DEVELOPMENT OF A COMPUTER-AIDED LANGUAGE LEARNING ENVIRONMENT FOR MANDARIN – FIRST STEPS

Hansjörg Mixdorff, Daniel Külls and Hussein Hussein

Beuth-Hochschule für Technik Berlin
mixdorff@bht-berlin.de

Summary: This contribution reports first activities towards the development of a computer-aided pronunciation environment for teaching Mandarin to Germans. This 3-year-project is funded by the German Ministry of Education and Research. Based on a contrastive analysis of Mandarin and German, a preliminary study of learner errors was conducted. We collected perception as well as production data from 19 German first-year students of Mandarin at FU Berlin in order to perform an evaluation of actual errors as perceived by native speakers of Mandarin. These outcomes are utilized to fine-tune the actual training systems a general outline of which is also presented in this paper.

1 Introduction

In a globalized world, the growing demand for foreign language competency stimulates activities towards computer-aided language learning. Within this area, 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. Since, however, pronunciation errors can cause a lot of frustration and the phonetic training only occupies a relatively small part within typical language courses, computer-based solutions are of great interest since they can provide assistance at the frequency, intensity and suitable time which the learner chooses. In a three-year project funded by the German Ministry of Educations and Research, we will develop a Mandarin training system for Germans and evaluate it within a university context. The current study reports on first experiments aimed at analyzing typical errors committed by German first-year students of Mandarin. This analysis is three-fold: (1) A narrow phonetic analysis by an expert for Mandarin (2) A performance and transcription analysis by native listeners of Mandarin (3) a Mandarin automatic speech recognition system.
Modern Mandarin (*Putonghua*) differs from German significantly on the segmental as well as the supra-segmental levels and poses a number of problems to the German learner.

1.1 Segments

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*. Many of the phonemes building initials and finals have exact or close counterparts in the German language. Therefore, German learners might occasionally be perceived by native listeners of Mandarin as speaking with an accent, but not generally wrong. Errors usually arise from phonemes of Mandarin without correspondences in German ([1], pp. 31-32).

Initials. Among the 21 initial consonants, the following yield the highest potential for errors (we provide Pinyin as well as IPA transcriptions). We will refer to Pinyin transcription indicated by italics.

One half of the problematic cases are formed by the five aspirated plosives and affricates *p*, *t*, *k*, *c*, und *ch* (see Table 1, left). Although approximate correspondences of these exist in German they are much more strongly aspirated in Mandarin, since aspiration is the only feature which distinguishes them from their counterparts *b*, *d*, *g*, *z* und *zh.* Since aspiration is not a distinctive feature of German, German learners tend to aspirate too weakly, causing possible confusion between the two groups of phonemes. This also applies to the aspirated palatal *q,* but in this case the situation is further aggravated by the existence of its inaspirated counterpart *j* as well as a third palatal consonant, *x,* which all do not exist in German. One therefore can expect confusions between *q*, *j* and *x,* as well as with the more remote, but similar phonemes *ch*, *zh* und *sh*.

Table 1-Problematic initials (left) and finals (right)-

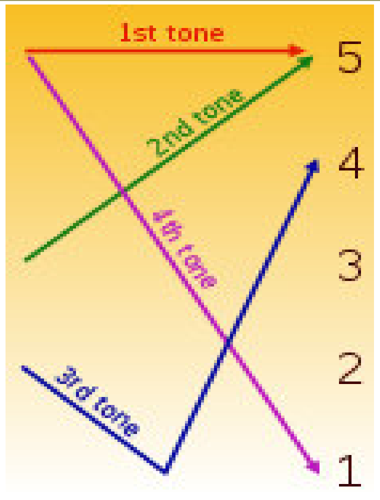
Problematic Initials			
Pinyin	IPA	Pinyin	IPA
P	pʰ	*q*	tɕ ʰ
T	tʰ	*j*	tɕ
K	kʰ	*x*	ɕ
C	tsʰ	*z*	tz
Ch	tʂ ʰ	*r*	ʐ

Problematic Finals	
Pinyin	IPA
e	ɤ
*(s)**i***	ɿ
*(sh)**i***	ʅ
eng	ɛ˜

Finals. As mentioned above, finals mainly consist of vocalic segments. The only consonants which may occur at the end of finals are r, n and ng. The status of finals [ɿ] und [ʅ] is somewhat disputed. Although the Pinyin transcription i suggests a vocalic quality, some publications (cf. [2], pp 35-36) treat them as syllabic consonants. As in the case of initials most problems are caused by vowels that do not exist in the German language. Germans often produce [ɿ] and [ʅ] with too much jaw opening and in the case of [ʅ] not enough retroflexed which might cause native speakers to perceive *e* [ɤ]. In addition, the slightly nasal [ɛ˜] of the final *eng* is often produced as [a], causing a percept of the final *ang,* or [ə], facilitating confusion with the final *en* (see Table 1, right).

1.2 Suprasegmentals and Tones

The segmental problems which Mandarin poses to German learners are certainly dwarfed by the complexity of its tonal distinctions. Mandarin has four syllabic tones, five including the neutral one:

Tone	Mark	Description	 1st tone 2nd tone 3rd tone 4th tone 5 4 3 2 1
1	***mā***	High and level.	
2	***má***	Starts medium in tone, then rises to the top.	
3	***mǎ***	Starts low, dips to the bottom, then rises toward the top.	
4	***mà***	Starts at the top, then falls sharp and strong to the bottom.	
neutral	***ma***	Flat, with no emphasis.	

The tonal contour of a syllable changes its meaning, i.e. the syllable *ma* means „mother", „hemp ", „horse ", „to scold" or is a question marker depending on the tone associated. When teaching these distinctions to Germans, tones are generally illustrated by analogies of sentence intonation: Straight „aaah" as in a medical examination of the throat for illustrating the first tone, echo-question „Ja?" for the second tone etc. Single tones can generally be acquired in a very short time. However, articulating a sequence of tones when reading poly-syllabic words or sentences appears to be much more difficult. If we consider the problem at the level of di-syllables, there are a total of 19 combinations[1]. Tone combinations 3-1, 3-2 and 3-4 tend to be the most difficult, since tone 3 is only realized half-way to the bottom of the tonal range and therefore differs from tone 3 produced in isolation. Germans tend to produce the rising movement of tone 3 as in isolated syllables which makes it confusable with tone 2. Another frequent error concerns the production of neutral tones since during their first weeks learners naturally focus on producing the right tonal contours and find it hard to realize a syllable lacking a clear tonal target.

2 Production and Perception Experiments

2.1 Corpus Design

The corpus recorded at FU Berlin consisted of 54 tokens. One half of these had been produced by a female native speaker of Mandarin and was imitated (shadowed) by the subjects. The other half was provided in Pinyin transcription and read aloud. Including both modes enabled us to examine potential differences in performance. Each part contained eight mono-syllabic and 19 di-syllabic words. By selecting these tokens we attempted to cover all initials, finals and tone combinations of Mandarin in a small set of words potentially unknown to the subjects, but adequate at their early stage of proficiency. Whereas the tokens of the imitation part were real words, the reading part contained nonsense words created by permutations of initials and finals of the real words to facilitate a better comparison. In addition to the 54 word tokens, we also recorded five short sentences which, however, were not included in the current study.

2.2 Data Collection and Participants

The 54 tokens were produced by 19 of a total number of 80 first-year students of Chinese Studies at the East Asia Seminar of Free University (FU) Berlin. At the time of the experiment they had completed 12 weeks of Mandarin language training using the text book „New Practical Chinese Reader 1". In addition to their regular classes, nine of the subjects (henceforth *WS*) (three male and six female) had attended a weekly seminar of two hours which was conducted by Külls. Roughly one half of the seminar was dedicated to phonetic exercises, the other half to grammar and translation. The phonetic exercises comprised the imitation and reading of mono- and di-syllables, contrastive exercises with minimal pairs of differing initials or finals, as well as slow reading from the text book, constantly monitored and corrected by the teacher. One objective of our experiment was to examine whether the additional training had resulted in tangible benefits to the participants (*WS*) by comparing their results to those from the group that had not taken part (henceforth *WOS*) (five male and five female students).

In the perception test participant were requested to write down using the Pinyin system eight mono-syllabic and 19 di-syllabic words produced once by a female native speaker. Most of the words were unknown to the subjects and contained all initials, finals, tones and tone combinations as in the production test.

[1] The neutral tone can only be the second in such a combination, and due to a tone Sandhi rule, 3-3 becomes 2-3.

2.3 Evaluation of Data

The data produced at FU Berlin was annotated, judged and processed three-fold:

(1) By Külls, a German teacher of Mandarin, from the expert and pedagogue's point of view (henceforth "expert"): His task was to provide useful feedback to the students afterwards and perform a critical, detailed analysis even of errors that were sub-phonemic.

(2) Ten female native speakers of Mandarin, all of them staff of *Iflytek Company*, Hefei, China (henceforth "native speakers"). They were between 20 and 30 years of age.

(3) An automatic speech recognition (ASR) system which is part of an automated proficiency test of Mandarin[3]. Results of this test are reported in [4].

Whereas the expert listened to all recordings several times and annotated errors with a high degree of detail, the native speakers were presented with each token 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 token 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.

3 Production Results

We evaluated the annotations by the native speakers in two steps. Initially we only examined the correctness of each token as a whole. Subsequently, we divided the syllables of the original token and its reproductions by the German students into initials, finals and tones in order to statistically evaluate all three components separately. The annotations produced by the expert served as a reference for judging native speakers' and ASR performances.

3.1 Comparison of Entire Tokens

The comparison between the annotations produced by the native speakers (without knowledge of the intended targets) and the original tokens yielded the following results:

1. For a total of 55.4% of presentations of tokens produced by the *WOS* group (2993 of 5400) and 61.2% (2974 of 4860) of the *WS* group, these were identified as the intended targets. This suggests a slightly better performance of the group that had participated in the phonetic seminar. We performed split-correlation reliability analysis on judgments of accent and intelligibility by dividing the utterance-wise judgments into two perceiver groups of five subjects each, yielding a cross-correlation between the two groups of .76 ($p<.001$) for the accent rating and of .83 ($p <.001$) for the intelligibility rating. This suggests that the judgments are more stable for the latter.

The mean accent and intelligibility ratings are 4.10 and 4.05 for the *WS* group and 3.95 and 3.83 for the *WOS* group, respectively. Independent samples T-tests suggest that these differences are highly significant ($T=-4.3$, $df=1024$, $p < .001$ for *accent*, $T=-4.1$, $df=1024$, $p < .001$ for *intelligibility*). The respective figures from the expert's judgments for *WOS* were 53.3% (288 of 540), and 60.3% (293 of 486) for *WS*, respectively, suggesting just a slightly more critical approach.

2. The comparison between shadowing and reading yielded the following result: In 66.3% (3402 of 5130) of cases, the shadowed tokens were correct, whereas the figure is only 50.0% (2565 of 5130) for the read tokens. This indicates a significantly better performance in the shadowing task as opposed to reading. These figures are supported by the mean values for

accent and intelligibility which are both 4.11 for the shadowing task, and 3.93 and 3.76 for the reading task, respectively. Again these differences prove to be highly significant. Similar results were reported by the expert: Shadowing yielded 63.0% correct (323 of 513), reading 50.3% (258 of 513), respectively

3.2 Analysis of Syllabic Components

By separately analysing initials, finals and tones we aimed to determine the most likely confusion partners of each "difficult" phoneme according to our prior contrastive analysis. Furthermore we wanted to calculate correlations between accent and intelligibility - being subjective measures of quality - and the objective errors annotated by the perceivers. Finally we were interested in the agreement of judgment between the expert, the native speakers and the ASR system.

3.2.1 Frequent Errors and Confusion Partners

It should be noted that we concentrate on those highly probable confusions which do not arise from insufficient knowledge of the Pinyin transcription system on the part of the German students. For instance, we do not consider confusions between Pinyins *y* and *j*, since this error certainly is not caused by the inability to produce *j* as [tɕ], but imperfect competence in the Pinyin writing system. Among the probable confusion partners we only considered those which reached a frequency of more than 2% of pooled realizations of that phoneme, in order to exclude idiosyncratic errors by a single subject.

Table 2 - Percentage correct (second column) and confusion partners of initials, native speakers.

ch	***ch***: 61.32	***zh***: 21.97	***sh***: 6.05	***x***: 2.89
c	***c***: 64.47	***z***: 21.58	***s***: 12.89	
q	***q***: 73.51	***j***: 17.37	***x***: 2.19 ***zh***: 2.19	***ch***: 2.02
j	***j***: 80.13	***q***: 5.66		
zh	***zh***: 84.47	***ch*** : 3.42	***z***: 2.89 ***j***: 2.89 ***c***: 2.89	

From Table 2 we can see that according to the native speakers' annotations the affricate group of initials was most problematic. The group of aspirated plosives appeared to be less difficult than expected (ratio correct: *p*: 99.7%, *t*: 90.0%, *k*: 90.3%). Even *r* reached 91.6%. In general, these results matched those by the expert with slight differences in the order of errors and of confusion partners.

In the case of finals (compare Table 2), we yielded partly unexpected results. Whereas our hypothesis regarding phonemes [ɿ] und [ʅ] - both represented by *i* in the Pinyin writing system – was confirmed, the final *e* caused fewer errors than predicted.

The highest frequency of errors, however, is found in syllables with consonant codas *n* und *ng* with *ing*, *an*, *uan*, ang being the most problematic finals. A large amount of confusion occurs between the nasal consonants, but also between the preceding vowel segments. These results matched those by the expert.

In line with our expectations, single tones were generally produced correctly (see Table 4). Notable confusions only occurred between tones 2 and 3. The expert identified more frequently erroneous third tones. Tonal combinations obviously posed greater problems for the German students. According to the annotations by the native speakers the tonal combinations listed in Table 5 were those produced with the highest frequency of errors.

Table 3 - Percentage correct (second column) and confusion partners of finals (native speakers).

ing	***ing***: 71.32	***in***: 28.42		
an	***an***: 78.68	***en***: 8.42	***ang***: 7.37	***eng***: 2.76
uan	***uan***: 78.95	***uang***: 5.26	***eng***: 3.16 ***ua***: 3.16	***en***: 2.89
(sh)i	***(sh)i***: 79.26	***e***: 13.58	***ue***: 2.84	***ü***[2]: 2.53
ang	***ang***: 82.37	***an***: 10.79	***eng***: 4.47	
(s)i	***(s)i***: 83.95	***e***: 15.13		

Table 4 - Percentage correct (second column) and confusion partners of single tones, native speakers.

2	***2***: 88.68	***3***: 8.95
3	***3***: 89.34	***2***: 9.74
4	***4***: 96.32	***3***: 2.76

Table 5 - Percentage correct (second column) and confusion partners of some tonal combinations.

4-3	***4-3***: 44.74	***4-2***: 43.42	***4-1***: 3.95	***4-0***: 2.63
3-0	***3-0***: 47.11	***3-1***: 26.32	***2-0***: 5.79	***2-1***: 5.26
2-0	***2-0***: 53.68	***2-1*** : 9.74	***1-0***: 8.16	***2-4***: 6.58
1-3	***1-3***: 57.37	***1-2***: 35.79	***4-3***: 2.63	***4-2***: 2.37
3-2	***3-2***: 60.00	***2-2***: 10.26	***3-1***: 6.58	***3-3***: 5.00
3-1	***3-1***: 65,79	***2-1***: 14,21	***4-1***: 10,79	***3-2***: 2,63

3.2.2. Correlations

In order to weight the degree of error found in a particular token we applied the following metric: For each trial (perceiver-token combination) each syllable was assigned one point for a correct onset, one for a correct final and one for a correct tone, respectively, and points were added for all syllables and divided by the number of syllables. The resulting scores were then correlated with the corresponding judgments of accent and intelligibility and yielded values of .60 and .71. Corresponding scores were determined for the separate initial, final and tonal components of each token and once again correlations with accent and intelligibility calculated. Results are .44/.56 for the initial, .41/.49 for the final and .33/.37 for the tone, respectively.

[2] The letter *ü* is used to denote the final [y].

4 Perception Results

In line with the evaluation of the production part we determined the main confusion partners of initials, finals, single tones and tonal combinations with contributions of at least 5%. For space reasons we only list the five most problematic initials (Table 6).

Table 6 - Percentage correct (second-left column) and confusion partners of initials.

q	***q***: 31.60	***x***: 14.04 ***zh***: 14.04	***ch***:10.53	
c	***c***: 42.11	***s***: 15.79	***sh***: 13.16	***ch***: 10.53 ***q***: 10.53
r	***r***: 47.37			
ch	***ch***: 50.00	***sh***: 18.42	***zh***: 13.16	***q***: 5.26 ***t***: 5.26
n	***n***: 57.89	***m*** : 42.11		

If we only consider initials with a low correct rate, we only find two significant new entries: *r* and *n*. In the case of *n* there is only one confusion partner, namely *m*. Interestingly, confusions in the opposite direction are rare. The case of *r* is more complicated. The original syllables *ruan* and *re* containg the intial *r* were recognized as *wan* in 34.21% of cases and as yue / ye in 13.16% of cases, respectively. Since Pinyin *w* and *y*, strictly speaking, are not considered as initials, but are used to indicate finals starting with *u* (*w*) and *i/ü* (*y*) this means that the initial *r* was simply not perceived.

Table 7 - Percentage correct and confusion partners of finals.

iong	***ong***: 47.37	***iong***: 26.32		
eng	***ong***: 52.63	***eng***: 31.58		
e	***e***: 43.86	***uo***: 26.32	***i***: 21.05	***ue***: 5.26
ün	***ün***: 47.37	***üan***: 15.79	***uan***: 10.53	
ang	***ang***: 57.89	***ao***: 26.32	***iang***: 10.53	

As is the case in the production part, finals ending in *n* und *ng* are problematic (Table 7). Confusions, however, mostly concern the vocalic part of the final, not the nasal coda. It is remarkable, that in the case of finals *iong* und *eng* the confusion partner is recognized more frequently than the intended form. This phenomenon was not observed in the production part.

Analysis of single tone results shows that the error rate is the lowest. Tone 2 seems to be the most likely confusion partner (Table 8).

Table 8 - Percentage correct and confusion partners of single tones.

2	***2***: 78.95	***3***: 18.42	
1	***1***: 81.58	***2***: 10.53	***3***: 7.89
3	***3***: 84.21	***2***: 7.89 ***4***: 7.89	
4	***4***: 89.47	***2: 10.53***	

Table 9 - Percentage correct and confusion partners of some tonal combinations.

3-2	***2-4***: 21.05	***1-2***: 10.53 ***1-3***: 10.53 ***2-0***: 10.53 ***2-1***: 10.53 ***2-2***: 10.53		
3-0	***3-0***: 21.05 ***2-1***: 21.05	***3-4***: 15.79	***3-1***: 10.53 ***3-2***: 10.53	
1-0	***1-4***: 63.16	***1-0*** : 26.32	***1-2***: 10.53	
2-0	***2-0***: 26.32 ***2-4***: 26.32	***1-0***: 10.53 ***1-4***: 10.53 ***3-0***: 10.53		
3-4	***2-4***: 36.84	***3-4***: 26.32	***4-4***: 15.79	***1-4***: 10.53

We witnessed the highest error rate in tonal combinations (Table 9). Those with a leading Tone 3 or a trailing tone 0 cause the most problems. In several cases the most likely confusion partner is perceived more frequently than the intended form, with 3-2 being the most extreme. This combination was not recognized even in a single case.

5 Considerations for the Training System

A computer-aided language learning system should be able to detect the deviation from the native standard and to provide instructions for improving the pronunciation. Therefore, a specialized recognizer is required to detect the kind and place of deviation. The recognizer operates in a forced alignment mode providing segment boundaries for the utterance to be analysed. Once the syllabic segmentation has been performed, a tone recognizer can be used to determine the syllabic tones.

The development of the training system is based on speech data collection, construction of recognizer components and creation of appropriate exercises for typical problems. It will provide the following functionality:

- Recording and playback of utterances and auditory-visual feedback, e.g. visualisation of speech signal, fundamental frequency (F0) contour and intensity contour in real-time.
- A specialized recognizer trained with native Mandarin models optimized using knowledge of probable German learner errors
- Indication of probable pronunciation errors by measuring the deviation between the learner and the native utterance.
- Detection of errors and generation of instructions (verbal or visual) for improving the pronunciation.
- Auxiliary materials for clarifying the error, e.g. video films or animated sagittal sections of vocal tract.
- Generation of corrected learner utterances (resynthesis).

The system can be divided into two levels [5] as shown in Figure 1:

1. Recognizer: It contains the determination of the individual components of recognizer.

2. Application: It manages the interaction with the user. The interface to the user takes the form of visual elements.

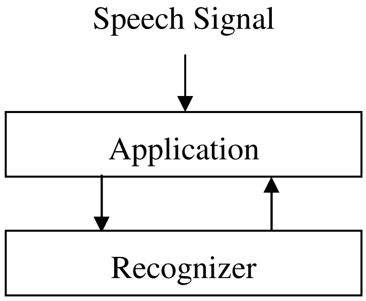


Figure 1 - Framework of computer-aided language learning system

The Chinese acoustic models, dictionary and networks, which were trained by our partner (National Taipei University of Technology - NTUT, were used as the recognizer components. The acoustic models were trained with the TCC300 database [6]. The TCC300 is a microphone speech database for Mandarin. It contains recordings of 300 speakers with a sampling frequency of 16 kHz and a resolution of 16 bit.

6 Discussion and Conclusions

Although case-by-case judgments suggest a relatively high phoneme-wise identification rate of the native speakers, they do not necessarily agree with each other. Therefore the pooled token-wise correct rates are much lower than those of the expert. As expected, affricates are the greatest sources of errors whereas plosives seem much less problematic, *r* was flagged only in the perception test. Pinyin *e* is less often mispronounced than predicted and finds its likely confusion partners in the finals of syllabic consonants (*sh)i* and *s(i).* An unexpected finding is the relatively high error rate in finals with nasal endings. In the tonal confusions tones 2 and 3 are the expected partners. In tonal combinations, tone 3 in leading position as well as tone 0 in trailing position are the most likely causes of errors. This result is found in the production as well as in the perception test. In addition, a trailing tone 3 often becomes tone 2, possibly because the rising part is exaggerated by the learners.

7 Acknowledgements

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

8 References

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

[2] Hunold, C., “Chinesisch”, *Hirschfeld, U. / Kelz, H.P. / Müller, U. (Hg.): Phonetik International – von Albanisch bis Zulu*, 2005.

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

[4] Mixdorff, H. et al.: Towards a Computer-aided Pronunciation Training System for German Learners of Mandarin. Proceedings of SLaTE 2009, Wroxall Abbey, England.

[5] Odell, J., Kershaw, D., Ollason, D., Valtchev, V., Whitehouse, D., and Wood, D.: A description of the HTK Application Programming Interface. September 1998.

[6] http://www.aclclp.org.tw/use_mat.php#tcc300edu