A MULTI-SPACE DISTRIBUTION (MSD) APPROACH TO SPEECH RECOGNITION OF TONAL LANGUAGES

Huanliang Wang[2] *Yao Qian*[1] *Frank K. Soong*[1] *Jianlai Zhou*[1] *Jiqing Han*[2]

[1]Microsoft Research Asia, Beijing, China

[2]Department of Computer Science and Technology, Harbin Institute of Technology

{yaoqian, frankkps, jlzhou}@microsoft.com, f-hlwang@msrchina.research.microsoft.com, jqhan@hit.edu.cn

Abstract

Tone plays an important role in recognizing spoken tonal languages like Chinese. However, the F0 contour discontinuity between voiced and unvoiced segments has traditionally been a bottleneck in modeling tone contour for automatic speech recognition and synthesis and various heuristic approaches were proposed to get around the problem. The Multi-Space Distribution (MSD) was proposed by Tokuda *et.al.* and applied to HMM-based speech synthesis, which models the two probability spaces, discrete for unvoiced region and continuous for voiced F0 contour, in a linearly weighted mixture. We extend the MSD to tone modeling for speech recognition applications. Specifically, modeling tones in speaker-independent, spoken Chinese is formulated and tested in a Mandarin speech database. The tone features and spectral features are further separated into two streams and stream-dependent models are built to cluster the two features into separated decision trees. The recognition results show that the ultimate performance of tonal syllable error rate can be improved from toneless baseline system to the MSD based stream-dependent system, 50.5% to 36.1% and 46.3% to 35.1%, for the two systems resulted from using two different phone sets. The absolute tonal syllable error rate improvement of the new approach is 5.5% and 6.1%, comparing with the conventional tone modeling.

Index Terms: speech recognition, MSD, tone modeling, LVCSR, Mandarin speech recognition

1. Introduction

Tones are essential for lexical access in tonal languages like Chinese. In Chinese, each character, the basic written unit, is pronounced as a tonal monosyllable. Tonal syllable recognition is critical to name entity identification and other scenarios that strong contextual information is not available. It also provides a direct examination for the precision of acoustic model at phonetic level by isolating the impact of language model in LVCSR. Moreover, it has many other possible applications, e.g. PPT (Mandarin proficiency test).

Tone is carried by perceived pitch in the voiced part of a syllable. Unlike the spectral features, no F0 is observed in unvoiced region. The discontinuity between voiced and unvoiced segments has traditionally made tone modeling difficult. Many ad hoc approaches have been proposed to interpolate F0 in unvoiced segments to bypass the discontinuity problem [1-4]. The interpolated F0s are generated from a quadratic spline function [1], an exponential decay function towards the running F0 average [2], or a probability density function (pdf) with a very large variance [3-4]. Despite their heuristic nature, these approaches are reasonably effective in incorporating F0 as extra components in the short-time acoustic features. As a result, the concatenated spectral and pitch features can be used into one-pass decoding. However, the artificial F0 has not any contributions to identify tones, or even incurs bias in the models. Furthermore, the spectral features essentially represent the vocal tract information, while the pitch features characterize the vibration of vocal cord. They are, to the first order, independent of each other. By using two data streams we can model spectral and pitch features independently [5-6].

Other approaches model tone and spectral information separately [7-8]. The tones are usually derived from force-aligned syllable boundaries in a post processing stage after the 1st-pass recognition. A longer time window can then be used that explicitly to take neighboring tone information into account [9]. As to integrate the tone model into the search process, rescoring lattice or N-best lists output from the recognition is usually adopted.

In this paper, we apply a multi-space distribution (MSD) based tone modeling to speech recognition of tonal languages. The MSD models the discontinuous pitch contours in a statistical compact and rigorous manner. The MSD was originally proposed by Tokuda *et.al.* and successfully applied to HMM-based speech synthesis [10]. We extend the model to speaker-independent Mandarin ("Putonghua") tone recognition. The resultant model is integrated naturally in the one-pass viterbi decoding of continuous speech recognition. The tone features and spectral features are separated into two streams and stream-dependent models are constructed to cluster the corresponding two features into separated decision trees.

2. MSD for Tone Modeling

Multi-Space Probability Distribution (MSD) was proposed by Tokuda *et.al.*[10]. It assumes that the observation space Ω of an event is made up of G sub-spaces. Each sub-space Ω_g has its prior probability $p(\Omega_g)$ and $\sum_{g=1}^{G} p(\Omega_g) = 1$. An observed vector, o, in each sub-space is randomly distributed according to an underlying pdf, $p_g(o)$. The dimensionality of the observation vector can be variable, i.e. different from one sub-space to the other. The observation probability of o is defined by

$$b(o) = \sum_{g \in S(o)} p(\Omega_g)\, p_g(o) \qquad (1)$$

where $S(o)$ is the index set of the sub-spaces that o belongs to. It is determined by feature extractor at each observation. A mixture of K Gaussians can be seen as a special case, i.e. K-

[2]Intern in Speech Group, Microsoft Research Asia

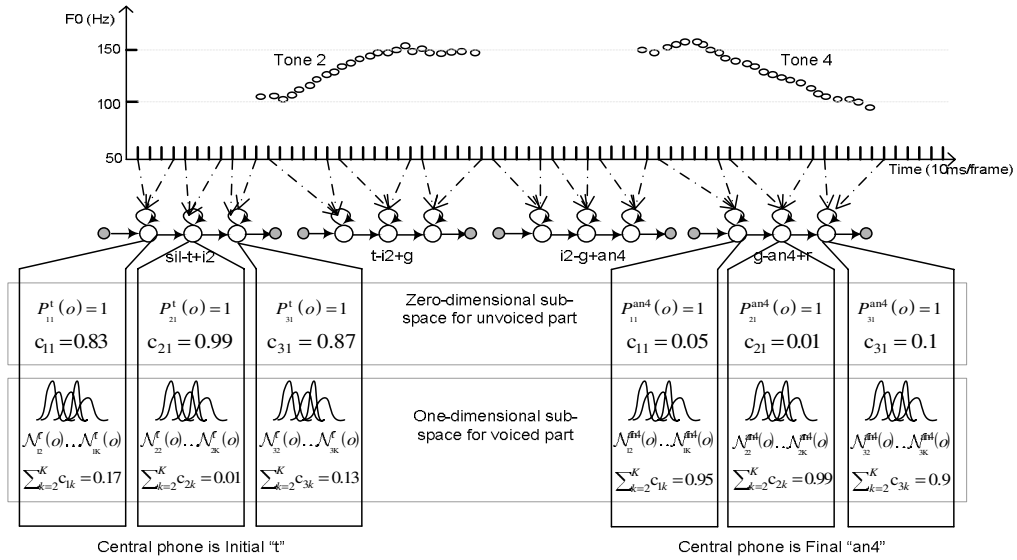

Fig. 1 F0 contour of tonal syllable "ti2 gan4" and a schematic representation of using MSD for tone modeling

subspace of MSD with the same dimensionality and Gaussian distribution in each sub-space. The mixture weight associated with *k*th Gaussian component c_k can be regarded as the prior probability of *k*th sub-space $c_k = p(\Omega_k)$.

F0 is a common feature of tone pattern used in tonal speech recognition. But F0, a continuous variable, only exists in the voiced region of speech. In the unvoiced region, only a discrete variable, or just the unvoiced symbol exists. Fig.1 shows two tonal syllables "ti2 gan4" (the numerical labels denote their tone type: tone 2 and tone 4.) in their triphone representation form and their F0 contours only span across voiced segments: t-i2+g and g-an4+r. The discontinuity of F0 between voiced and unvoiced segments used to make the conventional modeling difficult. MSD provides an almost perfect solution to model F0 without any heuristic assumptions. In the voiced region, F0 can be regarded as one-dimensional observation generated from several one-dimensional sub-spaces, while in the unvoiced region F0 can be treated as a symbol whose dimensionality is zero. In the implementation, we still can use mixture Gaussian output distribution, which is commonly used in current LVCSR and estimated by the Baum-Welch algorithm. It assumes that the output pdf of the zero-dimensional, unvoiced sub-space is a Kronecker delta function and the one-dimensional sub-space of the voiced sub-space has Gaussian distribution.

Fig.1 also gives a schematic representation of using MSD for tone modeling. For the unvoiced Initial 't', the weigh of mixture component in each state which represents unvoiced sub-space is close to one, while the weight summation of other mixture components describing the voiced sub-spaces approximates to zero, and vice versa for voiced Final 'an4'. Tone modeling in this way does not need any preprocessing for F0 feature. It directly models the original F0 feature and avoids any errors potentially incurred by F0 interpolation in unvoiced region.

3. Stream-dependent State Tying

In LVCSR, context-dependent phone models, e.g. tri-phone models, are commonly used to capture the acoustic co-articulation between neighboring phones. To deal with the data sparseness problem of context-dependent phone in the estimation process, model parameters are usually tied together, e.g. state tying based on decision-tree clustering method is widely used in current LVCSR.

Spectral features like MFCC represent essentially the vocal tract information. Tone features reflect the vibration frequency of vocal cord. They can be modeled through two independent data streams, thus it also can avoid the output likelihood being dominated by spectral feature since the dimensionality of spectral feature is much larger than that of tone feature. Moreover, the co-articulation effects of spectral feature and tone feature, or their context dependencies, are different. Accordingly, it is more reasonable to make state tying in two streams independently. We design two question sets corresponding to tonal and phonetic context dependence, respectively. Then decision-tree based clustering method is used to automatically find appropriate cluster for state tying.

An example of stream-dependent state tying based on decision-tree clustering is shown in Fig 2, which illustrates state tying process performed on state 2 of all tri-phones with central phone "y". Two decision trees are grown for this state by using their own question sets. From the top several questions which are used to split the data samples (states), we find that pitch feature stream mainly depends on the questions about tonal context, while the questions for spectral feature stream are about segmental context.

4. Tonal Syllable Recognition

The recognition process of tonal syllables can be rewritten as,

$$\hat{M} = \arg\max \prod_t P(q_t \mid q_{t-1}) \cdot \left[\sum_k^M c_{kq_t}^s \mathcal{N}(o_t^s; \mu_{kq_t}^s, \Sigma_{kq_t}^s) \right]^{\alpha} \cdot \left[\sum_k c_{kq_t}^p \mathcal{N}(o_t^p; \mu_{kq_t}^p, \Sigma_{kq_t}^p) \right]^{\beta} \quad (2)$$

where M represents tonal syllable sequence, q_t is the state at time t, and o_t is divided into two streams: o_t^s for the spectral feature and o_t^p for the pitch feature. α and β are the weights for streams separately and can be adjusted according to the training or development data to optimize, say, recognition performance. We set them equal to 1 in this study. $\sum_k c_{kq_t}^s \mathcal{N}(o_t^s; \mu_{kq_t}^s, \Sigma_{kq_t}^s)$ is a mixture of Gaussians trained by the spectral features, where $c_{kq_t}^s$ is the kth mixture weight; while $\sum_k c_{kq_t}^p \mathcal{N}(o_t^p; \mu_{kq_t}^p, \Sigma_{kq_t}^p)$ is a MSD trained by the pitch features, where $c_{kq_t}^p$ is the mixture weight. At the state q_t, spectral feature and pitch feature access their own decision trees to obtain state parameters, but they share the same state transition probability.

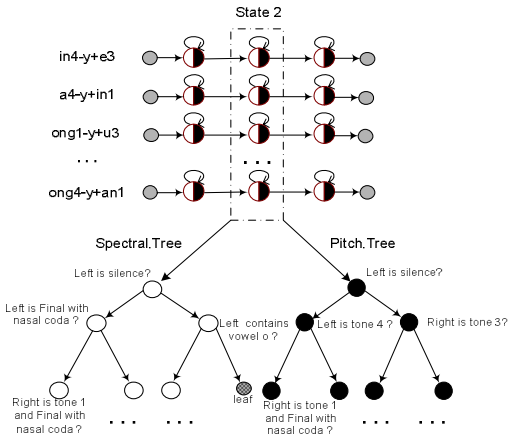

Fig. 2 An example of stream-dependent state tying based on decision-tree clustering

5. Experimental Results and Analysis

5.1 Experimental Setup

The recognition experiments are performed on a speaker-independent, gender-dependent database of read speech. Training set contains about 80 hours' data (about 50k utterances) from 250 male speakers. Testing set consists of 25 male speakers and 500 utterances. We tried two phone sets named Ph97 and Ph187, which are commonly used in Mandarin speech recognition and described as follows:

- Ph187: each tonal syllable is decomposed into a syllable Initial and a tonal Final.
- Ph97: each tonal syllable is divided into a consonant followed by two consecutive tonal sonorant segments [11].

The acoustic features are in two streams: spectral feature stream is 39-dimensional MFCC, consisting of 12-dimensional cepstral coefficients, logarithmic energy and their first and second order derivatives; and pitch feature stream is a 5-dimensional vector, consisting of logarithmic F0, its first and second order derivatives, and pitch duration and long-span pitch [12]. Each phone model is a cross word tri-phone HMM with three emitting states. HMM parameter size used for the configuration of the experiments is nearly 5000*16 Gaussians. Free tonal syllable loop (without language model) is employed in the decoding.

5.2 Experimental Results

Ten recognition experiments based on two phone sets are carried out. All experimental results are shown in Fig 3, where

- 39D-1S: 39 MFCC in one stream
- 44D-1S: 39 MFCC and 5 pitch in one stream; F0 interpolation for unvoiced segments
- 44D-2S: 39 MFCC and 5 pitch in two streams; F0 interpolation for unvoiced segments
- MSD-2S: 39 MFCC and 5 pitch in two streams; MSD used for tone modeling instead of F0 interpolation
- MSD-SD: 39 MFCC and 5 pitch in two streams; MSD used for tone modeling; and stream-dependent state tying employed.

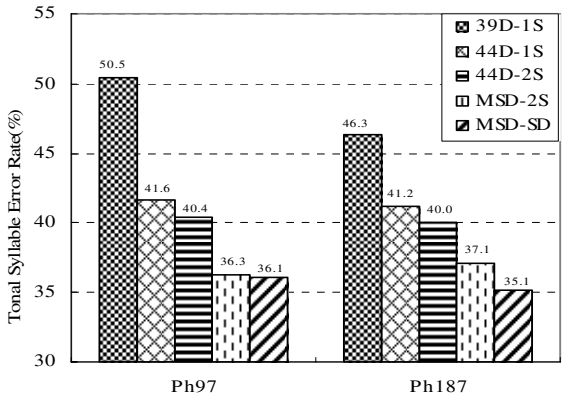

Fig. 3 Recognition performance in tonal syllable error rate (TSER)

39D-1S and 44D-1S are our two baseline systems, which give the comparison between with and without pitch feature for tonal syllable recognition [12]. The performance of tonal syllable error rate (TSER) can be improved from without pitch feature baseline (39D-1S) to the stream-dependent system (MSD-SD), 50.5% to 36.1% and 46.3% to 35.1%, for the two systems using different phone set, Ph97 and Ph187. Compared with the conventional baseline with pitch feature (44D-1S), MSD-SD can reduce the absolute TSER by 5.5% and 6.1% in Ph97 and Ph187. A breakdown of the results in Ph187 shows that using two streams (44D-2S), MSD (MSD-2S) and stream-dependent state tying (MSD-SD) can result in absolute TSER reduction of 1.2%, 2.9% and 2.0%, respectively. The similar results can be observed in Ph97 except that of MSD-SD. We think Ph97 uses 2-scale (H and L) to identify five tone types so that the questions used in state-tying for pitch stream have lower tone discrimination than that of Ph187.

The above experiments all use Maximum Likelihood (ML) criterion for context clustering. The total model size is manually controlled by the tradeoff between recognition speed and accuracy. Minimum description length (MDL) criterion can automatically control model complexity during state tying procedure [13]. In order to investigate the potential impact of MSD-SD on the performance of tonal syllable recognition, we employ MDL as the stop criterion in growing decision tree. The TSERs are reduced to 33.8% and 32.9%, i.e. absolute TSER reduction of 2.3% and 2.2%, in Ph187 and Ph97 with

triple increasing the parameter size.

5.3 Results Analysis

The error rates of base-syllable (toneless) and tone in Ph187 are further analyzed and shown in Fig 4, where we find that using two streams and stream-dependent state tying not only can improve the performance of tone, error rate reduction from 29.9% to 24.4%, but also that of base syllable, from 23.6% to 20.9%. In addition, when the tone and pitch features are augmented into one stream (44D-1S), despite its tonal syllable error rate is much lower than that of the baseline (39D-1S), the error rate of base-syllable is increased by 0.8%, compared with that of not using pitch features (39D-1S).

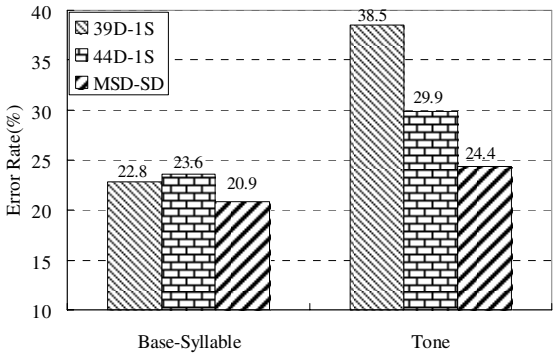

Fig. 4 Recognition performance in the error rate of base-syllable and tone

We also analyze the mixture weight values of unvoiced sub-space in the states of unvoiced and voiced phones. Their mean values are given in Fig. 5, in which we find the values of state 1 and state 3 in unvoiced phone model are lower than that of state 2, and opposite phenomena are observed in voiced phone model. We think that state 1 and state 3 are in a transition between unvoiced and voiced segments so they are less distinct than the central states in term of their voiced/unvoiced characteristic.

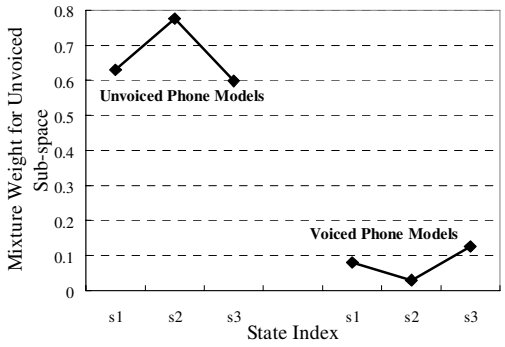

Fig. 5 Mean of mixture weight for unvoiced sub-space in the states of unvoiced and voiced phone model

6. Conclusions

We propose to use MSD, two streams and stream-dependent state tying for tone modeling in tonal syllable recognition. These approaches have a custom-made design for the characteristics of tone features: 1) modeling the original F0 features without any artificial interpolation for discontinuous regions; 2) separating state tying based on decision-tree clustering for tone and spectral features. It achieves a significant improvement of tonal syllable recognition performance. In the future, spectral and pitch stream will be further weighted according to their contributions to speech recognition applications.

7. Acknowledgements

The authors are grateful to Prof. Keiichi Tokuda and Dr. Heiga Zen, Department of Computer Science, Nagoya Institute of Technology, Japan for providing us MSD training tool HTS on the website: http://hts.ics.nitech.ac.jp/ .

8. References

[1] Hirst, D. and Espesser, R., "Automatic Modeling of Fundamental Frequency Using a Quadratic Spline Function", Travaux de l'Institut de Phonétique d'Aix 15, 71-85, 1993.

[2] Chen, C. J., Gopinath, R. A., Monkowski, M. D., Picheny, M. A., and Shen, K., "New Methods in Continuous Mandarin Speech Recognition", In Proc. Eurospeech 1997, 1543-1546, 1997.

[3] Chang, E., Zhou, J. L., Di, S., Huang, C., and Lee, K-F., "Large Vocabulary Mandarin Speech Recognition with Different Approach in Modeling Tones", In Proc. ICSLP 2000, 983-986, 2000.

[4] Freij, G. J. and Fallside, F., "Lexical Stress Recognition Using Hidden Markov Models", in Proc. ICASSP 1988, 135-138, 1988.

[5] Ho, T. H., Liu, C. J., Sun, H., Tsai, M. Y., and Lee, L. S., "Phonetic State Tied-Mixture Tone Modeling For Large Vocabulary Continuous Mandarin Speech Recognition", In Proc. EuroSpeech 1999, 883-886, 1999.

[6] Seide, F. and Wang, N. J. C., "Two-stream Modeling of Mandarin Tones", In Proc. ICSLP 2000, 495-498, 2000.

[7] Qian, Y., Soong, F. K., and Lee, T., "Tone-enhanced Generalized Character Posterior Probability (GCPP) for Cantonese LVCSR", In Proc. ICASSP 2006.

[8] Lin, C. H., Wu, C. H., Ting, P. Y., and Wang, H. M., "Frameworks for Recognition of Mandarin Syllables with Tones Using Sub-syllabic Units", Journal of Speech Communication, 18(2):175-190. 1996.

[9] Qian, Y., "Use of Tone Information in Cantonese LVCSR Based on Generalized Character Posterior Probability Decoding", PhD. Thesis, CUHK, 2005.

[10] Tokuda, K., Masuko, T., Miyazaki, N., and Kobayashi, T., "Multi-space Probability Distribution HMM", IEICE Trans. Inf. & Syst., E85-D(3): 455-464, 2002.

[11] Huang, C., Shi, Y., Zhou, J. L., Chu, M., Wang, T., and Chang, E., "Segmental Tonal Modeling for Phone Set Design in Mandarin LVCSR", In Proc. ICASSP 2004, 901-904, 2004.

[12] Zhou, J. L., Tian, Y., Shi, Y., Huang, C., and Chang, E., "Tone Articulation Modeling for Mandarin Spontaneous Speech Recognition", In Proc. ICASSP 2004, 997-1000, 2004.

[13] Shinoda, K. and Watanabe, T., "Acoustic Modeling Based on The MDL Principle for Speech Recognition", In Proc. EuroSpeech 1997, 99-102, 1997.