TONE RECOGNITION FOR CONTINUOUS ACCENTED MANDARIN CHINESE

Jiang Wu, Stephen A. Zahorian, Hongbing Hu

Department of Electrical and Computer Engineering, Binghamton University,
Binghamton, NY 13902, USA
{jiang.wu, zahorian, hongbing.hu}@binghamton.edu

ABSTRACT

In this paper, the ability of human listeners to recognize tones from continuous Mandarin Chinese is evaluated and compared to the accuracy of automatic systems for tone classification and recognition. All tones used for experimentation were extracted from the RASC863 continuous Mandarin Chinese database. The human listeners are native speakers of Mandarin and the automatic methods consist of tone classification using neural networks and tone recognition using Hidden Markov Models. Features used for the automatic methods are a combination of spectral/temporal features, energy contours, and pitch contours. When very little context is used (i.e., vowel segments only) the human and machine performance is comparable. However, as the context interval is increased, the human performance is much better than the best machine performance obtained.

Index Terms— tone recognition, continuous Mandarin Chinese, human listeners, neural networks, HMMs

1. INTRODUCTION

Typically a Mandarin Chinese syllable contains 3 acoustic elements: the syllable initial, the syllable final and the tone. Tones are important characteristics of Mandarin Chinese for conveying lexical meaning and distinguishing different characters. Thus tone recognition (especially for the explicit tone modeling technique) is required for automatic recognition of Mandarin.

Most literature on machine recognition of tones is based on syllables spoken in isolation [1] or high quality speech such as TV broadcast news [2]. This is likely due to the fact that recognizing tones from syllables extracted from conversational speech is difficult even for humans: some linguistic research suggests that human listeners require long duration acoustic cues in order to recognize tones correctly [3]. The perception of tones also varies depending on the listener's native experience with the tonal system of his/her own language [4].

This paper explores the recognition ability of humans for lexical tones in light-accented continuous Mandarin Chinese for different conditions. Then two automatic methods for monotone recognition will be discussed and compared with both human's accuracy and other tone modeling techniques [5] [6].

The Shanghai region data from RASC863 (Regional Accented Speech Corpus) [7] was utilized for all experiments reported in this paper, since it provides phonetically labeled transcriptions and the accent from speakers in this region is considered the "lightest" of all 4 regions included in the RASC863 database.

2. TONE RECOGNITION BY HUMAN LISTENERS

2.1. Test Participants

3 male and 3 female college students whose native language is Mandarin Chinese were selected as the experimental subjects. All participants were trained and tested by the first author of this paper and they each appeared to have a good understanding of lexical tones in Mandarin (High, Rising, Dipping and Falling) and were able to correctly identify these tones from careful listening of well pronounced tones.

Note that throughout this paper, for all experiments, the neutral tone was not considered since it occurs relatively infrequently.

2.2. Objective

The main object for this experiment was to investigate the capability of native listeners to recognize tones, for each of four cases, with varying amounts of context (i.e., length). For each case, listeners were given approximately 800 speech segments to listen to and were asked to identify tones, based on a single playback of each speech sample. The segments (approximately 3200 in total) were directly extracted from the continuous speech data portion of the RASC863 database. The type of speech segment for each case is listed in Table 1.

Table 1. Syllable segment cases

Case	Description	Label in this paper
Case 1	Syllable final (vowel part only)	Vowel
Case 2	Complete syllable (consonant and vowel)	1 SLB
Case 3	Two syllable segment	2 SLB
Case 4	Three syllable segment	3 SLB

The syllable segments for Cases 3 and 4 are consecutive syllables extracted from random positions in a sentence. In some cases the syllable strings were "words," but not in all cases, as discussed later.

2.3. Experimental Protocol and Test Software

Listeners used a PC for playing sound tokens and recording their answers with interaction with the computer via a Graphic User Interface (GUI) as shown in Fig. 1. Listeners simply

marked tone type (High, Rising, Dipping and Falling) according to what they just heard. Listeners were comfortably seated and used high quality headphones.

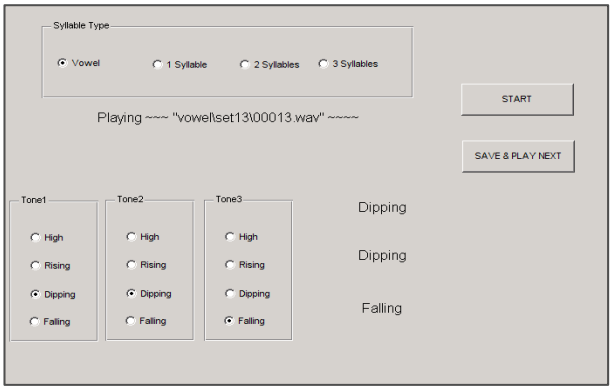

Fig. 1. GUI for listening experiment

Listeners were only allowed to listen once to each token, but could decide when to begin the playback of each token. Typically listeners required about 90 minutes for each of the four cases.

2.4. Listening Test Results and Analysis

2.4.1 General results

The overall tone recognition results from the 6 listeners (labeled as 'M1' to 'F3', with M for male, and F for female) and for the 4 cases (labeled 'Vowel' to '3 SLB'), are given in Table 2.

The average recognition rate for multiple syllable cases 3 and 4 is noticeable higher (78.2% and 79.5%) than for the single syllable cases 1 and 2 (61.9% and 58.0%)

Table 2. Tone recognition accuracies for human listeners

(%)	M1	M2	M3	F1	F2	F3	Avg.
Vowel	59.4	51.1	70.0	67.3	59.9	63.6	61.9
1 SLB	58.7	53.8	64.7	53.6	57.4	59.8	58.0
2 SLB	75.7	74.4	81.3	80.8	80.3	76.8	78.2
3 SLB	80.7	67.8	86.1	83.1	80.8	78.8	79.5
Avg.	68.6	61.8	75.5	71.2	69.6	69.8	69.4

2.4.2 Confusion matrix for tones based on "Vowel" only

For the most part, the tone is only associated with voiced speech. Thus most research on automatic Mandarin recognition, either implicitly or explicitly, models tones based on acoustic features that are extracted from vowels only.

A confusion matrix was generated from the tone recognition based on the case "Vowel," and is given in Table 3Table 3 as percentages. The High tone is recognized most accurately, whereas the Dipping and Rising tones are recognized least accurately. These results are consistent with [3], where it was argued that shorter acoustic cues are more suitable for recognizing High tones and Falling tones.

Table 3. Confusion matrix for tones based on vowels only

(%)	High	Rising	Dipping	Falling
High	70.4	10.9	5.4	13.3
Rising	21.7	59.6	8.6	10.2
Dipping	16.9	25.9	44.2	13.0
Falling	18.4	7.2	11.5	63.0

2.4.3 Tone Recognition in "words" vs. "not-a-word" segments

For the case of 2 SLB, listening results were sorted according to whether or not 2 SLB segments formed a word. Due to the random way that each segment was selected from a sentence, a syllable pair may or may not form a lexical word. (Note that most Chinese words are comprised of 2 syllables). In the data used, approximately half the two syllable segments were words, and half were not linguistically meaningful. As expected, when recognizing tones from segments which are "words," the linguistic knowledge of the listeners helps them to recognize tones much more accurately than when the segments are "not-words."

Tone recognition results of listeners for syllables in words or not-words are shown in Fig. 2.

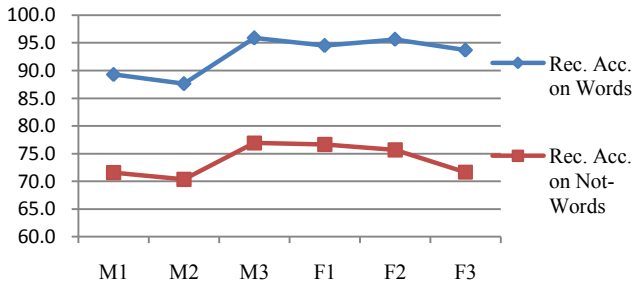

Fig. 2. Tone recognition accuracies of listeners for syllables in words or not-words

2.4.4 Gender effects

Results were also sorted by gender of both listeners and speakers and averages given in Table 4. These results do not show any large gender-dependent effect for either listener or speaker. For the small group of listeners used, the females were approximately 5% more accurate at tone identification than males. However, the small number of listeners used makes it doubtful that this difference is statistically meaningful.

Table 4. Tone recognition accuracy sorted by gender

Spks/Lsns(%)	Male (Lsn)	Female(Lsn)	Average
Male (Spk)	75.1	78.5	76.8
Female (Spk)	75.8	81.6	78.7
Average	75.4	80.0	78.2

2.4.5 Recognition for each syllable in a "3 SLB string"

As shown in Table 5, the results for the 3 SLB case have highest overall accuracy for the middle syllable. This is probably because the middle syllable has both right and left

context, whereas the first syllable has only right context and the final syllable has only left context.

Table 5. Tone recognition accuracy for 3 SLB case

(%)	M1	M2	M3	F1	F2	F3	Avg.
1st SLB	76.3	72.9	80.4	79.5	76.2	73.5	76.5
2nd SLB	78.4	80.4	90.9	88.3	86.2	83.5	84.6
3rd SLB	87.3	69.9	86.8	81.3	79.8	79.5	80.8

3. TONE RECOGNITION USING AUTOMATIC SYSTEMS

Because tones exhibit strong co-articulation phenomenon, most researchers have modeled tones explicitly in a bi-tone structure [5] or implicitly with other phones [6]. However, it is very difficult to perform phonetic level tone recognition in continuous speech this way because it is very hard to identify co-articulation boundaries. Therefore, in our work we classified or recognized tones as "monotones," essentially using techniques that would be used for phone classification and recognition in English, except there are only the four tones, and the features selected for the classifier and recognizer were selected to indicate changes in spectral energy concentrations over time.

3.1. Speech Data

Both automatic methods utilized the phonetically-balanced subset from RASC863 (e.g. All sentences that begin with the letter 's' and which have phonetic transcriptions), which includes a total of 2209 sentences from 20 speakers. (1548 sentences for training and 661 sentences for test).

3.2. Acoustic Features

3.2.1 DCTC/DCSC features

Spectral/temporal features have shown to be effective in past research on phonetic recognition for continuous English [8]. Although pitch (F_0) contours are most typically used as acoustic correlates for tone (and are used in this study), inspection of the low frequency parts of spectrograms, shows that the tonal characteristics are often apparent from the general shift in energy concentration over time, especially in the low frequency (below 1000 Hz) part of the spectrogram. Fig. 3 illustrates this idea.

As used in our work on phonetic recognition [8], Discrete Cosine Transform Coefficients (DCTCs) and Discrete Cosine Series Coefficients (DCSCs) were used as features for encoding the general shape of the spectrum. Experimentally, it was found that a small number of DCTC/DCSC terms (such as 6 DCTCs, each represented by 4 DCSCs), extracted from a low frequency range (75~700Hz) were most effective.

3.2.2 Pitch contour features

The pitch contour is widely accepted as the most effective feature for tone modeling ([1] [4] [5] [6] [7]). The research reported in this paper made use of the fundamental frequency tracking algorithm called YAAPT [9] for pitch. Pitch contours were represented with Discrete Cosine Series Coefficients terms (4-5) in the same manner as DCTC trajectories are encoded (i.e. as in [8]). The black dots placed on the spectrograms shown in Fig. 3 are pitch tracks computed by YAAPT.

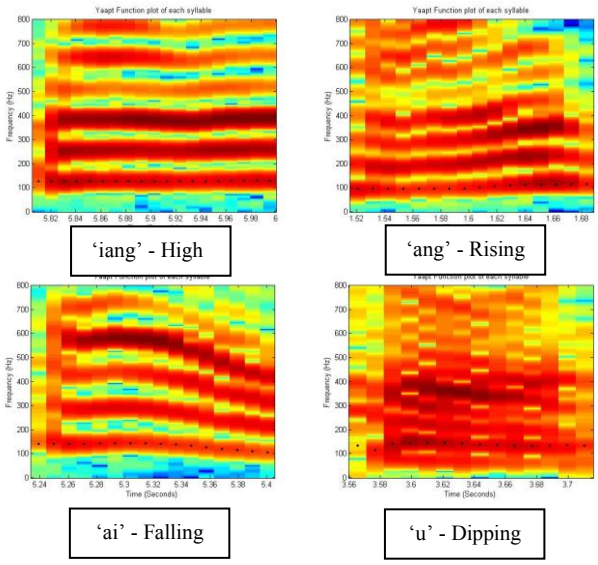

Fig. 3. Spectrograms and pitch for 4 tonal syllables.

3.3. Classification Using Neural Networks

For these experiments, segments were extracted from the data base using the supplied labels for vowels. Segment durations were varied from 100 ms to 500 ms in steps of 50 ms. For each segment length, DCTCs and/or pitch was computed and then represented with DCSC terms. Three conditions were tested:

a. DCTCs + Pitch: 4 DCTCs encoded with 5 DCSC terms each, pitch encoded with 5 DCSC terms (25 total features)
b. Pitch only: pitch feature encoded with 5 DCSC terms (5 features in total)
c. DCTCs only: 6 DCTCs encoded with 4 DCSC terms each, resulting in a total of 24 features.

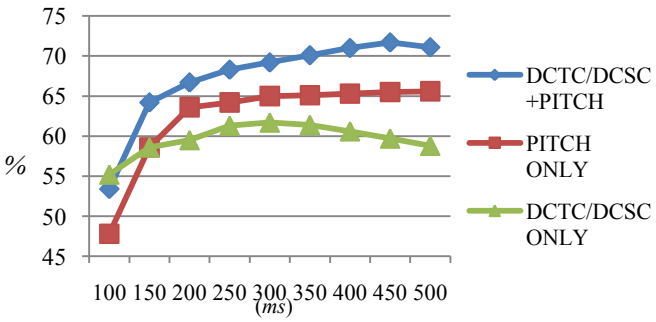

Fig. 4. Tone accuracy based on neural network classification

These features were classified with a neural network classifier having two hidden layers (100 hidden nodes, 25

hidden nodes) and an output layer of 4 nodes. Test results, for each of the three features sets, as a function of segment length are shown in Fig. 4. The confusion matrix is given in Table 6.

Table 6. Confusion matrix for neural network classification

(%)	High	Rising	Dipping	Falling
High	78.2	4.9	6.5	10.4
Rising	6.2	75.5	5.7	12.5
Dipping	19.1	13.8	61.6	5.5
Falling	14.7	15.6	4.2	65.5

3.4. Recognition Using Hidden Markov Models

The other automatic method is a recognizer built with the HTK toolkit (Ver3.4). This toolkit allows complete flexibility in terms of the number of mixtures, number of states, types of transitions, and provides for language modeling.

5 tones (High, Rising, Dipping, Falling and Neutral), silence, and all syllable initials (consonants) were modeled by 7 left-to-right 3-state HMMs with no skip states allowed. A bigram language model was also built from the statistical data of all transcriptions. The feature sets tested were:

a. 48 DCTC/DCSCs: 8 DCTCs each encoded with 6 DCSCs
b. DCTC/DCSCs + Pitch: 40 DCTC/DCSCs (8 DCTCs × 5 DCSCs) + Log energy + Pitch + Pitch's 6 DCSCs

As shown in Table 7 and 8, the best results were obtained with 48 acoustic features combining DCTC/DCSC and pitch features, modeled 48 Gaussian mixtures. All DCSCs were computed using a block length of 200 ms. Note that several other conditions were also tested; results are given for the "best" case found to date.

Table 7. Best result for the HMM recognition

	DCTC/DCSCs	DCTC/DCSCs + Pitch
Rec. Rate (%)	53.7	60.3

Table 8. Confusion matrix from HMM recognition experiment

(%)	High	Rising	Dipping	Falling	Neutral
High	81.0	7.0	3.1	6.8	2.2
Rising	10.0	69.8	9.7	5.6	4.9
Dipping	4.3	9.6	71.1	7.7	7.3
Falling	8.0	3.3	5.9	76.7	6.1
Neutral	5.0	5.0	10.4	8.3	71.4

Note that the results in Table 7 are based on 7 categories whereas the results in Table 8 are based on 5 tone categories only, hence the difference of the accuracies.

4. CONCLUSIONS

This paper compares the ability for recognizing tones from continuous Mandarin Chinese between human native speakers and two automatic methods.

The experimental results show humans need context to recognize tones very accurately. Nonnative speakers find this task nearly impossible. Without context, machine recognition and human recognition have about same accuracy, but different patterns of errors.

The most interesting and potentially significant result of this work is that reasonably accurate tone classification and recognition can be obtained without using a pitch feature. Best tone classification accuracy (71.7%) was obtained using both the spectral/temporal features and pitch trajectories.

Similarly, for the more difficult recognition case, most accurate results were obtained with the combined feature set versus either set alone (approximately 5% accuracy improvement with pitch added to DCTC/DCSC features).

Note that "raw" pitch was used for all pitch features. Presumably, as noted in other studies, higher accuracies would have been obtained if the pitch contours had been normalized to reduce speaker dependent effects.

5. ACKNOWLEDGEMENTS

This material is based on research sponsored by the Air Force Research Laboratory under agreement number FA8750-10-2-0160. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Research Laboratory or the U.S. government.

6. REFERENCES

[1] O. Kalinli, "Tone and Pitch Accent Classification Using Auditory Attention Cues," *IEEE ICASSP*, pp. 5208-5211, 2011.

[2] X. Lei, M. Siu, M.Y. Hwang, M. Ostendorf and T. Lee, "Improved Tone Modeling for Mandarin Broadcast News Speech Recognition," INTERSPEECH 2006.

[3] Y. Lai and S.H. Wu, "The Effect of Segmental Makeup on Mandarin Word and Tone Recognition," *meeting of Acoustical Society of America*, Kansas City, 2012.

[4] R. Wayland, D. Laphasradakul, E. Kaan, R. Cao, "Perception of Pitch Contours among Native Tone Listeners," INTERSPEECH *2012*.

[5] S. Promo-on, F. Liu, and Y. Xu, "Post-low Bouncing in Mandarin Chinese: Acoustic Analysis and Computational Modeling," *Journal of Acoustical Society of America*, pp. 421-432, 2012.

[6] J. Zhou, T. Ye, Y. Shi, C. Huang and E. Chang, "Tone Articulation Modeling For Mandarin Spontaneous Speech Recognition," *IEEE ICASSP*, 2004.

[7] A. Li, Z. Yin, T. Wang, Q. Fang and F. Hu, "RASC863 - A Chinese Speech Corpus with Four Regional Accents," report of Chinese Academy of Sciences.

[8] S. A. Zahorian, H. Hu, Z. Chen and J. Wu, "Spectral and Temporal Modulation Features for Phonetic Recognition," INTERSPEECH *2009*, pp. 1071-1074, 2009.

[9] S. A. Zahorian and H. Hu, "A Spectral/Temporal Method for Robust Fundamental Frequency Tracking," *Journal of Acoustical Society of America*, pp. 4559-4571, 2008.