MANDARIN TONE CLASSIFICATION WITHOUT PITCH TRACKING

Neville Ryant, Jiahong Yuan, and Mark Liberman

Linguistic Data Consortium, University of Pennsylvania

ABSTRACT

A deep neural network (DNN) based classifier achieved 27.38% frame error rate (FER) and 15.62% segment error rate (SER) in recognizing five tonal categories in Mandarin Chinese broadcast news, based on 40 mel-frequency cepstral coefficients (MFCCs). The same architecture scored substantially lower when trained and tested with F_0 and amplitude parameters alone: 40.05% FER and 22.66% SER. These results are substantially better than the best previously-reported results on broadcast-news tone classification [1] and are also better than a human listener achieved in categorizing test stimuli created by amplitude- and frequency-modulating complex tones to match the extracted F_0 and amplitude parameters.

Index Terms— speech recognition, Mandarin, tone modeling, deep neural networks

1. INTRODUCTION

As is well known, Mandarin Chinese has lexical tone, whereby words can be differentiated from one another by changes in fundamental frequency (F_0) contour and perhaps other prosodic features. The tonal possibilities in standard Mandarin are conventionally numbered from one to four. In terms of the International Phonetic Alphabet's Chao Tone Letter system [2], where tonal patterns are schematized in terms of a sequence of points scaled from 1 (lowest) to 5 (highest), the four basic lexical tones are represented as:

Tone number:	1	2	3	4
Description:	high	low-rising	low-falling	high-falling
Tone letter:	˥	˧˥	˨˩	˥˩
	(55)	(35)	(21)	(51)

Mandarin also has a neutral or zero tone (also sometimes called fifth tone) which is perhaps best considered to be lack of lexically specified tone, with the F_0 contour determined by context [3, 4, 5].

In continuous speech, the F_0 contours for the five tonal categories are subject to many sorts of variation. In "third tone sandhi", a closely-associated sequence of Tone3+Tone3 may become Tone2+Tone3. More generally, there is extensive tonal coarticulation, so that e.g. between the (high F_0) end of Tone1 and the (low F_0) start of the Tone2 pattern, a fall will occur. In addition, overall pitch range will vary substantially across speakers and a given speaker's pitch range will vary within and across phrases due to phrasal downtrends, variable emphasis, topic-shift effects, and so on [6, 7, 8, 9, 10]. These sources of variation mean that the recognition of tonal categories in continuous speech is not a trivial task.

Although many Chinese speech-recognition systems have included tonal features in order to improve performance in the integrated task of recognizing tonally-specified segments [11, 12, 13, 14], there are relatively few documented attempts to evaluate the automated recognition of tonal categories alone in continuous speech [15, 1, 16, 17]. [15] use decision trees and a segmental representation based on the fitting of polynomials to the F_0 contour to achieve 27.8% SER for continuous speech. For broadcast news, [1] achieve 23.8% SER using MLPs and contextual information. Most recently, [17] achieved 21% SER, albeit for command-and-control utterances, with the incorporation of biologically inspired auditory features. All save [17] perform explicit pitch tracking, though even [17] includes parameters that are probably an excellent proxy for F_0 slope. Notably, we improve substantially on these results with our system, despite the lack of explicit F_0 information

Given the well-documented advantages of deep neural network (DNN) algorithms on such tasks [18, 19], we were not surprised to find that a DNN-based system performed well, scoring better than previously reported results on roughly comparable tasks. We were, however, intrigued to see this level of performance in a system trained on a set of acoustic parameters in which F_0 is not explicitly represented and to find that a system trained on explicit F_0 contours had substantially worse performance.

All of these results are subject to confirmation and elaboration, but they suggest some non-obvious ideas about approaches to modeling prosodic features in general.

2. DATA

Testing and training sets were constructed using the 1997 Mandarin Broadcast News Speech corpus [20]. We ex-

This material is based upon work supported by the National Science Foundation under Grant No. IIS-0964556.

tracted all "utterances" (the between-pause units that are time-stamped in the transcripts) from the corpus and manually excluded those containing background noise or music. Utterances from speakers whose names were not tagged in the corpus or from speakers with accented speech were also excluded. In total 7,849 utterances from 20 speakers were selected. From these we randomly selected 50 utterances from each of six speakers to compose a test set, with the remaining 7,549 utterances reserved for training. The 300 test utterances were manually labeled and segmented into initials and finals by a native Mandarin speaker. Tones were marked on the finals, including Tone1 through Tone4, and Tone0 for the neutral tone. The total number of utterances, segments, and hours of speech are detailed in Table 1.

	Hours	Utterances	Segments	TBUs
Train	6.05	7,549	196,330	96,697
Test	0.22	300	7,189	3,464

Table 1: Train/test set composition. TBU = tone-bearing unit, defined as the syllable final.

3. TONE CLASSIFICATION WITH MFCCS

We propose attacking the problem of explicit tone classification as follows:

1) Train a DNN to classify each frame of speech as one of six classes: Tone0, Tone1, Tone2, Tone3, Tone4, No-tone.

2) Compute "tonal features" for each segment, defined as the mean of the outputs of the DNN over all frames contained within that segment. These are similar to the articulatory features of [16].

3) Use these "tonal features", along with segment duration and contextual features, to classify the tone-bearing units (TBUs).

3.1. Feature extraction

Forty mel frequency cepstral coefficients (MFCCs) were extracted every ms using the following analysis parameters: i) 0.97 preemphasis factor; ii) 25 ms Hamming window; iii) 1024-point DFT; iv) 40 filter mel-scale filterbank[1]. Cepstral mean-variance normalization was applied on a per-utterance basis. Neither the fundamental frequency nor the overtone series is transparently present in this representation.

[1]Our MFCCs may be reproduced using *melfcc* from [21] with the following parameter values: wintime=0.025, hoptime=0.001, nbands=40, numcep=40, lifterexp=-22, sumpower=0, minfreq=0, maxfreq=8000, dcttype=3.

3.2. Network training

We trained a DNN [22] to classify frames of the signal as one of six targets: Tone0, Tone1, Tone2, Tone3, Tone4, or No-tone. Input to the DNN consisted of an 840-dimensional feature vector derived by concatenating the MFCCs for the 21 frames with offsets of -100 ms, -90 ms, ..., +90 ms, +100 ms relative to the center frame. Training targets were derived by forced alignment of the HUB-4 training utterances using an HMM-based forced aligner built on the training utterances with the CALLHOME Mandarin Chinese Lexicon [23] and HTK. The aligner employed explicit phone boundary models [24] and achieved 93.1% agreement within 20 ms compared to manual segmentation on the test set. Additionally, we checked 100 training utterances on the tone labels automatically generated by the aligner. Among the 1,252 syllables in the 100 utterances, 15 syllables had a wrong tone, an error rate of 1.2%.

The full network topology consisted of: i) an 840 unit input layer; ii) 4 hidden layers, each consisting of 2000 rectified linear units (ReLUs) [25]; iii) an output layer consisting of 6 softmax units. The network was trained for 140 epochs (each epoch consisting of 250,000 examples) using stochastic gradient descent with a mini-batch size of 128, 20% dropout [26] in the input layer, 40% dropout in the hidden layers, and a cross-entropy objective. Learning rate was kept constant within epochs and followed the schedule $\eta(n) = \eta(0)\frac{500}{n+500}$, where $\eta(0) = 0.5$, while momentum was kept constant at 0.5 throughout training. No L_2 weight decay was used, but the incoming weight vector at each hidden unit was constrained to have a maximum L_2-norm of 3.

3.3. Segment-level classification

We consider four approaches to tone classification. Firstly, we consider a simple baseline which assigns each TBU to the the tone with the highest posterior probability according to the tonal features of that segment. We also consider three supervised methods: logistic regression with L_2 regularization, support vector machines (SVMs) using a radial basis function (RBF) kernel, and neural networks. Input features consist of the tonal features of the segment, duration (in seconds) of the segment (as determined by the forced alignment boundaries), and tonal features and durations of the two immediately preceding and two immediately following segments. The latter we add with the hope that the additional context may allow for modeling of coarticulation effects.

Hyperparameters (regularization strength, gamma, stopping criteria, etc.) for logistic regression and SVM training were set via grid-search using 5-fold cross-validation on the training set. The neural network contained a single hidden layer of 128 ReLUs and was trained for 100 epochs (epoch=100,000 instances) using stochastic gradient descent with minibatches of size 128, 50% dropout in the hidden layer, a learning rate of 0.5, and momentum of 0.9. The in-

coming weight vector at each hidden unit was constrained to have a maximum L_2-norm of 2.

3.4. Results

As an initial evaluation of the quality of the representation learned by the network, we consider its frame error rate (FER), defined as the percentage of frames incorrectly classified. As seen in in Table 2, overall FER is quite good at only 16.36%, a 3.34% absolute and 16.95% relative reduction from the previously best reported FER for broadcast news [27].

However, a score that includes non-tonal regions is problematic, because silences and other unvoiced regions are relatively easy to recognize in material of this kind (see Figure 1) and, therefore, a FER that includes such regions will depend on the amount of silence that is included in the test set. Our test set includes only the regions marked as parts of utterances in the published corpus and while we include overall FER for comparison with other results reported in this way, we feel that it is more meaningful to exclude frames that do not correspond to a tone bearing unit in the gold standard segmentation. By this metric, our FER rises to 27.3%. Excluding neutral tones, FER improves to 26.60%.

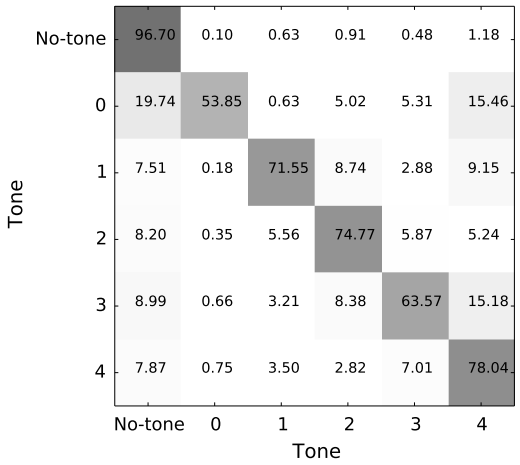

Fig. 1: Confusion matrix (%) for MFCC system.

	Overall	TBUs	Tones 1−4
MFCC	16.36	27.38	26.60
F0	24.22	40.05	37.78

Table 2: Frame error rates (%) on test set: overall, for TBUs, and for TBUs excluding Tone0.

Our primary metric, however, is segment error rate (SER), defined as the percentage of TBUs incorrectly classified. Using the tonal features derived from this DNN, our baseline system achieves an SER of 17.73%. Moving to a supervised approach substantially improves matters, as is seen in Table 3. Adding increasing amounts of context also helps, with SER lower across-the-board when the tonal features of surrounding segments are included as inputs. Interestingly, this contextual effect is seen even when only adding the immediately adjacent segments, which usually correspond to syllable initials. Our best SER, achieved with the combination of a 2 segment context and neural networks, is 15.62%, representing a 2.11% absolute (11.9% relative) reduction from the baseline. These numbers are the best reported in the literature outside the context of lab speech. Even our baseline represents a 6.07% absolute and 25.5% relative error reduction from [1], the best result we can find for broadcast news.

Features	Logistic	SVM	NN
tonal features and duration	16.98	16.98	16.57
+ 1 segment context	16.69	16.75	15.96
+ 2 segment context	16.31	16.43	15.62

Table 3: Segment error rates (%) for MFCC system under different combinations of segment-level features and classifier. The baseline achieves 17.73% SER.

4. TONE CLASSIFICATION WITH F_0 AND AMPLITUDE INFORMATION

The error rates reported in Section 3.4 were achieved without the inclusion of explicit F_0 information. Consequently, we decided to train a second system including F_0 to compare results. For this system we utilize the same DNN topology and training procedure, but replace MFCCs with F_0. F_0 was computed using RAPT [28] as implemented in ESPS's *get_f0* (parameters: wind_dur=0.01, min_f0=60, max_f0=650) and normalized to have mean 0 and variance 1 within voiced regions on a per-utterance basis. As a measure of amplitude, we also included log-energy, computed using a 25 ms Hamming window and 1024-point DFT and normalized to mean 0 and variance 1 on a per-utterance basis. Both F_0 and log-energy were extracted every ms and frames at offsets of -100 ms, -90 ms, ..., +90 ms, +100 ms concatenated as before, resulting in a 42-dimensional input to the neural network.

4.1. Results

Both FER and SER are substantially worse with the F_0-based system compared to the MFCC system. Overall FER rises from 16.36% to 24.22%, a nearly 50% increase, and FER on frames corresponding to TBUs rises from 27.38% to 40.05%. Similar performance degradation is seen in SER with the best performing combination of features/classifier achieving only 22.66%, a 4.93% absolute increase over the MFCC baseline and a 7.04% absolute (and 45.07% relative) increase over the best performing combination of features/classifier using MFCCs.

Features	Logistic	SVM	NN
tonal features and duration	30.31	29.36	29.27
+ 1 segment context	29.04	26.56	24.83
+ 2 segment context	27.51	26.33	22.66

Table 4: Segment error rates (%) for F_0 system under different combinations of segment-level features and classifier. The baseline achieves 31.64% SER.

4.2. Human performance

We were unable to find useful information on perception of tone from F_0 and amplitude alone in material of this kind, so we performed a simple pilot experiment to get an initial estimate for human performance. Tests of this kind often use low-pass filtered speech, but we found that even with a steep low-pass cutoff at 300 Hz, listeners were often able to guess what the original word sequence had been, so we used an approach based on synthesizing sounds from estimated F_0 and amplitude contours.

We produced F_0 and amplitude estimates for our test utterances using *get_f0* with min_f0=60, max_f0=350, wind_dur=0.01, and frame_step=0.005. We then created synthetic stimuli consisting of a fundamental and nine overtones at $\frac{1}{F}$ amplitudes, modulated with amplitude and frequency contours matching those extracted from the test utterances during voiced intervals, with amplitude set to 0 in unvoiced regions. Comparative listening verified that the resulting non-speech stimuli generally seemed to express the same subjective melody as the spoken originals, though in a few percent of cases there were problems due to voicing or pitch tracking errors.

We randomly selected 15 utterances from each of the six speakers in the test set, comprising 982 tones in all. One native speaker of Mandarin Chinese (the second author) categorized the tones in the re-synthesized versions of these 90 utterances. He used a waveform display with the syllabic segmentation indicated, listening carefully to the individual syllables, to the syllables in their immediate context, and to the entire utterance.

Subsequent checking determined that 39 of the 982 syllables had re-synthesis problems (most often failure to detect voicing), so we retained judgments for the remaining 943. The overall score was 71.79% correct. For the five (gold standard) tonal categories, the scores were T0: 22.22%; T1: 78.87%; T2: 75.33%; T3: 59.35%; T4: 77.68%. Without T0 (889 tones), the overall accuracy was 74.8%. (18 other tones were incorrectly identified as T0, so in a 4-way forced choice the results would have been slightly better.)

The overall confusion matrix is shown in Figure 2, with the gold-standard tones in the rows, and the listener's responses in the columns. To our surprise, the human score of 28.21% SER was substantially worse than our best MFCC system score of 15.62% and was much closer, though still worse, to our best F_0-based system score of 22.66% SER.

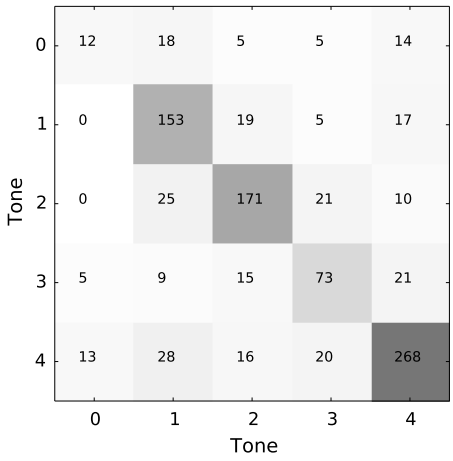

Fig. 2: Confusion matrix for perceptual experiment.

The human performance does not seem to reflect problems in the analysis or synthesis of the test stimuli. It may reflect difficulties in assigning non-speech melodies to tonal classes, though we would like to note the possibility that the lack of tone-associated acoustic features other than F_0 and amplitude may also play an important role.

5. CONCLUSION

The results of these experiments raise many questions, some of which are suggested by the following alternative ideas about the reasons for our results:

(1) Perhaps the tone-classification information is purely in the F_0 and amplitude contours, but the DNN system is getting this information more accurately from the MFCC parameters than it can from the ESPS pitch tracker. (In this context, it's important to note that we are retaining 40 MFCC spectral parameters, rather than the usual 12.)

(2) Perhaps other tone-class-related (or pitch-range-related) phonetic dimensions, such as voice quality, are providing additional information useful for tone classification, which the DNN system is able to extract from the MFCC parameters.

(3) Perhaps (in this test set) the tone classes are correlated with segmental features (vowel quality, nasality, etc.), which the DNN system is also able to extract from the MFCC parameters, even without any supervision for these features during the training phase.

If (1) or (2) are true even to some extent, then perhaps other kinds of prosodic analysis should start from richer representations of the speech signal.

6. REFERENCES

[1] X. Lei, M.-H. Siu, M.-Y. Hwang, M. Ostendorf, and T. Lee, "Improved tone modeling for Mandarin broadcast news speech recognition," in *INTERSPEECH*, 2006.

[2] Y. Chao, "A system of tone letters," *Le maître phonétique*, vol. 45, 1980.

[3] M. J. Yip, *The Tonal Phonology of Chinese*, Ph.D. thesis, Massachusetts Institute of Technology, 1980.

[4] C. Shih, "The phonetics of the Chinese tonal system," *AT&T Bell Labs technical memo*, 1987.

[5] J. van Santen, C. Shih, and B. Möbius, "Intonation," *Multilingual text-to-speech synthesis: the Bell Labs approach*, pp. 141–189, 1997.

[6] N. Umeda, "F_0 declination is situation dependent," *JASA*, vol. 68, pp. S70, 1980.

[7] C. Shih, "Declination in Mandarin," in *Intonation: Theory, Models and Applications*, 1997.

[8] Y. Xu, "Effects of tone and focus on the formation and alignment of f_0 contours," *Journal of Phonetics*, vol. 27, no. 1, pp. 55–105, 1999.

[9] J. Yuan, *Intonation in Mandarin Chinese: Acoustics, perception, and computational modeling*, Ph.D. thesis, Cornell University, Aug., 2004.

[10] J. Yuan and M. Liberman, "F_0 declination in English and Mandarin broadcast news speech," in *INTERSPEECH*, 2010, pp. 134–137.

[11] C. J. Chen, R. A. Gopinath, M. D. Monkowski, M. A. Picheny, and K. Shen, "New methods in continuous Mandarin speech recognition," in *Eurospeech*, 1997.

[12] E. Chang, J.-L. Zhou, S. Di, C. Huang, and K.-F. Lee, "Large vocabulary Mandarin speech recognition with different approaches in modeling tones," in *INTERSPEECH*, 2000, pp. 983–986.

[13] H. C.-H. Huang and F. Seide, "Pitch tracking and tone features for Mandarin speech recognition," in *Proceedings of ICASSP*, 2000, vol. 3, pp. 1523–1526.

[14] R. Sinha, M. Gales, D. Kim, X. Liu, K. Sim, and P. Woodland, "The CU-HTK Mandarin broadcast news transcription system," in *Proceedings of ICASSP*, 2006.

[15] W. Pui-Fung and S. Man-Hung, "Decision tree based tone modeling for Chinese speech recognition," in *Proceedings of ICASSP*, 2004, pp. 905–908.

[16] H. Chao, Z. Yang, and W. Liu, "Improved tone modeling by exploiting articulatory features for Mandarin speech recognition," in *Proceedings of ICASSP*, 2012, pp. 4741–4744.

[17] O. Kalinli, "Tone and pitch accent classification using auditory attention cues," in *Proceedings of ICASSP*, 2011, pp. 5208–5211.

[18] A. Mohamed, G. E. Dahl, and G. E. Hinton, "Deep belief networks for phone recognition," in *NIPS Workshop on Deep Learning for Speech Recognition and Related Applications*, 2009.

[19] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 30–42, 2012.

[20] S. Huang, J. Liu, X. Wu, L. Wu, Y. Yan, and Z. Qin, *1997 Mandarin Broadcast News Speech (HUB4-NE)*, Linguistic Data Consortium, 1998.

[21] D. P. W. Ellis, "PLP and RASTA (and MFCC, and inversion) in Matlab," 2005, online web resource.

[22] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[23] S. Huang, X. Bian, G. Wu, and C. McLemore, *CALLHOME Mandarin Chinese Lexicon*, Linguistic Data Consortium, 1997.

[24] J. Yuan, N. Ryant, M. Liberman, A. Stolcke, V. Mitra, and W. Wang, "Automatic phonetic segmentation using boundary models," in *INTERSPEECH*, 2013, pp. 2306–2310.

[25] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Procedings of ICML*, 2010, pp. 807–814.

[26] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[27] X. Lei, M.-Y. Hwang, and M. Ostendorf, "Incorporating tone-related MLP posteriors in the feature representation for Mandarin ASR," in *INTERSPEECH*, 2005, pp. 2981–2984.

[28] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, vol. 495, pp. 518, 1995.