INTERSPEECH 2009 BRIGHTON

A One-Step Tone Recognition Approach Using MSD-HMM for Continuous Speech

Changliang Liu, Fengpei Ge, Fuping Pan, Bin Dong, Yonghong Yan

ThinkIT Speech Lab. Institute of Acoustics, Chinese Academy of Science

{chliu,fge,fpan,bdong,yyan}@hccl.ioa.ac.cn

10.21437/Interspeech.2009-763

Abstract

There are two types of methods for tone recognition of continuous speech: one-step and two-step approaches. Two-step approaches need to identify the syllable boundaries firstly, while one-step approaches do not. Previous studies mostly focus on two-step approaches. In this paper, a one-step approach using Multi-space distribution HMM (MSD-HMM) is investigated. The F0, which only exists in voiced speech, is modeled by MSD-HMM. Then, a tonal syllable network is built based on the reference and Viterbi search is carried out on it to find the best tone sequence. Two modifications to the conventional triphone HMM models are investigated: tone-based context expansion and syllable-based model units. The experimental results proved that tone-based context information is more important for tone recognition and syllable-based HMM models are much better than phone-based ones. The final tone correct rate result is 88.8%, which is much higher than the state-of-the-art two-step approaches.

Index Terms: CALL, Speech recognition, Tone recognition, MSD-HMM.

1. Introduction

Chinese Mandarin is a tonal language and each character carries a tone out of five possible candidates. Tone recognition is very important for Computer Assisted Language Learning (CALL). Tone recognition on isolated speech has achieved good performance [1, 2]. However, it is not good enough in continuous speech. Generally, there are two types of methods in tone recognition of continuous speech: one-step and two-step approaches. One-step approaches, also known as embedded tone modeling, employ both spectral and pitch features for better HMM modeling of tonal phones. Two-step approaches, also known as explicit tone modeling, in which the syllable boundaries within an utterance are identified via forced alignment first, and tone recognition is performed on each segmented syllable.

In previous studies, most researches focused on two-step approaches and several techniques are proposed to improve its performance. Yao Qian proposed a supratone method to model the co-articulation between the two or more consecutive tones [3]. In order to find the best supratone model sequence in a global utterance, dynamic programming (DP) search algorithm was used to search the path with the highest likelihood for the given speech utterance in a supratone lattice. Jiang-Chun Chen proposed an algorithm called TRUES [4]. It extracted unbroken pitch contour from the speech and for each syllable segment, some extension to its preceding and succeeding syllables were used to capture the co-articulation information and impair the affection of the inaccuracy of syllable boundaries. Tri-tone-based HMM models were trained in each segment and DP search was carried out on a tri-tone lattice similar with that in [3].

The performance of two-step approaches depends highly on the accuracy of the syllable boundary detection, but the current forced-alignment is not precise enough, which is the most disadvantage of these approaches. Oppositely, the one-step approaches do not need that procedure. It may be a better scheme. The previous studies of one-step approaches are mostly based on large vocabulary continuous speech recognition (LVCSR) [5, 6] while hardly for tone recognition specially. In this paper, some methods specially for tone recognition based on one-step approaches will be investigated.

Multi-Space Probability Distribution HMM (MSD-HMM) is selected to model the F0 features. It is proposed by Keiichi etc. [7] and can model sequences which consist of both continuous vectors and discrete symbols, such as F0, without any heuristic assumptions. It is firstly used in speech synthesis [7] and begins to be used in tone recognition and evaluation [6, 8] in recent years. However, most those researches are also based on two-step approaches. MSD-HMM is applied in each segmented speech respectively. A one-step approach using MSD-HMM will be investigated in this paper.

In the rest of this paper, the new kind of HMM – MSD-HMM is introduced in section 2. Then, the baseline one-step system using MSD-HMM is described in section 3. In section 4 and 5, the data corpus is introduced and the efficiency of MSD-HMM is demonstrated. Two improvements specially for tone recognition are investigated in section 6. The one-step approach is compared with the two-step approach TRUES in section 7 and some conclusions are given in section 8.

2. Multi-Space Probability Distribution HMM

As we know, F0 only exists in voiced speech, but not in unvoiced speech, which is not continuous all over the utterance. The conventional HMM can not model the sequences which consist of both continuous and discrete values. The common solution is smoothing the F0 sequence based on some heuristic assumptions. However, it is not a theoretically reasonable scheme. MSD-HMM provides a perfect framework to cope with this problem.

Multi-Space Probability Distribution (MSD) assumes that the whole sample space Ω is made up of G sub-spaces:

$$\Omega = \bigcup_{g=1}^{G} \Omega_g, \tag{1}$$

where Ω_g is a sub-space with an index g and a prior probability $P(\Omega_g)$, where $\sum_{g=1}^{G}(P(\Omega_g)) = 1$. Each sub-space has its own

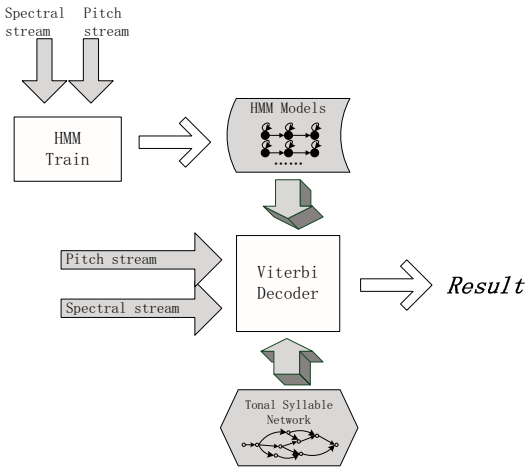

Figure 1: *The flow chart of one-step tone recognition*

probability distribution function (pdf) $\mathcal{N}_g(x)$, $x \in \Omega_g$, where $\int \mathcal{N}_g(x)dx = 1$. The dimensionality of each sub-space can be variable, i.e., different from one space from another. It also supports 0-dimensional sub-space. If the dimensionality of Ω_g is 0, we consider that it contains only one sample point and $\mathcal{N}_g(x) \equiv 1$.

Each observation event E is represented by a random vector o which consists of a set of space indices X and a continuous random variable $x \in \Omega_g$, that is,

$$o = (X, x), \tag{2}$$

where all spaces specified by X are the same dimensional. X is also a random variable, which will be determined by feature extractor. The observation probability of o is defined by

$$b(o) = \sum_{g \in X} P(\Omega_g)\mathcal{N}_g(x). \tag{3}$$

A mixture of K Gaussians can be seen as a special case of MSD, in which, the dimensionality of each sub-space is the same, $X \equiv \{1, 2, ..., G\}$ for each event E, and the pdf of each sub-space is a Gaussian distribution. The mixture weight associated with the kth Gaussian component k_c can be regarded as the prior probability of the kth sub-space $k_c = P(\Omega_k)$.

By using MSD, a new kind of HMM called MSD-HMM is defined. The output probability of each state is given by the multi-space probability distribution function as described above. When applying MSD-HMM to model F0, the F0 in voiced speech is considered from regular continuous sub-spaces while that in unvoiced speech only a symbol from 0-dimensional sub-spaces.

3. One-Step Tone Recognition Framework Using MSD-HMM

The one-step tone recognition is similar with the grammar-based speech recognition. MSD-HMM models are trained on two feature streams: spectral stream and pitch stream. In the recognizing procedure, a tonal syllable network is built based on the reference firstly and Viterbi decoding is carried out on it to generate a tonal syllable sequence. Then, the tone sequence is extracted from the tonal syllable sequence. The flow chart is shown in figure 1.

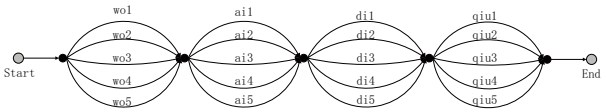

Figure 2: *A tonal syllable network for tone recognition of continuous speech*

An example of the tonal syllable network is shown in figure 2. In tone recognition, the base syllable sequence is always assumed to be known. Thus, the network only need to list all tones for each base syllable. In this example, the sentence is 我爱地球 ('I love the earth' in English). Its base syllable sequence is 'wo ai di qiu'. Assuming the tone is unknown, each base syllable has 5 possible tone candidates, deriving 5 tonal syllables. The network will be expanded according to the HMM model units and the context information in practice.

The features contain two streams: spectral and pitch feature stream. The pitch stream consists of logarithmic F0, its first and second order derivatives, pitch duration and long-span pitch [9], 5 dimensions in total. The spectral stream is 39-dimensional MFCC, which include 12-dimensional cepstrum, 1-dimensional energy and their first and second order derivatives. In each state of HMM, the spectral stream is represented by a conventional continuous probability distribution, while the pitch stream by MSD. As a baseline, the tri-phone model with three emitting states is constructed. We use HTK [10] modified by HTS [11] as our model training and recognizing tools.

4. Data Corpus

The data corpus in our experiments is from "Continuous Mandarin Speech Recognition Evaluation" held by the national "863" program of China. The training corpus contains about 80 hours' data from 84 females and 84 males. The testing corpus contains about 5 hours' data from 7 females and 7 males. There are about 8000 sentences in the testing corpus. Gender-independent models are trained.

5. Effectiveness of MSD-HMM

In this section, the performance of MSD-HMM for tone recognition is compared with the conventional HMM. Three kinds of systems as below are compared:

- MFCC/HMM: train conventional HMM models only using spectral feature.
- MFCC+Pitch/HMM: train conventional HMM models using spectral and pitch features in two streams.
- MFCC+Pitch/MSD: train MSD-HMM using spectral and pitch features in two streams.

The models in each system are cross-word tri-phone HMM with three emitting states. In the system MFCC+Pitch/HMM, the pitch feature is smoothed by interpolation, while in MFCC+Pitch/MSD, the pitch feature is represented by MSD. Stream-dependent state tying [12] is used in both MFCC+Pitch/HMM and MFCC+Pitch/MSD. Different question sets are designed for spectral and pitch streams respectively.

The experimental results of the three systems are illustrated in figure 3. Without pitch feature, the tone correct rate (TCR) is quite low, only 74.7%. After applying pitch feature, the conventional HMM can improve TCR to 79.2%. Furthermore, the best result is obtained by MFCC+Pitch/MSD, 84.1%. It improves

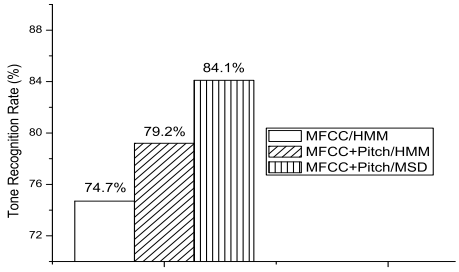

Figure 3: *Tone recognition rates of different models with different features*

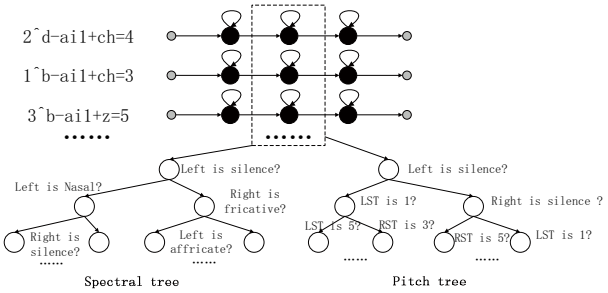

Figure 4: *Decision trees with LST and RST questions*

TCR 17.8% relatively compared with MFCC+Pitch/HMM, and 37.2% relatively compared with MFCC/HMM.

6. Modifications of Models Specially for Tone Recognition

6.1. Tone-based Context Expansion

F0 contours of tones vary extensively in continuous speech. The tonal variations are affected by many factors, such as prosodic structure, preceding and succeeding tones, syntax, pragmatics, emotions, etc. Researches have pointed that, only neighbored phone contexts are not enough to capture tonal variations [13]. In [13], syllable position in prosodic word, left tone and right tone are proved more important than left and right phones. Syllable position is not easy to be applied in the one-step framework. We firstly investigate the effect of left and right tones.

The conventional tri-phone is extended to a "quin-phone" which adds the contexts of tones in left and right syllables (LST and RST). New question sets are designed for LST and RST in tree-based state tying. From the decision tree growing in clustering, the importance of LST and RST for tone recognition can be observed. In the tree of pitch stream, many LST and RST related questions are more near from the root node than other questions, while in the tree of spectral stream, they are much farther away. An example is shown in figure 4. The "1^b-ai1+ch=3" denotes a quin-phone with a current phone "ai1", a preceding phone "b", a succeeding phone "c", a LST "1" and a RST "3".

Before decoding, the network shown in figure 2 is extended according to the new type of context-expansion to cover all possible quin-phone sequences.

The experimental result illustrated in figure 7 shows the improvement. The TCR is improved 6.3% relatively. Its baseline is MFCC+Pitch/MSD referred in section 2. This indicates that the tone-based contexts are more helpful for modeling tones than phone-based ones in the one-step tone recognition.

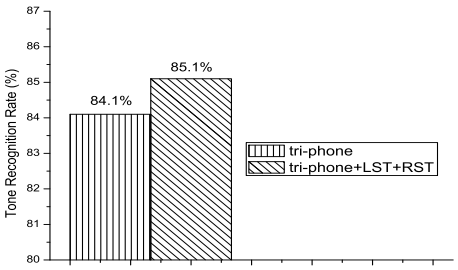

Figure 5: *Tone recognition rates of tri-phone models and quin-phone models with left and right tone contexts*

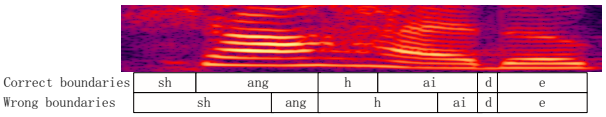

Figure 6: *An example of wrong final boundaries. The utterance is "上海的(shang hai de)"*

6.2. Syllable-based HMM Models

When checking the result of the above system, we find that many tone recognition mistakes occur with wrong final boundaries. A typical phenomena of mistake is as in figure 6. In the syllable "hai (h ai)", most speech is assigned to the initial, while only a little to the final. According to phonetics knowledge, the tone is mainly related with the final. Though the initial "h" occupies most speech, it is nearly unhelpful for tone recognition.

Phone-based HMM models is successful in LVCSR, but it may not be the best strategy in tone recognition. In [14], syllable-based HMM models are used to explore tone variations in Chinese dialects. Each model has four emitting states, expecting the first state to capture the unstable portion in the F0 curve, and the other three to capture the real F0 curve of the syllable. We test this method in the one-step tone recognition, hoping the larger model unit can reduce the impact of the boundaries between initials and finals. LST and RST are considered as contexts in this experiment.

There are about 400 base syllables in Chinese Mandarin. They are tied together by tree-based clustering. We have tested that there is no performance reduction by doing this and it can resolve the problem of data sparsity effectively.

The experimental result is shown in figure 7. Comparing with the system in section 6.1, the TCR is further improved 25% relatively, which proves the effectiveness of syllable-based model units.

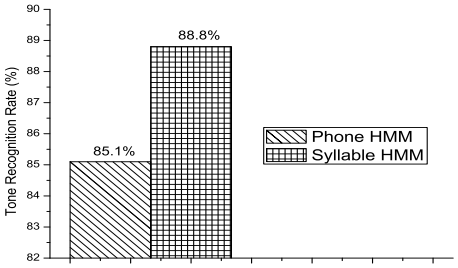

Figure 7: *Tone recognition rates of phone-based and syllable-based models*

Table 1: *Confusion matrix of the result of tone recognition by TRUES.(After lattice refinement. The average tone correct rate is 85.07%)(Copied from [4])*

	Tone 1	Tone 2	Tone 3	Tone 4	Tone 5
Tone 1	92.02%	4.82%	0.38%	2.71%	0.08%
Tone 2	8.00%	85.94%	2.94%	3.06%	0.07%
Tone 3	5.73%	23.56%	60.97%	9.50%	0.24%
Tone 4	7.44%	2.45%	1.32%	88.79%	0.00%
Tone 5	14.81%	7.41%	3.70%	14.81%	59.26%

Table 2: *Confusion matrix of the result of tone recognition by our system (Syllable-based MSD-HMM. The average tone correct rate is 88.8%)*

	Tone 1	Tone 2	Tone 3	Tone 4	Tone 5
Tone 1	90.63%	2.49%	1.85%	4.01%	1.00%
Tone 2	2.19%	88.30%	5.93%	2.53%	1.05%
Tone 3	0.76%	4.12%	91.03%	2.97%	1.12%
Tone 4	2.59%	2.02%	3.68%	89.40%	2.31%
Tone 5	1.18%	2.29%	1.92%	19.21%	75.39%

7. Comparison with Two-step Approaches

The final result of the one-step tone recognition approach using MSD-HMM is 88.8%. Among the state-of-the-art two-step systems, the best result is 85.07% obtained by TRUES in [4], which is 3.63% lower absolutely. Moreover, no lattice refinement (LR) is carried out in our system. In [4], LR is used to removes the illegal combinations of syllables and tones in the tone lattice based on phonological rules and language models. However, in CALL systems, the speech of language learners may not obey the rules extracted from the native languages. It is not appropriate to apply LR to tone recognition in CALL. If no LR, the result of TRUES is only 74.4%. Though the testing corpus of the two systems is different, the comparison is instructive. TRUES's corpus is from Tang Dynasty and read by students of TsingHua University, Taiwan. Our corpus is reading sentences from 863Test and each sentence has about 19 syllables on average. Our testing corpus is not easier than that of TRUES yet.

Tables 1 and 2 show the confusion matrixes of the tone recognition results by our system and TRUES [4]. Our result is better than TRUES in almost any tones, especially in tone 3 and tone 5. The performance of tone 3 in two-step method is always poor [1, 4] because its pitch pattern is the most complex among all tones in Chinese Mandarin. It is very confusable with tone 2 and tone 4 in continuous speech, especially when the syllable boundaries can not be detected precisely. The one-step approach using MSD-HMM do not need the pre-segmentation by the forced alignment and can search the best tone sequence in a larger search space. Therefore, it obtains better performance, even for the most difficult tone 3 and tone 5. The recognition rate of tone 5 is 75.39%, which is much higher than that of TRUES, 59.26%.

8. Conclusions

In this paper, a one-step tone recognition approach using MSD-HMM for continuous speech is investigated. MSD-HMM is a good framework for modeling the F0 sequence which consists of both continuous and discrete values in voiced and unvoiced speech respectively. It achieves 17.8% relative improvement compared with the conventional HMM. Two modifications to the conventional tri-phone HMM models are proposed to improve the tone recognition performance in the one-step framework — tone-based context expansion and syllable-based models. The investigation indicates that it is essential to re-consider the HMM model methods based on the properties of tones. Tone-based contexts, such as tones of left and right syllables, are more important than phone-based ones for tone recognition. Syllable-based model method achieve better performance than phone-based one because it model the syllable as a whole and the boundary of the initial and final can be coped with by the HMM itself automatically. The final result of the one-step system is 88.8%, which is much higher than the best two-step system [4]. The one-step approach is a very promising approach for tone recognition of continuous speech.

9. Acknowledgement

This work is partially supported by The National High Technology Research and Development Program of China (863 program, 2006AA010102), National Science & Technology Pillar Program (2008BAI50B00), MOST (973 program, 2004CB318106), National Natural Science Foundation of China (10874203, 60875014, 60535030).

10. References

[1] F. Pan, Q. Zhao, and Y. Yan, "Improvements in Tone Pronunciation Scoring for Strongly Accented Mandarin Speech," in *Fifth International Symposium on Chinese Spoken Language Processing*, 2006.

[2] S. Wei, H. Wang, Q. Liu, and R. Wang, "CDF-matching for automatic tone error detection in mandarin call system," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 4, 2007.

[3] Y. Qian, T. Lee, and F. Soong, "Tone recognition in continuous Cantonese speech using supratone models," *The Journal of the Acoustical Society of America*, vol. 121, no. 5 Pt1, p. 2936, 2007.

[4] J.-C. Chen and J.-S. R. Jang, "TRUES: Tone Recognition Using Extended Segments," *ACM Transactions on Asian Language Information Processing*, vol. 7, no. 5, pp. 10:1–10:23, 2008.

[5] F. Seide and N. J. Wang, "Two-stream modeling of mandarin tones," in *Sixth International Conference on Spoken Language Processing*, 2000, pp. 867–870.

[6] H. Wang, Y. Qian, F. Soong, J. Zhou, and J. Han, "A multi-space distribution (MSD) approach to speech recognition of tonal languages," in *Ninth International Conference on Spoken Language Processing*. ISCA, 2006.

[7] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Multi-space probability distribution HMM," *IEICE TRANSACTIONS on Information and Systems*, vol. 85, no. 3, pp. 455–464, 2002.

[8] Y.-B. Zhang, M. Chu, C. Huang, and M.-G. Liang, "Detecting Tone Errors in Continuous Mandarin Speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 5065–5068.

[9] J. Zhou, Y. Tian, Y. Shi, C. Huang, E. Chang, and C. Asia, "Tone articulation modeling for Mandarin spontaneous speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 2004.

[10] HTK, "http://htk.eng.cam.ac.uk/."

[11] HTS, "http://hts.sp.nitech.ac.jp/."

[12] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis," in *Sixth European Conference on Speech Communication and Technology*, 1999, pp. 2347–2350.

[13] Y. Hu, M. Chu, C. Huang, and Y. Zhang, "Exploring Tonal Variations via Context-Dependent Tone Models," in *Proc. Interspeech*, 2007.

[14] W. Guo and M. Chu, "Exploring Tone Variations in Chinese Dialects Using Context Dependent Tone Models," in *Sixth International Symposium on Chinese Spoken Language Processing*, 2008.