PHMM BASED ASYNCHRONOUS ACOUSTIC MODEL FOR CHINESE LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION

Hao Wu, Xihong Wu* and Huisheng Chi

Hearing Research Center
Key Laboratory of Machine Perception (Ministry of Education)
Peking University, Beijing, 100871, China
{wuhao,wxh,chi}@cis.pku.edu.cn

ABSTRACT

In this paper, we presented an asynchronous multiple stream based Chinese tonal acoustic modeling framework. In this framework, toneless phonetic units and tones are modeled separately with different acoustic features. During the training and decoding process, a set of models are coupled together with a product hidden Markov models (PHMM) to represent whole tonal phonetic units. Through this, a compound context dependent tonal model can be generated from a few simple models. Experiments show that such model scheme generates more compact and accurate model presentation and brings improvement on the performance for large vocabulary speech recognition tasks.

Index Terms— tonal language, multiple stream, PHMM

1. INTRODUCTION

Being an important part of Chinese phoneme system, tonal information is very important for solving the confusion between different words or Chinese characters. There has been continuous interest on incorporating tonal acoustic information in Chinese speech recognition.

Most tonal Chinese acoustic modeling schemes represent tonal information by tonal phonetic units. For example, several state-of-art Chinese speech recognition systems use initials and tonal finals as basic modeling units based on the observation that finals are the major part of a Chinese syllable to carry the tonal information, the initials are usually considered as toneless[1, 2]. However, using tonal units brings a dramatic increase of the number of the units, and the unevenness of the tonal units occurrence may brings inaccurate parameter estimation for those models with less training data. Moreover, because of its complexity, tonal phonetic unit based acoustic modeling scheme can hardly incorporate the cross syllable tonal context. Since the implementation of the tone is sensitive to the tones of neighboring syllables, overlooking such context information may brings degradation on the precision of tone modeling.

In Chinese, the suprasegmental information of tone is considered as being conveyed by glottal acoustic features such as pitch, while the segmental phoneme is supposed to be physically represented by vocal-tract related features, such as MFCC and PLP. Since these two types of features are relatively independent, two-stream HMM models can be used to model Chinese tonal phonetic units, including both tones and initial/finals (IFs). [3] reported superior results using the method over traditional approach. Also, relating the features of different streams to different units makes it possible to applying the parameters tying and unit context modeling separately on each stream, which gives great flexibility for generating more informative models with simple models.

When utilizing multiple stream models, synchronization of two stream is an important issue to be considered. Since movements of glottis and vocal tract are not in exact synchrony in most cases, allowing asynchronism between streams is helpful for increasing the precision of the model. To incorporate such asynchronism, a lot of renovate model structures can be used. Depending on the way states being used to present temporal evolvement, these models can be categorized into two main streams. The graphic model structure (GM), which is considered as a general and flexible structure for representing complicated multiple-observation streams, has been utilized in the area of automatic speech recognition in recent years, and been successfully applied to the Chinese toneme recognition task [4]. On the other hand, the Product Hidden Markov model (PHMM) being studied intensively in the area of audio-visual speech recognition, is another representative structure which is extended from traditional HMM structure by states coupling and transition matrices coupling on streams[5, 6].

Generally speaking, GM is more flexible and presentive than PHMM. However, being a direct extension of HMM, PHMM can be easily and efficiently implemented. Considering the complexity of the LVCSR task, we choose the convenient solution of transferring from traditional HMM based

*Correspondent author

LVCSR systems to PHMM based LVCSR system. In this LVCSR system, tonal syllable PHMMs are generated as basic models for acoustically representing the language model units. When generating these models, each syllable is decomposed down to the toneless phonetic units and the tone unit. The two types of units are physically represented with MFCC and pitch features respectively. While the parameter estimation and LVCSR decoding are performed on the jointed PHMMs , parameters of the models in two streams are clustered and refined separately.

The rest of this paper is organized as follows: the second section gives a detailed description of model specification of each individual stream, including model units definition, feature extraction and HMM topologies, etc. The third section illustrates the multiple stream framework, especially the product-HMM model topology and formulation. The experiment results and conclusions are given in the last two sections.

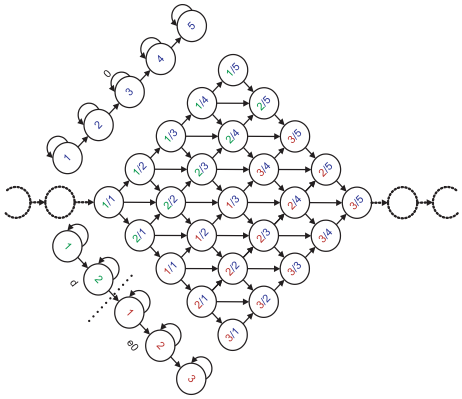

Fig. 1. Building PHMM for syllable "de0" from initial, final and tone models

2. MODEL STRUCTURES FOR THE STREAMS

For the MFCC stream, we took initials and finals (both toneless) as basic units. Each unit is modeled by an HMM containing 2 to 5 states,(the number of states is determined according to the phonetic structure of the unit). For the pitch stream, the tones of syllables are used as basic units, and each syllabic tone unit is modeled with a HMM. the multiple stream framework provides the flexibility of choosing different context modeling schemes for different streams. For both streams, cross-syllable context information were adopted.

2.1. MFCC stream modeling

In the "toneless" MFCC stream, the model set consists of totally 66 basic units including 37 toneless finals (35 standard finals and 2 allophone for "i"), 27 initials (23 standard initials and 4 dummy initials for zero-initial syllables) and 2 fillers. Each unit is modelled with left-to-right HMMs. For each unit, we use triphone like context dependent modeling which take the adjacent initials or finals as the contexts. This generates a relatively less amount of context dependent units compared with a complete context dependent tonal modeling scheme.

Conventional 39-dimension MFCCs (12 static MFCCs, log energy, and their first- and second-order time derivatives) are extracted as the feature for stream. Cepstral mean normalization is performed.

2.2. Pitch stream modelling

In the "tonal" stream, the model set consists of totally 16 basic units: except for filler unit SIL, each of 5 tones (including neutral tone) derives 3 models containing 6, 7 or 8 states, so that for each legal initial-final models concatenation, there is one corresponding tone model with same state number. Like in toneless stream, we uses the triphone like context dependent modeling which take tones of adjacent syllables as the model contexts. Models of the same tone with different states are regarded as the same context. This makes a small context dependent unit set, which contains 541 context dependent units.

Pitch was extracted using the method described in [3]. A patching and smoothing process is performed on the pitch sequence to make the contour continuous and smooth. Like in the MFCC stream, the 1st- and 2nd-order time derivatives of pitch are also calculated and concatenated with the pitch feature, which makes a 3-dimension pitch feature vector.

3. COUPLING OF TWO STREAMS

3.1. Multiple stream product hidden Markov model

When training or decoding with above mentioned models, the two streams are correlated with each other while they are not necessarily state synchronous. In PHMM, each state is built by merging an N-tuple (N = 2 here) of states from the model for each stream. The topologies of PHMM are defined so as to represent all the possible state paths given the initial HMM topologies for each stream. This model allows implementing independent search within syllable as well as synchrony constraints at syllable boundaries. Fig 1 shows how the PHMM of Chinese syllable de0 is build with a two-state initial model "d", a three state final model "e" and a five-state tone model "0".

In product HMM, the observation likelihood for each composite state is computed as:

$$\Pr[O^{(t)}|\Lambda] = \prod_{s\in\{MFCC,pitch\}} [\sum_{m=1}^{M_s} \omega_{sim} N(O^{(t)}; \mu_{sim}, \sigma_{sim})]^{\lambda_s} \quad (1)$$

The variable M_s represents the number of mixture components associated with state i, μ_{sim} and σ_{sim} are the mean and the diagonal covariance matrices corresponding to stream s given the state i and mixture component m, and ω_{sim} are the mixture weights corresponding to the state i. λ_s are the exponent stream weight. By assigning different stream weights, a particular stream can be emphasized in likelihood computation.

In our implementation, for each syllable, the initial and final models are first selected according to the context information and concatenate into a toneless syllable model, then the corresponding tone model are selected according to the state number of the toneless model conjunction and the given tonal context. The two models are then coupled into a syllable PHMM. As shown in Fig. 1, transition matrix between composite states were calculated as a product of the transition matrices of the two streams.

As the PHMM were introduced, there is an substantial increase in complexity in terms of time and computation due to the increase of the composite states. Also, allowing complete loose asynchronism between streams may brings excessive time stretch or compression to each stream. To make the training and decoding procedure feasible and efficient, we made a constraint that the state propagating asynchrony between the two streams is limited to certain range, which is refereed to as asynchronism width (with symbol τ in this paper), indicating how many states of asynchronism can be reached in the PHMM.

Following the expectation-minimum (EM) principle, HMM parameters for the two streams are trained jointly on PHMMs with a revised Baum-Welch algorithm [5]. For each utterance in training set, PHMMs are generated for each syllable in the transcription and then concatenated to represent the utterance. With the corresponding observation sequence, Backward-Forward algorithm for the 'E' step is then applied on this concatenated model to get the statistics from corpus. The renewed parameters, including means, covariances, mixture weights tied states and the decomposed transition matrices are then updated separately to the models for each stream for the 'M' step.

4. EXPERIMENTS AND RESULTS

In our experiments, the training corpus includes cleaned wideband male speech from three continuous mandarin speech database: the 863-1 standard database, the 863-2 accent database, and the Intel lab-recorded database. The total amount of data is more than 340 hours.

Hub-4NE Mandarin broadcast speech corpus was used for evaluation. By extracting wideband clean male speech data from the eval set, we got 238 spoken sentences for testing the acoustic models. Also, 649 sentences in similar spoken conditions were extracted from the HUB4-NE development set for decoder parameters optimization.

A HMM based Chinese LVCSR system is modified to adopt the PHMM acoustic model by revising the likelihood computation and adding model coupling module. A n-gram based language model with 60,000 words were used for all evaluating tasks. Chinese character error rate (CER) is used as the major performance index.

4.1. Baseline models

A baseline single stream acoustic model and a traditional two stream acoustic model are trained with same training set and feature specification, and were both used for comparison with the proposed models.

The baseline model is a traditional one-stream triphone based context dependent initial-tonal final model. The unit set contains 27 initials, 180 tonal finals and 2 fillers as basic units. The model is trained with standard Baum-Welch algorithm. Decision tree based parameter tying is applied and generated a state-tied context dependent model with 8,000 tied states. Each tied state in this model is represented by a 32 component gaussian mixture model.

A two-stream extension of the first baseline is also used for comparison. Except for the log-likelihood computation of the HMM state, the specification of this model is same as that of the first baseline.

4.2. PHMM models

Proposed models with different configurations are trained with the joint model training scheme. We adjusted two parameters as the main factor to control the constraints of the model as mention in 3.1:τ and λ. For τ, we applied the width of the asynchronism to 0, 1, 2 and 3. For λ, we tried configurations of $\lambda_{MFCC}/\lambda_{pitch} = 1.0/0.0, 0.8/0.2, 0.7/0.3$ and $0.5/0.5$.

Table 1 gives more information for each implemented model.

4.3. Results and discussion

The results of experiments are shown in figure 2

It can be observed from the comparison of the baseline and traditional two stream model that with multiple stream modeling, the model can give better performance on Chinese character rate (CER). Moreover, with the stream dependent parameter tying and inter-stream asynchronism being applied, the CER can be further improved. The results prove the effectiveness of the proposed modeling methods.

However, although the character substitution error rate keeps decreasing when the width of between stream asynchronism increases, a very high character insertion error rate was observed when the width of asynchronism increased to more than 2 states, which lead to a increasing to overall CER. The reason of such high insertion is still unclear and needs further exploration.

Table 1. Model specification

Model Name	Number of streams	Basic Unit for each stream	Feature (length) for each stream	Number of CI models	Number of CD models	Number of tied states
Baseline	1	Initial/tonal final	MFCC+Pitch (42)	209	344,241	8,000
Two stream 2	2	Initial/tonal final	MFCC (39) Pitch (3)	209	344,241	8,000
Proposed model	2	Tone Initial/final	MFCC(39) Pitch (3)	66 16	26,598 375	8,000 1,000

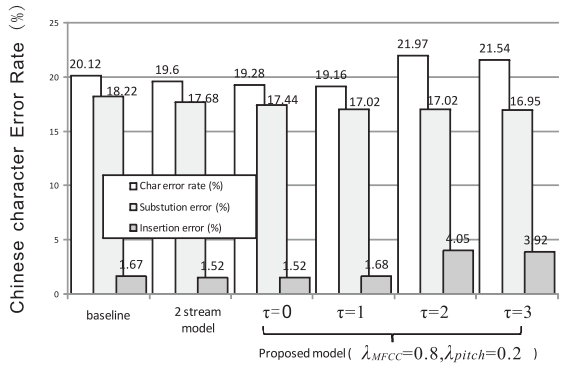

Fig. 2. Results of LVCSR experiment for different models

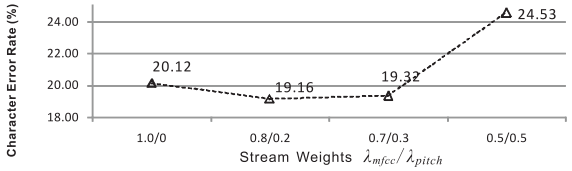

Fig. 3. Character error rate as a function of stream weight (asynchronism width been set to 1 state)

The stream weights λ_s also have influence on the performance (see fig 3). In our experiments, the best performance is achieved when the stream weights are set to $\lambda_{MFCC} = 0.8, \lambda_{pitch} = 0.2$. This may indicates that while tone information can improvement the result, the recognition of initials and finals is still the major contributor to the overall performance.

5. CONCLUSION

In this paper, we present our work on PHMM based asynchronous multiple stream acoustic modeling framework for large vocabulary continuous Chinese speech recognition. The experiments shown that the proposed model improves the CER of Chinese LVCSR result by a relative 4.7% compared with baseline, which proves the proposed method to be a promising one. However, there are several ways to further extend the method, the model structure refining, inter-stream dependency modeling, decoding algorithm optimization and corresponding discriminative training methods are considered and under intensive research.

6. ACKNOWLEDGEMENTS

The work was supported in part by the National Natural Science Foundation of China-NSFC (60605016), the National Key Basic Research Program of China-NKBRPC (2004CB318005, 2004CB318105), and the National High Technology Research and Development Program of China-NHTRDPC (2006AA010103).

7. REFERENCES

[1] J. LI, F. Zheng, J. Y. Zhang, and W. H. Wu, "Context dependent initial/final acoustic modeling for chinese continuous speech recognition," *J Tsinghua Univ (Sci & Tech)*, vol. 24, no. 1, pp. 61–64, Jan 2004.

[2] M.-Y.Hwang, "Acoustic modeling for mandarin large vocabulary continuous speech recognition," in *Advances in Chinese Spoken Language Processing*, Chin-Hui Lee, Haizhou Li, and Ren-Hua Wang, Eds., pp. 153–178. World scientific Publishing Co. Pte.Ltd, 2007.

[3] S. Ying, D. Willett, R. Brueckner, R. Gruhn, and D. Bhler, "Experiments on chinese speech recognition with tonal models and pitch estimation using the mandarin speecon data," in *INTERSPEECH*, 2006, pp. 1245–1248.

[4] Xin Lei, Gang Ji, Tim Ng, Jeff Bilmes, and Mari Ostendorf, "Dbn-based multi-stream models for mandarin toneme recognition," in *ICASSP*, 2005, pp. 349–352.

[5] A. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, "Dynamic bayesian networks for audio-visual speech recognition," *EURASIP, Journal of Applied Signal Processing*, , no. 11, pp. 1–15, 2002.

[6] M.J. Russell, M.J. Tomlinson, and N.M. Brooke, "Integrating audio and visual information to provide highly robust speech recognition," in *ICASSP*, 1996, pp. 821–824.