Development of a Computer-Aided Language Learning System for Mandarin – Tone Recognition and Pronunciation Error Detection

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

hussein@beuth-hochschule.de, siwei@iflytek.com, mixdorff@beuth-hochschule.de, kuells@ beuth-hochschule.de, shugong@mail.ustc.edu.cn, gphu@iflytek.com

Abstract

This paper reports on the continued activities towards the development of a computer-aided language learning system for teaching Mandarin to Germans. A method for f0 normalization based on maximum likelihood estimation and tone recognition was implemented. Furthermore, a method for detecting the pronunciation errors was tested by calculating the confidence distance between the first and second candidates of the recognition system. In the first experiments we used an Automatic Speech Recognition (ASR) system with an acoustic model trained on data of native speakers of Mandarin. The performance of the ASR system was too poor because it was not adapted to the errors expected from the German learners of Mandarin. In the current experiment we modified the ASR system by considering the most frequent pronunciation errors committed by the German learners using a well-targeted replacement list for every phoneme and adaptation of the acoustic model using the correct data from German learners of Mandarin. The modified ASR system performs better than the original one, but stills falls short of the performance of the human judges.

Index Terms: Computer-Aided Language Learning (CALL), tone recognition

1. Introduction

In a globalized world the growing demand for foreign language competence stimulates activities towards computer-aided language learning (CALL). 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 [1]. 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 Ministry of Educations and Research which started over a year ago.

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. 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. Errors usually arise from phonemes of Mandarin without correspondences in German [2].

Mandarin is a tonal language. Tone is very important to distinguish Mandarin syllables, i.e. the tonal contour of a syllable changes its meaning. The tone distinction in Mandarin is the most complex problem for German learners. Mandarin has four syllabic tones and a neutral tone. Mandarin tone can be represented by prototypical f0 contours [3] as shown in Figure 1 [4]. The acquisition of tonal patterns of poly-syllabic words is much more difficult than mono-syllabic words [1].

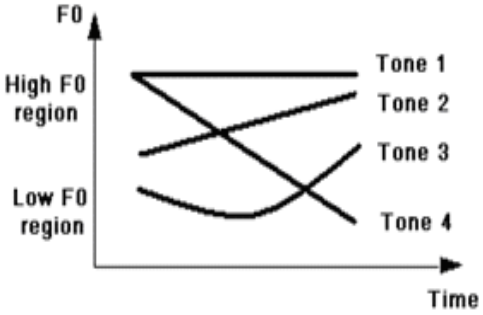

Figure 1: *Prototypical f0 contours of Mandarin tones.*

In the first experiments we analyzed the typical pronunciation errors committed by German learners of Mandarin [1][5]. We accomplished a contrastive analysis of syllabic components and determined the most probable confusion partners for initials, finals and tones. The Automatic Speech Recognition (ASR) system for Mandarin which was used in the first experiments used an acoustic model trained on data from native speakers of Mandarin. The results of this ASR system were not satisfying because it was not adapted to the errors expected from the German learners of Mandarin. Therefore, in this work we aimed to improve its performance by taking into account the most frequent pronunciation errors committed by the German learners, using a well-targeted replacement list for every phoneme and adapting the acoustic model using correct data from German learners of Mandarin. A method for detection of pronunciation errors was tested by calculating the confidence distance between the first and second candidates output by the ASR system. Furthermore we document the tone evaluation algorithm hitherto unpublished.

2. Experiment Method

This section describes the design of corpus and collection of data, the evaluation of data, the tone recognition algorithm, and the pronunciation error detection method based on the posterior probabilities of the first candidates output by the ASR system.

2.1. Corpus Design and Data Collection

The data used in this experiment is the same corpus used in [1] and [5]. The corpus recorded at Free University (FU) Berlin consisted of 54 tokens. One half of these had been produced by a female native speaker of Mandarin 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. 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.

The 54 tokens were produced by 19 first-year students (eight male and 11 female) of Chinese Studies at the East Asia Seminar of FU Berlin. At the time of the experiment they had completed 12 weeks of Mandarin language training. In addition to their regular classes, nine of the subjects had attended a weekly seminar of two hours as additional training. 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.

2.2. Data Evaluation

The collected data was annotated and processed threefold:

Expert (German teacher of Mandarin): The expert listened to the data several times and wrote down what he had perceived using Pinyin.

Ten female native speakers of Mandarin: The native speakers (listeners) were between 20 and 30 years of age. They were presented with each token one time and were requested to write down what they had perceived using Pinyin without prior knowledge of the intended target. Henceforth, we refer to both of expert and native speakers as human judges.

ASR system: The automatic speech recognition system which is part of an automated proficiency test of Mandarin [6] was used.

We used two versions of ASR system:

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. This version was used in the first experiments.

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 [1]. 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.

2.3. Tone Recognition

2.3.1. Characteristics of Mandarin Tone

Above all other features, the f0 contour reflects Mandarin tone. Judging from Figure 1, it seems easy to distinguish individual tones by their f0 contours. However, the task is much more difficult due to variations in speaker and style and most importantly, tonal coarticulation.

2.3.2. F0 Normalization Based On Maximum Likelihood Estimation

The most commonly used f0 normalization method is mean normalization, which is implemented via Eq. (1).

$$f' = f - \overline{f} \tag{1}$$

Where f and f' is the f0 value before and after normalization and $\overline{f}$ is the average f0 value of the person to be normalized.

Unfortunately, $\overline{f}$ could be biased if the words pronounced are not balanced. For example, if the syllables in a sentence mostly belong to tone1, then the estimation would be the mean of tone1, not the true f0 mean. In order to deal with this problem, we propose an f0 mean normalization based on maximum likelihood estimation, which is introduced in the following.

Suppose we get the tone models beforehand and at the same time we know the tone labels for the sentence (which is suitable in context-dependent pronunciation evaluation.). Then the mean normalization problem can be implemented via the following equation:

$$b = \arg\max_{b}(\log likelihood(O - b \mid M)) \tag{2}$$

Suppose the tone sequence from the given text to be $Tone_1^T$, the f0 feature sequence to be o_1^T. The tone models is set to be $\{M_i \quad 1 \le i \le N\}$, here N means the number of the tone model. At the same time, the mean normalization factor to be estimated is set to b, then using an expectation-maximization (EM) algorithm, we can get the auxiliary function corresponding to Eq. (2) as follows:

$$\begin{aligned} Q(b \mid \Lambda, o_1^T) &= \sum_j \sum_m \sum_{i=1}^{T} \gamma_{jm}(o_i) \log\left(\frac{1}{2\pi \mid \Sigma_{jm} \mid} \exp\left(-\frac{1}{2}(o_i - b - \mu_{jm})^T \Sigma_{jm}^{-1} (o_i - b - \mu_{jm}) \right) \right) \\ &= \sum_j \sum_m \sum_{i=1}^{T} \gamma_{jm}(o_i) \log\left(\frac{1}{2\pi \mid \Sigma_{jm} \mid} \right) - \frac{1}{2} \sum_j \sum_m \sum_{i=1}^{T} \gamma_{jm}(o_i)(o_i - b - \mu_{jm})^T \Sigma_{jm}^{-1} (o_i - b - \mu_{jm}) \\ &= constant - \frac{1}{2} \sum_j \sum_m \sum_{i=1}^{T} \gamma_{jm}(o_i)(o_i - b - \mu_{jm})^T \Sigma_{jm}^{-1} (o_i - b - \mu_{jm}) \end{aligned} \tag{3}$$

Here (μ_{jm}, Σ_{jm}) are the parameters from state *j* and Gaussian *m* of the tone model. $\gamma_{jm}(o_i)$ represents the posterior probability that the feature o_i was generated by the *jth* state

and the *mth* Gaussian from the tone model, which can be calculated with the Baum-Welch algorithm.
According to the auxiliary function, we can get the estimated f0 mean normalization factor *b* as follows:

$$\frac{\partial Q(b \mid \Lambda)}{\partial b} = \sum_j \sum_m \sum_{i=1}^{T} \gamma_{jm}(o_i)\Sigma_{jm}^{-1}(o_i - b - \mu_{jm}) = 0 \quad (4)$$

Then we get:

$$\left(\sum_j \sum_m \sum_{i=1}^{T} \gamma_{jm}(o_i)\Sigma_{jm}^{-1} \right) b = \sum_j \sum_m \sum_{i=1}^{T} \gamma_{jm}(o_i)\Sigma_{jm}^{-1}(o_i - \mu_{jm}) \quad (5)$$

Solving Eq. (5), we can get the optimal *b*, which is the estimator for f0 mean of a specific speaker.

2.3.3. Tone Recognition

After the optimal bias *b* is estimated, a hidden markov model (HMM) is used for model construction. Mono-tone models are used for monosyllabic words and bi-tone models for disyllabic words. Then tone recognition is done with these models.

2.4. Pronunciation Error Detection Based on the Posterior Probabilities

2.4.1. Phoneme-Level Posterior Probability Calculation

Phoneme-based forced-alignment was performed to determine segment boundaries. After that, phoneme-level posterior probabilities are calculated as follows:

$$P(M \mid O) = \frac{P(O \mid M)P(M)}{\sum_{M_i \in M_{set}} P(O \mid M_i)P(M_i)} \approx \frac{P(O \mid M)}{\sum_{M_i \in M_{set}} P(O \mid M_i)} \quad (6)$$

Here, O is the acoustic feature for phonemes (here: Mel-frequency cepstral coefficients - MFCCs), *M* is the given phoneme, and M_{set} is a predefined confusion matrix for phoneme *M*, which could contain all the phonemes or just be a compact confusion matrix of frequent mispronunciations by Germans.

2.4.2. Tone-Level Posterior Probability Calculation

In order to get accurate f0 means we combined eight monosyllabic words or four disyllabic words into a bigger token for every German student. Forced-alignment was performed to yield the syllable boundaries.
After that, tone-level posterior probabilities for monosyllabic words are calculated as follows:

$$P(T \mid O) = \frac{P(O \mid T)P(T)}{\sum_{T' \in T_{set}} P(O \mid T')P(T')} \approx \frac{P(O \mid T)}{\sum_{T' \in T_{set}} P(O \mid T')} \quad (7)$$

Here, O is the vector of tonal acoustic features (pitch, delta pitch and delta delta pitch), *T* is the given tone label, and T_{set} contains four tone models representing Mandarin tones.
For disyllabic words, posterior probabilities are calculated as follows:

$$P(T \mid O) = \frac{\sum_{T_i \in T_{set}^i} P(O \mid T_i)P(T_i)}{\sum_{T' \in T_{set}} P(O \mid T')P(T')} \approx \frac{\sum_{T_i \in T_{set}^i} P(O \mid T_i)}{\sum_{T' \in T_{set}} P(O \mid T')} \quad (8)$$

In this case, O is the vector of tonal acoustic features (pitch, delta pitch and delta delta pitch), *T* is the given tone label, and T_{set}^i contains four bi-tone models having the same central tone label *T*, and T_{set} contains all the bi-tone models.

2.4.3. Error Detection Method

After the posterior probability is calculated, error detection is performed as follows:

$$\begin{cases} Right & if\ Posterior \geq Thresh \\ Error & if\ Posterior < Thresh \end{cases} \quad (9)$$

$Thresh$ is the error detection threshold.
Eq. (9) is done for phoneme and tone level error detection with different threshold.

3. Results

In order to evaluate every syllable component individually the syllables of the original text, annotations of expert and the ten native speakers, and results generated by ASR systems were divided into initials, finals and tones. The annotations produced by the human judges served as a reference for judging the performance of the ASR systems.

3.1. Analysis of Syllabic Components

The annotations produced by the expert and the ten native speakers were compared with the results generated by the ASR systems and the target text. The results of the native speakers were averaged for each initial, final and tone. Each token was considered as correct if we found agreement between the expert or every native speaker, the ASR and the original text. The results of comparison are shown in Figure 2. The results of ASR2 are better than the results of ASR1 for initials and finals. The recognition of initials yielded better results than the recognition of finals. The results of tone recognition are the same, because no changes were made to the tone recognition algorithm.

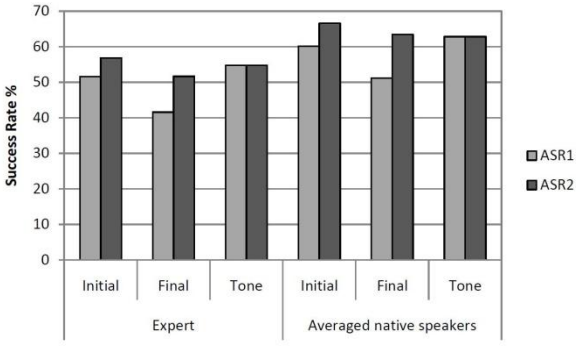

Figure 2: *Comparison of syllable components between the expert and the average of ten native speakers for ASR1 and ASR2.*

3.2. Pronunciation Error Detection

The annotations of human judges were used as a reference to evaluate the results of the ASR systems. The annotations of human judges were compared with the original text and the ASR results were compared with the annotations of human judges to verify the correctness. Our intention is ultimately to reproduce the assessment by the expert or the native listeners using the ASR system. To that effect we aim to keep the number of false hits - which would demotivate the learners - low, while maximizing the detection of true errors. In order to detect the correct results or errors from the ASR system we considered the following four cases: A (expert correct and ASR correct too), B (expert correct and ASR false), C (expert false and ASR correct) and D (expert and ASR false). The ratio of all tokens for the four cases by expert and averaged native speaker for ASR1 and ASR2 are shown in Figure 3. The case A represents the fully correct tokens and shows that more than half the tokens were correct. The correct tokens were improved in the ASR2 by the human judges. The cases B and C represent the cases in which the human judges and the ASR system are in disagreement. The number of tokens in the case of errors in ASR2 is smaller than in ASR1. The number of false tokens in the case B was minimized (5.09% and 7.25% by the expert and averaged native speakers respectively). In case D the evaluation of human judges and the result of ASR system are different. The ASR2 results are slightly greater than ASR1 in case D.

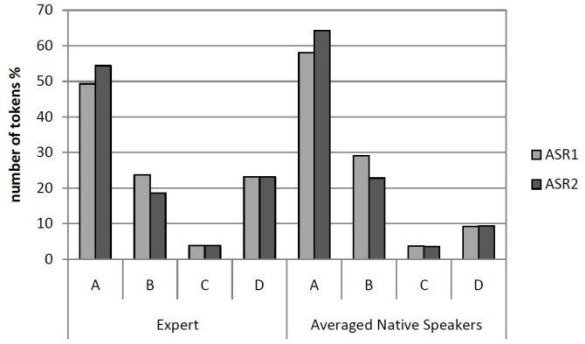

Figure 3: *The ratio of tokens for the four cases by the expert and the average of ten native speakers for ASR1 and ASR2.*

An analysis of the confidence distance between the first candidates of the ASR systems was performed 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. The mean values of confidence distance of initial, final and tone for every case are shown in Table 1.

Table 1. *The mean value of confidence distance between the first and second candidates of initial, final and tone for ASR systems.*

Reference	Case	ASR1	ASR2
Expert	A	0.5	0.7
	B	0.3	0.4
	C	0.4	0.6
	D	0.3	0.4
Averaged native speakers	A	0.5	0.7
	B	0.3	0.4
	C	0.4	0.6
	D	0.3	0.4

The results of expert and the average of ten native speakers are the same. The mean of confidence distance in the ASR2 are greater than in ASR1. The confidence distance between the first correct candidate and the second candidate (case A) is large (0.7). This means that the ASR system can be able to recognize the correct tokens as true and not as false. In the other cases the confidence distance between candidates is not small and shows that the recognition of falsely flagged tokens is not an easy task.

4. Conclusions

This paper reported on the continued activities towards the development of a CALL system for teaching Mandarin to Germans. The ASR system, which was used in the first experiment, was modified in this experiment to take into account the most frequent pronunciation errors committed by the German learners by adapting the acoustic model using the correct data from German learners of Mandarin and using a well-targeted replacement list for every phoneme. The results of both ASR systems were evaluated and compared with the annotations of a reference. The annotations produced by the expert and the native speakers used as a reference for judging the performance of the ASR systems. The annotations of native speakers were averaged. The experimental results showed an improvement in the performance of the modified ASR system. The ratio of correct initials is greater than finals. An analysis of the confidence distance between the first and second candidates output by the ASR systems was performed in order to detect the pronunciation errors. 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-CSC (Germany/China) and DAAD-NSC (Germany/Taiwan) project related travel grants for 2009/2010.

6. References

[1] Mixdorff, H., Külls, D., Hussein, H., Gong, S., Hu, G., 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.

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

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

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

[5] Mixdorff, H., Külls, D., 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.

[6] 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.