Language Teaching, Learning and Technology
6-7 Sep 2016, San Francisco, USA

Visualization of Mandarin Chinese Tone Production of Japanese L2 Learners for evaluation

Nicolas Loerbroks[1], *Yue Sun*[1], *Yoshinori Sagisaka*[1], *Jinsong Zhang*[2]

[1]Graduate School of Fundamental Science and Engineering, Waseda University, Japan
[2]College of Information Science, Beijing Language and Culture University, China

nicolas.loerbroks@posteo.de, yue.cherry.sun@gmail.com
ysagisaka@gmail.com, jinsong.zhang@blcu.edu.cn

Abstract

Aiming at automatic characterisation and evaluation of second language (L2) learners production of Mandarin Chinese tones, we applied robust F0-features previously used to characterize and visualize the tone control of native Chinese speakers. They consist of the average height and the average slope of the contour, which form a minimal set of F0-features to efficiently separate the four Mandarin tones. The resulting two-dimensional scatterplots represent L2 learner's characteristics of tone production very well, in particular the confusion between tone 2 and tone 3 which has been well known as a L2 learning problem of Chinese tone production. This analysis could be carried out completely automatically using open source tools. To further confirm that the information contained in those two features is sufficient to reflect native's perception, we used a neural network (NN) to predict natives evaluation of utterances of tone 2 and tone 3 based on the two features. The experiment showed a reasonable high correlation to natives subjective rating, which confirmed the availability of those two features to evaluate L2 proficiency of Chinese tone production.

Index Terms: L2, Mandarin Chinese, tone production, F0-features, evaluation, visualization

1. Introduction

For second language (L2) learners of Mandarin discriminating the four Mandarin tones, in perception as well as in production, is widely known to be a difficult task that requires intensive training which makes it an important problem in the field of Mandarin L2 education. In particular the discrimination of T2 and T3 is known to be the most difficult for, but not limited to, Japanese learners [2][3].

Having a method to automatically evaluate learners tone production capabilities is hence very desirable in order to easily give feedback to learners and decrease the workload for teachers. One common approach to this automatic evaluation is to train a machine learning system on a variety of F0 features in order to derive a score which has lead to very good results [4]. However, rather than automatically rating single utterances, we are aiming at analyzing and visulalizing the overall patterns of tone production of L2 learners in order to enable learners and teachers to interpret the results by themselves. This visualization could not only be used to assess the learner's level but also to understand how one might have to improve their tone production. For this purpose we need efficient extraction of F0-features that are small in number and easy to interprete.

The paper is structured as follows. First we give a detailed explanation of the employed F0-features in the following Section 2 before describing the methodology and the results of the visualization experiment in Section 3. We describe and discuss our verification approach of the feature-parameters through NNs in Section 4 before concluding the paper in Section 5.

2. Feature parameters for L2 tone control

The aim of this study is to propose a set of F0 features that are not only easily interpretable in terms of tone production characteristics but also separate the four Mandarin tones efficiently and are limited enough to allow visualization. Making use of a large amount of F0-features that contain all the pitch information as well as extracting a smaller set of features through an unsupervised learning approach is impractical for this purpose.

To efficiently characterize and visualize Chinese tones in this way two features, average F0 height and average F0 slope, have previously been proposed by Fujisaki et.al. (1990) and Peng (2006) for automatic tone recognition and phonetic analysis [5][6]. It can easily be seen why those features can be expected to be very robust in order to separate the four Mandarin tones. T2 is characterised by a positive and T4 by a negative average slope. T1 and T3 are both characterised by a slope around zero but different average F0 levels, and the four tones can therefore be expected to fall into 4 distinct groups (see Table 1). However, to the best of our knowledge, those features have not yet been applied to L2 speech. Nevertheless, it can be expected that they can not only be applied to investigate general patterns of tone production for L2 learners of different levels and language backgrounds, but can also give an insight into how an individual learner might be able to improve tone production. Afterwards, as the first step to confirm the validity of the approach, we further investigated the continuum between T2 and T3 by comparing natives perception to the performance of an artificial neural network that takes the two features as the input values.

Table 1: *Separation of the Mandarin tones based on Avg. F0 level and Avg. F0 slope*

	Avg F0 height	Avg F0 slope
Tone 1	high	zero
Tone 2	middle	positive
Tone 3	low	zero
Tone 4	middle	negative

http://dx.doi.org/10.21437/LTLT.2016-1

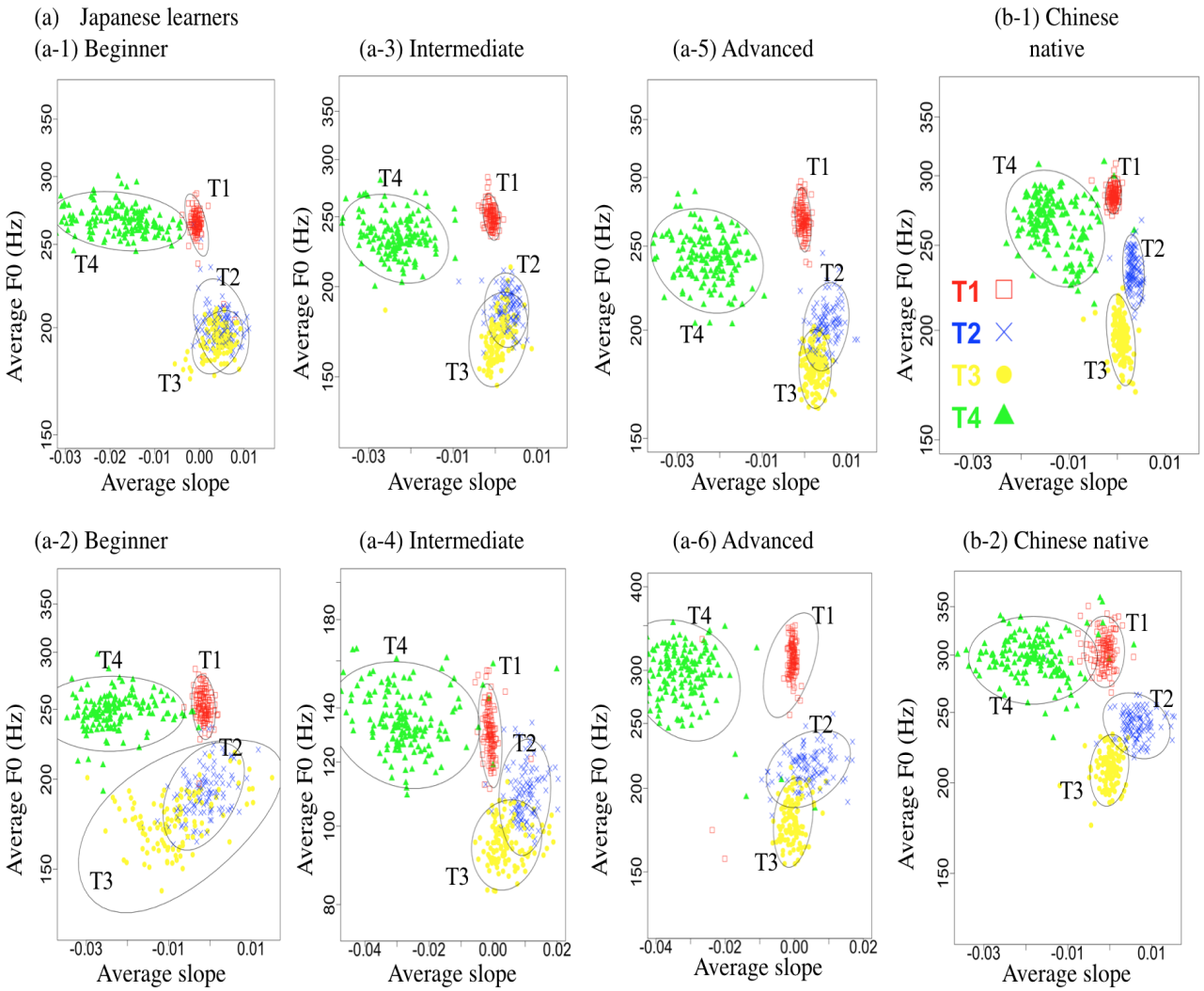

Figure 1: An example of distributional differences of average F0 height and average F0 slope between Japanese learners (beginner, intermediate, advanced) and Chinese natives. Each scatterplot represents a single speaker

3. Tone control visualization

In order to visualize the tone control of Japanese learner's of Mandarin we first extracted the features for monosyllabic utterances by multiple Japanese learners of Mandarin and visualised them using 2-dimensional scatterplots. The same was done for multiple Chinese natives for comparison. Confidence interval ellipses were added to each tone for further illustration and analysis.

3.1. Experimental setup

3.1.1. F0-feature extraction

First the F0 contour of all the monosyllabic utterances was automatically extracted at a sampling rate of 10ms using the Praat script ProsodyPro [7]. The F0-values were then transformed into log-scale and the first parameter, the average F0 height, was given by the average F0-value of the voiced part of the utterance in log-scale. The second parameter, the average slope was given by the slope value of the linear regression (again in log-scale). All the calculations were carried out in the open source statistical programming language R.

3.1.2. Calculation of confidence interval ellipse

Following the well established methodology of Peng (2006) and Zhou and Xu (2007), we created confidence interval ellipses around the 2 dimensional data points coming from each utterance for further analysis [6][8]. In order to calculate those tonal ellipses we performed a principal component analysis (PCA). The orientation of the ellipse was then given by the two principal components and the data points were projected onto the two axes. Their radii were then set to twice the standard deviation of the projections. Calculated in this way the ellipses should enclose 95% of utterances of the given tone by the given speaker. Figure 1 shows the resulting scatterplots for 2 Chinese native speakers as well as 6 Japanese learners as examples.

3.1.3. Speech samples

The subjects were 10 native speakers of Mandarin Chinese, 5 male and 5 female as well as 10 Japanese learners of Mandarin, 7 female and 3 male who were studying at Beijing Language

and Culture University (BLCU) at the time. Each subject was asked to utter the same 520 syllables consisting of all four tones.

3.2. Observation from the Scatterplots based on Average F0 height and Average F0 slope

Several differences of tone production characteristics between Chinese native speakers and L2 learners can be identified by visualising them using the two F0 features of average height and average slope. Those findings correlate very well with the commonly known problems of tone control by Japanese learners. Figure 1 shows the differentiability of the 4 tones based on the two features for 6 Japanese learners in the beginning, intermediate and advanced stages and 2 Chinese natives as examples.

3.2.1. Differences between natives and L2 learners

We can see that the four tones are clearly separated for the Chinese natives. Especially when looking at speaker b-1 we can see that the Chinese natives show a tendency towards smaller tone ellipses representing their consistent tone production. For the Japanese learners the distribution of the data points is clearly different compared to the Chinese natives. The Japanese learners can separate the tones clearly except for obvious severe problems between T2 and T3. This represents well known facts in the field of Mandarin L2 education and gives us a first indication that the chosen features can reflect the difficulties in tone production for Japanese learners.

3.2.2. Differences between the L2 learners

We can observe two reasons for the overlap between T2 and T3 from Figure 1. For some Japanese speakers the overlap of the tonal ellipses for T2 and T3 seems to be more caused by the large size of the ellipses representing an inconsistent tone production (e.g. a-2 and a-4). For other speakers the tonal ellipses for T2 and T3 are small but their center points are more close to one another (e.g. a-1 and a-3) which reflects the fact that the learners have accustomed themselves to a consistent but almost interchangeable tone production for T2 and T3. In the first case the learner can be expected to have understood the conceptual differences between the 2 tones but has to work on the consistency. In the second case the learner would have to be trained on the concepts of how to separate the two tones. Two different scores could therefore be derived to assess the focus of further training.

3.3. subconclusion

In this section we proposed a methodology to visualize L2 speakers tone control through two F0-features given by the average F0 height and the average F0 slope of the contour. The resulting scatterplots do not only seem to represent the general problematic patterns of L2 tone production well but can also give an insight into the characteristics of individual learners.

In the following experiment we are aiming at further verifying that these two features contain enough information of the F0-contour to reasonably reflect native's perception of tone production using a machine learning approach.

4. Feature verification

We want to investigate how well the two F0-features, average F0 height and average F0 slope, and the resulting scatterplots correlate with natives perception of the learner's tone production. It is a very well known fact that discriminating T2 and T3 is the most difficult part of tone acquisition of Mandarin Chinese. This is in particular the case for Japanese learners of Mandarin which is very well reflected in the scatterplots. It is for this reason that we focused our further analysis on the overlap between T2 and T3.

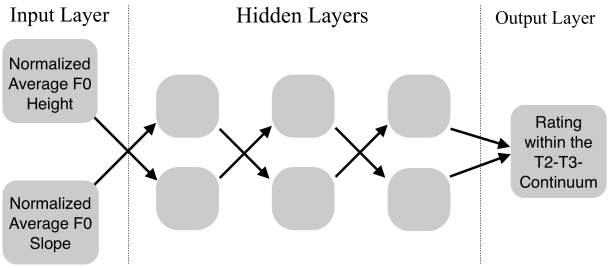

Figure 2: Architecture of the NN used for feature verification

Native Chinese listeners were asked to evaluate utterances of T2 and T3 by Japanese learners and a machine learning system was trained on this evaluations using the two features as the input. We used the performance of the system compared to one trained on 10 time-normalized F0-values to analyze the effectiveness of Average F0 height and average F0 slope to characterize the utterances of mandarin tones.

4.1. Analysis of the continuum between T2 and T3 of Japanese learners

Native Chinese subjects were asked to rate utterances on a scale from 1 to 5 representing how clearly it belonged to either the category of T2 or T3. The subjects can be expected to have a significant correlation since previous research indicates that native Chinese speakers perceive tones by the change of the F0-contour and that this perception is particularly sensitive in the boundary between the two categories of T2 and T3 [9]. We trained a NN on the same rating task using the two F0-features as the input. The correlation between the output of the NN and native's rating gives us an indication of how much information is contained in the two F0-features and therefore how well the scatterplots can represent a learner's tone production characteristics.

4.2. Experimental setup

4.2.1. Rating system within the categories of T2 and T3

We randomly chose 25 monosyllabic utterances of an intended T2 and 25 utterances of an intended T3 for all 10 Japanese subjects, 500 samples in total. We then asked 5 native speakers of Mandarin to listen to those utterances and rate them on a scale of one to five with one being an utterance perceived as a perfect native-like T2 and five being a perfect native-like T3. A rating of three was given if it was perceived as in between T2 and T3 and a rating of two or four was given if the perception was leaning more towards T2 or T3, respectively. If the utterance was perceived as either a T1 or T4 or not as a Mandarin tone the subjects were asked to mark the utterance. All utterances marked by at least one of the subjects were removed before further analysis. The Chinese subjects did not know which tone the learners had intended to produce.

4.2.2. *Chinese subjects for the rating experiment*

5 native Chinese subjects participated in the experiment. Two of the subjects were PhD-students with a background in phonetics and L2 teaching living in Tokyo. One of the other subject was also living in Tokyo being exposed to L2 Chinese speech by Japanese speakers on a daily basis. The last two subjects were living in Beijing with minimal exposure to foreign languages or L2 Chinese speech. The subject's average age was 28.4, three were female and two male.

4.2.3. *Training of the NN to predict Natives' Evaluation*

We normalised the two F0 features for the selected utterances into the interval [0,1] for each speaker and trained a feed-forward artificial neural network with the normalized features as the input and the rating of the two Chinese phonetic experts as the output. In order to get unbiased training data, 20 input and output pairs for each rating were randomly chosen, 100 training sets in total. The remaining rated utterances were left for testing. Only ratings on which the two Chinese experts had agreed were used for training. The NN was trained using the R package "neuralnet" with resilient backpropagation with weight backtracking and a logistic activation function [10]. The threshold was set to 0.1 and, the stepmax was set to one million and 50 repetitions with randomly initialized weights and different training/testing-sets were calculated. The architecture of the neural network is shown in Figure 2. The output of the NN was rounded to the closest integer to achieve the same rating system as for the human perception experiment. From here on we are only referring to the NN out of the 50 repetitions that showed the highest correlation on average to the ratings of the native Chinese subjects.

Another neural network was trained in the same fashion using ten time-normalized F0-values as the input (Full-Feature NN). The performance of those two neural networks was compared in order to analyze how well the limited Feature set, Average F0 height and Average F0-slope, performs in contrast to the more traditional approach.

Table 2: *Average Correlations of the ratings within the T2-T3-continuum between Native speakers and Neural Networks*

	Correlation
Between Chinese Phonetic experts	0.839
Between All Chinese listeners	0.701
Between Full-Feature NN and Chinese listeners	0.503
Between 2-Feature NN and Chinese listeners	0.422

4.3. Results and Discussion

Through this experiment we would like to confirm that the two features contain enough information to base accurate interpretations of a learner's tone production on them. Table 2 shows the average correlations of the rating within the T2-T3-continuum between the native Chinese listeners and the outputs of the neural networks. As expected the correlation between the native Chinese listeners is significantly high with the two phonetic experts reaching a correlation of 0.84, while the correlation to and between the other subjects is less than 0.7 on average, which confirms the validity of our verification approach. The neural network trained on the ratings of the two phonetic experts reaches a correlation to the other subjects of 0.45 and 0.42 on average while the neural network trained on 10 time-normalized F0-features reaches 0.55 on average.

We would like to point out that it is not possible for Chinese phoneticians to rate an utterance in the proposed way just by looking at the F0-contour, without any acoustic clues. Hence a correlation of 0.55 for a machine learning system based on F0-features is a very reasonable result. The neural network based on the two features, average F0 slope and average F0 height, reaching a correlation of 0.45, therefore provides us with a clear indication that statistical analysis based on those two features is a valuable approach to gain insight into L2 Chinese tone production.

5. General Conclusion and Future Work

In this paper we proposed a method to analyse L2 learners production of Mandarin tones. By applying two robust F0-features, average F0 height and average F0 slope, that had shown to be useful to characterize the tone production of native speakers of Mandarin, we described and visualized the tone production patterns of Japanese learners in a two-dimensional space. The analysis based on the resulting 2-dimensional scatterplots correlates very well with the common findings on problematic patterns of tone control of Japanese learners of Mandarin. In particular the problem of distinguishing tone 2 and tone 3 could be very well observed. This visualization does not only perform well at capturing the general patterns of tone production for Japanese learners of Mandarin but also has the potential to give an insight into assessing the level of an individual learner's tone production, and to give indications on how one might be able to improve.

All the analysis can be carried out completely automatically using the open source softwares Praat and R, and could therefore be carried out by teachers and learners by themselves, when made available in the future.

To reconfirm the explanatory power of Average F0 height and Average F0 slope for the visualization of L2 learner's tone control, we trained a NN to predict natives' ratings of utterances belonging to the continuum of tone 2 and tone 3 based on the two features. The NN showed up to 0.45 correlation to natives judgement compared to 0.55 for a NN with ten input features and 0.7 correlation among different native speakers which gives an indication that the features are sufficient to describe the problems of tone production for a Japanese learner of Mandarin.

In the future this research needs to be extended to the analysis of speech samples of L2 learners from different language backgrounds to confirm that the features also work well at capturing problems of distinguishing different tone pairs.

This research was carried out based on monosyllabic utterances and will need to be extended to multisyllabic utterances and running speech which usually poses as a much bigger challenge to L2 learners than monosyllabic utterances. However, in that case, the influence of adjacent syllables and the constraint of the declining phrase command on the average F0 height will have to be taken into consideration.

6. References

[1] Chao. Y.R, "A grammer of spoken Chinese", Berkeley: University of California Press. 1968

[2] A.H. Wang, "On Teaching Chinese Tone Features", Language Teaching and Linguistic Studies, vol.3, pp70-75, 2006

[3] Hussein Hussein, Hue San Do et al. "Mandarin Tone Perception and Production by German Learners", Slate 2011

[4] Jiang-Chun Chen, Jyh-Shing Roger Jang, and Te-Lu Tsai, "Automatic Pronunciation Assessment for Mandarin Chinese: Approaches and System Overview", Computational Linguistics and Chinese Language Processing, Vol. 12 pp. 443-458, 2007

[5] Changfu Wang, Hiroya Fujisaki, Keikichi Hirose, "Chinese four tone recognition based on the model for process of generating F0 contours of sentences", ICSLP 1990

[6] Gang Peng, "TEMPORAL AND TONAL ASPECTS OF CHINESE SYLLABLES: A CORPUS BASED COMPARATIVE STUDY OF MANDARIN AND CANTONESE", Journal of Chinese Linguistics Vol. 34, No.1, 134-154, 2006

[7] http://www.homepages.ucl.ac.uk/ uclyyix/ProsodyPro/

[8] Ning Zhou and Li Xu (2007), "Development and evaluation of methods for assessing tone production skills in Mandarin-speaking children with cochlear implants", J. Acoust. Soc. Am. 123, 1653- 1664

[9] Jinson Zhang, Yue Sun et al, "Improve Japanese C2L capability to distinguish Chinese tone 2 and tone 3 through perceptual training", Oriental COCOSDA 2013

[10] https://cran.r-project.org/web/packages/neuralnet/neuralnet.pdf