PROSODIC CORRELATES OF VOICE PREFERENCE IN MANDARIN CHINESE AND GERMAN: A CROSS-LINGUISTIC COMPARISON

Hongwei Ding[1,2], *Rüdiger Hoffmann*[2], *Oliver Jokisch*[3]

[1]*Institute of Cross-Linguistic Processing and Cognition, School of Foreign Languages, Shanghai Jiao Tong University, China,* [2]*Chair for System Theory and Speech Technology, TU Dresden, Germany,* [3]*Institute of Communications Engineering, HfT Leipzig, Germany*
hwding@sjtu.edu.cn, ruediger.hoffmann@tu-dresden.de, jokisch@hft-leipzig.de

Abstract: To find a pleasant voice for speech synthesis is of vital importance to the system. In this study we investigated the influence of prosodic parameters on voice preference from a cross-linguistic perspective. We conducted two voice preference tests. 24 German and 50 Chinese listeners took part in the first test on German speaker selection; while 25 Chinese and 10 German listeners participated in the second one on Chinese speaker selection. Then we employed Momel algorithm to extract 18 purely acoustic parameters, which are supposed to capture the prosodic pitch change patterns of these speakers. The results showed that there were strong correlations between the ranking scores of German and Chinese listeners on both German and Chinese speakers. However, the Chinese listeners showed more correlations between their voice preference and the melody metrics than the German listeners. The Chinese listeners preferred larger pitch changes in Mandarin Chinese. The experiment results suggested that a pleasant voice can have some prosodic characteristics, which makes an agreeable impression on both native and non-native listeners. But listeners may also rely on information of other prosodic parameters (such as duration, speech rate, intensity) as well as phonation types and formant spaces in the selection of a preferred voice, which will be investigated in the future. The current study may shed some light on talent voice selection in a speech synthesis system.

1 Introduction

This study was motivated by the voice casting for the multilingual speech synthesis system. Speaker casting is an indispensable procedure in inventory construction. We carried out the speaker casting in American English, British English, German, Chinese, Dutch, Spanish, Italian and French, and we have found that the favorite speakers usually possess some characteristics which are enjoyable for both native listeners and non-native listeners who have no knowledge of the concerned languages.

Listeners usually associate the voice with the personal features of the speaker, and the spoken utterances thus convey important information affecting the preference of the listeners. It has been demonstrated that several acoustic cues are related to voice attractiveness. Human voice pitch is one of the most important acoustic features that affect the impression of the personality of the speaker. Voice pitch is the perceptual correlate of fundamental frequency (F0), which reflects the rate of vocal fold vibrations. The vocal folds of men are longer and thicker than those of women, thus the F0 of men was found 50% lower than that of women [1]. Due to the fact that adult males have an average lower F0 than adult females and children, listeners usually associate pitch with sex, age and even personality. Many empirical studies have found

that women prefer lower-pitched men's voices, which are correlated with strength, competence and dominance [2, 3]; while men typically prefer higher-pitched women's voices, which are perceived as feminine and youthful [4]. However, attractive voices are distinct from pleasant ones [5]. Normally an agreeable and enjoyable other than an appealing voice is demanded for a text-to-speech (TTS) system. Contrary to the research findings on attractiveness of the voice, it has been reported that an average higher F0 of the female speakers might lead to a lower listening preference in TTS voice casting [6].

Moreover, not only the voice pitch but also pitch change patterns are associated with personalities of the speaker. On hearing a short utterance of a novel voice, listeners readily form personality impressions of the speaker based on the acoustic cues [7]. By changing prosodic parameters, such as pitch level, pitch range, articulation rate and loudness in synthetic speech, the personality features of the voice can be modelled [8].

In summary, previous studies have shown that the average pitch values and pitch change patterns can influence the preference of the voice [9]. Mean F0 values are mainly determined by speaker's anatomy and physiology and might be language independent, but prosodic pitch changes may be language- and culture-dependent. Regarding few researches have been devoted to the investigation of correlation of pitch and pitch changes with voice preference from a cross-linguistic perspective [6], this study was a preliminary endeavor towards this direction. Furthermore, it is also interesting to investigate this relationship between tone and non-tone languages in the current study.

2 Method

First we conducted voice preference tests with German and Chinese listeners on both German and Chinese speakers. Then we extracted the melody metrics from these sentences using Momel algorithm [10], which are purely acoustic parameters. And finally we correlated the preference rankings between German and Chinese listeners, and their preference scores with the pitch-related parameters.

2.1 Data collection

The data collection procedure includes speech data recording, preference-based ranking elicitation, melody metrics extraction, and calculation of related correlations.

2.1.1 Speech material

Both Mandarin Chinese and German speech data were taken from the recordings in the voice casting from candidate speakers for a multilingual speech synthesis. All the candidates were female speakers with the age between 22-39 years old. The Mandarin Chinese recordings consisted of three sentences in technical communication style and three sentences in fairy tales style from each seven speakers, resulting in 42 sentences. The German recordings consisted of two sentences of neutral speaking style from each nine speakers, resulting in 18 sentences.

2.1.2 Preference scores

The listeners in the preference test for Mandarin Chinese speakers were 25 Chinese and 10 German who had no knowledge of Chinese, and the male-female participants were balanced. Whereas listeners for German preference test were 24 German and 50 Chinese who had no knowledge of German, and the male to female ratio in both German and Chinese listeners was 1:1. All the Chinese and most of the German listeners were university students.

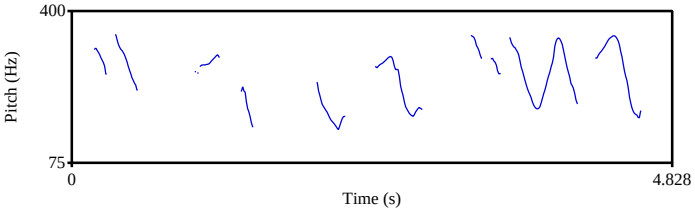

Figure 1 – F0 change patterns of a typical Chinese declarative sentence in the preference test

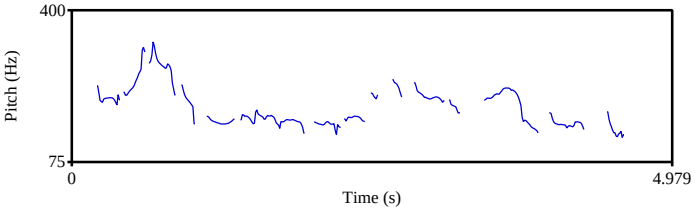

Figure 2 – F0 change patterns of a typical German declarative sentence in the preference test

The preference test was implemented by a Praat script, and carried out on the computer by listeners individually. Each time the same sentence from two speakers was played, the listener should choose which one was preferred by clicking a button of A or B. After the decision was made, the next sentence pair appeared.

If every sentence of each speaker were compared with that of all other speakers, too many pairs would be resulted, which might be confusing for the listeners. Thus we reduced the listening stimuli to reasonable numbers. For preference test on Chinese speakers, one sentence of each speaker should be compared with other two speakers. Therefore, 84 pairs of sentences were constructed. The same procedure was applied to the preference test on German speakers. Each speaker appeared twice, and each time the speaker was compared with another one; therefore, 18 pairs of sentences were resulted.

2.2 Acoustic data extraction

In this section we illustrate the different pitch change patterns of German and Mandarin Chinese visually, introduce the melody metrics extracted by Momel algorithm, and demonstrate the distinguishing ability of this algorithm.

2.2.1 Pitch change patterns

The different pitch change patterns between German and Mandarin Chinese can be perceived and displayed as well. As a tone language, Mandarin Chinese exhibits clear F0 contours in each lexical syllable in Figure 1, but the sentence intonation is not discernible here. As a non-tone language, German displays a typical F0 contours of hat pattern for a declarative sentence in Figure 2.

2.2.2 Prosodic parameters

The algorithm employed here is described by Hirst in [10]. The anchor points were scaled using the OMe (Octave-Median) scale with Formula (1) to reduce the inter-subject variability:

$$\mathrm{OMe} = \log_2(\mathrm{Hz}/\mathrm{Median}) \quad (1)$$

where *median* is the median value of F0 for the whole sentence.

From the anchor points the mean and standard deviation was calculated for:

1. Pitch value of all anchor points on the OMe scale (pitch_m, pitch_sd)
2. Pitch value of anchor points which are higher than the previous one (high_m, high_sd)
3. Pitch value of anchor points which are lower than the previous one (low_m, low_sd)
4. Interval absolute difference from previous point (interval_m, interval_sd)
5. Rise difference from previous point (rise_m, rise_sd)
6. Fall difference from previous point (fall_m, fall_sd)
7. Absolute difference from previous point divided by distance in seconds (slope_m, slope_sd)
8. Rise slope from previous point (rise_slope_m, rise_slope_sd)
9. Fall slope from previous point (fall_slope_m, fall_slope_sd)

The 18 parameters (9 means and 9 standard deviations) were collected for each sentence. The German recording contained 9 speakers and every speaker read 2 sentences, which resulted in 18 sets of values. The Chinese recording contained 7 speakers and every speaker read 6 sentences, which resulted in 42 sets of values. Since these values have been offset to the speaker's median F0 by Formula (1), they can be compared across speakers.

2.3 Linear discriminant analysis

It is interesting to test whether these 18 parameters can distinguish the German utterances from the Chinese ones. It was found that the discrimination rate was 100%. It has been reported that these melody metrics can not only distinguish different prosody patterns between languages but also differentiate different prosodic patterns among speakers of the same language.

3 Results

Results are presented in correlations in voice preference rankings and acoustic correlates of listeners' perceptual ratings.

3.1 Preference correlation

Results of preference tests are presented on German and Chinese speakers separately.

3.1.1 German speakers

The correlation between German and Chinese listeners was significant (r=0.73, p=0.03), which can be observed in Figure 3. The correlation between males and females was also significant among both Chinese listeners (r=0.81, p= 0.008) and German listeners (r=0.81, p= 0.009). (Note: in Figure 3 and Figure 4, the labels on x-axis represent the speaker identification, and the scales on y-axis show the ranking in percentage. While "scores_de" and "scores_cn" are the preference rankings evaluated by German and Chinese listeners separately.

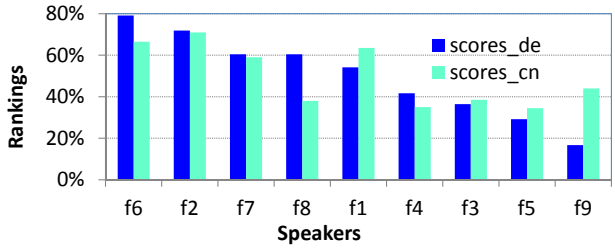

Figure 3 – Comparison of rankings on German speakers between German and Chinese listeners

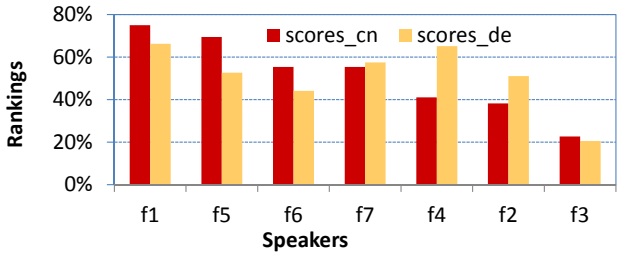

Figure 4 – Comparison of rankings on Chinese speakers between German and Chinese listeners

The mean F0 values for these nine speakers were between 165Hz-196Hz. F0 values of the preferred speakers for German listeners were 171Hz, 182Hz, 165Hz, and 193Hz; while those for Chinese listeners were 182Hz, 171Hz, 168Hz, and 165Hz. Among the four most preferred speakers, three were the same. They agreed on the best two speakers, but with a different order.

3.1.2 Chinese speakers

The correlation between German and Chinese preference rankings is also strong but not significant for both fairy tales (r=0.53, p=0.22) and technical communication (r=0.62, p=0.13). We averaged the ranking values across these two styles (r=0.531, p=0.22), which is illustrated in Figure 4.

The mean F0 values of these seven speakers ranged from 198Hz to 250Hz. The mean F0 values of the preferred speakers for the Chinese listeners were 240Hz, 198Hz, 242Hz, and 206Hz; while those for the German listeners were 240Hz, 250Hz and 206Hz. They agreed on the best speakers.

3.2 Acoustic correlation

For the convenience of comparison, in Table 1 and Table 2 the correlation coefficient is represented by "-" if it is smaller than 0.5 and is written in bold face if the correlation is significant.

3.2.1 German speakers

The relationship between the rankings and prosodic parameters on German speakers can be found in Table 1, where "Cn", "Cn-m", "Cn-f", "De", "De-m" and "De-f" represent "Chinese", "male Chinese", "female Chinese", "German", "male German" and "female German", respectively. Because the female-male ratio was exactly 1:1, we correlate the rankings of females and males separately for both Chinese and German listeners.

Table 1 – Correlation on German speakers

Metrics	Chinese listeners			German listeners		
	Cn	Cn-m	Cn-f	De	De-m	De-f
median	-0.60	-	**-0.74**	-	-	-
pitch_sd	-	-	-	-	0.52	-
high_m	-	-	-	0.52	0.59	-
high_sd	-	-	-	0.62	0.65	-
interval_m	**0.87**	**0.82**	**0.84**	-	-	0.55
slope_m	**0.68**	0.60	**0.67**	-	-	-

3.2.2 Chinese speakers

The correlation between the preference scores and prosodic parameters on Chinese speakers can be found in Table 2, where "tech. comm." represents "technical communication". We didn't separate males from females due to the small number of German listeners. But we distinguished the sentences of fairy tales from those of technical communications because the speaking styles were different, thereby resulting in different pitch change patterns.

Table 2 – Correlation on Chinese speakers

Metrics	Chinese listeners		German listeners	
	fairy tales	tech. comm.	fairy tales	tech. comm.
pitch_m	-	-0.58	-	-0.72
pitch_sd	0.63	0.71	-	-
high_m	**0.82**	-	-	-
low_m	0.61	0.63	-	0.60
low_sd	0.69	-	-	-
interval_sd	**0.83**	0.67	-	-
rise_m	**0.90**	0.66	-	-
rise_sd	-	-	-	0.55
fall_m	**0.73**	0.54	-	-
fall_sd	-	**0.76**	-	-
slop_sd	**0.87**	0.52	-	-
rise_slope_m	**0.85**	0.56	-	-
fall_slope_m	**0.85**	-	-	-
fall_slope_sd	0.63	-	-	-

From these two tables several results can be observed:

1. More correlations could be found in Chinese listeners than in German listeners between preference rankings and melody metrics.

2. For the Chinese listeners, more correlations could be identified in Chinese than in German, and more in fairy tales than in technical communication styles.

3. The Chinese listeners associated larger and quicker pitch changes with a higher preference of Chinese speakers, especially in reading styles of fairy tales, because higher correlation coefficients were found between their preference scores and rising and falling intervals and slopes in fairy tales.

4. For female Chinese listeners, higher F0 values were associated with a lower preference of German speakers.

4 Discussion

There is something in voice that can help naive non-native listeners to make similar selections of a pleasant voice to those of native speakers. For native speakers, a standard segmental pronunciation and an appropriate suprasegmental prosody should be essential for their selection of the preference. However, for non-native listeners who have no knowledge of the concerned language, neither the accuracy of segmentals nor the appropriateness of linguistic aspects of prosody can help. What impressed them is the paralinguistic or non-linguistic information conveyed by the prosody and voice quality of the speaker. And their criteria might further be influenced by their native language and culture. Despite all these discrepancies we still obtained strong correlations between German and Chinese listeners on the rankings both on German and Chinese speakers. Several conclusions can be drawn on the basis of our results:

1. Paralinguistic and non-linguistic prosody information is important to influence the preference of a speaker.

2. Compared with German listeners, Chinese listeners associate more pitch-related metrics to the preference of speakers, and more in tone languages such as Mandarin Chinese than in non-tone languages such as German.

3. Chinese listeners usually prefer speakers with larger pitch changes in Chinese. The reason might be that smaller pitch changes might be monotonous in Chinese but not in German.

4. Chinese female listeners prefer German female speakers with a lower-pitched voice, which is consistent with the findings in [6, 11].

5. It cannot be proved in this study that male listeners prefer high-pitched female voices because other prosodic parameters outweigh the average F0 values in the natural speech. But Chinese female listeners show a negative correlation between their preference and median F0 of German female speakers, but the males do not. This may indicate that males are not against female speakers with a higher F0.

Numerous factors are at play in the selection of a preferred voice. Generally speaking, segmentals, suprasegmentals, and phonation types [12] can all influence the preference of a voice. The current study only concerns the pitch-related parameters. In addition to pitch patterns, duration and intensity patterns can also influence the prosody of speech. All the prosodic parameters are integrated together to form the personality of the speakers. In the future investigation, other prosodic aspects and phonation types will also be considered.

5 Conclusion

This study investigated the relationship between voice preference and melody metrics from a cross-linguistic perspective. We have found that the paralingustic and non-linguistic information in prosody can help non-native listeners effectively select a talent voice in a foreign language, which they have no previous knowledge of.

6 Acknowledgements

The first author is sponsored by the Interdisciplinary Program of Shanghai Jiao Tong University (14JCZ03) and the Major Program of National Social Science Foundation of China (13&ZD189) for this research work.

References

[1] FITCH, W. T.: *Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques*. *Journal of the Acoustical Society of America*, 102, pp. 1213–1222, 1997.

[2] TIGUE, C., D. BORAK, J. O'CONNOR, C. SCHANDL, and D. FEINBERG: *Voice pitch influences voting behavior*. *Evolution and Human Behavior*, 33, pp. 210–216, 2011.

[3] KLOFSTAD, C., R. ANDERSON, and S. NOWICKI: *Perceptions of competence, strength, and age influence voters to select leaders with lower-pitched voices*. *PLoS ONE*, 10(8), pp. 2–14, 2015. doi:10.1371/journal.pone.0133779.

[4] FEINBERG, D., L. DEBRUINE, B. JONES, and D. PERRETT: *The role of femininity and averageness of voice pitch in aesthetic judgments of women's voices*. *Perception*, 37(4), pp. 615–623, 2008.

[5] PINTO-COELHO, L., D. BRAGA, M. SALES-DIAS, and C. GARCIA-MATEO: *On the development of an automatic voice pleasantness classification and intensity estimation system*. *Computer Speech and Language*, 27, pp. 75–88, 2013.

[6] HAIN, H.-U., O. JOKISCH, and L. COELHO: *Multilingual voice analysis: Towards prosodic correlates of voice preference*. In R. HOFFMANN (ed.), *Proc. Konferenz Elektronische Sprachsignalverarbeitung (ESSV) 2009*, vol. 53 of *Studientexte zur Sprachkommunikation*, pp. 215–221. Dresden, Germany, 2009.

[7] MCALEER, P., A. TODOROV, and P. BELIN: *How do you say 'hello'? personality impressions from brief novel voices*. *PLoS ONE*, 9(3), 2014. doi:10.1371/journal.pone.0090779.

[8] TROUVAIN, J., S. SCHMIDT, M. SCHRÖDER, M. SCHMITZ, and W. BARRY: *Modelling personality features by changing prosody in synthetic speech*. In *Proceedings of the Conference on Speech Prosody*. 2006.

[9] COELHO, L., H.-U. HAIN, O. JOKISCH, and D. BRAGA: *Towards an objective voice preference definition for the Portuguese language*. In *Proc. Iberian SLTech - Joint SIGIL/Microsoft Workshop on Speech and Language Technologies for Iberian Languages*, pp. 67–70. Porto Salvo, Portugal, 2009.

[10] HIRST, D. and H. DING: *Using melody metrics to compare English speech read by native speakers and by L2 Chinese speakers from Shanghai.* In *Interspeech*, pp. 1942–1946. 2015.

[11] ZHANG, J.: *A higher-than-average female voice can cause young adult female listeners to think about aggression more*. *Journal of Language and Social Psychology*, 35(6), pp. 645–666, 2016.

[12] XU, Y., A. LEE, W.-L. WU, X. LIU, and B. PETER: *Human vocal attractiveness as signaled by body size projection*. *PLoS ONE*, 8(4), pp. 1–9, 2013. doi:10.1371/journal.pone.0062397.