DBN-BASED MULTI-STREAM MODELS FOR MANDARIN TONEME RECOGNITION

Xin Lei, Gang Ji, Tim Ng, Jeff Bilmes, Mari Ostendorf

Electrical Engineering Dept., University of Washington, Seattle, WA 98105

{leixin,gang,tng,bilmes,mo}@ee.washington.edu

ABSTRACT

A toneme in Mandarin Chinese is a tonal phone which consists of a base phone (main vowel) and a tone. To capture both, most recognition systems use two feature streams: the standard MFCCs for the base phones, and pitch features for the tones. In this paper we propose the use of Dynamic Bayesian Networks for modeling the two streams in toneme recognition. We used the Graphical Model Toolkit to build and compare three different models: a standard HMM with concatenated features, and synchronous and asynchronous multi-stream systems. Stream-level model parameter tying is also exploited. The toneme recognition results show significant improvements by using the multi-stream models.

1. INTRODUCTION

Mandarin Chinese has the largest number of speakers in the world, and many studies on Mandarin speech recognition have been conducted, e.g. [1, 2, 3, 4]. Unlike English, Mandarin is a tonal language that benefits from explicitly modeling the five tones that are necessary to distinguish between ambiguous words. The five tones are characterized as high (1), rising (2), low (3), falling (4) and neutral (5). One popular way of modeling the tones [3] combines the main vowel with different tones as different phonemes, called *tonemes*. For example, a1, a2, a3, a4 and a5 are five different tonemes associated with the main vowel "a".

To distinguish the tonemes with the same main vowel, we need to add pitch information to our feature set. Therefore, the features for the toneme acoustic models consist of the standard Mel-scale cepstral coefficients (MFCCs) and their deltas as well as the new pitch features. The MFCCs model the main vowel part of the toneme, and the pitch features model the tone of the toneme. How best to integrate the multi-stream features has been a research topic for a long time. Previous approaches can be classified into three categories: feature fusion (or early integration), decision fusion (or late integration), and model fusion. Previous work [3, 4] adopted feature fusion, which simply concatenated the two streams of features into a single feature vector which is then modeled by a standard Gaussian mixture.

In this work we use multi-stream models expressed with a Dynamic Bayesian Network (DBN) to jointly represent both MFCC and pitch features, and show that it improves Mandarin toneme recognition. With a multi-stream model, we can use different stream exponents and we can also tie the MFCC or pitch model of a toneme separately. Using a DBN greatly simplifies statistical modeling issues, and the method can be easily extended to use asynchrony between streams or condition features across streams [5, 6, 7, 8].

In section 2, we describe the DBN based single stream and multi-stream models for training and decoding the tonemes. In section 3, the experiments and results are given. Finally, we summarize key points in section 4.

2. DBN-BASED MODELS FOR TONEME RECOGNITION

A Bayesian network is a statistical model that can be used to represent collections of random variables and their dependency relationships. A dynamic Bayesian network (DBN) is a Bayesian network with random variables for each time frame that share their underlying conditional distributions. A DBN can be thought of as a generalized HMM. DBN-based graphical models have been proposed to model continuous speech using the Graphical Models Toolkit (GMTK) in [9]. In this work, we use the GMTK framework to build a single-stream toneme recognition system (simple HMM), and then extend the baseline system to synchronous and asynchronous multi-stream systems.

2.1. Multi-stream models

A multi-stream model is a product model of the different feature streams. For S independent streams, the output distribution for state j using a Gaussian mixture is defined as:

$$b_j(o_t) = \prod_{s=1}^{S} \left[\sum_{m=1}^{M_s} c_{jsm} \mathcal{N}(o_{st}; \mu_{jsm}, \Sigma_{jsm}) \right]^{\gamma_s} \quad (1)$$

where M_s is the number of mixture components for stream s, c_{jsm} is the weight of the m-th component and $\mathcal{N}(\cdot; \mu, \Sigma)$ is a multivariate Gaussian with mean μ and covariance Σ [10]. The exponent γ_s is the weight for stream s.

There are several advantages to using a multi-stream model. First, since the two streams are quite different here, the factored distribution assumption is reasonable and efficient. Secondly, we can assign different stream weights to

emphasize particular streams, which can be optimized using brute force search or discriminative training [11]. Thirdly, it becomes quite easy to tie the mixture models in different streams using strategies that are specific to each stream. For example, tonemes with the same main vowel can share the same MFCC stream distribution, while tonemes with the same tones can share the same pitch stream distribution. Tying is especially useful in the case of limited training data and using context-dependent models. Finally, another key advantage of multi-stream models (particularly when implemented as a DBN) is that the streams may desynchronize. Specifically, it is possible to have higher level objects control lower level objects that can complete at different times.

2.2. DBN-based graphical models

The training and decoding graphs of the synchronous 2-stream DBN-based model are given in Figure 1 and Figure 2, respectively. This synchronous multi-stream model is similar to the HTK synchronous multi-stream model [10]. In the single stream case, there is only one observation node per frame. If there are more than two streams, we only need to add in more observation nodes which are emitted from the state node in the frame. More detailed information about the graphs is given in [5, 9]. Here we used a slightly different naming convention for our toneme recognition application. Each toneme has three states, as is most commonly used in HMM speech recognition systems. The major difference between the two graphs is that during training the phone sequence is known and indicated with the *Phone Counter* nodes, while during decoding we use phoneme-level language models (LMs). In the first frame, a phoneme-level unigram is used, since there is no parent for the phone variable. In subsequent frames, a phone bigram describes the random relationship between successive phone variables, enabled only when there is a phone transition in the specified frame, as shown in Figure 2 where the dashed arrow shows that a switching parent [12] controls the random relationship between successive phoneme nodes.

The synchronous models require that multiple observations are associated with the same state at each time step. In contrast, our asynchronous model (Figure 3) allows multiple hidden streams to desynchronize to a certain degree. A key advantage of our asynchronous system is its ability to easily explain speech variations by the differing degrees of asynchrony within the hidden streams. Representing such asynchrony using a standard HMM would require many more states, and training such an HMM with exactly the same constraints would be extremely difficult. Moreover, any increase in decoding costs associated with the larger state space under asynchrony can often be significantly reduced either by spending time searching for a better graph triangulation [13] or by using approximate inference.

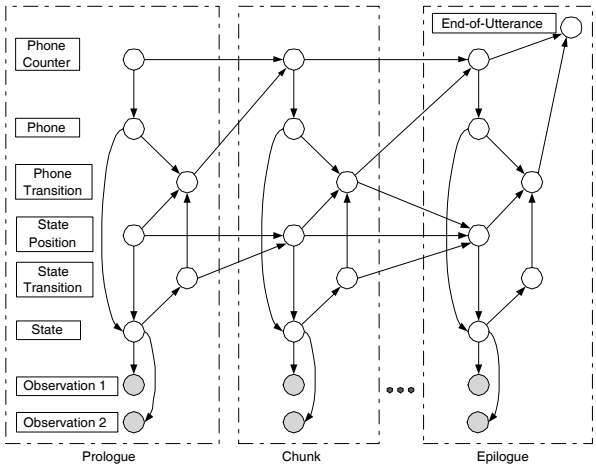

Fig. 1. *Synchronous multi-stream DBN model for training.*

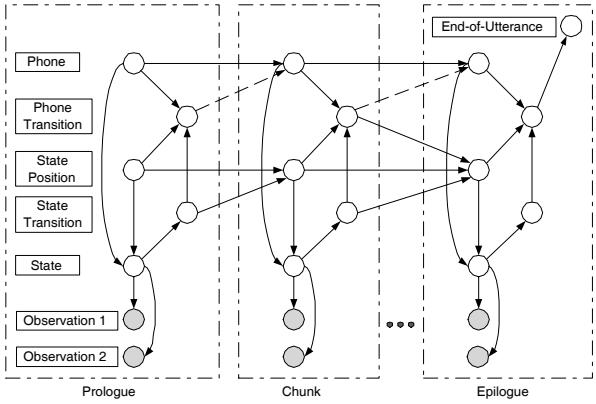

Fig. 2. *Synchronous multi-stream DBN model for decoding.*

3. EXPERIMENTS AND RESULTS

3.1. Task description

The experiments are based on the CallHome and CallFriend Mandarin corpora. The training set has roughly 35 hours of telephone speech from Mandarin speakers collected in the U.S. The testing set has 1,516 utterances with a total duration of about 1.5 hours. In this work, we recognize phones and tonemes, and report accuracy of recognizing these units to directly assess the acoustic modeling changes. We assume that gains will translate to character error rate, as improvements in phone accuracy typically translate to word accuracy in English (and because many sites report performance gains from improved tone modeling in Mandarin).

3.2. Feature extraction

The standard MFCC features are generated with the front-end of the SRI Decipher speech recognition system. The

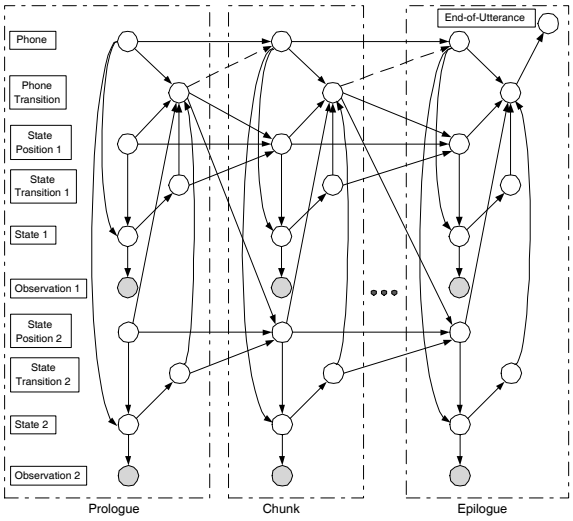

Fig. 3. *Asynchronous multi-stream DBN model for decoding.*

speech input was processed using a 25ms Hamming window, with a frame rate of 10ms. For each frame there are 13 MFCC features plus delta and double deltas, resulting in a 39 dimension feature vector.

The pitch features include the pitch, delta pitch and delta delta pitch. The pitch is extracted with ESPS *get_f0* and then processed by SRI pitch processing tool *graphtrack* [14] to eliminate halving and doubling errors. The pitch is then smoothed with an algorithm similar to [3]. The utterance mean is used for unvoiced pitch, and a moving average filter is used to smooth the pitch contour. Finally, the pitch features are mean and variance normalized per speaker.

3.3. Experiment setup

The phone set used here includes 65 non-tonal phones and tonemes as listed in Table 1. Since we compare the multi-stream models to single-stream models in context-independent phone recognition, and the pitch stream has little effect on the non-tonal phones, we focus on toneme recognition accuracy for evaluating system performance. For reference, we also provide results (accuracy) for the full phone set and for the 9 main vowels ignoring tone distinctions (E, EE, N, R, a, ey, i, o and u).

Table 1. *Phone set for Mandarin speech recognition.*

Category	Phones
Non-tonal phones	sp C S W Z b c d f g h j k l lau m n p q r rej s t w x y z
Tonemes	E1 E2 E3 E4 EE1 EE2 EE3 EE4 EE5 N1 N2 N3 N4 N5 R2 R4 a1 a2 a3 a4 a5 ey1 ey2 ey3 ey4 i1 i2 i3 i4 o1 o2 o3 o4 o5 u1 u2 u3 u4

In our baseline DBN system, we concatenated the MFCC and pitch features for each frame into a single feature vector. GMTK [12] was used to build all the DBN-based systems.[1] An HMM-based system was also built for performance comparison with the Hidden Markov Model Toolkit (HTK) [10]. The second experiment separated the two feature streams in a synchronous multi-stream DBN. In the third experiment, we tied the MFCC mixtures of different tonemes with the same main vowel, and tied the pitch mixtures of those tonemes with the same tones. Finally, we did experiments with asynchronous multi-stream models. In all cases, 32 Gaussian components were used for each mixture model, and decoding used a phoneme bigram.

Though the multi-stream models are trained without stream weights, different weights can be used for the two streams in decoding. We tried different pairs of weights and found that within some range a larger pitch stream weight improves the toneme accuracy but hurts the full phone set and main vowel accuracy. In all multi-stream experiments, we used 0.5 for the MFCC stream weight and 0.6 for the pitch stream weight to emphasize the tones. The LM weight is fixed, chosen to balance the insertion and deletion errors.

3.4. Results and discussions

The full phone set, main vowel and toneme recognition results are listed in Table 2. Note that accuracy on the baseline HMM system is low in part because of using context-independent phone models, but also because of the difficulty of the task. Conversational speech tends to have much more variability than read speech, and the crosstalk associated with telephone speech poses problems for extracting pitch features. As an indicator of difficulty, we note that the current best reported character error rate on this test set is 42.7% after combining multiple systems [15]. For the single-stream case, the GMTK system gives a significant improvement over the HTK system, as consistently observed in other experiments, probably due to its novel method of handling Gaussian mixture training via a splitting/vanishing algorithm [12].

The synchronous multi-stream system without tying provides 1.3% improvement in terms of main vowel accuracy and 3.4% improvement in terms of toneme accuracy. There are 65 phonemes, 3 states per phone, and 42-dimensional features. We used diagonal covariances and the feature dimension is 42. Therefore, the total number of parameters in the single-stream diagonal covariance HTK and GMTK single-stream systems is roughly 530k, which is the same as in the multi-stream system without tying. There are 36 unique base phones and 6 different tones (including "no-tone" for non-tonal phones) in the phone set, so the multi-stream system with tying has 108 tied MFCC

[1] We used a 2004 version of GMTK supporting LM penalties and scales.

Table 2. *Recognition accuracy results.*

System	Phone Acc.	Main vowel Acc.	Toneme Acc.	# of Params
HMM single-stream	25.8%	42.5%	20.0%	530k
DBN single-stream	28.5%	44.3%	21.9%	530k
DBN multi-stream	29.7%	45.6%	25.3%	530k
DBN multi-stream with tied mixtures	27.4%	43.6%	22.9%	277k
Asyn. DBN multi-stream	30.2%	46.0%	25.8%	530k
Asyn. DBN multi-stream with tying	27.4%	43.3%	22.7%	277k

mixtures and 18 tied pitch mixtures, resulting in a parameter size of 277k. Hence, the model size of the tied-mixture multi-stream system is nearly half that of the system without tying. Even with a much smaller model size, the tied-mixture multi-stream system still has better toneme recognition performance than the single-stream system. Its main vowel accuracy is worse than the single-stream system, because it has a smaller MFCC model size due to tying and the base vowels are entirely characterized by MFCC's. The overall accuracy loss shows that 35 hours of training data is enough for context-independent models, but tying may still be advantageous when using context-dependent models.

The performance of the synchronous and asynchronous multi-stream systems are very close. One possible reason is that in our asynchronous graph, we force the two streams to be synchronous at the phone rather than the syllable level, i.e. allowing state asynchrony only within phones. It could also be that we need further tuning of the stream weights.

4. CONCLUSIONS

In this paper we have described DBN-based single-stream and multi-stream models for Mandarin toneme recognition. The DBN-based models were implemented using GMTK to integrate pitch and cepstral features. The results show that the use of DBNs leads to significant improvement in toneme accuracy over a traditional HMM-based system, and that multi-stream models outperform single-stream models. Using multi-stream models also enables parameter sharing at the stream level, which will benefit larger size models in limited data conditions, but did not prove to be useful in these experiments based on context-independent models. Asynchronous multi-stream models are investigated but no significant gains are obtained.

There are several possible directions for future work. First, the stream weight estimation could be further studied. Currently we used a simple heuristic method to search for the best stream weights. For better estimation of the stream weights, discriminative training methods could be tried. Another future direction is data-driven parameter sharing, which will allow more complex tying schemes than the strict separation of tones and phones explored here.

Acknowledgments

The authors thank O. Cetin and C. Bartels for consultation on GMTK and multi-stream models. This work was supported by the Defense Advanced Research Projects Agency grant MDA972-02-C-0038. The opinions and conclusions are those of the authors and are not necessarily endorsed by the United States Government.

5. REFERENCES

[1] H.M. Wang et al., "Complete recognition of continuous Mandarin speech for Chinese language with very large vocabulary," *IEEE Trans. Speech & Audio Proc.*, vol. 5, pp. 195–200, 1997.

[2] Y.F. Liao and S.H. Chen, "A Modular RNN-based method for continuous Mandarin speech recognition," *IEEE Trans. Speech & Audio Proc.*, vol. 9, no. 3, pp. 252–263, 2001.

[3] C.J. Chen et al., "New methods in continuous Mandarin speech recognition," in *Proc. Eurospeech*, 1997, vol. 3, pp. 1543–1546.

[4] E. Chang et al., "Large vocabulary Mandarin speech recognition with different approaches in modeling tones," in *Proc. ICSLP*, 2000, vol. 2, pp. 983–986.

[5] Y. Zhang et al., "DBN based multi-stream models for speech," in *Proc. ICASSP*, 2003, vol. 1, pp. 884–887.

[6] J.N. Gowdy et al., "DBN based multi-stream models for audio-visual speech recognition," in *Proc. ICASSP*, 2004, vol. 1, pp. 993–996.

[7] O. Cetin and M. Ostendorf, "Cross-stream Observation Dependencies for Multi-stream Speech Recognition," in *Proc. Eurospeech*, 2003, pp. 2517–2520.

[8] J. Bilmes, "Buried Markov models for speech recognition," in *Proc. ICASSP*, 1999, vol. 2, pp. 713–716.

[9] J. Bilmes et al., "Discriminatively structured dynamic graphical models for speech recognition," Johns Hopkins University, CSLP 2001 Summer Workshop Final Report, 2001.

[10] S. Young et al., "The HTK Book (for Version 3.1)," 2001.

[11] H. Glotin et al., "Weighting schemes for audio-visual fusion in speech recognition," in *Proc. ICASSP*, 2001, vol. 1, pp. 173–176.

[12] J. Bilmes and G. Zweig, "The graphical models toolkit: An open source software system for speech and time-series processing," in *Proc. ICASSP*, 2002, vol. 4, pp. 3916–3919.

[13] J. Bilmes and C. Bartels, "On triangulating dynamic graphical models," in *Uncertainty in Artificial Intelligence: Proc. 19th Conference*, 2003, pp. 47–56.

[14] K. Sonmez et al., "Modeling dynamic prosodic variation for speaker verification," in *Proc. ICSLP*, 1998, vol. 7, pp. 3189–3192.

[15] R. Schwartz et al., "Speech recognition in multiple languages and domains: the 2003 BBN/LIMSI EARS system," in *Proc. ICASSP*, 2004, vol. 3, pp. 17–21.