DEVELOPMENT OF A COMPUTER-AIDED PRONUNCIATION TRAINING SYSTEM FOR TEACHING MANDARIN FOR GERMAN LEARNERS – PRONUNCIATION ERRORS

Hussein Hussein[1], *Hansjörg Mixdorff*[1], *Hue San Do*[1], *Si Wei*[2], *Qianyong Gao*[2], *Shu Gong*[2], *Hongwei Ding*[3] *and Guoping Hu*[2]

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

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

[3] *School of Foreign Languages, Tongji University, Shanghai, China*

{hussein, mixdorff, hsdo}@beuth-hochschule.de,

{siwei, qygao, shugong, gphu}@iflytek.com, hongwei.ding@tongji.edu.cn

Abstract: This paper reports on the continued activities towards the development of a computer-aided language learning (CALL) system for German learners of Mandarin. In this experiment the method for detecting the pronunciation errors which was presented in a previous experiment was tested on two different databases in order to study the effect of complexity of corpus on the results of pronunciation error detection. The first corpus is simple and consists of monosyllabic and disyllabic words and read from German students of Mandarin in the first year of language education. The second corpus is more complex and consists of whole sentences and read from German students from three different years of language education. The data are perceptually evaluated by human judges as well as processed by two Automatic Speech Recognition (ASR) systems. Acoustic model of the first ASR system trained on data of native speakers of Mandarin. The second ASR system used an adapted acoustic model that considers the errors expected from the German learners of Mandarin. The experimental results show that the performance of the modified ASR system is better. The ratings of strength of foreign accent and intelligibility are strongly correlated with the correctness of tones than with the correctness of initials and finals. The ratio of correct initials and finals in the complex corpus is greater than in the simple corpus, but the number of correct tones is lower in the complex corpus.

1 Introduction

A growing demand for foreign language competence stimulates activities towards computer-aided language learning (CALL). CALL is a tool to facilitate the individualized language learning process and can be used for pronunciation training. Therefore, many CALL systems were developed [1][2]. 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 [3]. In the current 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 which started since 2 years ago.

Modern Mandarin (Putonghua) differs from German significantly on the segmental as well as the suprasegmental level and poses a number of problems to the German learner. 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. Some of

phonemes building initials and finals have exact or close counterparts in the German language. Errors usually arise from phonemes of Mandarin without correspondences in German [4].

Mandarin is a tonal language. Tone is very important to distinguish Mandarin syllables, i.e. the tonal contour of a syllable changes its meaning. Mandarin has four syllabic tones and a neutral tone. However, the amount of syllables used in real speech is only about 1200 syllables with different lexical tones. Mandarin tone can be represented by prototypical f0 contours [5] as shown in Figure 1 [6]. Apart from certain affricate initials that do not exist as German phonems the tonal distinction in Mandarin is the most complex feature for German learners to acquire. The acquisition of tonal patterns of poly-syllabic words is much more difficult than mono-syllabic words [3].

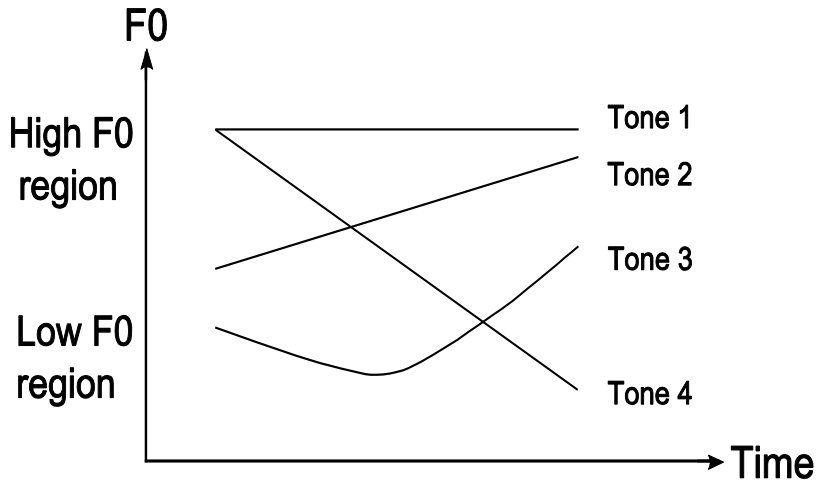

Figure 1 - Prototypical f0 contours of Mandarin tones

In the previous experiments of the system we conducted an analysis of typical errors committed by German students of Mandarin in the first year of language education for monosyllabic and disyllabic words [3][7] and a method for detecting the pronunciation errors is tested by calculating the confidence distance between the first and second candidates of the recognition system [8]. Furthermore, a contrastive analysis of prosodic features (rhythmic and intonational) of the Mandarin tones between native speakers and German learners of Mandarin was performed to identify the differences and similarities [9].

In the current experiment we aimed to study the effect of complexity of corpus on the results of the pronunciation error detection. Therefore, we tested the method for detecting the pronunciation error [8] on two different databases. The first corpus consists of monosyllabic and disyllabic words. The second corpus is more complex than the first one and consists of whole sentences. The data were perceptually evaluated by human judges (teaching expert for Mandarin and two groups of native speakers of Mandarin) as well as processed by two Mandarin Automatic Speech Recognition (ASR) systems. The acoustic model of the first ASR system is trained on data of native speakers of Mandarin. The second ASR system used an adapted acoustic model by taking into account the most frequent pronunciation errors committed by the German learners [3][7]. The annotations produced by the human judges used as a reference to evaluate the results of the ASR systems.

The database and the experiment method are described in section 2. The experimental results are given in section 3. Finally, Section 4 contains the conclusion.

2 Experiment Method

2.1 Corpus Design and Data Collection

The data used in this experiment consist of recordings from German students of Chinese Studies at the East Asia Seminar of Free University Berlin (FUB). The data was recorded with a sampling frequency of 16 kHz and a resolution of 16 bit. In addition to the regular three-hour classes of Mandarin language training, the German students had attended a weekly tutorial of two hours as additional training. About one half of the tutorial was dedicated to phonetic, the other half to grammar and translation exercises. The phonetic exercises comprised discrimination, identification and imitation of mono- and disyllables, contrastive exercises with minimal pairs of differing initials or finals as well as reading from the text book, constantly monitored and corrected by the teacher.

The data collected from German students of Mandarin at FUB consist of two parts:

The first part of German data (henceforth "DE1") is the same corpus used in the previous experiments [3][7][8][9]. 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. The corpus was produced by 19 first-year students (eight male and 11 female). At the time of the recording they had completed 12 weeks of Mandarin language training.

The second part of German data (henceforth "DE2") is the same corpus used in the last experiment [9]. The corpus consisted of 62 sentences, 22 sentences for the first- (henceforth "DE2_Y1") and 20 sentences each for the second- (henceforth "DE2_Y2") and third-year (henceforth "DE2_Y3") German students. The sentences presented to each group were chosen from six different types: declarative sentences, polar questions (yes-no-questions), constituent questions (wh-questions), rhetorical questions, imperative and exclamatory sentences. They contained both monosyllabic and disyllabic words, with a minimum of two and a maximum of 14 syllables. Furthermore, half of the sentences presented to each group were the same for all three groups. The sentences were provided in Chinese character and read aloud (reading mode). 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). At the time of recording they had completed 12 weeks, 36 weeks, and 60 weeks of Mandarin language training, respectively. The second part of German data was recorded after about one year from recording the first part.

2.2 Data Evaluation

The collected data was annotated and processed by different means:
(1) Expert (German teacher of Mandarin): The expert listened to the data several times and wrote down what she had perceived using Pinyin.
(2) Five native speakers of Mandarin, staff of iFlyTek company, Hefei, China (henceforth "*Hefei*") and five native speakers of Mandarin by the School of Foreign Languages, Tongji University, Shanghai, China (henceforth "*Shanghai*"): *Hefei* and *Shanghai* were between 20 and 30 years of age. They were presented with the data only twice. The first time, they were requested to write down what they had perceived using Pinyin without prior knowledge of the intended target. The second time, they were presented with the original data and had to rate strength of foreign accent and intelligibility on a scale from 1 to 5, five being the best score, that is, native-like competence. Henceforth, we refer to both expert and native speakers (listeners) of Mandarin as human judges. The human judges annotate only the data of German students.

(3) An ASR system: The ASR system which is part of an automated proficiency test of Mandarin [10] was used. We used two versions of ASR system as in [8]:
The first ASR system (henceforth "ASR1") used the original acoustic model trained on data from native speakers of Mandarin. This ASR system considers likely and unlikely confusion partners with respect to the German learners because it used the overall replacement list.

The second ASR system (henceforth "ASR2") used an adapted acoustic model. The correct phoneme and tone data from German participants according to the result of forced alignment and recognition was used in the adaptation of the acoustic model. A global maximum likelihood linear regression (MLLR) adaptation was performed first and then an MLLR and maximum a posteriori (MAP) adaptation was implemented in the phoneme model adaptation. In the tone model adaptation, an MLLR adaptation and MAP adaptation were also implemented. Only the most likely confusion partners were used to minimize the search space for the recognizer. Common pronunciation errors of German learners were detected by comparing the given text and the labeling of native speakers of Mandarin from [3]. Therefore, a well-targeted (small) replacement list for every phoneme was used in the second ASR system. If the probability of confusion between two phonemes was more than a threshold of 10% the phoneme was added into the well-targeted replacement list. The two ASR systems used the same tone models. This means that we have the same results on the tone-level.

3 Experimental Results

The annotations produced by the human judges used as a reference to evaluate the results of the ASR systems. The syllables of the original text, annotations of the expert and the native speakers of Mandarin, and the results of the ASR systems were divided into initials, finals and tones to evaluate every syllable component individually.

3.1 Correlation Analysis

We first examined the correlation between the judgments of the two groups of Chinese evaluators (*Hefei* and *Shanghai*) on the DE2. These are .801 for strength of foreign accent and .752 for intelligibility. On average, *Hefei* assigned better ratings for strength of accent which were also less at variance (mean/s.d.: 3.46/.64) than those of *Shanghai* (mean/s.d.: 2.90/.78). The differences were less marked for intelligibility (mean/s.d. *Hefei*: 3.73/.76, *Shanghai*: 3.70/.97). For the following analyses we pooled the scores of the two groups. As can be expected, ratings of strength of foreign accent and intelligibility are significantly correlated (ρ=.813). In order to examine which type of error had the most influence on these judgments, we assigned correctness scores to each utterance by counting the number of correct initial, final and tone components of each syllable and divided this value by the number of syllables in the utterance. As can be seen from Table 1, both ratings of accent and intelligibility are most strongly correlated with the correctness of tones, but in the case of intelligibility the contributions of initial and final correctness for native speakers are almost as strong as those of the tones. In the case of the teacher, however, generally speaking correlations are lower than for the native speakers and tonal correctness correlates best with both strength of accent and intelligibility. We have to bear in mind that strength of foreign accent and intelligibility ratings were made by the native speakers and therefore will inevitably be more strongly correlated with those same subjects' judgments on correctness than the teacher's. In contrast to our earlier study on mono- and disyllables with first year students, the correctness of the tone component seems to be more crucial, possibly because more advanced students will show fewer errors on the segmental level.

Table 1 – Correlations between ratings of strength of foreign accent and intelligibility assigned by the the native listeners and correctness scores assigned by native listeners and teacher.

	correctness initial (native)	correctness final (native)	correctness tone (native)	correctness initial (teacher)	correctness final (teacher)	correctness tone (teacher)
accent (native)	.246**	.219**	.553**	.138**	.106*	.483**
intelligibility (native)	.510**	.497**	.543**	.258**	.202**	.424**

(**. Correlation significant at the 0.01 level, *. Correlation significant at the 0.05 level.)

If we look at the correlations between the mean correctness scores of the native listeners and the teacher we find that they are relatively low (ρ=.399, .387 and .480, for initials, finals tones. The mean correctness scores suggest that the teacher generally marked as few errors with respect to initials and finals (means of .974 and .962) as the native subjects (.974 and .972). However, she seems to have been more critical with respect to tones than the native listeners (.725 vs. .804). We also found that the number of syllables in the utterance is slightly negatively correlated (ρ=-.161) with the ratings of accent, that is, the longer the utterance, the less accented.

3.2 Analysis of Syllable Components

The annotations of the human judges (expert and native listeners) were used as reference to evaluate the results of the ASR systems. In order to evaluate every syllable component individually the syllables of the original text, annotations of the expert and the native speakers of Mandarin, and the results of the ASR systems were divided into initials, finals and tones. Each syllable component was considered as correct if there was an agreement between the annotation of the expert or every native speaker, the ASR and the original text. The results of the *Hefei* and *Shanghai* were averaged for each initial, final and tone. The difference in the results of *Hefei* and *Shanghai* is very small (about 1%). Therefore, the results of *Hefei* and *Shanghai* were averaged.

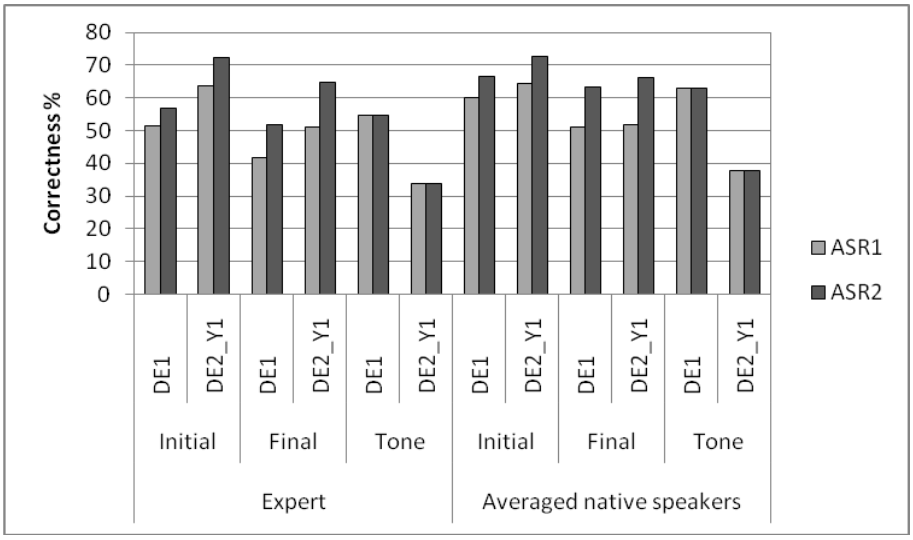

Figure 2 - Comparison of correctness of syllable components between the expert and the average of native speakers for ASR1 and ASR2 in DE1 and DE2_Y1

The results of the comparison are shown in Figure 2 for data of German learners of Mandarin in the same year of language education (DE1 and DE2_Y1). The results of ASR2 are better

than the results of ASR1 for initials and finals which are due to the adapted acoustic model for ASR2. The recognition of initials yielded better results than the recognition of finals for DE1 and DE2_Y1. The results of tone recognition in ASR1 and ASR2 are the same because no changes were made to the tone recognition algorithm. But the results for tone recognition in DE2_Y1 show lower correctness than in DE1 (more than 20% difference). The German learners are less adept to produce accurate and correct tones in sentences with subsequent syllables than when they are required to read out or imitate single mono- and disyllabic words. The German learners might not able to remember the tonal feature of the Chinese characters read aloud and to hit the accurate tone when the syllables appear in succession.

3.3 Pronunciation Error Detection

The intention is to reproduce the assessment by the expert or the native listeners using the ASR system. Therefore, we aim to keep the number of false hits - which would demotivate the learners - low, while maximizing the detection of true errors. The annotations produced by the human judges used as a reference for judging the performance of the ASR systems. The annotations of human judges were compared with the original text and the results of ASR systems were compared with the annotations of human judges to verify the correctness. For this reason to detect the correct results or errors from the ASR system we considered the following four cases: A (expert and ASR correct), B (expert correct and ASR false), C (expert false and ASR correct) and D (expert and ASR false). The ratio of all syllable components for the four cases for ASR1 and ASR2 for the data from the first year of language education (DE1 and DE2_Y1) is shown in Figure 3. The difference in the results between *Hefei* and *Shanghai* is less than 1.4%. Therefore, we averaged the results of all native speakers. The fully correct tokens in the case A were improved in the ASR2 by the human judges for DE1 and DE2_Y1. The number of correct tokens by DE2_Y1 is greater than by DE1 (see Figure 3) due to the large number of correct initials and finals by DE2_Y1 (see Figure 2). In the case of errors the human judges and the ASR system are in disagreement. The cases B and C represent the cases of error. The Figure 3 shows that the number of tokens in the case of errors in ASR2 is smaller than in ASR1. The number of correct tones in DE2_Y1 is smaller than in DE1 (see Figure 2). Therefore, the number of false tokens (cases B and C) in DE2_Y1 is greater than in DE1 (see Figure 3). The evaluation of human judges and the results of ASR system are different in the case D. The results of ASR2 are slightly greater than ASR1 in case D for both data groups.

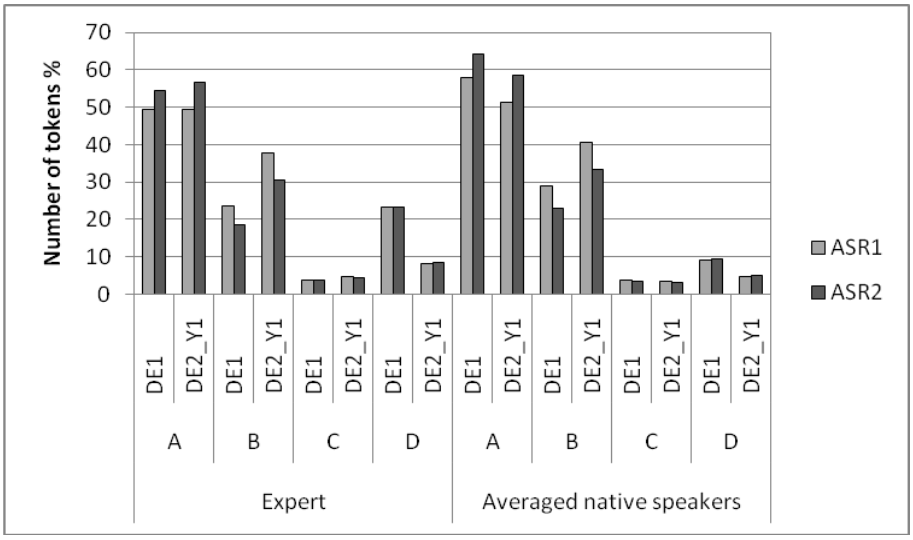

Figure 3 – The ratio of tokens for the four cases by the expert and the average of native speakers for ASR1 and ASR2 in DE1 and DE2_Y1

We performed an analysis of the confidence distance between the first candidates of the ASR systems for the cases A, B, C, and D. Therefore, the mean of the confidence distance between the first and second candidates of initials, finals and tones was computed for ASR1 and ASR2 for each case for the data DE1 and DE2_Y1. The mean values of confidence distance of initial, final, and tone for every case are shown in Table 2. The difference in the results of confidence distance between *Hefei* and *Shanghai* is less than (0.03). Therefore, we averaged the results of the ten native speakers. The results in each case for DE1 and DE2_Y1 by expert and the average of ten native speakers are nearly the same. The mean of confidence distance in the ASR2 are greater than in ASR1. In the case of correct results (case A) the confidence distance between the first correct candidate and the second candidate is more than (0.7). This means that the correct syllable components could be recognized by the ASR system as correct tokens and not as false. The mean value of confidence distance for correct tokens in DE2_Y1 is greater than in DE1 due to the great number of initials and finals (see Figure 2). The confidence distance in case of errors (B and C) is not small for DE1 and DE2_Y1. This showed that the recognition of falsely flagged tokens is not reliable.

Table 2 - The mean value of confidence distance between the first and second candidates of initial, final and tone for ASR systems in DE1 and DE2_Y1

Reference	Case	Data	ASR1	ASR2
Expert	A	DE1	0.5659	0.7142
		DE2_Y1	0.5983	0.7851
	B	DE1	0.3248	0.4466
		DE2_Y1	0.4139	0.5654
	C	DE1	0.4469	0.6128
		DE2_Y1	0.4491	0.4647
	D	DE1	0.3538	0.4787
		DE2_Y1	0.3882	0.4586
Averaged native speakers	A	DE1	0.5635	0.7117
		DE2_Y1	0.597	0.7801
	B	DE1	0.3247	0.4447
		DE2_Y1	0.4026	0.5434
	C	DE1	0.4885	0.6224
		DE2_Y1	0.552	0.5558
	D	DE1	0.3496	0.4759
		DE2_Y1	0.41	0.4624

4 Conclusions

This paper reported on the continued activities towards the development of the CALL system for teaching Mandarin to Germans. The method for detection of pronunciation errors which was presented in [8] was tested again on two databases to study the effect of complexity of corpus on the results of pronunciation error detection. The first corpus is simple and consists of monosyllabic and disyllabic words. The second corpus is more complex and consists of whole sentences. The data was evaluated by the expert and ten native speakers of Mandarin and processed by two versions of ASR systems. The first ASR system used an acoustic model trained on data of native speakers and the second ASR system used an adapted acoustic model. The annotations produced by the human judges used as a reference for judging the performance of the ASR systems. The experimental results showed an improvement in the performance of the modified ASR system. The ratings of accent and intelligibility are most strongly correlated with the correctness of tones than with the correctness of initials and finals. The ratio of correct initials is greater than finals. The correct tones in the complex corpus (DE2_Y1) is lower than in the simple corpus (DE1) because the German learners are less adept to produce accurate and correct tones in sentences than when they are required to

imitate single words. An analysis of the confidence distance between the first and second candidates output by the ASR systems was performed. Four cases were considered to detect the correct results or errors. The mean value of confidence distance showed that there is no significant difference between DE1 and DE2_Y1. The confidence distance in case of fully correct tokens is large and in case of errors is not small. This showed that in case of errors the recognition of falsely flagged tokens is not reliable.

5 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] LISTEN: *The LISTEN Project*. Carnegie Mellon University, Pittsburgh, Pennsylvania, USA. http://www.cs.cmu.edu/~listen/.

[2] EURONOUNCE: *The EURONONCE Project*. Dresden University of Technology, Dresden, Saxonia, Germany. http://www.euronounce.net/.

[3] 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.

[4] Hunold, C.: *Chinesische Phonetik. Konzepte, Analysen und Übungsvorschläge für den Unterricht Chinesisch als Fremdsprache*. Sinica, Vol. 17, Bochum, 2005.

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

[6] Zhou, J.-L., Tian, Y., Shi, Y., Huang, C.: *Tone Articulation Modeling for Mandarin Spontaneous Speech Recognition*. Proceedings of *ICASSP*, pp. 997-1000, 2004.

[7] Mixdorff, H., Külls, D. and Hussein, H.: *Development of a Computer-Aided Language Learning Environment for Mandarin - First Steps*. Proceedings of 20. Conference of *Elektronische Sprachsignalverarbeitung ESSV*, Dresden, Germany, September 2009.

[8] 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.

[9] 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 (in press).

[10] 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.