COMPARISON OF FUJISAKI-MODEL PARAMETERS BETWEEN GERMAN LEARNERS AND NATIVE SPEAKERS OF MANDARIN

Hussein Hussein[1, 2], *Hansjörg Mixdorff*[1], *Hue San Do*[1], *Marijo Mateljan*[1],
Qianyong Gao[3], *Guoping Hu*[3], *Si Wei*[3] *and Zhao Chao*[3]

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

[2] *Laboratory of Acoustics and Speech Communication, Dresden University of Technology,*
Dresden, Germany

[3] *Department of EEIS, University of Science and Technology of China,*
Hefei, Anhui, P.R. China

{hussein, mixdorff, hsdo, mateljan}@beuth-hochschule.de,
{qygao, gphu, siwei, chaozhao}@iflytek.com

Abstract: The current study reports on the continued activities to develop a computer-aided phonetic learning system for German learners of Mandarin. Two corpora were used in the current study. The first one consists of recordings from German learners of Mandarin. It was used to adapt the Hidden Markov Models (HMM) of phone and tone recognizers. The second corpus consists of recordings from German students and native speakers of Mandarin. The phone and syllable labels as well as the *F0* contour and Fujisaki-model parameters were manually corrected for the second corpus. The differences in the amplitude and duration of tone commands of the Fujisaki-model at the beginning and at the end of sentences between German and Chinese data for the four Mandarin tones were compared. The correctness of the phone and tone recognition systems with adapted models is better than with the original models. The duration of tone commands at the end of sentences is greater than at the beginning of sentences for both German and Chinese data and for all tones. The amplitude of tone commands of German data is greater than in the Chinese data. It was also found that for most utterances phrase commands of magnitude greater zero occurred, indicating that the phrase component should be taken into account when analyzing and synthesizing *F0* contour of Mandarin.

1 Introduction

The growing interest in speaking a foreign language in a globalized world stimulates activities towards computer-aided language learning (CALL). CALL is a tool to facilitate individualized language learning and pronunciation training, see, for example [1]. The pronunciation training might be the most difficult to be transferred to a computer because providing useful and robust feedback on learner errors is far from being a solved problem [2]. In this paper we report on the on-going development of a Mandarin training system for German learners within a three-year project funded by the German Federal Ministry of Education and Research [2][3][4][5][6][7].

Mandarin comprises a relatively small number of about 400 different syllables which are formed by combining 22 consonant initials (including glottal stop) and 38 mostly vocalic finals. It is commonly known that Mandarin is a tone language and hence the tonal contour of a syllable changes its meaning [8]. The most important acoustic correlate of tone is *F0*. 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 table 1), but when several syllables are connected, *F0* contours observed vary considerably due to tonal coarticulation. We observed that the acquisition of tonal patterns of poly-syllabic words is much more difficult than of mono-syllabic words [2]. German is a non-tone 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.

The well-known Fujisaki model is a parsimonious method for parameterizing *F0* contours in speech synthesis for intonation analysis and intonation generation [9]. The model reproduces a given *F0* contour by superimposing three components: a speaker-individual base frequency Fb, a phrase component and an accent component in stress-timed languages or a tone component of positive and negative polarities in tone languages. As previous studies showed Mandarin tone can be represented by prototypical *F0* contours [10] and requires tone commands of positive and negative polarity (see table 1). The phrase component results from responses to impulse-wise phrase commands associated with prosodic breaks. Phrase commands are described by their onset time *T0*, magnitude *Ap* and time constant α (time constant of the phrase control mechanism). The tone component results from step-wise tone commands associated with syllable-tones. In [10] tone commands are described by on- and offset times *T1* and *T2*, amplitude *At*, time constant β_p of the tone control mechanism for positive commands, time constant β_n of the tone control mechanism for negative commands and constants γ_p (relative ceiling level of positive tone components) and γ_n (relative ceiling level of negative tone components). The use of a common value for both constants β and γ (20/s and 0.9, respectively) irrespective of the polarity of tone command is acceptable [11]. Each syllable is associated with two independent tone commands, each command is jointed with its respective *T1*, *T2*, *At*, as well as β and γ. However, previous experiments suggested that tone 1 and tone 3 are typically associated with a single tone command of positive and negative polarity, respectively. Furthermore, depending on the underlying tones, commands of the same polarity have been observed to be stretching across at least two syllables, hence the syllabic tone commands are actually concatenated into longer tonal gestures [12].

Table 1 – Mandarin tones with prototypical tone command assignment [11].

Tone	*F0* contour	Tone commands assigned
1	high	positive
2	rising	negative/positive
3	falling-rising	negative
4	falling	positive/negative

The current paper presents the framework and components of the CALL-Mandarin system which was developed for teaching Mandarin for German learners. Tow corpora were used in the current paper. The Hidden Markov Models (HMM) of phone and tone recognition systems were adapted. We intended to study the differences in the amplitude and duration of tone commands of the Fujisaki-model at the beginning and at the end of sentences between German learners and native speakers of Mandarin for the four Mandarin tones. In the context of the Fujisaki model, we expect the absence of declination to result in phrase command magnitudes close to zero. Therefore, we wish to examine whether or not declination in Mandarin needs to be taken into account in *F0* contour synthesis.

2 Framework of CALL-Mandarin System

The CALL system can be divided into two levels [3]:

1. The automatic speech recognition system containing the individual components of a recognizer for syllable initials, finals and tones.

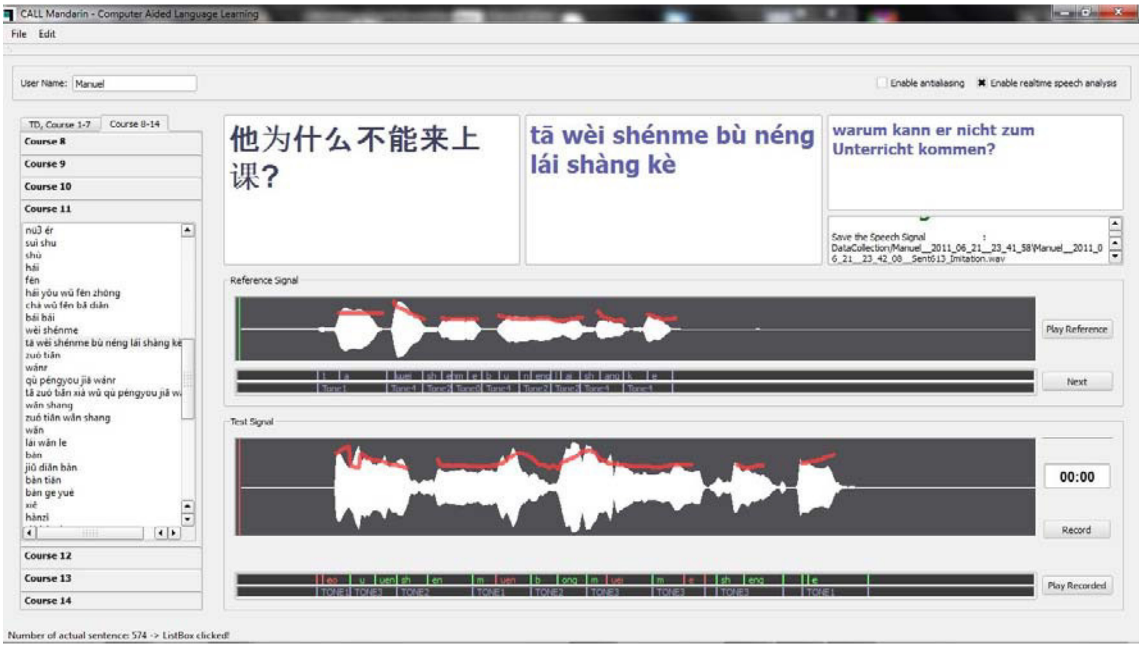
2. The training application manages the interaction with the user through a graphical user interface (GUI) using acoustic and visual feedback.

2.1 Phone and Tone Recognizers

The phone and tone recognizers of *iFlyTek* company, Hefei, China are used in current the CALL-Mandarin system. The acoustic features employed by the phone recognition system are mel-frequency cepstral coefficients (MFCCs), delta and delta-delta coefficients, yielding a total of 39 features for every frame. The segmental phonetic models consist of three valid states for consonant initials and five valid states for vocalic finals. The tone recognition system is based on the *F0*, delta and delta-delta features which were calculated for finals determined using the phone-based forced-alignment. The *F0* contour was calculated using the *Praat* [13] algorithm. The tone acoustic models consist of four valid states. Three sets of acoustic tone models were used: mono-tone for one-syllable words, bi-tone for two-syllable words and tri-tone for sentences. The Baum-Welch algorithm (using the HTK tool „HERest“ [14]) was used to train the HMM. The training data consists of speech signals from native speakers of Mandarin (164 female and 105 male speakers with a total time of 30 minutes for each speaker).

2.2 Application and Functionality of System

The figure 1 shows the GUI of the computer-aided pronunciation training system for German learners of Mandarin. It provides the user with a list of 15 lessons. The text is displayed in Chinese characters as well as in Pinyin transcription where the tones are shown by diacritical symbols and translation to German. The framework contains a set of reference which comprises utterances produced by native speakers of Mandarin. The *F0* contour (red contour) and the energy envelope curve of the audio signal are shown for the reference and the imitation signals. The speech signals are analyzed using an automatic speech recognition (ASR) system and an automatic tone recognition system (see Section 2.1). The results of phone and tone recognition are displayed for both signals.



CALL Mandarin - Computer Aided Language Learning

User Name: Manuel

□ Enable antialiasing ☒ Enable realtime speech analysis

他为什么不能来上课?

tā wèi shénme bù néng lái shàng kè

warum kann er nicht zum Unterricht kommen?

Save the Speech Signal :
DataCollection/Manuel__2011_06_21__23_41_58\Manuel__2011_06_21__23_42_08__Sent613_Imitation.wav

Course 11

nü3 ér
sui shu
shù
hǎi
fēn
hǎi yǒu wǔ fēn zhōng
chà wǔ fēn bā diǎn
bāi bāi
wèi shénme
tā wèi shénme bù néng lái shàng kè
zuó tiān
wánr
qù péngyou jiā wánr
tā zuó tiān xià wǔ qù péngyou jiā w
wǎn shang
zuó tiān wǎn shang
wǎn
lái wǎn le
bàn
jiǔ diǎn bàn
bàn tiān
bàn ge yuè
xiě
hànzi

00:00

Number of actual sentence: 574 -> ListBox clicked!

Figure 1 – Graphical User Interface of the CALL-Mandarin system.

The task of the user is to imitate the reference signal. Therefore, the learner can at first listen to the reference audio signal. Then s/he can start the imitation of the reference signal after clicking on the "Record" button in a training session. The software records the user's voice and shows the obtained data (*F0* contour and energy envelope) in real time as a visual feedback. The position of the imitation signal is adjusted in relation to the start position of the reference signal (after 0.5 sec from the beginning) in order to compare the reference and imitation signals. The training session must be started only one time. Then it will be continued indefinitely in order to facilitate the imitation of the reference signal. The recording is terminated by clicking on the "Stop" button, or it ends automatically when the user for a certain period of time (five times the length of the reference signal) does not start a new imitation attempt. The repetition of the imitation is controlled by a threshold of the energy contour. A new imitation attempt starts when the length of pause behind an imitated speech signal is more than 1.2 sec. The results of phone and tone recognition are shown below the imitation signal. The alignment of the imitation signal to the reference signal and the repetition process of imitation of the reference signal are intended to help the learner improve the quality of his/her pronunciation.

3 Experiment Method

3.1 Corpus Design and Data Collection

The data used in this study consists of recordings from German students of Chinese Studies at the East Asian Seminar of Free University Berlin (*FUB*) and Chinese native speakers. The data was recorded with a sampling frequency of 16 kHz and a resolution of 16 bit. In addition to the regular classes of Mandarin language training (eight hours per week), some German students had attended a weekly seminar (henceforth "*WS*") of two hours as additional training and used a computer-aided phonetic pronunciation training system at home (see Section 2). The rest of testing group (henceforth "*WOS*") had not. About two-third of the seminar was dedicated to phonetic, one-third to grammar and conversation exercises. Two parts of data were used in the current study:

1. The first part of data consists of only German data (henceforth "*DE1*") which is the same corpus used in the previous experiments [2][3][4][5][6]. The corpus consisted of 54 tokens. One half of these had been produced by a female native speaker and was imitated by the subjects (imitation mode). The other half was provided in Pinyin transcription and read aloud (reading mode). Each part contained eight mono-syllabic and 19 di-syllabic words. In addition to the 54 word tokens, five short sentences were recorded. The corpus was produced by 19 first-year students (eight male and 11 female), yielding 304 one-syllable words, 722 two-syllable words, 95 sentences. At the time of the recording they had completed 12 weeks of Mandarin language training.

2. The second part of data consists of German data (henceforth "*DE2*") and Chines data (henceforth "*CN2*") which were used in the previous experiment [7]. The *DE2* consists of recordings from 13 first year German students (seven seminar students (*WS*) and six non-seminar students (*WOS*)). At the time of the recording the students had completed 12 weeks of Mandarin language training. The *DE2* consists of four parts: perception test and three production tests. Only the second production test (henceforth "*Production 2*") was used in this paper. The students were asked seven questions in Chinese and they read seven sentences as answers aloud. Thereafter, the students asked three questions. The sentences and questions were presented to them on a computer screen both in Chinese characters and Pinyin transcription (reading mode). Each sentence contains both monosyllabic and disyllabic words, with a minimum of three and a maximum of nine syllables. The total number of speech signals of *DE2* used in the current study is 130 utterances. The *CN2* consists of ten sentences

(the same ten sentences as *Production 2*). The data collected from six native speakers of Mandarin from *Tongji* University, Shanghai, China (three females and three males), yielding a total of 60 utterances. The sentences were presented to the Chinese native speakers in Chinese characters and then read aloud without preceding questions (reading mode).

3.2 Adaptation of Acoustic Models of Phone and Tone Recognizers

In order to consider the most frequent pronunciation errors committed by the German learner of Mandarin, the original phone and tone models trained on data from native speakers of Mandarin were adapted. The correct phone and tone data from *DE1* according to the result of forced-alignment and recognition was used in the adaptation of acoustic models. A global maximum likelihood linear regression (MLLR) adaptation was performed first and then a MLLR and maximum a posteriori (MAP) adaptation was implemented in the phone model adaptation. In the tone model adaptation, an MLLR adaptation and MAP adaptation were also implemented.

3.3 Analysis Method

The *F0* contour reflects the tone on the syllable level. Therefore, the data (*DE2* and *CN2*) was forced-aligned on the syllable and phone-levels using the ASR system in a forced alignment mode. The ASR system is a part of an automated proficiency test of Mandarin [15]. The label files from the forced-alignment were converted to *Praat* TextGrid format and combined in a single TextGrid containing syllable and phone labels. The syllabe and phone boundaries were then hand-corrected.

The *F0* contours for German and Chinese data were calculated using the *Praat* algorithm with a step of 10msec and different standard settings of the minimum and maximum parameters of *F0* for male (100 and 350 Hz) and for female speakers (120 and 450 Hz). It was found by checking the *F0* contours that some speech signals do not have *F0* values in the voiced segments. Therefore, the *F0* contours for such signals were recalculated. The modified parameters of the minimum and maximum *F0* were (50 and 250 Hz) for male and (80 and 400 Hz) for female speakers. The *F0* contours were checked and corrected using the *Praat Pitch-Editor*. The Fujisaki parameters were estimated automatically using the algorithm [12] and if necessary corrected using the interactive tool *FujiParaEditor* [16]. Figure 2 shows an example of Fujisaki parameters of the sentence “ta1 xi3 huan0 he1 zhong1 guo2 cha2”-“He likes to drink Chinese tea” uttered by native speaker of Mandarin.

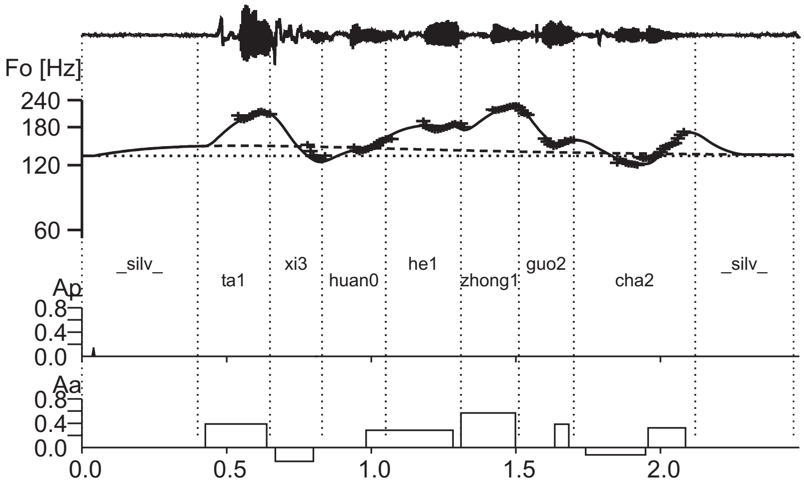


Figure 2 - Speech signal, *F0* contour, phrase commands and tone commands of the utterance: “ta1 xi3 huan0 he1 zhong1 guo2 cha2.” (“He likes to drink Chinese tea.”).

4 Experimental Results

4.1 Correctness of Phone and Tone Recognizers

The correctness of phone and tone recognizers using original and adapted models was calculated. The *DE1* was used as test data. The table 2 shows the evaluation results of phone and tone correctness with original and adapted models. The phone and tone recognition results of adapted acoustic models are better than original models.

Table 2 – Correctness of phone and tone recognizers with original and adapted models

Word Type	Phone Correctness (%)		Tone Correctness (%)	
	Original Models	Adapted Models	Original Models	Adapted Models
1-Syllable Word	73.40	74.00	65.80	73.70
2-Syllable Word	68.00	67.80	63.60	67.80
Sentence	78.50	81.30	50.50	51.70

4.2 Comparison of Tone Commands between *DE2* and *CN2*

Syllables at the beginning and end of an utterance are special cases as they are not preceded or followed by another syllable. We therefore calculated their properties separately.

Table 3 – Mean, standard deviation and number of tone command amplitude and tone duration according to the position of tone in the sentence for *DE2* and *CN2*.

Tone		*DE2*						*CN2*					
		Beginning of sentence			End of sentence			Beginning of sentence			End of sentence		
		At1	*At2*	Duration [sec]	*At1*	*At2*	Duration [sec]	*At1*	*At2*	Duration [sec]	*At1*	*At2*	Duration [sec]
1	Mean	0.25	-	0.18	0.39	-	0.26	0.20	-	0.19	0.19	-	0.24
	SD	0.19	-	0.06	0.22	-	0.11	0.07	-	0.03	0.07	-	0.05
	N	11			25			6			12		
2	Mean	-	-	-	-0.19	0.40	0.24	-	-	-	-0.13	0.22	0.27
	SD	-	-	-	0.08	0.14	0.06	-	-	-	0.05	0.07	0.12
	N	-			29			-			11		
3	Mean	-0.20	-	0.13	-0.19	-	0.15	-0.17	-	0.14	-0.24	-	0.11
	SD	0.16	-	0.06	0.10	-	0.07	0.07	-	0.05	0.16	-	0.05
	N	47			22			41			8		
4	Mean	0.48	-0.26	0.12	0.47	-0.24	0.19	0.21	-0.14	0.16	0.40	-0.16	0.22
	SD	0.26	0.10	0.02	0.30	0.10	0.08	0.09	0.01	0.05	0.19	0.04	0.05
	N	8			8			5			6		

Table 3 shows the mean, standard deviation (SD) of amplitude and duration of tone commands and the number of tones (N) according to the position either at the beginning or the end of the sentence for *DE2* and *CN2*, respectively. The duration of tone commands at the end of sentences is greater than at the beginning of sentences for *DE2* and *CN2* with the exception of tone 3 of *CN2*. The table 3 shows that the amplitude of tone commands between the same tones which are located at the beginning or the end of sentences of *DE2* is greater than of *CN2* with the exception that tone 3 of *CN2* at the end of sentences. The speech rate of German learners of Mandarin is slower than of native speakers of Mandarin [7]. Therefore, the *At* of German learners is larger. The Mann-Whitney U-tests of the absolute values of *At1* and *At2* of tones 1, 2 and 4 for both *DE2* and *CN2* suggest that these differences are highly significant (p < .001 and p < .002 for absolute values of *At1* and *At2*, respectively). Subsequent analysis yields that at least the absolute value of *At1* is correlated with syllabic duration (ρ=.155, p <.001). This suggests that the higher *At* in the German subjects might at least be partly due to their lower speech rate. For paired commands (tone 2 and 4) we find that the absolute values of *At1* and *At2* are negatively correlated (ρ=-.190, p <.001). This is an interesting finding as it indicates a compensatory effect between the tone commands associated with one and the same syllable.

4.3 Comparison of Phrase Commands between *DE2* and *CN2*

Three phrase commands at most were detected in the utterances. The ratio of sentences which have multi-phrase commands is 22.15 % and 15.49% for *DE2* and *CN2*, respectively. The table 4 shows that there are differences in the amplitude of phrase commands between female and male speakers for *DE2* and *CN2*. The *Ap* of male speakers of *DE2* is greater than of female speakers. But the *Ap* of female speakers of *CN2* is greater than of male speakers. The fact that values are greater than zero indicates that the phrasal contour must be taken into account when synthesizing *F0* contours of Mandarin. The correlation between the number of syllables in an utterance and *Ap* was found to be significant (ρ=.197, p <.05) for *DE2*, but not for *CN2*.

Table 4 – Mean, standard deviation and number of phrase command amplitude for *DE2* and *CN2*.

Ap	*DE2*		*CN2*	
	Female	Male	Female	Male
Mean	0.27	0.39	0.34	0.23
SD	0.18	0.25	0.17	0.12
N	92	75	39	32

5 Discussion and Conclusions

The paper presented the CALL-Mandarin system which was developed for teaching Mandarin to German learners. Two corpora were used in the paper. The acoustic models of the phone and tone recognition systems were adapted in order to consider the most frequent pronunciation errors committed by the German learner of Mandarin. The correct phone and tone of German data from the first corpus were used in the adaptation of acoustic models. The recognition results of adapted models were better than of the original models. The phone and syllable labels as well as the *F0* contour and Fujisaki-model parameters were manually corrected for the second corpus. The Fujisaki parameters for German and Chinese data were compared for the four Mandarin tones. The duration of tone commands at the end of sentences is greater than at the beginning of sentences for both German and Chinese data. The amplitude of tone commands of German data is greater than of Chinese data which might be partly due to their lower speech rate. The amplitudes of paired commands are negatively correlated indicating a compensatory effect. The results concerning phrase commands suggest that declination must be taken into account when synthesizing *F0* contours of Mandarin.

6 Acknowledgements

This work is funded by 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.

References

[1] EURONOUNCE: *The EURONONCE Project: Intelligent Language Tutoring System with multimodal feedback functions*. Dresden University of Technology, Dresden, Saxonia, Germany. http://www.euronounce.net/.

[2] Mixdorff, H., Külls, D., Hussein, H., Gong, S., Hu, G. and Wei, S.: *Towards a Computer-Aided Pronunciation Training System for German Learners of Mandarin*. Proceedings of *SLaTE Workshop on Speech and Language Technology in Education*, Wroxall Abbey Estate, Warwickshire, England, 3-5 September 2009.

[3] Mixdorff, H., Külls, D. and Hussein, H.: *Development of a Computer-Aided Language Learning Environment for Mandarin - First Steps*. Proceedings of 20st Conference of *Elektronische Sprachsignalverarbeitung (ESSV)* (Studientexte zur Sprachkommunikation Bd. 53), pp. 354-363, Dresden, Germany, September 2009.

[4] Hussein, H., Wei, S., Mixdorff, H., Külls, D., Gong, S. and Hu, G.: *Development of a Computer-Aided Language Learning System for Mandarin - Tone Recognition and Pronunciation Error Detection*. Proceedings of the *Speech Prosody 2010*, Chicago, Illinois, May 2010.

[5] Hussein, H., Mixdorff, H., Do, H. S., Wei, S., Gao, Q., Gong, S., Ding, H. and Hu, G.: *Development of a Computer-Aided Pronunciation Training System for Teaching Mandarin for German Learners – Pronunciation Errors*. Proceedings of 21st Conference of *Elektronische Sprachsignalverarbeitung (ESSV)* (Studientexte zur Sprachkommunikation Bd. 58), pp. 288-295, Berlin, Germany, September 2010.

[6] Hussein, H., Mixdorff, H., Do, H. S., Wei, S., Gong, S., Ding, H., Gao, Q. and Hu, G.: *Towards a Computer-Aided Pronunciation Training System for German Learners of Mandarin - Prosodic Analysis*. Proceedings of *Workshop on Second Language Studies: Acquisition, Learning, Education and Technology*. Tokyo, Japan. September 2010.

[7] Hussein, H., Do, H. S., Mixdorff, H., Ding, H., Gao, Q, Hue, G, Wei, S. and Chao, Z.: *Mandarin Tone Perception and Production by German Learners*. Proceedings of *SLaTE Workshop on Speech and Language Technology in Education*, Venice, Italy, August 2011 (in press).

[8] Wang, W. S.-Y.: *Phonological Features of Tone*. *International Journal of American Lingustics*, pp. 93-105, Vol. 33, 2, 1967.

[9] Fujisaki, H. and Hirose, K.: *Analysis of Voice Fundamental Frequency Contours for Declarative Sentences of Japanese*. *Journal of the Acoustical Society of Japan* (E), Vol. 5, 4, pp. 233-242, 1984.

[10] Fujisaki, H., Hirose, K., Halle, P., Lei, H.: *Analysis and Modeling of Tonal Features in Polysyllabic Words and Sentences of the Standard Chinese*. Proc. of *ICSLP*, pp. 841-844, 1990.

[11] Fujisaki, H.: *The Roles of Physiology, Physics and Mathematics in Modeling Prosodic Features of Speech*. Proc. of *Speech Prosody 2006*, Dresden, Germany, May 2006.

[12] Mixdorff, H., Fujisaki, H., Chen, G. P. and Hu, Y.: *Towards the Automatic Extraction of Fujisaki Model Parameters for Mandarin*. Proc. of *Eurospeech 2003*, Vol. 2, pp. 873-876, Geneva, Switzerland, 2003.

[13] Boersma, P. and Weenink, D.: *Praat doing Phonetics by Computer*. www.praat.org.

[14] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu,G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P.Woodland: *The HTK Book (Version 3.4).* Cambridge University Press, 2006.

[15] Wang, R. H., Liu, Q. F., and Wei, S.: *Putonghua Proficiency Test and Evaluation*. Advances in Chinese Spoken Language Processing, Chapter 18, Springer press, pp. 407-430, 2006.

[16] Mixdorff, H.: *FujiParaEditor*. http://public.bht-berlin.de/~mixdorff/thesis/fujisaki.html, 26.01.2011.