# Real-Time Tone Recognition in A Computer-Assisted Language Learning System for German Learners of Mandarin

*Hussein HUSSEIN*[1] *Hansjörg MIXDORFF*[2] *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
*hussein.hussein@mailbox.tu-dresden.de, mixdorff@beuth-hochschule.de*

ABSTRACT
This paper presents an evaluation of tone recognition systems integrated in a computer-assisted pronunciation training system for German learners of Mandarin. Both the reference tone recognition system as well as a recently redesigned tone recognition system contain monotone, bitone and tritone recognizers for isolated monosyllabic and disyllabic words and sentences, respectively. The performance of the reference system and the redesigned tone recognition systems was compared on data from German learners of Mandarin, while varying the feature vector to contain spectral as well as prosodic features. The redesigned tone recognition system matched or outperformed the reference system. For monosyllabic and disyllabic words it improved when spectral features were added to prosodic features. In contrast, results of tone recognition in sentences yielded better results based on prosodic features only.

KEYWORDS: Mandarin Chinese, Tone Recognition, Computer-Assisted Language Learning.

*Proceedings of the Workshop on Speech and Language Processing Tools in Education*, pages 37–42,
COLING 2012, Mumbai, December 2012.

## 1 Introduction

It is commonly known that Mandarin or standard Chinese is a tone language and hence tonal contours of syllables change the meaning. There are 22 consonant initials (including glottal stop) and 39 vowel finals. Mandarin comprises a relatively small number of syllables. The most important acoustic correlate of tones is $F_0$. Mandarin has four syllabic tones, that is, high, rising, low, and falling (Tones 1-4) and a neutral tone (Tone 0) in unstressed syllables. In citation forms of monosyllabic words the tonal patterns are very distinct as shown in figure 1, but when several syllables are connected, $F_0$ contours observed vary considerably due to tonal coarticulation. German is a stress-timed non-tonal language. Mandarin differs from German on the segmental level, but it is the tonal distinctions that pose serious problems to German learners, especially in the context of two or more syllables. Therefore tone display, recognition and correction are paramount features for a pronunciation training system. In the current paper we present results from a redesigned tone recognition system intended to bring improvement over the reference approach employed in the computer-assisted pronunciation teaching (CAPT) system so far.

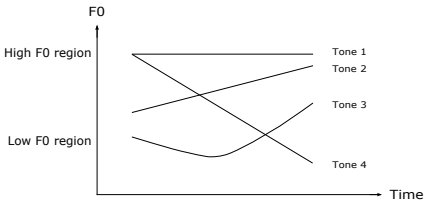


Figure 1: Typical $F_0$ patterns of Tones 1-4 in mono-syllables.

Robust feature extraction and tone modeling techniques are required for reliable tone recognition algorithms. Accuracy of tone recognition obtained on isolated words is typically high, but deteriorates on continuous speech. This implies that hitherto most speech recognition systems for tone languages only rely on spectral features, because they can be estimated more reliably than prosodic features. Many statistical methods for Mandarin tone recognition have been proposed, including Hidden Markov Models (HMM), Neural Networks (NN), Decision-tree classification, Support Vector Machines (SVM) and rule-based methods (Chen and Jang, 2008)(Liao et al., 2010). Most tone recognition algorithms use $F_0$ contours as basic features. The accuracy of tone recognition for Tones 1-4 is usually high, but much lower for the neutral tone, because $F_0$ features are not effective to discriminate the neutral tone. Energy, however, has been found to be an effective cue for tone perception when $F_0$ is missing.

Taking into account tonal coarticulation in the context of the Computer-Assisted Language Learning (CALL) system for German learners of Mandarin ("*CALL-Mandarin system*"), tone recognition systems consisting of monotone, bitone and tritone recognizers were integrated. Whereas a monotone model operates on isolated syllables, a bitone model takes into consideration the tone of the left neighboring syllable and a tritone model depends on the tones of both the left and right neighboring syllables. The reference tone recognition system of our project partner (iFlyTek company, Hefei, China) and the redesigned system consist of monotone and bitone recognizers for isolated monosyllabic and disyllabic words as well as a tritone-based continuous speech recognizer for sentences. They were evaluated on data from German leaners 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 ("*CN_Mono*", "*CN_Bi*" and "*CN_Sent*").

1. ***CN_Mono* - Isolated Monosyllabic Words**:
   The monotone recognizer was trained on isolated monosyllabic words. The monosyllables 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 with isolated disyllabic words. The disyllabic words were produced by the same native speakers as in *CN_Mono* with a total of 75000 disyllables (28.83 hours).
3. ***CN_Sent* - Sentences**:
   The tritone recognizer was trained on sentence data. The sentences were produced by 200 native speakers of Mandarin with 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 a paragraph is about 115 syllables.

### 2.2 German Data - L2

Three speech databases from German learners of Mandarin ("*DE_Mono*", "*DE_Bi*" and "*DE_Sent*") were used for the evaluation of tone recognizers by German learners of Mandarin in real-time. The amount of these data is rather small, but they include all available data which are not used in the adaptation process.

1. ***DE_Mono* - Isolated Monosyllabic Words**:
   *DE_Mono* consists of eight monosyllabic words produced by 5 German learners with a total of 40 utterances.
2. ***DE_Bi* - Isolated Disyllabic Words**:
   *DE_Bi* consists of 10 disyllabic words produced by 12 German students yielding a total of 120 utterances.
3. ***DE_Sent* - Sentences**:
   *DE_Sent* consists of 10 sentences produced by 12 German students with a total of 120 utterances.

## 3 Tone Recognition

In order for tone recognition to take place the utterance must be segmented on the syllable and phone levels. This task is performed by the phone recognizer of iFlyTek for both the reference and redesigned tone recognition system.

### 3.1 Reference Recognizer

The tone recognition system of iFlyTek is part of an automated proficiency test of Mandarin. (Wang et al., 2007). $F_0$ contours are calculated with the Praat default algorithm

(Boersma and Weenink, 2001). Tone models consist of four emitting states for monotone, bitone and tritone models. HMMs were employed for tone modeling. The training data, which is different from the data described in section 2.1, consists of utterances from native speakers of Mandarin (164 female and 105 male speakers, 30 minutes for each speaker).

## 3.2 Redesigned Recognizer

Different kinds of features, including spectral and prosodic-based features, were used. $F_0$ contours were calculated via the Robust Algorithm for Pitch Tracking (RAPT) (Talkin, 1995). RAPT was modified and integrated in the *CALL-Mandarin system* to work in real-time. The output of RAPT contains in addition to $F_0$ values, energy (RMS) and voicing (DoV) measures. Since $F_0$ contours are often affected by extraction errors and micro-prosody, and are interrupted for unvoiced sounds, the raw $F_0$ data is often pre-processed by applying interpolation and smoothing. In our case we applied a cubic spline interpolation and smoothing and filtered the resulting contour at a stop-frequency of 0.5 Hz yielding a high frequency component (HFC) and a low frequency component (LFC) as in (Mixdorff, 2000). Based on the fact that phrase components should be taken into account when analyzing and synthesizing $F_0$ contours of Mandarin, it was found that tone recognition results based on HFCs are better than results based on smoothed $F_0$ contours (Hussein et al., 2012). For the subsequent processing we only used the HFC, thus disregarding low frequency phrase level influences. High frequency contours and energy contours were normalized applying *z-score* normalization. The spectral features, 13 Mel-Frequency Cepstral Coefficients (MFCCs) and their deltas and delta-deltas were also used for tone recognition. All features were only extracted from the final segments. We compared the performance of several feature vectors:

- A: $F_0$-based features.
- B: $F_0$- and energy-based features.
- C: $F_0$-, energy-based and voicing features. These features refer to prosodic features.
- D: MFCC-based features.
- E: MFCC-, $F_0$-, energy-based and voicing features.

HMMs were employed for tone modeling. The tone models consist of three emitting 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 the training of the monotone, bitone and tritone models, respectively. Every database was divided into training data (90%) and test data (10%). 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 were robustly estimated. The number of Gaussian components in each mixture was increased iteratively during training. Six iterations gave the best results for cases A, B and C. 16 and 20 iterations gave the best results for cases D and E, respectively. Tone models were adapted by using German students' data labeled as correct by Chinese native listeners Maximum Likelihood Linear Regression (MLLR) was implemented for the adaptation of tone models.

## 3.3 Evaluation of Mandarin Tone Recognition

Two experiments were performed. First, we tested the redesigned recognition system on data *CN_Mono*, *CN_Bi* and *CN_Sent* from native speakers of Mandarin and compared the perfor-

mance on feature vectors A-E. This test was run outside the CAPT system. Second, we compared the reference system with two versions of the redesigned system after integrating them into the CAPT system, on data *DE_Mono*, *DE_Bi* and *DE_Sent* from German learners of Mandarin:

- R: Tone recognizer by iFlyTek (reference).
- C′: Tone recognizer using prosodic features (case C) and adapted tone models.
- E′: Tone recognizer using both spectral and prosodic features (case E) and adapted tone models.

## 4 Experimental Results

The correctness of the three tone recognizers trained on feature sets A to E is displayed in table 1. The table shows that adding energy and voicing features to $F_0$ features (case C) improved the tone recognition results. The combination of spectral and prosodic features (case E) improved the tone correctness in comparison to individual features, especially in tone recognition of sentences. Tone correctness for monotone, bitone and tritone recognizers is 99.50%, 98.86% and 77.03% using both spectral and prosodic features for monosyllables, disyllables and sentences, respectively. The result of the bitone recognizer only concerns the recognition of Tones 1-4, since the data *CN_Bi* did not contain neutral tone.

| Feature | *CN_Mono* | *CN_Bi* | *CN_Sent* |
|---|---|---|---|
| A | 81.53 | 94.69 | 63.79 |
| B | 96.28 | 96.90 | 66.68 |
| C | 97.39 | 97.31 | 67.37 |
| D | 97.36 | 93.90 | 58.68 |
| E | 99.50 | 98.86 | 77.03 |

Table 1: Tone correctness of monotone, bitone and tritone recognizers by using different kinds of features and normalized HFCs on data by native speakers of Mandarin (in %).

Figure 2 presents the tone evaluation results of monotone, bitone and tritone recognizers for the reference tone recognizer (R) and the redesigned tone recognizers when applying prosodic features (C′) and combined spectral and prosodic features (E′). The monotone, bitone and tritone recognizers were tested in the *CALL-Mandarin system* by using the data *DE_Mono*, *DE_Bi* and *DE_Sent*, respectively. The tone correctness of monosyllables in R and E′ is the same. The tone recognition in monosyllabic words using both spectral and prosodic features improved the results significantly in comparison to prosodic features only. On disyllabic words, the bitone recognizer based on the combination of spectral and prosodic features outperformed the other presented algorithm. In contrast, for sentence recognition, the tritone recognizer based only on prosodic features outperformed the other algorithms. This result suggests that the variation in the MFCCs which is mostly due to segmental and not tonal variations affects the tritone models more than the monotone and bitone models.

## Conclusions

This study compared redesigned monotone, bitone and tritone HMM-based Mandarin tone recognizers for our CALL system for German learners of Mandarin with a pre-exisiting reference. During development different spectral and prosodic features were tested. Of the $F_0$

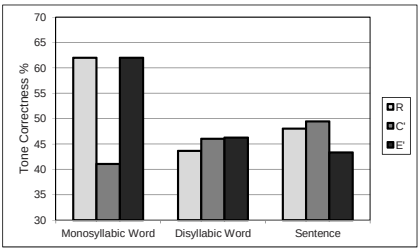


Figure 2: Tone correctness of monotone, bitone and tritone recognizers for reference (R) and redesigned tone recognizers (C′ and E′) on data by German learners of Mandarin in *CALL-Mandarin system*.

contour we only employed the HFC, hence suppressing phrasal contributions. The tone models were adapted by using correct data from German learners of Mandarin. Tone recognition of mono- and disyllabic words using both spectral and prosodic features yielded the best results. In contrast, for sentence recognition the tritone recognizer based on only prosodic features performed best. Overall, the redesigned tone recognition system matched or surpassed the performance of the reference system and will therefore henceforth be employed in the CALL system. Including the new features slightly increases the computation time of the system which, however, as informal tests have shown, is still short enough to provide online feedback.

## Acknowledgements

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

## References

Boersma, P. and Weenink, D. (2001). Praat: doing phonetics by computer.

Chen, J.-C. and Jang, J.-S. R. (2008). TRUES: Tone Recognition Using Extended Segments. *ACM Transactions on Asian Language Information Processing*, 7(3).

Hussein, H., Mixdorff, H., Liao, Y.-F., and Hoffmann, R. (2012). HMM-Based Mandarin Tone Recognition - Application in Computer-Aided Language Learning System for Mandarin. In *Proc. of ESSV*, pages 347–354, Cottbus, Germany. TUDpress.

Liao, H.-C., Chen, J.-C., Chang, S.-C., Guan, Y.-H., and Lee, C.-H. (2010). Decision Tree Based Tone Modeling with Corrective Feedbacks for Automatic Mandarin Tone Assessment. In *Proc. of Interspeech*, pages 602–605, Makuhari, Chiba, Japan.

Mixdorff, H. (2000). A Novel Approach to the Fully Automatic Extraction of Fujisaki Model Parameters. In *Proc. of ICASSP*, volume 3, pages 1281–1284, Istanbul, Turkey.

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

Wang, R.-H., Liu, Q., and Wei, S. (2007). *Advances in Chinese Spoken Language Processing*, chapter Putonghua Proficiency Test and Evaluation, pages 407–429.