HMM-Based Mandarin Tone Recognition - Application in Computer-Aided Language Learning System for Mandarin

Hussein Hussein[1]*, *Hansjörg Mixdorff*[2], *Yuan-Fu Liao*[3] *and Rüdiger Hoffmann*[1]

[1] *Chair for System Theory and Speech Technology, Dresden University of Technology, Dresden, Germany*

[2] *Department of Computer Sciences and Media, Beuth University of Applied Sciences, Berlin, Germany*

[3] *Department of Electronic Engineering, National Taipei University of Technology, Taipei, Taiwan*

hussein.hussein@mailbox.tu-dresden.de, mixdorff@beuth-hochschule.de, yfliao@ntut.edu.tw, ruediger.hoffmann@tu-dresden.de

Abstract: The current paper reports our study on automatic Mandarin tone recognition towards the integration of tone recognition system in a computer-aided language learning (CALL) system for German learners of Mandarin. Three HMM-based tone recognition systems were developed including monotone, bitone and tritone recognizer for isolated monosyllabic, bisyllabic words and sentences, respectively. Different kinds of features, including prosodic and spectral-based features, were used in order to study its quality for tone recognition. The F_0 contour was decomposed according to the Fujisaki model to its components which contain phrase components and tone components. In order to test the tone recognition systems on data from German learners of Mandarin, the tone models were adapted using correct data from the German students. The combination of prosodic and spectral-based features yielded better results than individual features. The results indicated that the proposed monotone and bitone recognizers outperform existing state-of-the-art algorithms. The tone correctness of adapted acoustic models was better than original models.

1 Introduction

Mandarin (standard Chinese) is a tone language and hence the tonal contour of a syllable changes its meaning. The traditional syllable structure of Mandarin is called initial-final. There are 22 initials (including glottal stop) and 39 finals. Mandarin comprises a relatively small number of syllables (range from 403 to 413 in both speech recognition and speech synthesis). The most important acoustic correlate of tone is the F_0 contour. Mandarin has four syllabic tones and a neutral tone in unstressed syllables. In citation forms of monosyllabic words the tonal patterns are very distinct (see figure 1), but when several syllables are connected, F_0 contours observed vary considerably due to tonal coarticulation. German is a non-tonal language. Mandarin differs from German significantly on the segmental as well as the suprasegmental level and poses a number of problems to the German learners, especially the tonal distinction [1].

Accurate Mandarin tone recognition plays an important role in automatic speech recognition. High accuracy of tone recognition were obtained only in isolated words in comparison to con-

*The author developed a large part of this work during his work at the Beuth University of Applied Sciences.

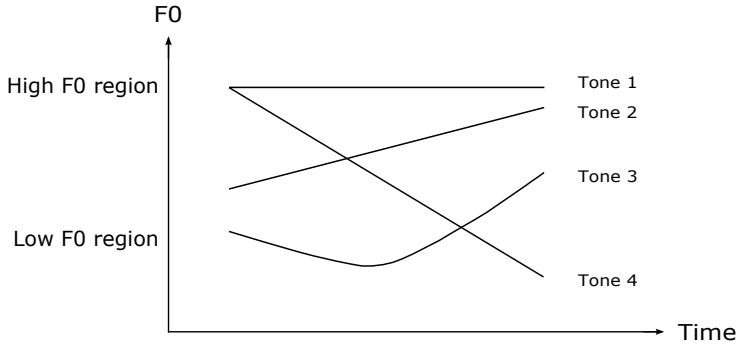

Figure 1 - Typical F_0 patterns of four basic lexical tones.

tinuous speech [2]. Many methods for Mandarin tone recognition have been proposed in the literature including Hidden Markov Models (HMMs), Gaussian Mixture Models (GMMs), Neural Networks (NNs), Decision tree classification, Support Vector Machines (SVMs) and rule based methods, e.g. [3][2][4].

A robust feature extractin and robust tone modeling techniques are needed for good performance of the tone recognition system [5]. Most of tone recognition systems use F_0 features. The accuracy of tone recognition for the four basic tones are high, but it is low for the neutral tone. Because the F_0 features are not effective to discriminate the neutral tone. The energy features is an effective cue for tone perception when F_0 is missing [6]. The spectral features were used usualy to recognize initials and finals. Therefore, different kinds of features, including prosodic and spectral-based features, were used in this study in order to overcome the changes in F_0 contour of syllables and to study its quality for tone recognition. Three tone recognition systems (monotone and bitone recognizer for isolated words as well as tritone based continuous speech recognizer) were developed using continuous density HMMs during the development of the CALL System for German learners of Mandarin (henceforth "*CALL-Mandarin system*") [1][7].

The well-known Fujisaki model is a super-positional model for parameterizing F_0 contours in speech synthesis for intonation analysis and intonation generation [8]. The Fujisaki model in tone languages reproduces a given F_0 contour by superimposing three components: a speaker-individual base frequency Fb, phrase components and tone components. It was found in our previous experiment that for most utterances phrase commands of magnitude Ap greater 0 occurred, indicating that the phrase component should be taken into account when analyzing and synthesizing of F_0 contour of Mandarin [9]. Therefore, in order to investigate the effect of phrase components on the tone recognition, two feature sets were constructed using the smoothed F_0 contour on one hand and the high frequency contour of tone components (removing the phrase components and Fb) on the other hand. The tone models were trained using speech data from native speakers of Mandarin. Before the integration of tone recognition systems in the *CALL-Mandarin system*, the tone models were adapted using correct data from the German learners of Mandarin.

2 Speech Material

2.1 Chinese Data - L1

The experiments of speaker-independent tone recognition were carried out using three read speech databases from native speakers of Mandarin (henceforth "*CN_Mono*", "*CN_Bi*" and "*CN_Sent*").

1. ***CN_Mono* - Isolated Monosyllabic Words**:
   The monotone recognizer was trained using isolated monosyllabic words. It were uttered by 29 female and 27 male native speakers of Mandarin, yielding a total of 45000 monosyllables (14.83 hours).
2. ***CN_Bi* - Isolated Disyllabic Words**:
   The bitone recognizer was trained using isolated disyllabic words. The disyllabic words were produced by 29 female and 27 male native speakers of Mandarin with a total of 75000 disyllables (28.83 hours).
   The iFLYTEK company, Hefei, China provided the speech data and segmentation on syllable and phone-levels for both *CN_Mono* and *CN_Bi*. The data was recorded with a sampling frequency of 16 kHz and a resolution of 16 bit.
3. ***CN_Sent* - Sentences**:
   The tritone recognizer was trained using sentences which contain mono-, di- and polysyllabic words. A part of the Mandarin speech corpus TCC300 [10] was used. The TCC300 database was produced by three universities. The speech data of each university was recorded by 100 speakers (50 males and 50 females). The recordings contain read speech in an original binary format (sampling frequency of 16 kHz and a resolution of 16 bit). The speech data from two universities (200 speakers) was used only in the experiments. It contains a total of 2023 utterances (18.60 hours). Each utterance contains a recording of one paragraph composed of several long sentences with a minimum of 11 and a maximum of 231 syllables. The average length of utterances is about 115 syllables. A two-stage sample-based phone boundary detector using segmental similarity features was used for segmentation on phone-level [11].

The distribution of the Mandarin tones in the Chinese databases shows that tone 4 occurs the most frequently and tone 0 the least frequently. We want to mention that the Chinese database of disyllabic words (*CN_Bi*) did not contain neutral tones.

2.2 German Data - L2

The tone recognition systems were also tested on three read speech databases from German learners of Mandarin (henceforth "*DE_Mono*", "*DE_Bi*" and "*DE_Sent*"). The *DE* was collected during the development of the *CALL-Mandarin system*.

1. ***DE_Mono* - Isolated Monosyllabic Words**:
   The first part of German data consists of eight monosyllabic words. The tokens was provided in Pinyin transcription and read aloud (reading mode). It was produced by 14 first-year German learners of Mandarin (seven male and seven female).
2. ***DE_Bi* - Isolated Disyllabic Words**:
   The second part of German data consists of 19 disyllabic words. The *DE_Bi* was recorded in reading mode by the same students who produced *DE_Mono*.
3. ***DE_Sent* - Sentences**:
   The third part of German data consists of 62 sentences. The average length of utterances is about 7 syllables, with a minimum of two and a maximum of 14 syllables. They were produced by ten first-year students (two male and eight female), three second-year students (one male and two female), and eight third-year students (two male and six female).

3 Tone Recognition System

The basic tone recognition system consists of features extraction as well as model training and tone recognition/evaluation.

3.1 Feature Extraction and Normalization

The height and shape of the F_0 contour are critical for Mandarin tone recognition. The F_0 contours were calculated using the RAPT algorithm with a step of 10msec [12]. The output of RAPT contains in addition to F_0 values the energy- (Root Mean Square - RMS) and degree-of-voicing-measures (DoV) (*ESPS/waves+* format). The DoV value ranges from 0 (fully unvoiced) to 1 (fully voiced segment). The extracted F_0 contours were preprocessed by detection the minimum and maximum parameters of F_0 for male (70 and 350 Hz) and for female speakers (90 and 450 Hz). The automatic approach for extraction of Fujisaki model parameters was used to decompose the F_0 contour to its components [13]. Therefore, the F_0 contour is first interpolated and smoothed using a quadratic spline stylization. A high-pass filter is then applied to extract the High Frequency Contour (HFC) from the smoothed F_0 contour. The HFC contains tone commands. The HFC is subtracted from the smoothed contour, yielding a Low Frequency Contour (LFC) from which phrase commands are extracted (see figure 2, c). In order to investigate the effect of phrase components on the tone recognition, two feature sets were constructed using the smoothed F_0 contour on one hand and the HFC of tone components on the other hand. Since the smoothed F_0 contour and HFC are continuous unbroken pitch contour, they provide pitch information over voiced regions for tone recognition. This provides some transition information from the preceding and to the succeeding syllables [4]. The Mel-Frequency Cepstral Coefficients (MFCC) (39 features) were used also for tone recognition.

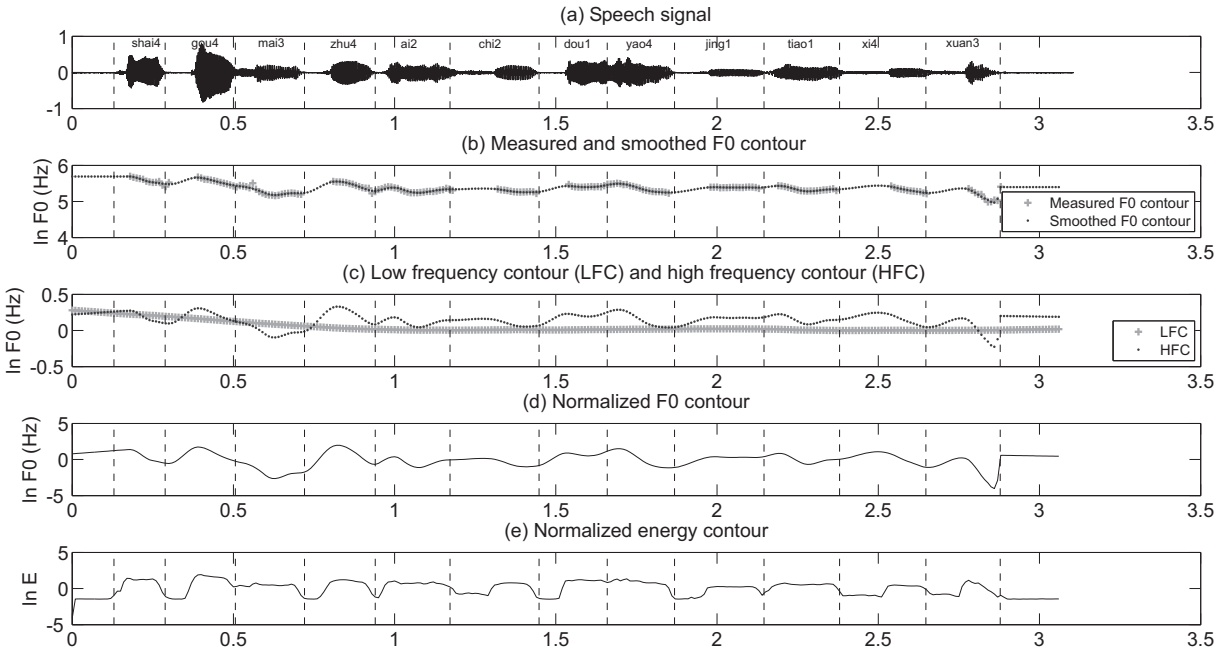

Figure 2 - Speech signal, measured and smoothed F_0 contours, low and high frequency contours, normalized F_0 and energy contours. The vertical lines show the syllable boundaries.

The range of F_0 contour depends on the speaker's gender and age. Therefore, F_0 was further normalized across speakers to enable a meaning comparison of F_0 values between different

speakers. *Z-score* normalization was applied on the smoothed F_0 contour as well as on the HFC (the mean and variance of F_0 contour normalized to zero and one, respectively). The normalized F_0 value is given by:

$$Y_i = \frac{X_i - \mu}{\sigma} \tag{1}$$

where Y_i is a normalized F_0 value, X_i is a value of F_0 contour, μ is the mean F_0 for a given speaker and σ is the standard deviation of F_0 contour (see figure 2, d). The energy contour was also normalized using *z-score* normalization (see equation 1) (see figure 2, e).

Tone sandhi is the change of syllable tones when forming words. Therefore, according to the rule of tone sandhi in Mandarin, tones 3-3 were manually changed to 2-3 and tones 3-3-3 were labeled as 3-2-3. About 1.90% and 0.07% syllables were changed from tone 3 to tone 2 in the *CN_Sent* and *DE_Sent*, respectively. There were some cases which contain more than three consecutive tones 3 in the *CN_Sent*. Therefore, it was corrected manually by tone perception from native speaker of Mandarin. Syllables with a duration of less than 40 msec or more than 600 msec were removed from the data set. It was considered to be erroneous cases in forced alignment [4].

3.2 Tone Recognition Based on HMMs

In Mandarin syllable, the initial is a consonant and the final is a vowel. Hence, features were only extracted from the final segments for tone recognition of single syllable. The F_0 features were usually used in tone recognition. In order to investigate the quality of different kinds of features, many feature vector sets were used:

- A: Pitch-based features (F_0, ΔF_0, $\Delta\Delta F_0$)
- B: Pitch- and energy-based features (F_0, RMS, ΔF_0, ΔRMS, $\Delta\Delta F_0$, $\Delta\Delta$RMS)
- C: Pitch-, energy-based and DoV features (F_0, RMS, DoV, ΔF_0, ΔRMS, $\Delta\Delta F_0$, $\Delta\Delta$RMS)
- D: MFCC-based features
- E: MFCC-, pitch-, energy-based and DoV features (MFCC, F_0, RMS, DoV, ΔF_0, ΔRMS, $\Delta\Delta F_0$, $\Delta\Delta$RMS)

Continuous density HMMs was employed for tone modeling (one model for each tone). Three sets of acoustic tone models were used: monotone for isolated monosyllabic words, context-dependent bitone for isolated disyllabic words and context-dependent tritone for sentences. The tone models consist of three valid states for monotone, bitone and tritone models. 64 mixtures were used for cases A, B and C and 512 mixtures for cases D and E. The data *CN_Mono*, *CN_Bi* and *CN_Sent* were used for training monotone, bitone and tritone models, respectively (Every database was divided into training data (90%) and test data (10%).). The Baum-Welch training algorithm (using the HTK tool "HERest") was used to train the HMMs. Since there will be insufficient data associated with many of the states, similar acoustic states within bitone or tritone sets were tied to ensure that all state distributions can be robustly estimated. The number of Gaussian components in each mixture was increased iteratively in the training algorithm. Six to twenty iterations gave the best results for cases A to E.

3.3 Adaptation of Acoustic Models for L2 Speakers

The three tone models were trained using data of native speakers of Mandarin. Non-native database is required to optimize the recognition tasks to adapte parameters of acoustic models

for non-native speech signals [14]. A Maximum Likelihood Linear Regression (MLLR) was implemented for adaptation of tone models using correct data of *DE_Mono*, *DE_Bi* and *DE_Sent* for monotone, bitone and tritone models, respectively. The correct pronunciation were chosen by comparison of original text and annotations of an expert (German teacher of Mandarin). The expert listened to the data several times and wrote down what she had perceived using Pinyin. If all tones in isolated mono-, disyllabic word or sentence were falsly pronunced, it were removed from the adaptation experiment. The speech data which contains correct pronunciation of initial, final and tone in isolated mono- and disyllabic words was chosen as adaptation data. When the correctness of tones in a sentence is more than 90%, the sentence was used in the adaptation data. The remaining data was chosen as test data for adaptation.

4 Experimental Results

The correctness of the three tone recognition systems using different feature sets (cases A to E) is shown in table 1. The table shows that the adding of energy and DoV features to the F_0 features improved the tone recognition results. It shows also that the combination of MFCC and prosodic features (case E) improved the tone correctness in comparison to the individual features, especially in tone recognition of sentences. There were no large difference between results based on feature extraction from the smoothing of F_0 contour or HFC for monotone and bitone recognizers. In general, the tone correctness using smoothed F_0 contour has a small improvement comparable to the HFC for monotone and bitone recognizers. But the tone correctness for the tritone recognizer using the HFC is better than using the smoothed F_0 contour (the results using MFCC features (case D) are the same for the smoothed F_0 contour and HFC). This indicated that the removing of phrase components in isolated monosyllabic and disyllabic words does not play a role for tone recognition, since the phrase command has very small amplitude or does not exist in the isolated words. But the removing of phrase components in sentences improved the tone recognition results obviously. With the developed methods of case E and using HFC, correct tone recognition rate of 99.50%, 98.86% and 77.03% was achieved for monosyllables, disyllables and sentences, respectively (the results of bitone recognizer are for the four basic tones only, since the data *CN_Bi* did not contain neutral tone). In the literature, it was found that tone correctness of 98.50% was obtained for monosyllables [15]. The tone correctness for disyllables of 98.16% for four basic tones [5] and 94.50% for the five Mandarin tones [15] was obtained. The tone correctness of tritone recognizer is 82.55% using latent prosody model in continuous speech [3] and 85.07% using extended segments in poems [4]. This indicated that the proposed monotone and bitone recognizers outperform existing state-of-the-art algorithms. But the developed tritone recognizer still need optimization to improve the performance of tone recognition.

Table 2 shows the results of tone correctness for original and adapted models of monotone, bitone and tritone recognition systems by using spectral and prosodic feature (case E). The tone recognition results of adapted tone models are better than original models by using feature sets based on the smoothed F_0 contour or HFC. Therefore, the adptated tone models can be integrate in the *CALL-Mandarin system*. The tone correctness of monosyllabic and disyllabic words is better than in sentences before and after adaptation. This indicated that the acquisition of tonal patterns of polysyllabic words is much more difficult than of isolated words [1].

5 Conclusion

The paper presented the developing of three HMM-based Mandarin tone recognition systems (monotone, bitone and tritone recognizer using isolated monosyllabic and disyllabic words and

	Smoothed F_0			HFC		
	CN_Mono	*CN_Bi*	*CN_Sent*	*CN_Mono*	*CN_Bi*	*CN_Sent*
A	82.76	95.49	60.34	81.53	94.69	63.79
B	97.34	97.91	64.20	96.28	96.90	66.68
C	98.59	98.10	66.29	97.39	97.31	67.37
D	97.36	93.90	58.68	97.36	93.90	58.68
E	99.42	99.04	75.78	99.50	98.86	77.03

Table 1 - Correctness of monotone, bitone and tritone recognition systems using different kinds of features by normalization of smoothed F_0 contour as well as HFC for native speakers of Mandarin.

	Smoothed F_0			HFC		
	DE_Mono	*DE_Bi*	*DE_Sent*	*DE_Mono*	*DE_Bi*	*DE_Sent*
Original Models	61.54	49.38	39.26	61.54	51.54	40.10
Adapted Models	65.38	51.23	42.43	69.23	53.09	42.19

Table 2 - Correctness of monotone, bitone and tritone recognition systems using MFCC and prosodic features (case E) with original and adapted tone models for German learners of Mandarin.

sentences, respectively) during the development of the CALL system for German learners of Mandarin. Different kinds of features, including prosodic and spectral-based features, were tested. The F_0 contour was decomposed according to the Fujisaki model to phrase components and tone components. The tone models were adapted using correct data from the German lerners of Mandarin to test its performance before the integration in the *CALL-Mandarin system*. The combination of prosodic and spectral-based features yielded better results than individual features. Tone correctness in sentences using features based on HFC were better than results based on the smoothed F_0 contour. The results indicated that the proposed monotone and bitone recognizers outperform existing state-of-the-art algorithms for isolated monosyllabic and disyllabic words. The tone correctness after adaptation of tone models was better than original models for data from German students. In the future, we try to optimize the tritone recognizer by using features from neighbouring syllables to compensate the coarticulation effect. Thereafter, we want to integrate tone recognition systems in the *CALL-Mandarin system*.

6 Acknowledgements

This work is funded by the German Ministry of Education and Research grant 1746X08 and supported by DAAD-NSC (Germany/Taiwan) and DAAD-CSC (Germany/China) project related travel grants for 2009/2010.

We thank our colleagues Qianyong Gao, Si Wei and Guoping Hu from iFLYTEK company for providing the speech data and segmentations of *CN_Mono* and *CN_Bi* as well as segmentations of German data *DE*. Many thanks to Prof. Wang from National Chiao Tung University (NCTU), Taiwan for providing the phone boundaries for the data *CN_Sent*.

References

[1] H. Mixdorff, D. Külls, H. Hussein, G. Shu, H. Guoping, and W. Si. Towards a computer-aided pronunciation training system for german learners of mandarin. In *Proc. of the*

Workshop (SLaTE), Wroxall Abbey, Warwickshire, England, September 2009.

[2] Y. Qian, T. Lee, and F. K. Soong. Tone recognition in continuous cantonese speech using supratone models. *Journal of the Acoustic Society of America*, 121(5):2936–2945, May 2007.

[3] C.-Y. Chiang, X.-D. Wang, Y.-F. Liao, Y.-R. Wang, S.-H. Chen, and K. Hirose. Latent prosody model of continuous mandarin speech. In *Proc. of the ICASSP*, volume 4, pages IV–625 – IV–628, Honolulu, Hawaii, USA, April 2007.

[4] J.-C. Chen and J.-S. R. Jang. Trues: Tone recognition using extended segments. *ACM Transactions on Asian Language Information Processing*, 7(3), August 2008.

[5] J.-S. Zhang and K. Hirose. A robust tone recognition method of chinese based on sub-syllabic f0 contours. In *Proc. of ICSLP*, Sydney, Australia, November 1998.

[6] J.-S. Zhang and H. Kawanami. Modeling carryover and anticipation effects for chinese tone recognition. In *Proc. of Eurospeech*, pages 747–750, Budapest, Hungary, 1999.

[7] H. Hussein, H. Mixdorff, H. S. Do, M. Mateljan, Q. Gao, G. Hu, S. Wei, and Z. Chao. Comparison of fujisaki-model parameters between german learners and native speakers of mandarin. In *Proc. of the ESSV*, pages 146–153, Aachen, Germany, 2011.

[8] H. Fujisaki and K. Hirose. Analysis of voice fundamental frequency contours for declarative sentences of japanese. *Journal of the Acoustical Society of Japan (E)*, 5(4):233–242, 1984.

[9] H. Hussein, H. Mixdorff, H. S. Do, and R. Hoffmann. Quantitative analysis of tone coarticulation in mandarin. In *Proc. of Interspeech*, Florence, Italy, August 2011.

[10] Tcc-300edu. `http://www.aclclp.org.tw/use_mat.php#tcc300edu`, Last Accessed: April 24, 2012.

[11] Y.-R. Wang. A two-stage sample-based phone boundary detector using segmental similarity features. In *Proc. of the Interspeech*, pages 413–416, Florence, Italy, August 2011.

[12] D. Talkin. *Speech Coding and Synthesis*, chapter A Robust Algorithm for Pitch Tracking (RAPT), pages 495–518. Elsevier Science, New York, USA, 1995.

[13] H. Mixdorff. A novel approach to the fully automatic extraction of fujisaki model parameters. In *Proc. of the ICASSP*, volume 3, pages 1281–1284, Istanbul, Turkey, June 2000.

[14] E. Atwell, P. Howarth, C. Souter, P. Baldo, R. Bisiani, D. Pezzotta, P. Bonaventura, W. Menzel, D. Herron, R. Morton, and J. Schmidt. User-guided system development in interactive spoken language education. *Natural Language Engineering Journal*, 6(3-4):229–241, February 2000.

[15] X. Hu and K. Hirose. Recognition of chinese tones in monosyllabic and disyllabic speech using hmm. In *Proc. of ICSLP*, pages 203–206, Yokohama, Japan, September 1994.