Towards a Computer-aided Pronunciation Training System for German Learners of Mandarin

Hansjörg Mixdorff[1], *Daniel Külls*[1], *Hussein Hussein*[1], *Gong Shu*[2], *Hu Guoping*[2], *Wei Si*[2]

[1]Department of Informatics and Media, BHT University of Applied Sciences, Berlin, Germany
[2]Dept. EEIS, University of Science and Technology of China, Hefei, Anhui, P.R.China
mixdorff@bht-berlin.de, kuells@bht-berlin.de, hussein@bht-berlin.de,
shugong@mail.ustc.edu.cn, gphu@iflytek.com, siwei@iflytek.com

Abstract

The current paper discusses first investigations aimed to lay the groundwork for the development of computer-aided pronunciation training for teaching Mandarin to Germans. We conducted a contrastive analysis of the two languages leading to a set of tokens for a production and perception experiment involving German first-year students of Mandarin. Their data were perceptually evaluated by a teaching expert for Mandarin, native speakers of Mandarin as well as processed by a Mandarin automatic speech recognition system.

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.

Pinyin	IPA	Pinyin	IPA
P	p^h	*q*	$tɕ^h$
T	t^h	*j*	tɕ
K	k^h	*x*	ɕ
C	ts^h	*z*	tz
Ch	$tʂ^h$	*r*	ʐ

One half of the problematic cases are formed by the five aspirated plosives and affricates *p*, *t*, *k*, *c*, und *ch*. 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*.

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. These are displayed in the following table:

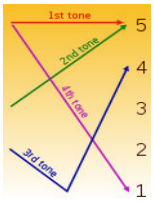
Pinyin	IPA
e	ɤ
*(s)**i***	ɿ
*(sh)**i***	ʅ
eng	ɛ̃

Germans often produce [ɹ] and [l] with too much jaw opening and in the case of [l] not enough retroflexed which might cause native speakers to perceive *e* [ɤ]. In addition, the slightly nasal [ɛ̃] of the final *eng* is often produced as [a], causing a percept of the final *ang,* or [ə], facilitating confusion with the final *en*.

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
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. Perceptual Experiment

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).

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].

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. Perceptual 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

[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.

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 1: *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 1 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.

Table 2*: 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***[1]: 2.53
ang	***ang***: 82.37	***an***: 10.79	***eng***: 4.47	
(s)i	***(s)i***: 83.95	***e***: 15.13		

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.

Table 3*: 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

In line with our expectations, single tones were generally produced correctly (see Table 3). 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 4 were those produced with the highest frequency of errors.

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

Table 4*: 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.

4. Results from the ASR System and Comparison with Human Judges

The recognition results generated by the ASR system were separated into initial, final and tone components and compared with the original tokens. Space limitations prevent us from presenting them in detail.

Table 5*: Percentage correct (second column) and confusion partners of finals (analysis of ASR System).*

ün	***ün***: 31.58	***in***: 21.05	***en***: 13.16 ***ing***: 13.16	***ue***: 7.89
ue	***ue***: 36.84	***i***: 21.05	***ui***: 13.16	***ing***: 5.26
an	***an***: 38.16	***en***: 13.16 ***un***: 13.16	***eng***: 10.53	***ang***: 7.89
ua	***ua***: 39.47	***uo***: 18.42	***uang***: 13.16	***a***: 10.53
üan	***üan***: 39.47	***en***: 13.16	***eng***: 7.89 ***ü***: 7.89	***e***: 5,26 ***i***: 5,26 ***ian***: 5,26 ***ün***: 5.26
eng	***eng***: 44.74	***uang***: 15.79	***en***: 7.89	***e***: 5.26 ***iao***: 5.26 ***ou***: 5.26 ***un***: 5.26

In general, the percentages correct are much lower than those assigned by the humans. However, the principle errors and confusion partners are very much in line with the annotations of the human perceivers. In addition to the expected confusion partners, the initials *z* and *r* exhibit conspicuously high error rates. *q* appears as the most frequent confusion partner of such unlikely phonemes as *c* and *z*. In the case of finals, the results diverge even more from the judgments of the native speakers as Table 5 illustrates. Tonal confusions detected are comparable to those found by the human perceivers, however, once again with generally much lower correct rates. In order to facilitate a comparison between the judgments of the native speakers on the one hand and the expert, as well as the ASR system on the other, we applied the same metric as in section 3.2.2 and calculated a score for each token. The scores of the native speakers were averaged for each of the tokens. The resulting percentages correct (maximum score of 3) are 55.8 for the expert, 23.0 for the averaged scores of the native speakers and only 16.7 for the ASR system. We performed correlation analysis of the scores and found a figure of .60 between averaged native speakers and expert, whereas the figure is as low as .15 between the ASR system and the expert, as well as the ASR system and the averaged native speakers, respectively. All these results are highly significant ($p < .001$).

5. 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 as erroneous only by the ASR system. 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. In addition, a trailing tone 3 often becomes tone 2, possibly because the rising part is exaggerated by the learners.

The results from the ASR system suggest a higher sensitivity and tendency of false hits. The agreement between the human judgment and that of the machine is surprisingly low, though more detailed analysis of deviations remains to be performed. In this context, recognition of Pinyin *q* appears to be especially problematic because it is often the most frequent confusion partner even in such unlikely cases as *c* and *z*. The high rate of confusion in finals indicates that these are more difficult to identify than initials or tones. Overall we have to bear in mind that the ASR system was not adapted to the likely set of errors expected from the German learners. Fine-tuning based on the result of this study will certainly improve its robustness and selectivity.

6. Acknowledgements

This work is funded by German Ministry of Education and Research grant 1746X08 as well as a DAAD-CSC project-related travel grant for 2009/2010.

7. 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.