perceptual linear prediction (PLP) coefficients. (We will assume without loss of generality that the PLP coefficients are the standard features.) The tandem approach has given significant performance improvements for large vocabulary speech recognition of English conversational telephone speech (CTS) [15], and Arabic and Mandarin broadcast news and conversations [11, 14]. The tandem approach also seems to have some cross-domain and cross-language generalization [12].

The majority of the previous tandem approaches have been limited to phone-based posterior estimation. In this work, we propose an alternate approach based on articulatory features. The articulatory feature-based tandem features are extracted from the posterior estimates of a set of MLPs trained for articulatory feature classification, one for each feature. The main motivations for the proposed AF-based tandem approach instead of the traditional phone-based approach are as follows. First, the AFs in general can more accurately and parsimoniously characterize the pronunciation and acoustic variability associated with conversational speech [8]. Second, AF classification is simpler, involving multiple classification problems with a small number of classes each, instead of a single phone classifier with a large number of classes. Third, while not explored in this work, AFs are more language universal than phones, and therefore they can better generalize and be easier to adapt to new languages [13].

As described earlier, the usual approach of using posterior features in ASR systems is to concatenate them with some standard features and model the concatenated feature vector with a single HMM output distribution. The standard and posterior features are forced to have the same hidden mixture and state-tying structure, even though they are likely to have quite different statistical properties, being derived from two opposite paradigms, prior knowledge and heuristics vs. data-driven statistical learning [12]. The large dimensionality of the concatenated vector also gives highly concentrated Gaussian probability estimates. In this work, we develop an factored modeling approach, where each component is separately modeled at the HMM output distributions, avoiding these problems. Our approach is similar to [9] which also proposed an AF-based tandem method, but we directly combine the posterior features with the standard features at the HMM outputs, and experiment with new observation models.

Feature	Values
Place	labial, labio-dental, dental, alveolar, post-alveolar, velar, glottal, rhotic, lateral, none, silence
Degree/manner	vowel, approximant, flap, fricative, closure, silence
Nasality	+, -, silence
Glottal state	voiced, voiceless, aspirated, silence
Rounding	+, -, silence
Vowel	aa, ae, ah, ao, aw1, aw2, ax, ay1, ay2, eh, er, ey1, ey2, ih, iy, ow1, ow2, oy1, oy2, uh, uw, not-a-vowel, silence
Height	very high, high, mid-high, mid, mid-low, low, nil, silence
Frontness	back, mid-back, mid, mid-front, front, silence

Table 1. The articulatory feature set.

2. EXPERIMENTAL PARADIGM

We use SVitchboard, a set of reduced-vocabulary tasks derived from Switchboard1 [7]. In particular, we use one of the SVitchboard 500-word tasks, which has predefined training (A, B, and C), cross-validation (D), and testing (E) sets, and includes a total of 6.4 hours of speech. The vocabulary is closed at 500 words with no out-of-vocabulary words.

We use 13-dimensional PLP coefficients and their first- and second-order differences as input features to both the MLP classifiers and the HMM acoustic models. Mean subtraction and variance normalization are performed on a per-speaker basis. All recognition systems are trained and tested using the Graphical Models Toolkit (GMTK) [2]. For context-dependent modeling, we use a new GMTK clustering tool, `gmtkTie`, allowing for clustering by an arbitrary set of user-defined questions and variables. Decoding is first-pass using a standard bigram language model estimated from the transcripts of the A, B, and C sets. The dictionary allows up to three pronunciations per word. For each experiment reported below, the language model scaling factor and insertion penalty as well as the number of mixture components in the observation models are optimized to minimize the word error rate (WER) on a subset of the D set.

3. MLP ARTICULATORY CLASSIFIERS

The AF set that we used in our experiments is given in Table 1. A separate MLP for each feature has been trained. The MLPs are standard three-layer feedforward networks, classifying each frame into one of the values of the corresponding feature. The inputs to the MLP are the PLPs from the current frame as well as those from the four frames forward and backward in time, a total of 351 values. The MLPs are gender-independent, and trained using a total of 1776 hours of data from Fisher and Switchboard2, excluding Switchboard1 (thus SVitchboard). The number of the MLP hidden units are selected so as to have a roughly $1000{:}1$ ratio of the number of training frames to the number of parameters. The targets for MLP training are obtained from a deterministic phone-to-feature mapping of forced phonetic alignments from a SRI CTS system. The training took around ten days per MLP, on a 2.2GHz Sun v40z machine using highly optimized software (e.g., multi-threading). The frame level accuracy of the trained MLPs ranged from 70% to 90%. See [5] for more details.

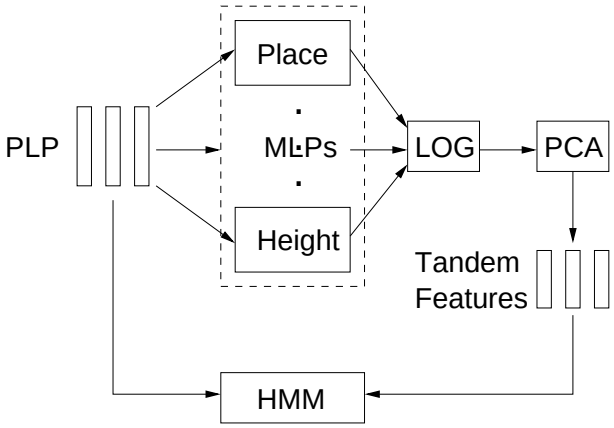

Fig. 1. Articulatory feature-based tandem processing.

We also trained two sets of MLPs on the SVitchboard training data, for comparing AF- and phone-based tandem approaches: (1) a set of AF MLPs identical to the Fisher-trained MLPs above except that the numbers of hidden units were scaled down according to the amount of data, (2) a phone MLP with 46 outputs corresponding to the SRI phone set.

4. ARTICULATORY FEATURE-BASED TANDEM OBSERVATIONS

4.1. Tandem Processing

We extract the AF-based tandem features as follows (cf. Figure 1). For each time frame, the posterior probability estimates from the each of the eight AF MLPs are joined together, a total of 64 values. Their logarithm is taken to expand the dynamic range of the posterior probabilities to the real axis. Roughly speaking, the logarithm undoes the effect of the final softmax nonlinearity at the MLP outputs, making them more amenable to Gaussian modeling. After the logarithm, the principal component analysis (PCA) is applied to reduce the dimensionality to 26, which is determined to contain the 95% of the total variance. The PCA decorrelates the features, and eliminates redundancies among the posterior probabilities form different MLPs, as the AF set that we use is not orthogonal. The PCA transform is estimated on the MLP training set. Finally, per-speaker mean subtraction and variance normalization are applied. The resulting 26 dimensional vectors along with 39-dimensional PLP vectors are used as acoustic observations in HMM.

Feature	WER
PLP	67.7
PLP + Phone tandem (SVBD)	63.0
PLP + AF tandem (SVBD)	62.3
PLP + AF tandem (Fisher)	59.7
PLP + AF tandem (Fisher) + Factoring	59.1

Table 2. WERs (%) for the monophone systems using PLP, and PLP in combination with various tandem features, on the SVitchboard (SVBD) 500-word E set. The corpus name in parentheses refers to the MLP training set. All tandem systems except the last one concatenates PLP and tandem features, and "factoring" refers to factored observation modeling (cf. Section 5).

Feature	# of states	WER
PLP	477	59.2
PLP + AF tandem (Fisher) Feature concatenated	880	55.0
PLP + AF tandem (Fisher) Observation factored	467 / 641	53.8

Table 3. WERs (%) for the various triphone systems on the SVitchboard 500-word E set. The number of states refers to the number of decision-tree clustered triphone states; the pair for the observation factored model is the number of states for the PLP and the tandem, respectively, factors. See Table 2 caption for the notation.

4.2. Experiments

Using the procedure in Figure 1, we extracted two sets of 26-dimensional AF-based tandem observations using the MLPs trained on Fisher, and trained on SVitchboard. For comparison, we also generate a set of 26-dimensional phone-based tandem observations using the phone MLP trained on SVitchboard. In Table 2 (the first four lines), we report the WER for the baseline monophone system using PLPs, and the monophone systems using various kinds of tandem features: the AF-based tandem features from the Fisher AF MLPs, the AF-based tandem features from the SVitchboard AF MLPs, and the phone-based tandem features from the SVitchboard phone MLP. In Table 3 (the first two lines), we report the WERs for the cross-word triphone systems using PLPs, and using the concatenated PLP and the tandem features from the Fisher AF MLPs. The number of components per Gaussian mixture was 128 for the monophone systems, and 64 for the triphone systems. The number of components was separately optimized for each system on the SVitchboard 500-word D set. (As an aside, we note that the WERs are uniformly high, mainly due to the sparse training data, less than four hours, and the fact that SVitchboard utterances often contain frequent words that tend to have wide pronunciation and acoustic variability.)

A few observations about the results of Tables 2 and 3 are in order. First, all of the systems using any type of tandem feature (phone or AF-based, or SVitchboard or Fisher trained) in both monophone and triphone models significantly improve the performance over the baseline PLP system. Second, the AF-based tandem features are as effective as the phone-based tandem features (the difference is not statistically significant). Third, the Fisher MLPs are significantly better than the SVitchboard MLPs for the purposes of AF-based tandem processing. This is expected given that the Fisher MLPs were trained on two orders of magnitude more training data. Fourth, the AF-based tandem observations are also very effective in triphone modeling, even though the relative improvement is lower (12% for monohones vs. 7% for triphones). All pairs of results in Tables 2 and 3, except the phone- vs. AF-based tandem pair in Table 2, are statistically significant according to matched pairs sentence-segment word error test ($p < 0.01$). Overall the AF-based tandem processing seems to be highly effective.

5. FACTORED OBSERVATION MODELING

5.1. Model

In the basic tandem approach described in Section 4 and also in most of the previous work (e.g., [6, 9, 4, 15]), the tandem observation vectors were simply concatenated to PLPs, and then jointly modeled using the same hidden mixture and state-tying structures in HMMs with diagonal-covariance mixture distributions. The HMM output distributions of these concatenated features can be represented as

$$p(x, y|q) = \sum_t p(t|q)\, p(x|t, q)\, p(y|t, q) \qquad (1)$$

where x and y denote the PLP and tandem, respectively, vectors, q denotes the HMM state, and t denotes the hidden mixture component. See Figure 2(a) for a graphical model depiction (there is no arrow between x and y due to the diagonal-covariance assumption).

The PLP and the tandem vectors are generated by two very different philosophies, prior knowledge and heuristics vs. data-driven statistical learning. Their statistical characteristics are likely to be quite different, and enforcing the same mixture structure and decision-tree state-tying scheme could be inefficient for learning their distributions from sparse data. In addition, the large dimensionality of the concatenated vector gives highly concentrated Gaussians, which in previous work has been dealt with using a heuristic weighting factor [15]. Instead, here we propose a principled approach based on factored modeling of the PLP and tandem observation vectors at the HMM output distributions, allowing for separate hidden mixture and state-tying structures for each vector. In particular, the HMM output distributions for the PLP and tandem vectors are factored as follows, as in multi-stream ASR models [3],

$$\begin{aligned} p(x, y|q) &= \Big(\sum_z p(z|q)\, p(x|z, q)\Big) \\ &\quad \times \Big(\sum_w p(w|q)\, p(y|w, q)\Big) \end{aligned} \qquad (2)$$

where w and z are the hidden mixture variables for x and y, respectively. See Figure 2(b) for a graphical model depiction. The product of sums in Equation 2 has a smoothing effect.

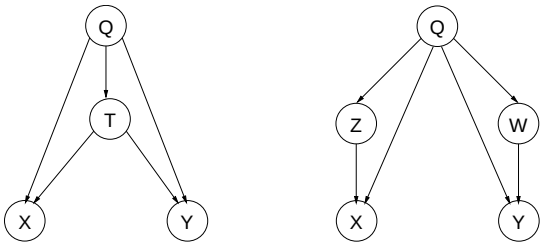

Fig. 2. (a) Feature concatenation (assuming diagonal covariance Gaussian modeling), and (b) factored modeling.

Notice that in Equation 2 the PLP and tandem vectors are assumed to be conditionally independent given the HMM state, whereas in Equation 1, they are indirectly coupled via the hidden mixture variable t. However, in general, neither form is subsumed by the other one due to Gaussian parameterization, and there are distributions that can be represented by one and not the other one, and vice versa. The factored modeling has the advantage that it can more accurately characterize each of the PLP and tandem observation vectors by modeling each of them separately. On the other hand, it could be at a disadvantage if the PLP and tandem features are highly correlated even when conditioned on the HMM state.

5.2. Experiments

Using the AF-based tandem observations from the Fisher MLPs, we have compared factored modeling to feature concatenation. The results with the monophone models are in Table 2, and those with the triphone models are in Table 3. The decision tree-based state clustering for the triphone system with the factored observation model was performed exactly the same way it was performed for the feature-concatenated system, except that the clustering procedure was invoked twice, one for each factor in Equation 2.

We can draw the following conclusions from Tables 2 and 3. First, factored observation modeling significantly improves over feature concatenation (differences are statistically significant). Therefore, the benefit from the more accurate characterization of the PLP and tandem vectors individually seems to outweight the loss from the conditional independence assumption (and/or, dependencies are weak). Second, the clustering results demonstrate that the factored approach not only gives better results, but it is also more parsimonious. The system with factored observation models has about 40% fewer parameters. The decision trees for the PLP and tandem observations differ in both structure and size, corroborating the earlier claim that the statistical properties of the PLPs and tandem features are quiet different.

6. CONCLUSIONS

This paper has proposed the use of the processed output of AF MLPs as features in a tandem-based system. Furthermore, a factored observation model has been proposed to model the acoustic and tandem features with separate HMM output distributions. Two main conclusions that can drawn from the initial experiments conducted on SVitchboard are: (1) AF-based tandem features are as effective as phone-based tandem features, and (2) factored observation models, apart from resulting in models with fewer parameters, outperform the feature concatenation approach, suggesting that the distributions of acoustic feature and tandem features are better modeled independently rather than jointly. Ongoing work focuses on highly factored observation models, one factor per AF. An interesting future research direction is to relax the conditional independence assumption in the factored model by sparse cross-observation dependencies [1]. We also intend to explore the applications to multilingual ASR.

Acknowledgments This material is based upon work supported by the NSF under Grant No. 0121285. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF. This work was conducted at the 2006 JHU Workshop, as part of a project on articulatory feature-based speech recognition [10]. It was supported in part by the Swiss NSF through IM2.

7. REFERENCES

[1] J. Bilmes, "Data-driven extensions to HMM statistical dependencies," in *Proc. ICSLP*, pp. 69–72, 1998.

[2] J. Bilmes and G. Zweig, "The Graphical Models Toolkit: An open source software system for speech and time-series processing," in *Proc. ICASSP*, pp. 3916–3919, 2002.

[3] H. Bourlard, S. Dupont, and C. Ris, "Multi-stream speech recognition," Technical Report *IDIAP-RR 96-07*, IDIAP, 1996.

[4] D.P.W. Ellis, R. Singh and S. Sivadas, "Tandem acoustic modeling in large-vocabulary recognition," in *Proc. ICASSP*, pp. 517-520, 2001.

[5] J. Frankel et al., "Articulatory feature classifiers trained on 2000 hours of telephone speech," submitted to *ICASSP*, 2007.

[6] H. Hermansky, D.P.W. Ellis, S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. ICASSP*, pp. 1635–1638, 2000.

[7] S. King, J. Bilmes, and C. Bartels, "SVitchboard 1: Small-vocabulary tasks from Switchboard 1," in *Proc. INTERSPEECH*, pp. 3385-3388, 2005.

[8] S. King et al., "Speech production knowledge in automatic speech recognition," submitted to *JASA*, 2006.

[9] 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, 2000.

[10] K. Livescu et al., "Articulatory feature-based methods for acoustic and audio-visual speech recognition: Summary from the 2006 JHU Summer Workshop," submitted to *ICASSP*, 2007.

[11] X. Lei, M.-Y. Hwang, and M. Ostendorf, "Incorporating tone-related MLP posteriors in the feature representation for Mandarin ASR," in *Proc. INTERSPEECH*, pp. 2981–2984, 2005.

[12] A. Stolcke et al., "Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptrons," in *Proc. ICASSP*, pp. 321–324, 2005.

[13] S. Stueker, F. Metze, T. Schultz, and A. Waibel, "Integrating multilingual articulatory features into speech recognition," in *Proc. EUROSPEECH*, pp. 1033-1036, 2003.

[14] J. Zheng et al., "Combining discriminative feature, transform, and model training for large vocabulary speech recogniton," submitted to *ICASSP*, 2007.

[15] Q. Zhu, A. Stolcke, B.Y. Chen, and N. Morgan. "Incorporating tandem/HATs MLP features into SRI's conversational speech recognition system," in *Proc. DARPA RT Workshop*, 2004.