View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Edinburgh Research Archive

# MONOLINGUAL AND CROSSLINGUAL COMPARISON OF TANDEM FEATURES DERIVED FROM ARTICULATORY AND PHONE MLPS

*Özgür Çetin*[1] *Mathew Magimai-Doss*[2] *Karen Livescu*[3]
*Arthur Kantor*[4] *Simon King*[5] *Chris Bartels*[6] *Joe Frankel*[5]

[1]Yahoo!, Inc., Santa Clara, USA
[2]IDIAP Research Institute, Martigny, Switzerland
[3]Massachusetts Institute of Technology, Cambridge, USA
[4]University of Illinois, Urbana-Champaign, USA
[5]University of Edinburgh, Edinburgh, UK
[6]University of Washington, Seattle, USA

## ABSTRACT

In recent years, the features derived from posteriors of a multilayer perceptron (MLP), known as tandem features, have proven to be very effective for automatic speech recognition. Most tandem features to date have relied on MLPs trained for phone classification. We recently showed on a relatively small data set that MLPs trained for articulatory feature classification can be equally effective. In this paper, we provide a similar comparison using MLPs trained on a much larger data set—2000 hours of English conversational telephone speech. We also explore how portable phone- and articulatory feature-based tandem features are in an entirely different language—Mandarin—without any retraining. We find that while the phone-based features perform slightly better in the matched-language condition, they perform significantly better in the cross-language condition. Yet, in the cross-language condition, neither approach is as effective as the tandem features extracted from an MLP trained on a relatively small amount of in-domain data. Beyond feature concatenation, we also explore novel observation modeling schemes that allow for greater flexibility in combining the tandem and standard features at hidden Markov model (HMM) outputs.

***Index Terms***— Speech recognition, feedforward neural networks, hidden Markov models.

## 1. INTRODUCTION

The so-called *tandem* acoustic modeling refers to a data-driven feature extraction method using MLPs [1, 2, 3]. In tandem modeling, the transformed posterior probabilities of an MLP are used as observations in HMMs, usually in combination with some standard feature such as mel-frequency cepstral coefficients (MFCCs), or perceptual linear prediction (PLP) coefficients. The tandem processing is simple, and integrable into an existing recognizer with virtually no change in the statistical back-end. This simplicity and modularity make tandem features attractive for large-vocabulary continuous speech recognition (LVCSR). In recent years, tandem features have produced impressive word error rate (WER) reductions in state-of-the-art systems in multiple languages, e.g., English, Mandarin, and Arabic, and in different domains, e.g., conversational telephone speech (CTS), broadcast news (BN), and multiparty meetings, and in tasks that are small and large [4, 5, 6, 7, 8, 9].

Most tandem approaches to date have used phone posteriors for deriving features. While it can be argued that features optimized for phone discrimination will better couple with phonetic modeling units used in HMMs, there is nothing inherent in tandem processing that will prevent an alternative partitioning of the acoustic space and the posteriors from that space, being the basis of tandem processing. Articulatory features (AFs) can provide one such alternative. AFs have a long history in ASR proposals; see, e.g., [10, 11, 2, 12, 13]. Among the arguments for the use of AFs in ASR are (1) they can better account for pronunciation and acoustic variability than phones, (2) AF classification is simpler, involving multiple small classification problems, and (3) AFs are more language universal than phones, and therefore they can better generalize and adapt to new languages. In recent work [14], we showed that AF-based tandem features indeed can be as effective as phone-based tandem features on a subset of the Switchboard database, where the amount of MLP training used in comparisons was limited (five hours).

In this paper, we report comparisons between AF- and phone-based tandem features, derived from MLPs trained on a large amount of data (2000 hours of English CTS), on a number of tasks. First, the AF- and phone- based approaches are compared for English CTS using a subset of Switchboard. Second, the language portability of AF- and phone-based tandem features is addressed. Similar to an acoustic model, the tandem features are language dependent, because the under-

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

**Table 1**. The articulatory feature set.

lying MLP is tuned to a particular language and task. The AF- and phone-based features from English-trained MLPs are compared for Mandarin LVCSR, which in turn are compared to a set of features extracted from an MLP trained on in-domain data.

Parallel to the AF vs. phone comparison, we also explore new observation models for systems using tandem features, continuing our initial work in this area [14]. As mentioned earlier, the usual method of incorporating tandem features in ASR systems is to concatenate them with some standard feature, and then tie the hidden mixtures and context-dependent state clusters for the tandem and standard features together. This restriction could be inefficient, because the standard and tandem acoustic features are likely to have different statistical properties, being derived from two opposite paradigms, knowledge-based signal processing vs. data-driven statistical learning [7]. Instead, a factored approach is explored here, where each feature is allowed to have its own mixture and tying structure. For AF-based tandem features, a fully factored approach is also explored, where there are multiple tandem vectors corresponding to each AF category.

## 2. RECOGNITION SYSTEMS

In this section, we describe the English CTS and Mandarin BN speech recognition systems used in our experiments.

### 2.1. English CTS

SVitchboard, a set of reduced-vocabulary tasks derived from Switchboard 1 [15], is used for English CTS experiments. In particular, we use one of the SVitchboard 500-word tasks, which includes a total of 6.4 hours of speech, and which has been partitioned into training (A, B, and C), development (D), and testing (E) sets.

All recognition systems including triphone systems are trained and tested using the Graphical Models Toolkit (GMTK) [16]. 13 PLP coefficients and their first- and second-order derivatives are used as standard acoustic features. Mean subtraction and variance normalization are performed on a per-speaker basis. Decoding is first-pass using a bigram LM estimated from the training transcripts. The vocabulary is closed to 500 words without any out-of-vocabulary word; the dictionary allows up to three pronunciations per word. The LM scales and penalties as well as the number of mixture components in the observation models are optimized on the development set to minimize WER.

### 2.2. Mandarin BN

About 97 hours of LDC Mandarin Hub4 and TDT4 data, released as part of the DARPA GALE program, are used for acoustic model training. The TDT4 closed captions were filtered with flexible alignment [17]. The 2004 GALE Mandarin Rich Transcription development (RT04-dev) and evaluation sets (RT04-eval) are used for system development and final testing, respectively. RT04-dev and RT04-eval include about half an hour and one hour, respectively, of BN speech. The Mandarin BN speech has a bandwidth of 8 kHz, whereas the English CTS data on which the AF MLPs and English phone MLP are trained has a bandwidth of 4 kHz. Therefore, the Mandarin BN data was downsampled from 16 kHz to 8 kHz for consistent evaluation of Mandarin systems in all experiments.

SRI's Decipher LVCSR system is used for Mandarin BN experiments. 13 MFCCs plus pitch, and their first- and second-order derivatives are used as standard features. Vocal tract length normalization, mean subtraction and variance normalization are performed on a per-cluster basis (the clusters are automatically deduced). Decoding is first pass using a trigram LM, with a lexicon consisting of about 49000 words. Decipher includes a mechanism to smooth Gaussian probabilities using an exponential weight, which was found to be particularly helpful in tandem systems that use large-dimensional feature vectors. The Gaussian weights as well as the LM scales and penalties are optimized on RT04-dev, and the final results are reported on RT04-eval. See [8] for more details about the Mandarin system.

## 3. MLP CLASSIFIERS

We have trained a number of MLPs for AF and phone classification using about 2000 hours of speech from Fisher and Switchboard 2 corpora. (Note that while the domain is similar, these MLP training data have no overlap with the SVitchboard data, cf. Section 2.1.) The AF set used in our experiments is given in Table 1. A separate gender-independent MLP for each feature is trained. The MLPs are standard feed-forward networks, with input, hidden, and output layers. The

| **MLP Classifier** | **# of units** | **Accuracy** |
|---|---|---|
| English AF | | |
| Place | 1900 / 10 | 76.2 |
| Degree | 1600 / 6 | 77.8 |
| Nasality | 1200 / 3 | 90.5 |
| Glottal state | 1400 / 4 | 87.1 |
| Rounding | 1200 / 3 | 87.7 |
| Vowel | 2400 / 23 | 73.3 |
| Height | 1800 / 8 | 75.4 |
| Frontness | 1700 / 7 | 75.8 |
| English phone | 4800 / 46 | 61.4 |
| Mandarin phone | 2000 / 65 | 73.1 |

**Table 2**. The number of hidden units / output units, and CV accuracy (%) for various phone and AF MLPs trained on English and Mandarin.

inputs to the MLP are the PLP coefficients from the current frame as well as those from the four frames of left and right time context, a total of 351 values. The number of hidden units are set to have an approximate $1000:1$ ratio between the number of training frames and parameters. The AF targets for MLP training are obtained from a deterministic phone-to-AF mapping of forced phonetic alignments from an SRI CTS system. See [18] for more details about the AF MLPs. An MLP for phone classification, using a 46 dimensional phone set, has been trained on the same data set as well.

To gauge the effectiveness of the English-trained MLPs on Mandarin, we also trained an MLP for phone classification using the Mandarin BN training data, cf. Section 2.2. This MLP is similar to the English phone MLP except that it uses 65 Chinese phones, which also encode lexical tone. This Mandarin MLP was originally developed as part of the 2006 GALE Mandarin evaluations [8].

The number of MLP hidden units and the frame-level classification rates for the various MLPs are reported in Table 2. The cross-validation (CV) accuracy is measured against the forced-aligned labels, on a 10% subset of the data that were set aside during MLP training. While Mandarin has a significantly larger phone set, it is recognized more accurately than English, possibly due to the generally lower error rates for BN than for CTS.

## 4. ENGLISH CTS EXPERIMENTS

We have performed a number of experiments comparing the AF-based tandem features to the phone-based ones, and the factored observation models to the popular feature concatenation approach, for English CTS.

### 4.1. Tandem Processing

Extraction of the tandem features from the AF MLPs is similar to the standard tandem processing [1, 3, 4]. For each time frame, the posterior outputs from all AF MLPs are joined together to form a 64-dimensional vector. Their logarithm is taken,[1] and principal component analysis (PCA) is applied. The logarithm and PCA expand the dynamic range of the posteriors, and makes them more amenable to Gaussian modeling. The PCA transform is estimated on the MLP CV set, cf. Section 3; the number of principal components was 26, which was found to account for the $95\%$ of the total variance. The resulting 26-dimensional vectors after mean subtraction and variance normalization are used as acoustic observation vectors in the HMMs.

Extraction of the tandem features from the phone MLP is similar, except that instead of the concatenated outputs from multiple AF MLPs, the outputs of the phone MLP are used. The first 24 principal components were sufficient to account for $95\%$ of the total variance.

Finally, a third set of tandem features were extracted from the concatenated outputs of all AF MLPs and the phone MLP to evaluate how much complimentary information is provided by the phone MLP and the AF MLPs. The number of principal components were set to 37 using the aforementioned variance criterion.

### 4.2. Observation Modeling

In most previous work using tandem features, the tandem features are concatenated with some standard acoustic features, for example, PLP coefficients, which are then fed into HMMs. These HMM outputs with mixture of diagonal-covariance Gaussian distributions can be expressed 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 PLP and tandem, respectively, vectors, $q$ denotes the HMM state, and $t$ denotes the mixture component. (Note that the tandem and PLP vectors appear in two separate factors *inside* the summation because of the diagonal covariance modeling.) The tandem features are constrained to have the same mixture and tying structures as PLP coefficients, and vice versa.

While feature concatenation is convenient from a system design perspective, it could be inefficient for statistical modeling. A transformed posterior probability and a PLP coefficient are likely to have different statistical properties, and they could be better modeled if they are allowed to have separate mixture and tying structures [14]:

$$p(x,y|q) = \sum_z p(z|q)\, p(x|z,q) \sum_w p(w|q)\, p(y|w,q). \qquad (2)$$

As compared to Equation 1, the tandem and PLP vectors in Equation 2 appear in two separate factors without a joint summation: the two vectors are assumed to be conditionally independent. The factored model can better model each of the

[1] It is also possible to use the MLP outputs before the final nonlinearity instead of taking the logarithm; this method gives similar results.

| | Feature | WER |
|---|---|---|
| 1 | PLP | 67.7 |
| 2 | Phone tandem | 61.4 |
| 3 | PLP + Phone tandem (Concatenated) | 58.2 |
| 4 | AF tandem | 61.1 |
| 5 | PLP + AF tandem (Concatenated) | 59.7 |
| 6 | PLP × AF tandem (Factored) | 59.1 |
| 7 | PLP × AF tandem (Fully factored) | 63.8 |
| 8 | PLP + Phone-AF tandem (Concatenated) | 59.8 |

**Table 3**. WERs (%) for various monophone systems using PLP coefficients, and phone- and AF-based tandem features, on the SVitchboard 500-word E set. We use + to denote feature concatenation, and × to denote observation factoring.

PLP and tandem vectors. On the other hand, if the two vectors are highly dependent even when conditioned on the HMM state, the factored approach could suffer.

Within AF-based tandem processing, one can extend the factored model so that each AF category has its own factor, which we refer to as the fully factored model. In this model, a separate tandem vector is extracted from each AF MLP, using the procedure described in Section 4.1. After applying separate PCAs to keep the $95\%$ of the total variance within each AF category, the number of tandem features was 4 for place, 4 for degree, 2 for nasality, 2 for glottal state, 2 for rounding, 13 for vowel, 5 for height, and 5 for frontness. Note that the total dimensionality (37) is larger than the dimensionality (26) from the jointly concatenated approach, cf. Section 4.1. This is expected given that AF categories are overlapping and redundant in the acoustic space. We note that because the fully factored model loses the benefit of joint optimization, it is expected to suffer when used with the AF-tandem features.

### 4.3. Results

To compare the performances of AF- and phone-based tandem features, and of factored modeling and feature concatenation, we have conducted a number of experiments on SVitchboard using monophone and triphone models. In Table 3, we report the WERs for monophone systems using PLP coefficients, phone and AF tandem features both alone and in combination with PLP coefficients, and factored and fully-factored models using various AF tandem features. In Table 4, the key comparisons are repeated using triphone systems. In order to separate the benefits of factoring and of the factor-specified state tying, an experiment is devised, where the tandem and PLP features are still factored, but they are forced to share the state-tying structure from the concatenated model. Line 8 in Table 3 and line 6 in Table 4 use the tandem features extracted using the outputs of both the AF MLPs and the phone MLP, cf. Section 4.1. (Some of the AF tandem results reported in Table 3—lines 1, 5, and 6—first appeared in [14], and reproduced here for the sake of AF- vs. phone-based tandem features, and feature concatenation vs. factored vs. fully-factored modeling comparisons. Also, the triphone experiments in this paper use a relative likelihood improvement criterion for state clustering, which is found to better scale across systems using varying number of features than an absolute likelihood improvement criterion as used in [14]. No triphone AF- vs. phone-based tandem features, and factored vs. factored-and-tied modeling comparisons appeared in [14].)

| | Feature | # of states | WER |
|---|---|---|---|
| 1 | PLP | 675 | 61.7 |
| 2 | PLP + Phone tandem (Concatenated) | 441 | 54.9 |
| 3 | PLP + AF tandem (Concatenated) | 426 | 55.4 |
| 4 | PLP × AF tandem (Factored) | 689 / 302 | 54.4 |
| 5 | PLP × AF tandem (Factored & Tied) | 426 / 426 | 54.9 |
| 6 | PLP + Phone-AF tandem (Concatenated) | 376 | 55.3 |

**Table 4**. 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 models is the number of states for the PLP and tandem, respectively, factors. See the Table 3 caption for the notation.

A few observations about Tables 3 and 4 are in order. First, in Table 3, either the phone-based tandem features or the AF-based ones alone significantly improve over the baseline PLP features (lines 2 and 4 in Table 3), and they perform comparably to each other. Second, after concatenation with the PLP coefficients, the phone-based tandem features perform better than the AF-based tandem features in both monophone and triphone systems (lines 3 and 5 in Table 3, and lines 2 and 3 in Table 4). Third, the tandem features extracted using all the MLPs fail to provide any improvement over the tandem features extracted using either set of MLPs (lines 3 and 5 vs. 8 in Table 3, and lines 2 and 3 vs. 6 in Table 4). Therefore, the phone MLP and the AF MLPs seem to provide no complimentary information to each other, as utilized by the feature concatenation framework. Fourth, there is a consistent gain from factored modeling over feature concatenation (lines 5 and 6 in Table 3, and lines 3 and 4 in Table 4). The fully-factored approach is significantly inferior (line 7 in Table 3), which is expected given that the AF set used here is not orthogonal. Finally, we observe that the context-dependent clusterings for the two factors are significantly different in both size and structure (lines 3 and 4 in Table 4). The tandem features require less than half of the tied states required for the PLP coefficients (lines 4 in Table 4). Constraining them to use the same clustering degrades performance (lines 4 and

| Feature | WER |
|---|---|
| MFCC | 21.5 |
| MFCC + Phone tandem (Mandarin) | 19.5 |
| MFCC + Phone tandem (English) | 21.2 |
| MFCC + AF tandem (English) | 21.9 |

**Table 5**. WERs (%) for the various systems using MFCCs, and various tandem features on Mandarin RT04-eval set. The language on which the MLP is trained is given in parentheses. All tandem systems employ feature concatenation.

5 in Table 4). The factored model seems to equally benefit from the use of separate mixture and tying structures.

## 5. MANDARIN BN EXPERIMENTS

As opposed to PLP coefficients or MFCCs, which contain a handful of adjustable parameters, the tandem features in effect contain millions of parameters, by way of the MLPs from which these features are derived. These free parameters allow the optimization of the front-end signal processing to a particular task, and more generally, to a particular language, as shown by significant WER improvements in Section 4. However, at the same time, such a high degree of adaptability could easily become a burden, if the tandem features do not generalize well, especially for tasks and languages where the amount of training data is small. While languages can have radically different phone sets, for example, 46 English phones vs. 65 Mandarin phones with tone, AFs are more likely to be shared across languages. Therefore, one would expect that the AF distinctions learned in one language would better generalize to another language, and that the AF-based tandem features would be more language-portable than the phone-based tandem features [12, 7]. In previous work, it was shown that the phone-based tandem features exhibit significant cross-task and -language portability [7].

To test the hypothesis that the AF-based features would generalize better than the phone-based features, we used the English-trained MLPs for tandem feature extraction for Mandarin. The tandem processing is identical to the procedure described in Section 4.1 except that new PCA transforms are estimated on the Mandarin training data, reducing dimensionality to 29 and 25 for the AF- and phone-based, respectively, tandem features (again using the $95\%$ total variance criterion). A third set of 32-dimensional tandem features is extracted using a Mandarin phone MLP trained on the Mandarin training data, cf. Section 3. The WERs for the system using MFCCs, and the systems using various sets of tandem features are given in Table 5. All tandem systems concatenate the tandem features with the MFCCs.

Table 5 contains a number of interesting results. First, we find that the Mandarin-trained phone tandem features bring gains as impressive as the gains from English tandem features in Section 4.3 (the relative WER improvements are around $10\%$). Second, while the English-trained phone tandem features bring a small WER reduction, the English-trained AF features actually degrade performance. Third, overall it is more advantageous to use a relatively small amount of in-domain data to tune tandem features to a particular language, rather than transporting them across languages.

Contrary to the hypothesis that the learned AF distinctions would generalize better, the English-trained AF features actually degraded the performance for Mandarin BN, which could be due to a number of factors. First, the AF learning in our setup was restricted by the lack of ground truth AF labels. We used the deterministic phone-to-articulatory mappings for creating AF training targets, which could be inaccurate. Embedded training can improve results [19]. Second, all of the AF MLPs in this study were trained with the same standard acoustic features (39 PLP coefficients). For AF tandem features to generalize across different languages, it may be important to also use acoustic features that are specific to the AF set, in addition to standard acoustic features. For instance, specific acoustic-phonetic features like fundamental frequency, voicing, voice-onset time, glottalization, burst related impulses, intensity can be helpful [20]. Third, in addition to the language mismatch, the domain mismatch (CTS vs. BN) probably tampers with generalization as well.

## 6. CONCLUSIONS AND FUTURE WORK

In this paper, we compared AF-based tandem features to phone-based ones, and factored observation modeling to feature concatenation, on a number of monolingual and crosslingual tasks using MLPs trained on 2000 hours of English data. We found that while the AF-based tandem features are comparable to the phone-based ones when the MLPs are trained and tested on the same language, the phone-based approach is significantly better on a new language, without retraining. In crosslingual studies, we found that the tandem features from an MLP trained on a small amount of in-domain data performed the best. Furthermore, in the AF tandem studies, we found that there is consistent benefit from a limited form of factoring in AF-based tandem observation models, but not from fully factoring.

To the best of our knowledge, this kind of monolingual and crosslingual comparison of AF- and phone-based tandem features, extracted from competitively trained MLPs does not appear in previous work. Our results suggest a number of interesting future research directions. Iterative, embedded training of AF MLPs that can provide more accurate AF targets, which could be initialized by deterministic phone-to-articulatory mappings as in the present work, could be necessary to fully exploit the power of AF representations. Methods of transfer learning between languages, e.g., MLP retraining and adaptation [21], is a largely unexplored area. It is also necessary to repeat these cross-language studies for other pairs of languages with varying degrees of acoustic-phonetic similarity. The negative results with the fully factored observation models suggest that relaxing the conditional independence assumption in the factored model by cross-factor dependencies could be beneficial [22].

## 7. ACKNOWLEDGMENTS

The authors gratefully acknowledge International Computer Science Institute (ICSI) for the use of their computers. This work was conducted in part while the first two authors were at ICSI, and during the 2006 JHU Workshop as part of a project on articulatory feature-based speech recognition [13]. This material is based upon work supported in part by the NSF under Grant No. 0121285, by the DARPA under Contract No. HR0011-06-C-0023, by the Swiss National Science Foundation under the National Center of Competence in Research (NCCR) on Interactive Multimodal Information Management (IM2) program, and by the European Union Augmented Multiparty Interaction (AMI) program. 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 supporting agencies.

## 8. REFERENCES

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

[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, 2000.

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

[4] Q. Zhu, A. Stolcke, B. Y. Chen, and N. Morgan, “Incorporating Tandem/HATs MLP features into SRI’s conversational speech recognition system,” in *Proc. EARS RT-04F Workshop*, 2004.

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

[6] M. Karafiat *et al.*, “Robust heteroscedastic linear discriminant analysis and LCRC posterior features in meeting recognition,” in *Proc. MLMI*, 2006.

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

[8] M.-Y. Hwang *et al.*, “Advances in Mandarin broadcast speech recognition,” in *Proc. INTERSPEECH*, 2007.

[9] J. Zheng *et al.*, “Combining discriminative feature, transform, and model training for large vocabulary speech recognition,” in *Proc. ICASSP*, 2007, pp. 633–636.

[10] T. Schultz and A. Waibel, “Language independent and language adaptive large vocabulary speech recognition,” in *Proc. ICASSP*, 1998, pp. 1819–1822.

[11] M. Ostendorf, “Moving beyond the “beads-on-a-string” model of speech,” in *Proc. ASRU*, 1999, pp. 79–84.

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

[13] K. Livescu *et al.*, “Articulatory feature-based methods for acoustic and audio-visual speech recognition: Summary from the 2006 JHU summer workshop,” in *Proc. ICASSP*, 2007, pp. 621–624.

[14] Ö. Çetin *et al.*, “An articulatory feature-based tandem approach and factored observation modeling,” in *Proc. ICASSP*, 2007, pp. 645–648.

[15] S. King, J. Bilmes, and C. Bartels, “SVitchboard: Small-vocabulary tasks from Switchboard,” in *Proc. INTERSPEECH*, 2005, pp. 3385–3388.

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

[17] A. Venkataraman *et al.*, “An efficient repair procedure for quick transcriptions,” in *Proc. ICSLP*, 2004, pp. 1961–1964.

[18] J. Frankel *et al.*, “Articulatory feature classifiers trained on 2000 hours of telephone speech,” in *Proc. INTERSPEECH*, 2007.

[19] M. Wester, J. Frankel, and S. King, “Asynchronous articulatory feature recognition using dynamic bayesian networks,” in *Proc. IEICI Beyond HMM Workshop*, 2004.

[20] J. P. Hosom, “Automatic phoneme alignment based on acoustic-phonetic modeling,” in *Proc. ICSLP*, 2002, pp. 357–360.

[21] X. Li and J. Bilmes, “Regularized adaptation of discriminative classifiers,” in *Proc. ICASSP*, 2006, pp. 237–240.

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