

Incorporating Tone-related MLP Posteriors in the Feature Representation for Mandarin ASR

Xin Lei, Mei-Yuh Hwang, Mari Ostendorf

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

{leixin,mhwang,mo}@ee.washington.edu

Abstract

Tone has a crucial role in Mandarin speech in distinguishing ambiguous words. In most state-of-the-art Mandarin automatic speech recognition systems, tonal acoustic units are used and F_0 features are appended to the spectral features (MFCC/PLP). However, a tone depends on the F_0 contour of a time span much longer than a frame. Ideally, systems would compute the frame-level likelihood of a tone using more than the F_0 and derivative values at the current frame. Inspired by the tandem approach, we propose to extract tone-related features for each frame by using longer acoustic context information in a multi-layer perceptron (MLP). The extracted tone-related posteriors are then appended to the spectral feature vector to form a new feature vector for back-end HMM systems. Results show that significant improvement can be achieved by adding these tone-related MLP posterior features in a Mandarin conversational telephone speech recognition task. In one configuration, the character error rate was reduced from 35.7% to 33.2%.

1. Introduction

Many studies on Mandarin speech recognition have been conducted in the recent years [1, 2, 3, 4, 5]. Quite different from English and some other Western languages, Mandarin Chinese is a tone language which benefits from explicitly modeling five tones to distinguish ambiguous words. The five tones are characterized by syllable-level pitch or F_0 contour patterns [6]: high-level (tone 1), high-rising (tone 2), low-dipping (tone 3), high-falling (tone 4) and neutral (tone 5). The neutral tone often occurs in word-end or sentence-end contexts in continuous speech and does not have a stable F_0 contour.

In most state-of-the-art Mandarin speech recognition systems, tonal acoustic units such as tonal syllables [2] and tonemes [1] are used in HMM-based systems. The feature vector is obtained by appending F_0 and its derivatives to the conventional MFCC/PLP spectral features. This approach achieves significant improvement in various Mandarin ASR tasks [2, 3, 4]. However, a tone depends on the F_0 contour at the syllable-level, which is much longer than a frame. The F_0 of a frame and its derivatives contain some information about the tone but not enough to depict the whole F_0 contour of the current syllable. Therefore, in this work we investigate the use of tone-related features that contain more information than

frame-level F_0 values, in combination with HMM tonal acoustic units.

We propose to use a multi-layer perceptron (MLP) to generate tone-related posterior features for our Mandarin conversational telephone speech (CTS) recognition task. In [7], Hermansky *et al.* proposed the tandem approach which uses MLP posterior outputs as the input features for Gaussian mixture models of a conventional speech recognizer. Researchers at ICSI have found significant improvement in English large vocabulary speech recognition by using variations of MLP-based features [8, 9]. In this paper, we explore the use of MLP-based tone posterior and toneme posterior features and their combinations with conventional spectral features.

The rest of the paper is organized as follows: In Section 2, we describe the architecture of tone-related posterior feature extraction and its usage in Mandarin ASR. In Section 3, the experimental results and findings are presented. Finally, we summarize key points and propose future work in Section 4.

2. Tone-related Posterior Features

The overall configuration of our feature extraction stage is illustrated in Figure 1. Three different features, including their first order and second order derivatives, are extracted from the input speech: MFCC, F_0 and PLP. Both MFCC and PLP front ends are used to exploit the cross-system benefits. The F_0 features (smoothed F_0 plus the first two derivatives) are appended to the MFCC features to form a new feature vector for each frame. By concatenating the feature vectors from neighboring frames, we form a high dimension feature vector and feed it into the MLP to classify tone-related targets. Because the MLP output posterior has a very non-Gaussian distribution (between 0 and 1 by the sigmoid operation), we take the log of the posterior to make it more Gaussian-like [8]. After that, principal component analysis (PCA) is performed to decorrelate and reduce the dimensions of the posterior feature vector. The resulting tone-related features are then appended with PLP and optionally F_0 features to form the final feature vector for the back-end HMM-based SRI DECIPHER recognizer.

In the training phase, the discriminative MLP is

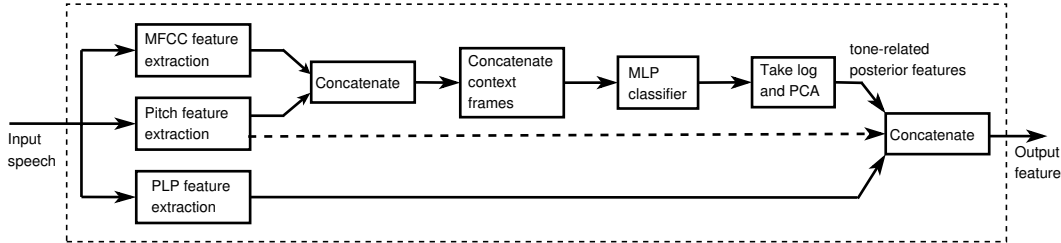


Figure 1: Block diagram of the feature extraction stage.

trained with tone-related targets using the error back propagation algorithm with a minimum cross-entropy criterion [9]. In this paper we consider two different types of targets: tones and tonemes.

2.1. Tone posterior feature

There are five lexical tones in Mandarin speech. For silence or unvoiced regions, there is no tone pattern. Therefore, we train a tone MLP classifier with six targets: five tones and a no-tone target. This MLP classifier is actually an extended tone classifier since it not only classifies five tones but also does voicing detection (for the no-tone decision). The MLP is trained to distinguish the six categories according to the input MFCC and F_0 features of the current and neighboring frames. All the results reported in this paper use single hidden layer MLPs and a 9-frame context window. For each frame, we extract 39+3 dimension MFCC+ F_0 features. Therefore, the input size of the MLP is 378 for MFCC+ F_0 features. In training, the MLP output units have target values of 1 for the tone associated with the current frame and 0 for others. The target labels of the training data are assigned automatically using Viterbi alignment with an existing set of HMMs. After the MLP is trained, it is used to generate tone posteriors. The log tone posteriors are optionally transformed with PCA to 3-dimension.

In the tone posterior system, we want to explore two questions. First, since the tone MLP classifier is trained with spectral and F_0 features from a much longer time span than a single frame, we want to find out whether the tone posterior features perform better than using frame-level F_0 features. Second, given that the dimension of the tone posteriors is small, we want to find out whether further PCA dimension reduction is helpful at all.

2.2. Toneme posterior feature

A toneme is defined as a phoneme consisting of a specific tone in a tone language [1]. For example, a1, a2, a3, a4 and a5 are five different tonemes associated with the same main vowel “a”. Consonants can be regarded as special tonemes without tones. In our Mandarin speech recognition system [4], we have 62 speech tonemes plus one silence phone, one for laughter and one for all other

nonspeech events. The 62 speech phones consists of 27 non-tonal phones and 35 tonal phones.

In the toneme posterior system, we train an MLP to classify the 64 sub-word units (except the one for all other nonspeech events). PCA is performed on the log of the output MLP features and the first 25 principle components are taken, as suggested in [8]. This system is quite similar to the PLP/MLP feature based system in [9], except that we are using F_0 features combined with MFCC features to classify tone-dependent acoustic units.

Finally, we compare the system performance with different combinations of F_0 , PLP, tone posterior and toneme posterior features.

3. Experiments and Results

For all of the experiments reported in this paper, we use the Mandarin CTS data collected by Hong Kong University of Science and Technology (HKUST) in 2004. The training set, *train04*, consists of 251 phone calls, totaling 57.7 hours of speech and silence, made within China and Hong Kong by mostly college students.¹ The testing set, *dev04*, consists of 24 phone calls and has a total length of 2 hours. More specifically, the audio data of *dev04* are manually segmented according to the timemarks in the transcription. The conversation sides are given, and each conversation side has exactly one speaker.

3.1. Feature extraction

The MFCC and PLP features are generated with the front-end of the SRI DECIPHER speech recognizer. Speech input is processed using a 25ms Hamming window, with a frame rate of 10ms. For each frame there are 13 MFCC/PLP coefficients, plus their derivatives. Cepstral mean and variance normalization and vocal tract length normalization are applied per speaker.

The F_0 is extracted with ESPS’s *get_f0* and then passed to a lognormal tied mixture model [10] to alleviate pitch halving and doubling problems. Then a smoothing algorithm similar to [1] is applied and derivatives are

¹In previously published results [4], we also used the CallHome (CH) and CallFriend (CF) corpora as training data. Here, to simplify MLP training, we use only *train04* data. The performance improvement by adding CH&CF data is around 1% absolute.

computed. Finally the F_0 features are also mean and variance normalized per speaker to decrease the speaker dependency and to be normalized to the same range as spectral features.

3.2. MLP training

For tone and toneme MLP classifier training, we first randomize the order of the training utterances lest the MLP training fall into a local optimum. A portion (10%) of the training data is held out as a cross validation set in MLP training. The tone and toneme training targets are generated from forced alignment with the recognizer using an existing set of triphone HMMs. We tune the number of context frames and hidden nodes for the best frame classification accuracy. Frame accuracy is defined as the ratio of the number of correctly classified frames to the total number of frames, where classification is deemed to be correct if the highest output of the MLP corresponds to the correct target. This is a good preliminary indicator of system performance and provides an efficient way to tune the parameters without running the whole system.

It is found that for both tone and toneme classification, a 9-frame window gives satisfactory results, although a longer time window provides a marginal gain in frame accuracy. In the tone MLP classifier, 900 hidden nodes are enough.² In the toneme MLP classifier, 1500 hidden nodes provide good performance. The frame accuracy scores of the tone and toneme MLP classifiers on the cross validation set are listed in Table 1. The toneme frame accuracy is slightly better than the published English phoneme frame accuracy results on a similar CTS task [9], where 46 phoneme targets are used.

Table 1: *Frame accuracy of tone and toneme MLP classifiers on the cross validation set.*

Targets	Cardinality	Frame Acc.
tone	6	80.3%
toneme	64	68.8%

3.3. HMM system training and decoding

After the MLP classifiers are trained, we use them to generate tone-related posterior features for a back-end HMM system as described in Section 2. The posteriors are also mean and variance normalized per speaker. All HMM systems here are maximum likelihood trained using decision-tree state clustering. The pronunciation phone set includes consonants and tonal vowels, with a total of 65 phones. All triphones of the same base phone with different tones are in the same tree. Categorical questions include tone questions, in addition to other phone classes and individual phone questions, state ID, etc. Unless noted, all systems use within-word triphones.

²The number of hidden nodes is large for 6-tone classification, since the input size is also large.

The decoding lexicon consists of 11.5K multi-character words. The language model is a trigram model trained from training data transcriptions and text data collected from the web [4]. To evaluate various feature combinations efficiently, we run only the last stage of a multi-pass decoding system [4], rescoring a lattice of word hypotheses using a 7-class MLLR adaptation. The adaptation is unsupervised based on the recognition hypotheses from an earlier pass. The results are evaluated in terms of character error rate (CER).

3.4. Tone posterior results

The CER results from tone posterior systems are listed in Table 2. As we can see, the system with PLP+(tone posterior) features outperforms the PLP+ F_0 system by 0.3%. The system with both F_0 and tone posterior features has the best performance and outperforms the PLP+ F_0 system by 0.5% absolute. This shows that the tone posterior offers additional information beyond using F_0 features directly. However, as in tandem experiments the best results are obtained using both. We also find that PCA on the small dimension (6) is not necessary, though it slightly reduces the computation and memory requirements.

Table 2: *CER of systems using tone posteriors.*

Feature	Dim.	CER
PLP	39	36.8%
PLP+ F_0	42	35.7%
PLP+(tone posterior)	45	35.4%
PLP+PCA(tone posterior)	42	35.6%
PLP+ F_0 +(tone posterior)	48	35.2%
PLP+ F_0 +PCA(tone posterior)	45	35.2%

3.5. Toneme posterior results

Table 3 shows the CER results from toneme posterior and combined posterior systems. In all experiments reported here, PCA reduces the MLP log posteriors down to 25 dimensions. The PLP+PCA(toneme posterior) feature system has 2.0% absolute improvement in CER over the baseline PLP+ F_0 system. Since the toneme posterior contains discriminative information for both phone units (as in English experiments) and tones, the significant performance improvement is reasonable and consistent with the English results reported in [9]. Similar to the tone posterior experiments, adding F_0 features to the system provides a further 0.5% improvement.

We then try to combine the tone and toneme posterior features. The performance is essentially the same as the PLP+ F_0 +PCA(toneme posterior) system. Finally, we combine all features together (PLP, F_0 and PCA of tone and toneme posterior features) in a single system but

no further improvement is obtained. The last two experiments probably indicate that the information provided by the tone posterior is covered by the combination of F_0 and toneme posterior; or alternatively the F_0 information is covered by the combination of tone posterior and toneme posterior.³ That is, given the toneme posterior features, adding either plain F_0 features or the tone posteriors is useful but not both.

Table 3: CER of systems using toneme posteriors.

Feature	Dim.	CER
PLP+ F_0	42	35.7%
PLP+PCA(toneme posterior)	64	33.7%
PLP+ F_0 +PCA(toneme posterior)	67	33.2%
PLP+PCA(tone, toneme posterior)	64	33.3%
PLP+ F_0 +PCA(tone, toneme posterior)	67	33.3%

3.6. Cross-word triphone systems

We have also trained cross-word triphone systems based on the best feature combination, with results shown in Table 4.

Table 4: CER of cross-word triphone systems.

Feature	Dim.	CER
PLP+ F_0	42	35.0%
PLP+ F_0 +PCA(toneme posterior)	67	33.0%

The performance improvement of 2.0% absolute is consistent but smaller than the 2.5% gain in the within-word triphone system. It may be that, to model cross-word coarticulation effects, we might want to use a longer time window to generate the posteriors.

4. Conclusions and Future Work

In this work, we have tried different approaches to incorporate tone-related MLP posteriors in feature representation for a Mandarin conversational telephone speech recognition task. More specifically, tone posteriors, toneme posteriors and their combinations with F_0 and PLP features are explored. We found that tone posteriors and the plain F_0 features complement each other, though with some overlap. Much more significant improvement is achieved by using toneme posterior features, which is probably in part because of incorporating segmental cues, known to be important from other work [9]. By combining toneme posteriors with either F_0 features or tone posteriors, we have reduced CER by 2-2.5% absolute (or 6-7% relative).

³We increased the output dimension after PCA, but it did not help.

For future work we will investigate context-dependent tone (“tri-tone”) classification and the use of tri-tone posteriors in the feature representation. Tri-tone modeling has the advantage of modeling tone sandhi [6] implicitly. We plan to investigate tri-tones at both the phonetic level and the syllabic level. Since tri-tones model tone coarticulation and the 6^3 MLP output units can provide more discrimination than the simple 6 tone posteriors, we expect bigger gains from the tri-tone posteriors. It will be interesting as well to study how tri-tones interact with toneme posteriors.

5. Acknowledgments

The authors would like to thank Q. Zhu and B. Chen at ICSI for providing technical help with the MLP system. Thanks also to A. Stolcke and V. Gadde at SRI for their help on the SRI DECIPHER recognition system. This work was supported by the Defense Advanced Research Projects Agency, grant MDA97202-C-0038. Distribution is unlimited. The views herein are those of the authors and do not reflect the views of the funding agencies.

6. References

- [1] C.J. Chen et al., “New methods in continuous Mandarin speech recognition,” in *Proc. Eur. Conf. Speech Communication Technology*, 1997, vol. 3, pp. 1543–1546.
- [2] E. Chang et al., “Large vocabulary Mandarin speech recognition with different approaches in modeling tones,” in *Proc. ICSLP*, 2000, vol. 2, pp. 983–986.
- [3] H.C. Huang and F. Seide, “Pitch tracking and tone features for Mandarin speech recognition,” in *Proc. ICASSP*, 2000, vol. 3, pp. 1523–1526.
- [4] M. Hwang et al., “Porting DECIPHER from English to Mandarin,” in *Proc. DARPA 2004 Rich Transcriptions Workshop*, 2004.
- [5] M.J.F. Gales et al., “Development of the CUHTK 2004 Mandarin conversational telephone speech transcription system,” in *Proc. ICASSP*, 2005, vol. 1, pp. 841–844.
- [6] C. Wang, *Prosodic Modeling for Improved Speech Recognition and Understanding*, Ph.D. thesis, Massachusetts Institute of Technology, 2001.
- [7] H. Hermansky, D. Ellis, and S. Sharma, “Tandem connectionist feature extraction for conventional HMM systems,” in *Proc. ICASSP*, 2000, pp. 1635–1638.
- [8] N. Morgan, B. Chen, Q. Zhu, and A. Stolcke, “TRAPping conversational speech: Extending TRAP/Tandem approaches to conversational telephone speech recognition,” in *Proc. ICASSP*, 2004, vol. 1, pp. 537–540.
- [9] B. Chen, Q. Zhu, and N. Morgan, “Tonotopic multi-layered perceptron: a neural network for learning long-term temporal features for speech recognition,” in *Proc. ICASSP*, 2005, vol. 1, pp. 945–948.
- [10] M.K. Sonmez, L. Heck, M. Weintraub, and E. Shriberg, “A lognormal model of pitch for prosody-based speaker recognition,” in *Proc. Eur. Conf. Speech Communication Technology*, 1997, vol. 3, pp. 1391–1394.