DETECTING TONE ERRORS IN CONTINUOUS MANDARIN SPEECH

Yan-Bin Zhang[1*] *, Min Chu*[2]*, Chao Huang*[2]*, Man-Gui Liang*[1]

[1] Information technology institute, Beijing Jiao tong University, Beijing, China
[2]Microsoft research Asia, Beijing, China
y-ybzhang@hotmail.com,{minchu,chaoh}@microsoft.com,mgliang@bjtu.edu.cn

ABSTRACT

This paper proposes a new approach for detecting tone errors in continuous Mandarin speech. In the training phase, tone variations are modeled with context-depended MSD-HMM which considers six contextual factors instead of two in traditional triphone HMM. In the evaluation phase, the goodness of tone pronunciation is measured by Kullback-Leibler Divergence (KLD) between the expected tone model and the most representative tone model. When the KLD between the two models is larger than a threshold, the tone is detected as a pronunciation error. In the ROC curve, we get the equal error rate at 2.6%.

Index Terms—Context Depended Tone Model (CDTM), Kullback-Leibler Divergence (KLD), Tone Error Detection

1. INTRODUCTION

In recent years, much progress has been made in the area of computer-assisted language learning (CALL) system[1][2][3], in which pronunciation evaluation plays an important role. Yet, only a few works have been done in evaluating Mandarin pronunciation and most of them are on segmental goodness[4][5][6]. Since Mandarin is a tonal language, it is very important to pronounce tone precisely in live communication, and therefore, detecting tone errors is crucial for a Mandarin CALL system. In this paper, we proposed to detect tonal errors by measuring the Kullback-Leibler Divergence (KLD) between the expected tone model and the most representative tone model (the tone model that matched real speech the most). And we proposed to model tone variations by Context-Dependent Tone Model (CDTM), which considers six contextual factors instead of two in traditional triphone models.

In a previous work, Zhang et. al. [7] used log-posterior probability as a measure of goodness of tone pronunciation. In a monosyllabic corpus, they got about 90% accuracy allowing 4% false acceptance rates. Wei [8] used a similar approach but with F0 after CDF-matching normalization as the feature to detect tone errors. The Cross-Correlation between human experts and automatic tone error detection system is close to 0.79. Both works modeled tones with triphone Hidden Markov Models (HMMs) and achieved promising results on isolated syllables.

In this paper, we focus on tone error detection in continuous speech and propose to model tone variations with more contextual factors. For a continuous speech segment, a sequence of expected CDTMs is derived from the corresponding script and a sequence of most representative CDTMs is generated by model selection against the speech. We propose to measure the goodness of tone pronunciation by the KLD between the expected model and representative model.

This paper is organized as the follows. In Section 2, CDTM training process and KLD-based tone error detection are introduced. Experiments setups and the results and analysis are described in Section 3. Conclusions are drawn in Section 4.

2. TONE ERROR DETECTION BY KLD BETWEEN CDTMS

The work contains three parts: tone modeling, KLD calculation (between HMMs) and KLD-based tone error detection. Details are introduced in the following sub-session 2.1 to 2.3, respectively.

2.1. Tone modeling

Context-dependent phone models, such as tri-phone HMMs, have been successfully used in speech recognition. However, for Mandarin, only considering neighbored phones seems not enough to capture tonal variations. In[9], supra-tone models that are trained from out-line features of two or three succeeding tones were used to re-score the tonal-syllable lattice and achieved significant improvement in tonal-syllable accuracy. In this paper, we take more factors into consider and model tone variations with CDTMs.

2.1.1. CDTM training

The flowchart of CDTM training is shown in Fig.1. It is similar to the training of tri-phone HMM. In both processes, tonal mono-phone models are trained first. The main differences between them are in the context expending part and model tying part.

Multi-Space Distribution (MSD) HMM, proposed by Tokuda *et. al*[10], can model both discrete and continuous features at the same time. When it is used to model pitch patterns, no voice/unvoice decision is needed. MSD-HMM has shown benefit in Chinese tone recognition [11]. In this work, MSD-HMM is used for all tone models.

[*]Join the work as an intern at Microsoft Research Asia

1) Context expansion

Since our goal is to model tone, the tone to be modeled is referred as current tone, CT in short. Six contextual factors are considered, as listed below. (Each syllable contains two phones, initial and final, in Mandarin. Finals carry tones and initials don't.)

- Left Tone (LT): tone of the syllable before the current syllable
- Right Tone (RT): tone of the syllable after the current syllable
- Current Phone (CP): the initial or final of the current syllable
- Left Phone (LP): the initial of the current syllable or the final of syllable before the current syllable
- Right Phone (RP): the initial of the syllable after the current syllable or the final of the current syllable
- Syllable Position (SP): the position information of the current syllable in prosodic word or phrase.

The definition of Syllable Position is given in [12].

Since the combination of six factors may result a large number of context models for a given tone and many of them don't have enough training data. We expand the context step-by-step. The tonal monophone model can be viewed as a tone model in context CP, denoted as CT-CP model. Then, CT-CP model is expanded by the factor SP to CT-CP-SP model. Next, the LP and RP factors are added. Finally, the LT and RT factors are expanded. In each step, the expanded models are re-trained to get more precise start point for next expansion.

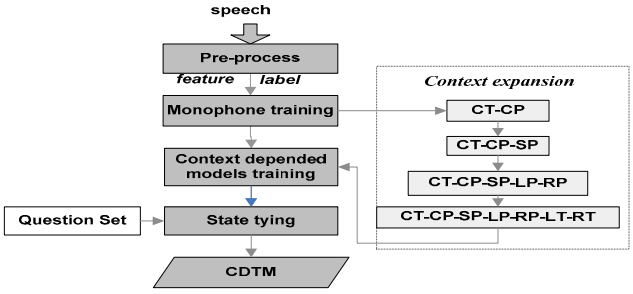

Fig.1 CDTM training

2) State tying

In traditional tri-phone HMM training, models with the same central phone are tied with Classification and Regression Tree (CART). In our tonal modeling, models with the same central tone are tied with the same tree and one tree for each states. Therefore, there will be 5*3 trees for initial models and final models, respectively, in the case of five tones and three states per model.

2.1.2. Search for the most representative model sequence

In order to evaluate the tone pronunciation, we need to find the model sequence that generate the highest likelihood for a given speech segment. To simplify the problem, we assume that segmental pronunciation is correct. Therefore, we know the base syllables from the script for reading. From the syllable sequence, we derive the CP, LP and RP for each model in the best sequence. Then, we enumerate all possible combinations of CT, LT, RT and SP, which gets up to 480 CDTM candidates for each phone. Each CDTM output a likelihood of the observations between the given boundaries of the phone. We merge the likelihood of initial and final of the same syllable as the likelihood of the syllable. Finally, Viterbi search is used to find the best CDTM path that complies with the constraints between neighbored models (transition-constraints) and constraints between position and tone of each model (self -constraints).

Transition-constraints include: CT of the current syllable should be the same as the RT of the syllable before it and LT of the current syllable should equal to CT of preceding syllable etc.

Self-constraints includes: if the SP of the current phone is phrase end, its RT should be silence; if the SP of the current phone is a monosyllable, both RT and LT of it should be silence etc.

2.2. KLD between HMMs

Kullback-Leibler Divergence (KLD)[13] is a quantity measure of the difference between two probability distributions. It has been applied successfully to various applications, such as distortion measure and model clustering. If M represents the true distribution of a random variable x and M^* is an estimated distribution of it, KLD defined as eq.1, measures how much the estimated distribution M^* can be distinguished from M.

$$D(M^*/M)=\int_{R^N} P(x/M)\log\frac{P(x/M)}{P(x/M^*)}dx \tag{1}$$

In order to get a symmetrical KLD, the integrals in two directions are summed as in eq. 2.

$$D_s(M/M^*)=D(M/M^*)+D(M^*/M) \tag{2}$$

When M is normal distribution, there is a close form solution for KLD. However, when the probability function is as complicated as a Gaussian Mixture Model (GMM), no closed form expression exists. Commonly, Monte Carlo method is used to numerically approximate the integral. Do[14] proposed a fast approximation of KLD upper bound between two HMMs. Base on this work, Liu[15] made a further simplification.

Given two HMMs $\mathcal{H}$ and $\tilde{\mathcal{H}}$ with parameter sets of $\boldsymbol{\theta}=\{\boldsymbol{\pi},\mathbf{A},\mathbf{B}\}$ and $\tilde{\boldsymbol{\theta}}=\{\tilde{\boldsymbol{\pi}},\tilde{\boldsymbol{A}},\tilde{\boldsymbol{B}}\}$, respectively. Their KLD upper bound is given by eq. (3) and (4).

$$D_{\mathrm{S}}\left(\mathcal{H}\,\|\tilde{\mathcal{H}}\right)\le\sum_{i=1}^{J-1}\Delta_{i,i} \tag{3}$$

$$\Delta_{i,j}=\frac{D\left(b_i\|\tilde{b}_j\right)}{1-a_{jj}}+\frac{D\left(\tilde{b}_j\|b_i\right)}{1-\tilde{a}_{jj}}+\frac{\left(a_{ii}-\tilde{a}_{jj}\right)\log\left(a_{ii}/\tilde{a}_{jj}\right)}{\left(1-a_{ii}\right)\left(1-\tilde{a}_{jj}\right)} \tag{4}$$

When the distribution functions are all Gaussian Mixtures, the KLD between two GMMs is given by

$$D\left(b\|\tilde{b}\right)\approx\frac{1}{2N}\sum_{m=1}^{M}w_m\sum_{k=1}^{2N}\log\frac{b\left(\boldsymbol{o}_{m,k}\right)}{\tilde{b}\left(\boldsymbol{o}_{m,k}\right)} \tag{5}$$

2.3. KLD-BASED TONE ERROR DETECTION

The flowchart of tone error detection is shown in Fig. 2. For a speech utterance, an expected CDTM sequence is derived from the corresponding script and a sequence of most representative CDTMs are generated by model selection against speech as described in session 2.1.2. Then, KLD between each pair of models are calculated and compared with a threshold, once a KLD is larger than the threshold, the corresponding tone is marked as a reading error.

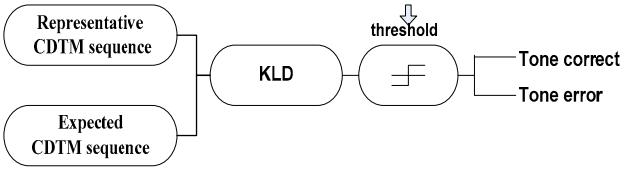

Fig.2 Error Detection Framework

Here is an example:

The script a user wants to read is: "无不为之动容 / wu2 bu4 wei2 zhi1 dong4 rong2". The expected CDTMs derived from it are shown in Table 1. And the most representative CDTMs sequence are shown in Table 2. The KLD between the expected models and the representative models are listed in Fig. 3. It is seen that the KLDs for the initial and final of the third syllable are much larger than others. From the speech, we find that the rising tone is read as the falling tone in this case.

Table 1.The expected CDTM sequence

	wu2		**bu4**		**wei2**		**zhi1**	
Phone	w	u2	b	u4	w	ei2	zh	i1
LT	0	0	4	4	4	4	2	2
RT	4	4	2	2	1	1	4	4
CP	w2	u2	b4	u4	w	ib2	zh1	ib1
LP	sil	w	u2	b	u4	w	ib2	zh
RP	u2	b	u4	w	ib2	zh	ib1	R
SP	1	1	4	4	2	2	4	4

Table 2. The representative CDTM sequence

	wu2		**bu4**		**wei2**		**zhi1**	
phone	w	u2	b	u4	w	ei2	zh	i1
LT	0	0	4	4	4	4	2	4
RT	4	1	2	3	2	4	4	3
CP	w2	u2	b4	u4	w	ei4	zh1	ib1
LP	sil	w	in	ku	i4	b	an4	ch
RP	u4	ii	u2	m	ei2	zhu	ib1	ku
SP	1	1	4	4	2	2	4	4

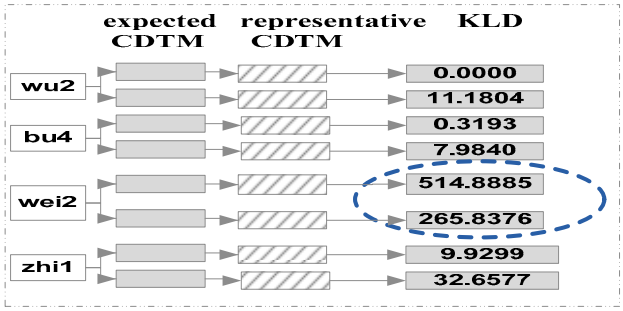

Fig.3 KLD compared result

3. EXPERIMENT

3.1. CDTM evaluation

Before doing tone error detection experiment, we first perform a tone reorganization experiment. The goal is to check whether CDTMs capture better tone information than traditional tri-phone models. Since our focus in on tone, we assume that we know all the base syllables in each utterance in the testing. The only thing we don't know is the tone of each syllable. Therefore, all models with the same CP, LF and RF are enumerated for all base syllables and Viterbi search is used to find out the model sequence that generated the highest likelihood. We used both CDTMs and tradition tri-phone HMMs to do the model selection. The tone error is 6.79% for tri-phone models and 5.32% CDTMs, respectively. 21.6% error reduction is achieved. Therefore, we can conclude that CDTM models better tone variations than tri-phone models.

3.2. Tone error detection

3.2.1. Corpus

In order to exclude the differences between different people, a large Mandarin speech corpus read by single speaker is used in our experiments. It is designed for speech synthesis and covers rich phonetic and prosodic variations. The corpus contains 14476 sentences, totally 181588 syllables, 14376 of which are used for training and 1000 sentences for testing. Parameters are tuned with the first 500 sentences in the testing set as developing set and the remaining 500 sentences as test set.

Since we don't have a testing corpus with tone errors labeled, we simulate erroneous test set by changing the script of the recorded speech, i.e. we modified the tone of syllables in the Pinyin transcription to cause inconsistent between the script and the actual speech. A testing sentence is generated by replacing the tone of one syllable with another tone and only one error is generated each time. The same operation is performed three times for each syllable and is repeated for all syllables (except those neutral syllables) in a sentence. For example, from the sentence "我来自北京/wo3 lai2 zi4 bei3 jing1 (I come from beijing)", 15 testing sentences are generated, the Pinyin transcription for them are like: "***wo1*** lai2 zi4 bei3 jing1", "***wo2*** lai2 zi4 bei3 jing1", "***wo4*** lai2 zi4 bei3 jing1", "wo3 ***lai1*** zi4 bei3 jing1", etc. With this method, we generate a developing set containing 28401 sentences which included 620367 syllables and 28401 tone errors and a testing corpus containing 28467 sentences which included 590688 syllables, 28467 tone errors.

3.2.2. Evaluation

After tone detection, there may be two types of errors: false positive, which means tone errors haven't been detected; false negative, which means the correct pronunciations are judged as the wrong. The performance is represented by its Receiver Operating Characteristic (ROC). The ROC curve is the plot of a False Negative Rate (FNR) with respect to a False Positive Rate (FPR) at each threshold value. It is used to determine the rejection threshold on developing set. Normally, we choose the threshold that generate equal-error-rate (EER), i.e. FNR=FPR.

3.2.3. Methods

The KLD between the expected CDTM sequence derived from the corresponding script and representative CDTM sequence generated by model selection against the speech is calculated and it is compared with the threshold to judge the pronunciation right or wrong. The comparison is carried on both syllable finals and syllable as whole:

- Syllable final only: For each syllable, KLD between its expected final CDTM and its representative CDTM is calculated and is used to make correct/wrong decision directly.
- Syllable: For each syllable, KLDs are calculated for both initial models and final models. The sum of initial KLD and final KLD is used to represent the similarity of the syllable. Correct/wrong decision is made based on the syllable KLD.

3.2.4 Results

FNR and FPR in developing set are calculated. The result is given in Figure 4 below.

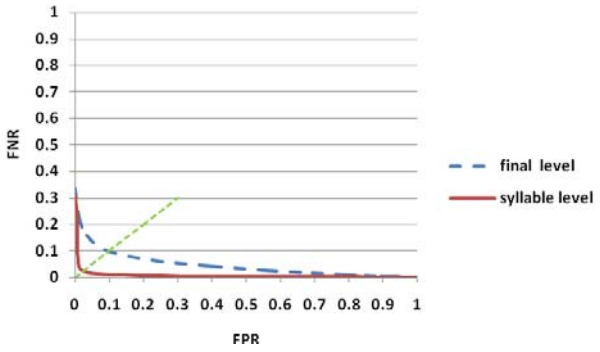

Fig.4 ROC curve

We can see that the syllable level decision is much better than phone decision. We used the threshold that generates EER, which is 2.4%.

When the threshold applied on test set, we got FNR = 2.6%, FPR=2.5% in test set.

4. CONCLUSION

In this paper, we proposed to model Mandarin tones with contextual dependent model and measure the goodness of pronunciation of tones with the KLD between the expected tone model and the tone model represents the actual pronunciation the best.

Our experiments show that the CTDM captures more precise tone variations in continuous speech and results 21.6% relatively error reduction in tone recognition task, comparing with traditional tri-phone model.

The KLD based error detection model can generate 2.6% FNR at the equal error rate conditions.

In current experiment setup, a large speech corpus from a single speaker is used for training and testing. In the next step, we will extend it to multi-speaker testing.

5. REFERENCES

[1] Truong, K., Neri, A., Cucchiarini, C. & Strik, H. "Automatic pronunciation error detection: an acoustic-phonetic approach", *Proceedingsof InSTIL/ICALL2004--NLP and Speech Technologies in Advanced Language Learning Systems--Venice*, 17, 19, 2004.

[2] Liang, M. Hong, Z. Lyu, R. & Chiang, Y. "Data-Driven Approach to Pronunciation Error Detection for Computer Assisted Language Teaching". *Advanced Learning Technologies, 2007. ICALT 2007. Seventh IEEE International Conference on*, 359-361, 2007.

[3] Ito, A. Lim, Y. Suzuki, M. & Makino, S. "Pronunciation Error Detection Method based on Error Rule Clustering using a Decision Tree" *Proc. Eurospeech*, 173-176, 2005.

[4] Chen J.-C., Jang J.-S. R., Li J.-Y. and Wu M.-C.: Automatic Pronunciation Assessment for Mandarin Chinese. in *Proc. ICME*, pp.1979-1982, 2004

[5] Wei S., Liu Q.S., Hu Y., Wang R.H.Automatic Pronunciation Assessment for Mandarin Chinese with Accent, *NCMMSC8*, pp. 22-25, 2005 (In Chinese)

[6] Dong B., Zhao Q.W., Yan Y.H.: Analysis of Methods for Automatic Pronunciation Assessment, *NCMMSC8*, pp.26-30, 2005, (In Chinese)

[7] Zhang, L. Huang, C. Chu, M. Soong, F. Zhang, X. & Chen, Y. "Automatic Detection of Tone Mispronunciation in Mandarin" *Proc.ISCSLP2006, LNAI 4272*, 590-601,2006.

[8] Si Wei, Hai-Kun Wang, Qing-Sheng Liu, Ren-Hua Wang "CDF-matching for automatic tone error detection in mandarin call system".*proc. ICASSP,* 205-208,2007.

[9] Wang, H. L., Qian, Y., Soong, F. K., Zhou, J. L. and Han. J. Q., "Improved Mandarin Speech Recognition by Lattice Rescoring with Enhanced Tone Models", *Proc. of ISCSLP-2006, Singapore* 736-747, 2006.

[10] Tokuda, K., Masuko, T., Miyazaki, T. and Kobayashi, T., "Hidden Markov Models based on Multi-Space probability Distribution for Pitch Pattern Modeling", *Proc. of ICASSP,* 1999.

[11] Wang, H., Qian, Y., Soong, F. K., Zhou J. and Han J., "A Multi-Space Distribution Approach to Speech Recognition of Tonal Languages", *Proc. of Interspeech*, 2006.

[12] Yue-Ning Hu, Min Chu, Chao Huang, et al., "Exploring Tonal Variations via Context-Dependent Tone Models" , *Proc.Interspeech*, 2007.

[13] Cover, T. M. and Thomas, J. A., "Elements of Information Theory", *Wiley Interscience, New York, NY*, 1991.

[14] Do, M. N., "Fast Approximation of Kullback-Leibler Distance for Dependences Trees and Hidden Markov Models", *IEEE signal Proc. letters*, Apr 2003.

[15] Peng Liu, Frank K. Soong, Jian-Lai Zhou, "Effective Estimation of Kullback-Leibler Divergence between Speech Models", *Microsoft Research Asia, Technical Report*, 2005.