ARTICULATORY FEATURE-BASED METHODS FOR ACOUSTIC AND AUDIO-VISUAL SPEECH RECOGNITION: SUMMARY FROM THE 2006 JHU SUMMER WORKSHOP

Karen Livescu[1], *Özgür Çetin*[2], *Mark Hasegawa-Johnson*[3], *Simon King*[4], *Chris Bartels*[5], *Nash Borges*[6], *Arthur Kantor*[3], *Partha Lal*[4], *Lisa Yung*[6], *Ari Bezman*[7], *Stephen Dawson-Haggerty*[8], *Bronwyn Woods*[9], *Joe Frankel*[2,4], *Mathew Magimai-Doss*[2], *Kate Saenko*[1]

1. MIT 2. ICSI 3. U. Illinois, Urbana-Champaign 4. U. Edinburgh 5. U. Washington
6. Johns Hopkins U. 7. Dartmouth College 8. Harvard U. 9. Swarthmore College

ABSTRACT

We report on investigations, conducted at the 2006 Johns Hopkins Workshop, into the use of articulatory features (AFs) for observation and pronunciation models in speech recognition. In the area of observation modeling, we use the outputs of AF classifiers both directly, in an extension of hybrid HMM/neural network models, and as part of the observation vector, an extension of the "tandem" approach. In the area of pronunciation modeling, we investigate a model having multiple streams of AF states with soft synchrony constraints, for both audio-only and audio-visual recognition. The models are implemented as dynamic Bayesian networks, and tested on tasks from the Small-Vocabulary Switchboard (SVitchboard) corpus and the CUAVE audio-visual digits corpus. Finally, we analyze AF classification and forced alignment using a newly collected set of feature-level manual transcriptions.

Index Terms— Speech recognition, speech processing

1. INTRODUCTION

Articulatory features have a long history in proposals for automatic speech recognition (ASR) techniques (e.g., [1, 2, 3, 4]). Some motivations are that (1) such models should help account for coarticulation effects, (2) certain aspects of articulation can be more robustly detected than others, and (3) several classifiers, each with a small number of classes, may make better use of sparse training data than a single phone classifier. Approaches using articulatory features (AFs) have had some success, for example in noisy conditions [5], for hyperarticulated speech [6], or in multilingual settings [7]. Improvements have also been obtained in lexical access experiments using models of articulatory asynchrony and reduction [8].

Our ultimate goal is to build complete continuous speech recognizers using AFs at all levels, including multiple feature-specific state streams and observations or observation models tailored to those streams. We assume in this work that the task is first-pass Viterbi decoding to find the jointly most likely string of words w^* and set of state assignments q^*:

$$\{w^*, q^*\} = \arg\max_{w,q} p(o|q)p(q|w)p(w),$$

where o are the observations. In hidden Markov model (HMM)-based recognition, q is the phonetic state sequence. In our case, q can be any collection of hidden variables corresponding to the sub-word representation; e.g., it may be the assignments for a set of hidden state streams. We refer to $p(o|q)$ as the *observation model* and to $p(q|w)$ as the *pronunciation model*. This is a non-standard definition of the pronunciation model and refers to the entire probabilistic mapping from words to sub-word structure. The remaining term $p(w)$ is the language model, which we assume to be fixed.

We describe several approaches for observation and pronunciation modeling. For observation modeling, we use the outputs of multilayer perceptron (MLP) AF classifiers in two ways: to estimate $p(o|q)$ (a "hybrid" approach [9]); and as part of the observation vector after post-processing (a "tandem" approach [10]). We investigate "embedded training" of the MLPs, in which training data is aligned using an AF-based recognizer and the MLPs are retrained [11]. For pronunciation modeling, we test a model consisting of multiple loosely-synchronized hidden AF streams, for both audio-only and audio-visual recognition. The motivation for the audio-visual case is that articulatory dynamics can explain the observed asynchrony between the audio and video signals. In this work, we do not combine the new observation and pronunciation models; for observation modeling experiments, the hidden states are a single phonetic state stream, and for pronunciation modeling experiments we use Gaussian mixture observation models. All models are implemented as dynamic Bayesian networks (DBNs) [12]. We have also collected a small set of manual transcriptions at the AF level, and we compare AF classifier outputs and alignments against these. Due to space limitations, we will summarize only briefly many aspects of this work. Additional details will appear separately (e.g. [13, 14, 15]).

This material is based upon work supported by the National Science Foundation under Grant No. 0121285. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. This work was partly supported by the Swiss National Science Foundation through the research network IM2. We are grateful to the JHU Center for Language and Speech Processing for facilitating this research.

2. DATA AND BASELINES

For audio-only experiments, we use a subset of SVitchboard 1, a set of small-vocabulary tasks from Switchboard 1 [16]. We use the 10-word and 500-word vocabulary tasks, and use the "ABC" sets for training, a subset of "D" for development, and "E" for final testing. We use the 10-word task for preliminary tests and for experiments with the more compute-intensive models. We have found a high correlation between 10- and 500-word results. The observations are speaker-normalized perceptual linear prediction (PLP) coefficients.

Table 1 shows several baseline performance results. All systems, except ones marked "HTK", are trained and tested with the Graphical Modeling Toolkit (GMTK) [17]. The GMTK triphone system uses a new state tying tool, gmtkTie, which generalizes triphone decision tree clustering to distributions with arbitrary conditional parents and user-defined questions. The monophone model has been trained both with and without word alignments (from Mississippi State [18]). All of the new baselines, except the 10-word monophone without alignments, outperform previously published baselines [16], and the GMTK triphone model compares favorably to a similarly trained HTK model.

Model	10-wd WER	500-wd WER
HTK whole-word + MFCCs [16]	20.8	70.8
Monophone, no alignments	24.5	67.7
Monophone, with alignments	19.6	65.0
HTK triphone	-	61.2
GMTK/gmtkTie triphone	-	59.2

Table 1. *Baseline word error rates (%) on the 10- and 500-word vocabulary test sets. HTK is the HMM Toolkit and the whole-word baseline uses mel-frequency cepstral coefficients (MFCCs). All systems besides whole-word use PLPs. Triphone systems were not tested on the 10-word task.*

3. PRONUNCIATION MODELING VIA MULTIPLE HIDDEN STREAMS OF ARTICULATORY FEATURES

Our pronunciation models are based on the approach of [8]. Each AF is represented as a separate hidden stream, with soft synchrony constraints between streams. This is based on the motivation that the articulators can move in a semi-independent way, and that this accounts for many coarticulatory pronunciation effects. The soft (i.e. probabilistic) synchrony constraints between AF streams are modeled via "asynchrony variables", whose distributions represent the probability of one AF being ahead or behind another by a given number of states. For example, an asynchrony of one state between nasality and the lips or tongue can produce effects such as epenthetic stop insertion and vowel nasalization. This type of pronunciation model has shown promise in lexical access experiments, in which the articulatory features are given and there are no acoustic observations [8]. Here we use the model for first-pass decoding, for both audio-only and audio-visual speech. For the current work, we assume that the observation depends jointly on all of the AFs via a Gaussian mixture distribution. Fig. 1(a) shows the main components of the model, assuming three AFs.

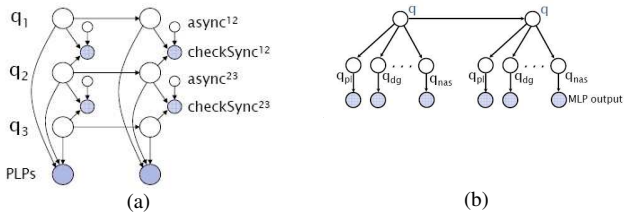

Fig. 1. *Two frames of DBNs for recognizers using (a) a multistream AF-based pronunciation model (b) an AF-based hybrid observation model. q_F is the hidden state of feature F; q is the phonetic state; $async^{ij}$ is the current degree of asynchrony between streams i and j; and the $checkSync^{ij}$ variables enforce the synchrony constraints.*

3.1. Experiments

We use features based on articulatory phonology [19], consisting of the lip, tongue, glottis, and velum states. We group all lip features into one stream, tongue features into another stream, and the glottis and velum into a third stream. We allow at most two states of asynchrony between streams, with complete synchronization at word boundaries.

3.1.1. Audio-only experiments on SVitchboard

We trained AF-based models analogous to the monophone baseline, referred to as "monofeat" models, by conditioning the observation vector on the current AF states only. We consider 1-state and 3-state versions, analogously to 1-state and 3-state monophones. In the 3-state version, each sub-phone state has different AF values. Word alignments were used in training. A selection of the error rates is shown in Table 2. In "3-state monofeat, sync", all AF streams are completely synchronized, which we expect to be worse than baseline since there are not unique AF values for all phone states. The asynchronous model slightly underperforms the synchronous one, indicating that the asynchronous states do not help in this case. We have found that many states corresponding to asynchronous configurations have extremely low occupancies and therefore poorly trained Gaussians, a possible cause of the impaired performance. Ongoing work is focused on improved initialization approaches and state tying to alleviate low-occupancy issues. Work is also ongoing to model the effects of AFs straying from their target values [8] and to account for the effects of linguistic context, as well as to relax the word boundary synchronization constraint.

Model	10-wd WER	500-wd WER
Monophone	19.6	65.0
1-state monofeat	28.5	74.8
3-state monofeat, sync	20.7	65.2
3-state monofeat, async	21.3	67.4

Table 2. *Test set word error rates of the baseline and "monofeat" models.*

3.1.2. Audio-visual experiments on CUAVE

We apply a similar model to audio-visual speech recognition (AVSR). A common approach is a "phoneme-viseme" model,

with one stream for the "audible" state and one for the "visible" state [20]. AFs have previously been proposed for use in AVSR [21], but to our knowledge have not been used for continuous word recognition. The task here is recognition of read digit strings from a portion of the CUAVE corpus [22]. Noise is added to the test (but not training) data. The acoustic observations are MFCCs, and the visual observations are discrete cosine transform coefficients of a region of interest around the mouth [1]. It has been shown that, on this task, adding the visual signal improves recognition and asynchronous phoneme-viseme models outperform synchronous ones [23].

We use the same AF-based model as in Sec. 3.1.1, but with two observation vectors. We assume that the audio and video are independent given the AFs, and that each depends on all three AF streams. Table 3 shows development set results. The phoneme-viseme models are also new, in that we use a 2-stream version of the structure in Fig. 1(a). The AF model does not differ significantly in performance from the phoneme-viseme ones, but it makes different mistakes, and a system combination (ROVER) including an AF model outperforms one using only phoneme-viseme systems. We plan to apply these models to more complex tasks, where the asynchrony may be more pronounced.

Model	WER
Phoneme-viseme, 1 state async	22.6
Phoneme-viseme, 2 state async	21.8
AF-based	22.1
ROVER, best 3 phoneme-viseme	20.1
ROVER, best 3 including AF	19.4

Table 3. *Word error rates on the CUAVE AVSR task, averaged across development sets at several SNRs from clean to -4 dB.*

4. CLASSIFIER-BASED OBSERVATION MODELING

We now consider the use of AF classifiers in observation modeling, in both a hybrid approach (Sec. 4.1) and a tandem approach (Sec. 4.2). For this section, we assume that the hidden structure is a single stream of phonetic states as in the baseline models. The AF set here includes place and degree of constriction, nasality, rounding, glottal state, and vowel quality (see [14] for more details), and the classifiers are MLPs. Two versions of the MLPs were trained: one on the Fisher and Switchboard 2 databases minus Switchboard 1 (1776 hours); and one on only the SVitchboard training set. The rationale for the former is that MLPs could be trained on a large database, then ported to a data-poor domain. Initial training labels were produced from forced phone alignments converted to AF labels [2].

4.1. Experiments with hybrid AF-based models

In hybrid models, $p(o|q)$ is replaced with a scaled likelihood estimated from the MLP outputs [9]. In contrast to standard hybrid approaches, we have multiple state variables $q_f, f \in \{place, degree, ...\}$. We use a non-deterministic (learned) mapping from the phonetic state q to the AF states q_f, and a distribution $p(o|q_f)$ for each AF given by the MLPs. The DBN for this model is shown in Fig. 1(b). AF-based hybrid models have been used previously [5], although with a deterministic mapping between phones and AFs.

Results obtained with a monophone version of this model, using the Fisher-trained MLPs, on the 10-word test set are shown in Table 4. The hybrid model alone is far behind the HMM baseline. When this model is used to align the SVitchboard training set and the MLPs are retrained on these alignments, performance improves drastically, although still remaining behind the baseline. The hybrid approach may hold promise for cross-domain or cross-lingual work; a domain in which there is little data may benefit from classifiers trained on a data-rich domain, and AFs may be more domain- and language-independent than phones.

Model	WER
Monophone baseline	20.0
Hybrid	30.1
Hybrid + embedded training	24.3

Table 4. *Word error rates of hybrid models on the 10-word test set.*

4.2. Experiments with AF-based tandem observations

In the tandem approach, the MLP outputs are post-processed and appended to the acoustic observation vector, and the combined vector is modeled as usual with Gaussian mixtures. This approach has been used in state-of-the-art large-vocabulary systems [24]. We experiment with variants of this approach using the AF MLPs [13]. The idea is similar to the approach of Kirchhoff [5], although we use different ways of combining the MLP outputs with the PLPs. We also experiment with factoring the observation model into two factors, one over the PLPs and one over the MLP outputs, which can reduce training data needs and allows for different state clusterings for the two factors. Results on the 500-word task are shown in Table 5. No training word alignments were used.

The main conclusions are: the Fisher-trained MLPs significantly outperform the SVitchboard ones; the AF-based tandem monophone slightly outperforms the standard phone-based tandem monophone, though not significantly; and factoring gives a significant performance improvement. The triphone model with factored AF tandem observations gives the lowest error rate to date on this test set. Ongoing work is focused on other factorizations, such as one factor per AF.

5. ANALYSIS USING MANUAL TRANSCRIPTIONS

One of the obstacles in AF-based recognition research is the lack of ground-truth AF values. To begin remedying this, we have collected a small set of utterances labeled manually at the AF level, including 78 SVitchboard utterances, for testing accuracies of classifiers and alignments. The feature set used

[1] Thanks to Amar Subramanya for providing the visual observations.

[2] We are grateful to SRI for providing the phone alignments.

Model	WER
1: Monophone w/PLPs (baseline)	67.7
2: Monophone w/phone tandem, SVB-trained	63.0
3: Monophone w/AF tandem, SVB-trained	62.3
4: Monophone w/AF tandem, Fisher-trained	59.7
5: (4) + factoring	59.1
6: Triphone w/PLPs (baseline)	59.2
7: Triphone w/AF tandem, Fisher-trained	55.0
8: (7) + factoring	53.8

Table 5. *Word error rates for tandem models on the 500-word test set.*

for labeling is a slightly more detailed version of the observation modeling features; see [14] for more information.

Analysis of classifier and AF alignment accuracies have yielded some interesting observations. For example, comparing the 1-state monofeat and hybrid models, we find that AF alignments computed with the monofeat model are closer to human labels than the hybrid alignments are, and MLPs retrained on these alignments outperform ones trained on hybrid alignments. On the other hand, the 1-state monofeat has very poor recognition performance, and the hybrid alignments do improve recognition when used for embedded MLP training. These results highlight the fact that different models may be appropriate for different purposes; for example, the monofeat models may be useful for data transcription and analysis.

6. CONCLUSIONS

We have summarized the main aspects of our investigations of AF-based speech recognition. The most encouraging recognition results thus far come from the tandem observation experiments. Hybrid models, while currently behind other models in terms of accuracy, require very little training data beyond the MLP training, and may therefore hold greater promise in multilingual scenarios. The multistream models we have tested on both audio-only and audio-visual speech are close to, but not outperforming, phone-based models. When used to produce forced AF alignments of training data, however, the new pronunciation models are able to improve the accuracy of the MLP AF classifiers, suggesting an alternative use for such models in data transcription and analysis. Additional contributions of this work are: new SVitchboard baselines; a set of manual feature-level transcriptions; and several tools, including gmtkTie for generalized state tying in GMTK, site-independent parallel training and decoding scripts for GMTK, and a tool for visualizing GMTK Viterbi paths. The transcriptions, classifiers, and tools will be available for download from `http://people.csail.mit.edu/klivescu/WS06AFSR`.

Beyond the ongoing work mentioned above, a longer term goal is to combine classifier-based observation models with multistream pronunciation models. To achieve state-of-the-art performance, it may be necessary to model effects such as articulatory substitution, cross-word asynchrony, and context dependence. Another direction is in the area of automatic AF transcription for analysis. If we can accurately align a large data set automatically, we can address scientific questions about articulatory feature behavior, and the answers should in turn serve to improve AF-based recognition models.

7. REFERENCES

[1] R. C. Rose, J. Schroeter, and M. M. Sondhi, "An investigation of the potential role of speech production models in automatic speech recognition," in *ICSLP*, 1994.

[2] M. Ostendorf, "Moving beyond the 'beads-on-a-string' model of speech," in *IEEE ASRU Workshop*, 1999.

[3] L. Deng, G. Ramsay, and D. Sun, "Production models as a structural basis for automatic speech recognition," *Speech Communication*, vol. 33, pp. 93–111, 1997.

[4] S. King et al., "Speech production knowledge in automatic speech recognition," *To appear in JASA*, 2007.

[5] 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.

[6] H. Soltau, F. Metze, and A. Waibel, "Compensating for hyperarticulation by modeling articulatory properties," in *ICSLP*, 2002.

[7] S. Stueker et al., "Integrating multilingual articulatory features into speech recognition," in *Eurospeech*, 2003.

[8] K. Livescu and J. R. Glass, "Feature-based pronunciation modeling with trainable asynchrony probabilities," in *ICSLP*, 2004.

[9] N. Morgan and H. Bourlard, "Continuous speech recognition: An introduction to the hybrid HMM/connectionist approach," *IEEE Signal Processing Magazine*, vol. 12, no. 3, pp. 25–42, 1995.

[10] H. Hermansky, D.P.W. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *ICASSP*, 2000.

[11] M. Wester, J. Frankel, and S. King, "Asynchronous articulatory feature recognition using dynamic Bayesian networks," in *IEICI Beyond HMM Workshop*, 2004.

[12] K. Murphy, *Dynamic Bayesian Networks: Representation, Inference and Learning*, PhD dissertation, U. C. Berkeley, Berkeley, CA, 2002.

[13] O. Cetin et al., "An articulatory feature-based tandem approach and factored tandem observation modeling," in *ICASSP*, 2007.

[14] K. Livescu et al., "Manual transcription of conversational speech at the articulatory feature level," in *ICASSP*, 2007.

[15] K. Livescu et al., "Articulatory feature-based methods for acoustic and audio-visual speech recognition: JHU Summer Workshop Final Report," Technical report, Johns Hopkins University Center for Language and Speech Processing, 2007, in preparation.

[16] S. King, C. Bartels, and J. Bilmes, "SVitchboard 1: Small vocabulary tasks from Switchboard 1," in *Interspeech*, 2005.

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

[18] N. Ganapathiraju et al., "Resegmentation of SWITCHBOARD," in *ICSLP*, 1998.

[19] C. P. Browman and L. Goldstein, "Articulatory phonology: An overview," *Phonetica*, vol. 49, pp. 155–180, 1992.

[20] G. Potamianos et al., "Automatic recognition of audio-visual speech: Recent progress and challenges," *Proc. IEEE*, vol. 91, no. 9, 2003.

[21] P. Niyogi, E. Petajan, and J. Zhong, "Feature based representation for audio-visual speech recognition," in *AVSP*, 1999.

[22] E. Patterson et al., "CUAVE: A new audio-visual database for multimodal human-computer interface research," in *ICASSP*, 2002.

[23] J. Gowdy et al., "DBN based multi-stream models for audio-visual speech recognition," in *ICASSP*, 2004.

[24] Q. Zhu et al., "Incorporating tandem/HATs MLP features into SRI's conversational speech recognition system," in *EARS RT-04F Workshop*, 2004.