Towards a Computer-Aided Pronunciation Training System for German Learners of Mandarin - Prosodic Analysis

Hussein Hussein[1], *Hansjörg Mixdorff* [1], *Hue San Do*[1], *Si Wei*[2], *Shu Gong*[2],
Hongwei Ding[3], *Qianyong Gao*[2] *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, shugong, qygao, 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 system for German learners of Mandarin. In this experiment we used a complex corpus which consists of whole sentences and read from German students from three different years of language education and native speakers of Mandarin. 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. We aimed to study the effect of learning time of Mandarin on the development of learner's level. Therefore, the rhythmic and intonational features of tones were compared between German lerners according to every year of language education. German students tend to exaggerate the F0 contours to discriminate the tones better and learn to adapt these to the tones of native speakers with increasing learning time. The syllable duration depending on the tone by German learner is longer than by native speakers and the changes of F0 parameter of Mandarin tones by German students are greater than by native speakers of Mandarin.

Index Terms: Computer-Aided Language Learning (CALL), Mandarin, prosodic analysis

1. Introduction

A growing demand for foreign language competence stimulates activities towards computer-aided language learning (CALL) [1][2] . CALL is a tool to facilitate the individualized language learning process and can be used for pronunciation training. 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 over 20 months 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 the 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].

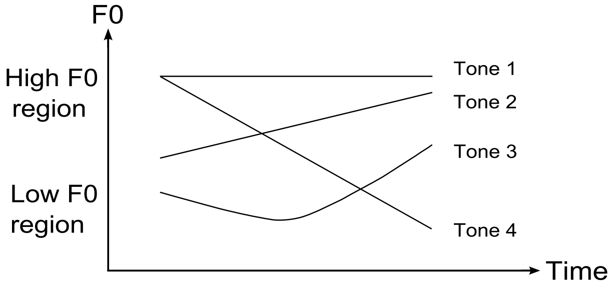

Figure 1: *Prototypical f0 contours of Mandarin tones.*

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. In order to study the effect of learning time of Mandarin on the development of learner's level the rhythmic and intonational features of tones were compared between German lerners according to every year of language education. The data were perceptually evaluated by human judges (teaching expert for Mandarin, two groups of native speakers of Mandarin) as well as processed by a Mandarin automatic speech recognition (ASR) system. The annotations produced by the human judges used as a reference for judging the correctness of syllable components produced by German students.

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

2. Experiment Method

This section describes the design of corpus and collection of data, the analysis of data and extraction of prosodic parameters for a contrastive analysis between German learners and native speakers of Mandarin, and the evaluation of data.

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) and native speakers of Mandarin by iFlyTek company, Hefei, China. 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 first experiments [3][7][8]. 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 were 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*") consists 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.

We also collected data from the native speakers of Mandarin (henceforth "*CN*") by iFlyTek company, Hefei, China. Every native speaker of Mandarin read all 62 sentences which were used in *DE2*. The sentences were produced by 20 native speakers (ten male and ten female).

2.2. Data Analysis and Data Evaluation

In order to compare the prosodic properties for German learners and native speakers of Mandarin we calculated the rhythmic and intonational features of the Mandarin tones on the syllable level. Therefore, the data was labeled automatically on the syllable and phone-levels using the ASR system in a forced alignment mode. The ASR system provided also the posterior probabilities for tones and phonemes. The F0 contour was calculated using the *Praat* [9] algorithm with a step of 10ms and different standard settings of the minimum and maximum parameters of F0 for male (100 and 350 Hz) and (120 and 450 Hz) for female speakers.

The F0 contour reflects the Mandarin tone (see Figure 1). However, the task is much more difficult due to variations in speaker and style and most importantly, tonal coarticulation. In order to reduce the variation of the speaker's F0 range the F0 was normalized. The most commonly used F0 normalization method is the mean normalization, which is implemented as follows:

$$F0_{\grave{i}} = F0_i - \overline{F0} \qquad (1)$$

where $F0_i$ and $F0_{\grave{i}}$ is the F0 value before and after normalization and $\overline{F0}$ is the average F0 value of the person to be normalized.

The minimum, maximum, mean, standard deviation, slope, and range of the normalized F0 subcontour of tones were calculated. The slope of F0 was estimated from the start and end of F0 subcontour (in Hz) and the start and end time of syllable (in second) as follows:

$$F0_{slope} = \frac{F0_{end} - F0_{start}}{t_{end} - t_{start}} \qquad (2)$$

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 by the iFlyTek company, Hefei, China (henceforth "native speakers 1") and five native speakers of Mandarin by the School of Foreign Languages, Tongji University, Shanghai, China (henceforth "native speakers 2"): The native speakers 1 and native speakers 2 (henceforth "native speakers") 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 intelligibility and strength of foreign accent 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. The ASR system processed both the German and Chinese data. The ASR system used the original acoustic model trained on data from native speakers of Mandarin.

3. Results

The syllable duration was calculated from the generated labels by the ASR system in forced alignment mode and the tone parameters were calculated from the normalized F0 contour for the contrastive analysis of Mandarin tones. The annotations produced by the human judges used as a reference for judging the correctness of syllable components produced by German students.

3.1. Analysis of Prosodic Features of Mandarin Tone

We performed a contrastive analysis of the prosodic properties on the rhythmic and intonational features of the Mandarin tones on the syllable level to identify the differences and similarities between native speakers and German learners of Mandarin.

3.1.1. Analysis of Syllable Duration Depending on the Mandarin Tone

Table 1 displays the mean and standard deviation (SD) of duration (in second) of the Mandarin tones for *DE2* and *CN*. The mean and standard deviation of syllable duration depending on the tone produced by the German learners is longer than the duration of the native speakers of Mandarin. This confirms the hypothesis that learners of a language speak more slowly. If we compare the duration of tones according to every year of language education for German learners, we notice the decrease of syllable duration depending on the tone from first- to second- and to third-year German students (see table 2). This indicates that the German students are able to speak faster the longer they learn Chinese.

The mean normalized pair-wise inter-variability index *npvi* [11] is 37.61 for Chinese as compared with 62.48 for the German subjects (the *npvi* for Year3, Year2, and Year1 is 53.5, 57.4, and 70.4, respectively). This indicates that German speakers are much more at variance with respect to syllabic durations as the Chinese. Looking more closely at the rhythmic patterns of individual sentences we correlated the syllabic durations in one realization of a sentence with the syllabic durations in all the other realizations of the same sentence. The advantage of this approach is that the effect of the speech rate on this measure is rather small. It was previously used for evaluating the quality of a duration-predicting model in text-to-speech synthesis and also in an earlier study on Australian English [12] spoken by Vietnamese learners. Results indicate that the Chinese realizations (mean ρ=.77) are more similar in their rhythmic structure (more highly correlated) than the German ones (mean ρ=.50). The mean inter-group correlation is only .37. If we examine this figure for years 1, 2 and 3 separately, Year 3 students exhibit a higher correlation with the Chinese controls (.47) than Year 2 (.35) and Year 1 (.33).

Table 1: *Mean and standard deviation of syllable duration depending on the tone for DE2 and CN.*

Rating	*DE2*					*CN*				
	Tone 0	Tone 1	Tone 2	Tone 3	Tone 4	Tone 0	Tone 1	Tone 2	Tone 3	Tone 4
mean	0.25	0.34	0.28	0.30	0.31	0.19	0.21	0.19	0.19	0.20
SD	0.12	0.12	0.11	0.13	0.13	0.09	0.07	0.07	0.07	0.07

3.1.2. Analysis of F0 Contour of Tone

The parameters of the normalized F0 contour of Mandarin tones (in Hz) for German learners and Chinese native speakers, and for German learners according the years of language education are presented in the tables 3 and 4. The table 3 shows that the mean F0 of tones by German learners of Mandarin is smaller than by native speakers, but the standard deviation and F0 range are greater by German learners of Mandarin for all tones compared to the Chinese native speakers. The tone 2 by native speakers has an increasing slope, but by German learners do not have a slope. The slope of tone 4 by native speakers and German students is positive and do not agree with the tone 4 in the Figure 1. The slope can not be used on the tone 3 due to the falling and rising contour of this tone.

Table 2: *Mean and standard deviation of syllable duration depending on the tone for German learners according to the years of language education.*

Data	Rating	Tone 0	Tone 1	Tone 2	Tone 3	Tone 4
DE2_Y1	mean	0.28	0.41	0.34	0.37	0.35
	SD	0.12	0.11	0.11	0.13	0.14
DE2_Y2	mean	0.27	0.36	0.28	0.30	0.31
	SD	0.13	0.13	0.10	0.12	0.14
DE2_Y3	mean	0.22	0.28	0.22	0.23	0.27
	SD	0.10	0.09	0.08	0.09	0.10

The table 4 shows that the beginner learners of Mandarin (*DE2_Y1*) display a wide F0 range and a large standard deviation for all tones compared to the Chinese native speakers while the F0 range and standard deviation produced by third-year students (*DE2_Y3*) are small and show much more similarity with their Mandarin counterparts. First-Year students tend to exaggerate tone contours to discriminate the tones better and learn to adapt these with increasing study time.

Table 3: *Minimum, maximum, mean, standard deviation, slope, and range of F0 subcontour of Mandarin tones for DE2 and CN.*

Rating	*DE2*					*CN*				
	Tone 0	Tone 1	Tone 2	Tone 3	Tone 4	Tone 0	Tone 1	Tone 2	Tone 3	Tone 4
min	-26.91	-6.68	-25.56	-35.65	-19.97	-31.39	1.31	-18.93	-37.25	-14.50
max	25.90	50.39	29.27	22.80	53.91	5.42	50.11	22.70	19.42	43.74
mean	-3.17	21.60	-4.02	-12.25	15.25	-13.15	29.62	-1.86	-11.95	16.94
SD	16.09	16.08	17.40	18.11	23.49	11.68	15.28	14.43	19.62	19.09
slope	35.00	43.56	-0.01	-54.65	23.32	27.33	174.02	19.96	-165.34	33.15
range	52.82	57.08	54.84	58.45	73.88	36.82	48.80	41.64	56.67	58.25

Table 4: *Minimum, maximum, mean, standard deviation, slope, and range of F0 subcontour of Mandarin tones for German learners according to the years of language education.*

Data	Rating	Tone 0	Tone 1	Tone 2	Tone 3	Tone 4
DE2_Y1	min	-29.48	-2.50	-32.87	-43.51	-21.80
	max	34.22	60.62	35.71	28.10	71.58
	mean	-1.60	28.5153	-7.10	-14.69	22.98
	SD	19.7	17.77	21.45	21.59	30.93
	slope	71.23	38.91	-7.22	-34.28	33.89
	range	63.71	63.13	68.58	71.62	93.39
DE2_Y2	min	-29.62	-12.60	-24.88	-37.40	-23.77
	max	30.70	47.66	30.06	22.21	50.7
	mean	-1.21	20.07	-2.6	-11.41	10.88
	SD	18.11	17.14	17.47	18.41	22.64
	slope	26.73	18.89	7.85	-31.13	22.44
	range	60.33	60.27	54.94	59.61	74.47
DE2_Y3	min	-24.24	-8.21	-18.1	-27.32	-17.36
	max	18.48	43.58	22.23	17.77	40.52
	mean	-4.9	16.92	-1.23	-10.11	10.26
	SD	12.89	14.55	13.11	14.58	17.7
	slope	11.88	53.4	5.05	-82.51	15.0
	range	42.72	51.79	40.33	45.1	57.89

3.2. Analysis of Correctness of Syllable Components

In order to evaluate every syllable component individually the syllables of the original text, annotations of human judges 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 and the original text. The results of the native speakers were averaged for each initial, final and tone. The results of the correctness are shown in Figure 2 for the data *DE2_Y1*, *DE2_Y2*, and *DE2_Y3*.

The Figure 2 shows that there is no significant difference in the results of initials and finals between *DE2_Y1*, *DE2_Y2*, and *DE2_Y3*. But the corerctness of tones by third-year students are better than first- and second–year students. The correctness of tones show lower success rates than initials and finals. The German learners are less adept to produce accurate and correct tones in sentences with subsequent syllables when they are required to read out or imitate it. 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.

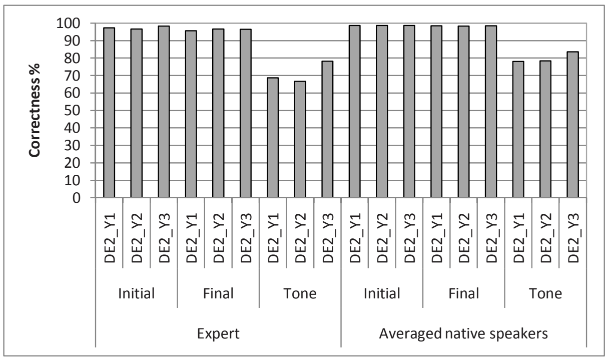

Figure 2: *Comparison of correctness of syllable components between the expert and the average of native speakers for German data from three different years of language education.*

3.3. Comparison of Entire Utterance

We analyzed the utterance-wise judgments of accent and intelligibility of the German data. The table 5 shows the mean accent and intelligibility for *DE1*, *DE2_Y1*, *DE2_Y2*, and *DE2_Y3*. The *DE1* was evaluated by 10 native speakers of iFlyTek company in the first experiment [3] and not evaluated by native speakers 2. Native speakers scored the corpus of *DE2_Y1* lower regarding accent than *DE1* which means that a stronger accent was perceived for syllables as a string in a complex sentence than for single mono- and disyllabic words. The accent of *DE2_Y3* is greater than of *DE2_Y1* and *DE2_Y2*. The accent perceived could be related to the higher tonal accuracy produced by the *DE2_Y3* speakers as shown in Figure 2. Intelligibility of *DE2_Y3* was also rated higher than *DE2_Y1* and *DE2_Y2* which might be due to the reasons mentioned above.

Table 5: *Mean of accent and intelligibility for DE1, DE2_Y1, DE2_Y2, and DE2_Y3 by the average of native speakers.*

Data	Accent	Intelligibility
DE1	3.93	3.76
DE2_Y1	3.35	4.05
DE2_Y2	3.26	4.03
DE2_Y3	3.64	4.24

4. Discussion and Conclusions

A contrastive analysis of rhythmic and intonational features of the Mandarin tones was performed to identify the differences and similarities between native speakers and German learners of Mandarin. The syllable duration depending on the tone by German learner is longer than by Chinese native speakers and the syllable duration depending on the tone decreases from first- to second- and to third-year German students. The changes of F0 contour of Mandarin tones are greater by German students than by native speakers. German students tend to exaggerate the F0 contour to discriminate the tones better and they learn to adapt the tones to these of native speakers with increasing study time. The third-year students can pronounce the Mandarin tones more accurate than the first- and second-year students. The accent and intelligibility of third-year students were rated higher than beginners due to the higher tonal accuracy produced by third-year speakers.

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.

6. References

[1] LISTEN, "The LISTEN Project", Carnegie Mellon University, Pittsburgh, Pennsylvania, USA. `http://www.cs.cmu.edu/~listen/`.

[2] EURONOUNCE, "The EURONOUNCE 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., and Chang, E., "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 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] Boersma, P. and Weenink, D., "Praat doing Phonetics by Computer", version 5.0.42, 15 April 2010, `www.praat.org`.

[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.

[11] White, L. and Mattys, S.L., "Calibrating rhythm: First language and second language studies", Journal of Phonetics, 35, 501-522, 2007.

[12] Mixdorff, H. and Ingram, J., "Prosodic Analysis of Foreign-Accented English", In Proceedings of Interspeech 2009, Brighton, England, 2009.