VOICE PREFERENCE IN GERMAN: A CROSS-LINGUISTIC COMPARISON OF NATIVE AND CHINESE LISTENERS

Hongwei Ding[1], *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: In this study we investigated whether native and non-native speakers demonstrate a similar preference in the selection of a pleasant voice, and how their preference rankings correlate with the voice quality and the prosodic features. 50 Chinese without prior knowledge of German and 10 native German listeners participated in a pair comparison test to choose the best voice out of eight candidate speakers for a German speech synthesis. The results showed that the ranking scores of German and Chinese listeners were highly correlated. A further investigation revealed though both German and Chinese listeners preferred a breathy voice, Chinese listeners also voted for speakers who exhibited faster falling pitch movements. The decisions of German listeners relied mainly on voice quality, while those of the Chinese listeners relied more on pitch movements. The findings suggest that speakers of a tone language may associate their voice preference with local pitch movements, especially when they have no knowledge with the concerned language.

1 Introduction

Listeners usually associate a voice with certain personal features, and the spoken utterances thus convey important information affecting the preference of the listeners. It has been demonstrated that several acoustic features are related to voice attractiveness, among which human voice pitch is one of the most important that affects the impression of the listeners. Voice pitch is the perceptual correlate of fundamental frequency (f0), which reflects the rate of vocal fold vibrations. 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 [1]; while men typically prefer higher-pitched women's voices, which are perceived as being feminine and youthful [2]. However, attractive voices are different from pleasant ones [3]. Normally for a text-to-speech (TTS) system, an agreeable and enjoyable voice is preferred rather than an appealing voice. Contrary to the research findings on attractiveness of the voice, it has been reported that a higher f0 of the female speaker might lead to a lower listening preference in a TTS voice casting [4]. Moreover, not only the voice pitch but also pitch change patterns are associated with personalities of the speaker. By changing prosodic parameters, such as pitch level, pitch range, articulation rate, and loudness in synthetic speech, the personality features of the voice can thus be modified [5].

Apart from prosody, voice quality is another important acoustic feature which affects the pleasantness of a voice. Voice quality is usually described in terms of phonation types, which are generally classified into modal, creaky, and breathy [6]. It has been demonstrated that breathy

voice of female speakers are preferred by male listeners [7]. However, it should be noted that the use of prosody or voice quality and its relation to emotional or attitudinal aspects seems to be language and culture dependent.

In summary, previous studies have shown that the average pitch values, prosody, and voice quality are supposed to be the main features which can influence the attractiveness of a voice [8]. Mean f0 values are mainly determined by speaker's anatomy and physiology, which are related to body size [7]; while the utilisation of prosody and voice quality may be language-dependent. Since less attention has been devoted to the investigation of acoustic correlates of voice preference from a cross-linguistic perspective [4], this study aims to provide some preliminary understanding. Furthermore, it is also interesting to compare the correlates between listeners of a tone and a non-tone language. We chose German and Chinese listeners to vote for their preferred German voice. Since the Chinese listeners had no previous knowledge of German, they mainly relied on the paralinguistic and non-linguistic information for their selections.

2 Method

First we conducted a voice preference test with German and Chinese listeners on German speakers to select a pleasant voice for TTS (Text-to-speech). Then we extracted prosody and voice quality parameters from these sentences. And finally we correlated the preference rankings between German and Chinese listeners, and their preference scores with these purely acoustic parameters.

2.1 Data collection

The data collection procedure includes recording, preference-based ranking elicitation, prosodic and voice quality parameters extraction, and calculation of related correlations.

2.1.1 Speech material

German speech data were taken from the recordings in the voice casting from eight candidate speakers for a German speech synthesis. All the candidates were female speakers between 22-39 years old. The speech material consisted of two sentences read by each eight speakers.

2.1.2 Preference scores

50 Chinese listeners who had no knowledge of German and 10 native German listeners participated in the preference test. The male to female ratio in both German and Chinese listeners was 1:1. The Chinese and German listeners aged from 18 to 20 and 21 to 29 respectively. All of them were university students.

These two sentences of each speaker were compared to those of other seven speakers. 28 pairs were constructed. The preference test was implemented by a Praat's Multiple Forced Choice (MFC) experiment script [9], and was carried out on a computer by listeners individually. Each time the same sentences from two speakers were played in a random order, 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.

2.2 Parameter extraction

Acoustic parameters include parameters of prosody and voice quality, and prosodic parameters are further classified into pitch change patterns and speech rate.

2.2.1 Melody metrics

Though all the candidates speak standard German, every speaker displayed her own particular pattern. In order to compare such different pitch change patterns or melody metrics, the algorithm described by Hirst in [10] was employed. The anchor points were scaled using the OMe (Octave-Median) with Formula (1) to reduce the inter-subject variability:

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

where *median* is the median value of f0 for the whole sentence.

The mean on the OMe scale were calculated for:

1) differences between each anchor point and the previous point (interval_m),
2) rise differences from the previous point (rise_m),
3) fall differences from the previous point (fall_m),
4) rise and fall slope from the previous point (slope_m),
5) rise slope from the previous point (rise_slope_m), and
6) fall slope from the previous point (fall_slope_m).

The 6 parameters were collected for each speaker. Since these values have been offset to the speaker's median f0 by Formula (1), they can be compared across speakers.

2.2.2 Speech rate

Speech rate was obtained by dividing the number of syllables by the duration in seconds of the whole sentence. Pauses between sentences were not included, but pauses between phrases were also regarded as the sentence duration.

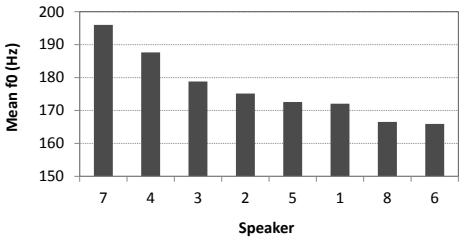
2.2.3 Voice quality

Voice quality is analyzed in terms of phonation types. There are a number of acoustic properties that potentially distinguish breathy phonation from the modal or creaky phonation. Though Different languages use different set of properties to distinguish these phonation types, there is generally some agreement between languages in how phonation contrasts are signaled. The acoustic characteristics defining phonation differences usually include periodicity, spectral tilt, fundamental frequency, formant frequencies, intensity, and duration [11].

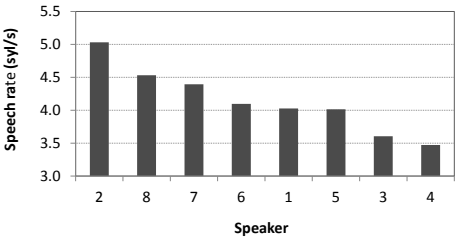
Detecting non-modal phonation in speech requires reliable f0 analysis. STRAIGHT pitch-tracker has the advantage of producing a smooth f0 and harmonic amplitude estimation by reporting f0 every millisecond [12], STRAIGHT algorithm was thus selected to perform all voice quality analyses with VoiceSauce [13]. Values were only taken from labeled vowels, and the acoustic measures were then averaged over the entire vowel's duration. We took a number of measurements commonly used to characterize voice quality, as listed in the following:

1) H1*–H2*: Difference between amplitudes of the first two harmonics,
2) H2*-H4*: Difference between amplitudes of the second and the fourth harmonics,
3) H1*-A1*, H1*-A2*, H1*-A3*: Difference in amplitudes of the first harmonic and the harmonic nearest F1, F2, and F3, respectively.

As the speaker's voice goes from breathy to pressed, all the above measurements show decreasing values, which indicates an overall reduction of spectral tilt.

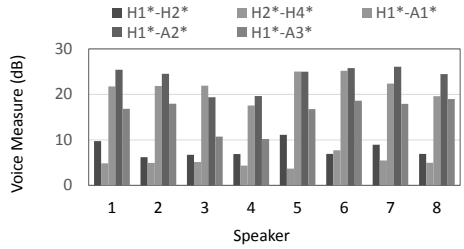


(a) Comparison of mean f0

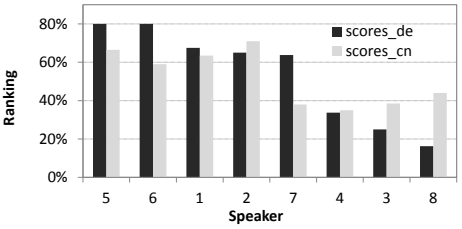


(b) Comparison of speech rate

Figure 1 – Comparison of mean f0 and speech rate of the speakers.



(a) Comparison of voice quality of the speakers



(b) Comparison of rankings of the listeners

Figure 2 – Comparison of voice quality of speakers and rankings of German and Chinese listeners.

3 Results

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

3.1 Acoustic parameters

Because the melody metrics calculated by Momel were numerous, only the average f0 is presented here, which can be observed in Figure 1a. Speech rate and voice quality measures are illustrated in Figure 1b and Figure 2a, respectively.

3.2 Correlation

3.2.1 German-Chinese preference correlation

The preference scores between the German and Chinese listeners are highly correlated (r=0.73) and the correlation is significant (p=0.048), which can be observed in Figure 2b. The correlations between males and females within the same language were also significantly high for German listeners (r=0.85, p=0.0069) and for Chinese listeners (r=0.86, p= 0.0065). In Figure 2b 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 the German and the Chinese listeners separately. If a speaker was selected each time when compared to any other speaker, she would get 100%.

The mean f0 values for these eight speakers were between 165Hz-196Hz. F0 values of the preferred speakers for the German listeners were 173Hz, 166Hz, 172Hz, and 175Hz; while those for the Chinese listeners were 175Hz, 173Hz, 172Hz, and 166Hz. The German and the Chinese listeners agreed on the best four speakers, but with a slightly different order.

3.2.2 Preference-acoustics correlation

If we correlated the preference score with each single parameter, the results 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. We also correlated the rankings of females and males separately for both Chinese and German listeners to show gender difference. In Table 1 the coefficient is written in boldface if the correlation is significant, printed in regular if the correlation is moderate and not significant, and represented by "-" if the correlation is low.

Table 1 – Correlation between preferences and acoustic features

Acoustic features	**German listeners**			**Chinese listeners**		
	de	**de-m**	**de-f**	**cn**	**cn-m**	**cn-f**
pitch related parameters						
median	-	-	-	-0.59	-0.52	-0.69
interval_m	-	-	-	-0.66	-0.60	**-0.74**
slope_m	-	-	-	**-0.76**	**-0.70**	**-0.83**
rise_slope_m	-	-	-	-0.53	-0.56	-
duration related parameters						
speech rate	-	-	-	-	0.52	-
voice quality related parameters						
H1*-H2*	0.53	-	0.65	-	-	-
H1*-A1*	**0.77**	0.65	**0.84**	0.58	0.59	0.52
H1*-A2*	0.68	0.62	0.69	0.57	0.52	0.64
H1*-A3*	-	-	-	0.55	-	0.61

Because *interval_m* represents the mean pitch differences in octave between each anchor-point and the preceding point, and *slope_m* shows the slope of intervals, both these values can be positive and negative. If *interval_m* is negative, it means that the sum of all pitch fall intervals exceeds that of all pitch rise intervals. And the same applies to the slope of intervals (*slope_m*). Because the sentences are not interrogative, these two values are normally negative. A negative correlation with the preference scores means that the larger or the faster the pitch falls, the higher preference score the speaker can achieve. Several results can be observed from Table 1:

1. More correlations could be found for the Chinese listeners than for the German listeners between preference rankings and melody metrics (all correlation coefficients are "_").

2. For the Chinese listeners, a higher f0 mean of German voice was associated with a lower preference, especially for female listeners (r=-0.69 between *cn-f* and *median*).

3. For the Chinese listeners, faster and larger falling pitch movements were associated with a higher preference (r=-0.66 between *cn* and *interval_m*, r=-0.76 between *cn* and *slope_m*).

4. For male Chinese listeners, faster rising pitch movements were associated with a lower preference (r=-0.56 between *cn-m* and *rise_slope_m*).

5. For Chinese male listeners, a faster speech rate was associated with a better preference (r=0.52 between *cn-m* and *speech rate*).

6. For both German and Chinese listeners, more breathiness is associated with a higher preference, because most breathiness measures are positively correlated with the preference.

In order to demonstrate which variable is more important between measures of voice quality and pitch movements, we selected the variables which demonstrated significant correlations with preference scores (i. e. *H1*-A1** and *slope_m*) , and entered them into a multiple linear regression in R [14]. The results are presented in Table 2. The outputs for German and Chinese listeners are *F=7.808 (p=0.0059)* and *F=7.464 (p=0.0069)* respectively, both indicting that these parameters have collective effects on the voice preference.

Table 2 – Coefficients between relevant parameters of the speakers and preference scores of the listeners (Significance codes: 0'***' 0.001'**' 0.01'*' 0.05'.' 0.1' ' 1)

Parameters	German listeners	Chinese listeners
H1*-A1*	0.00268 **	0.1646
slope_m	0.87640	0.0176 *

The results show that the German listeners relied more on voice quality (*H1*-A1**), while the Chinese listeners relied more on pitch movements (*slope_m*) for their choices of a preferred voice. And the correlation for the German listeners is more significant.

4 Discussion

Voice preference is a perceptual measurement, which is related to anatomy and physiology of vocal organs, languages, and also cultural aspects. 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 of German 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. Because of native tone language, Mandarin listeners may be more sensitive to local pitch changes than German listeners.

2. Compared with German listeners, Chinese listeners associate more pitch-related metrics to their preference of speakers.

3. Chinese listeners usually prefer speakers with faster and larger pitch falls. The reason might be that smaller pitch falls might sound monotonous for Chinese listeners.

4. Chinese listeners do not seem to prefer female speakers with a higher pitch, which confirms the findings in [4], while Chinese female listeners prefer female speakers with a lower-pitched voice, which is consistent with the findings in [15].

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.

6. Compared with Chinese listeners, German listeners associate their preference more with voice quality than local f0 movements, which confirms the results found in [16].

Numerous factors are at play in the selection of a preferred voice. Generally speaking, segmentals, suprasegmentals, and phonation types can all influence the preference of a voice. By conducting a cross-linguistic investigation, we have found that both German and Chinese listeners prefer female breathy voices. Besides, Mandarin Chinese listeners also vote for speakers with faster falling pitch movements, which is usually considered as vivid speech in Mandarin Chinese [17]. These statistical findings on naturally produced speech can be further tested experimentally with synthesized stimuli by varying one variable and holding the others constant in the future investigations.

5 Conclusion

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

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] 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.

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

[3] 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.

[4] 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.

[5] 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.

[6] GERRATT, B. R. and J. KREIMAN: *Toward a taxonomy of nonmodal phonation*. *Journal of Phonetics*, 29, pp. 365–381, 2001.

[7] 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.

[8] 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.

[9] BOERSMA, P. and D. WEENINK: *Praat: doing phonetics by computer [computer program]*. 2017. URL `http://www.praat.org`. Version 6.0.24, retrieved 31 Januray, 2017.

[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] GORDON, M. and P. LADEFOGED: *Phonation types: a cross-linguistic overview*. *Journal of Phonetics*, 29, pp. 383–406, 2001.

[12] KAWAHARA, H., I. MASUDA-KATSUSE, and A. DE CHEVEIGNE: *Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction*. *Speech Communication*, 27, pp. 187–207, 1999.

[13] SHUE, Y.-L., P. KEATING, C. VICENIK, and K. YU: *Voicesauce: A program for voice analysis*. In *Proceedings of the International Congress of Phonetic Sciences (ICPhS 17)*, pp. 1846–1849. Hong Kong, 2011.

[14] R DEVELOPMENT CORE TEAM: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL `http://www.R-project.org`. Version 3.3.3, retrieved on 1 June, 2017.

[15] 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.

[16] KREIMAN, J. and B. R. GERRATT: *Perceptual sensitivity to first harmonic amplitude in the voice source*. *Jounal of Acoustical Society of America*, 128, pp. 2085–2089, 2010.

[17] DING, H., R. HOFFMANN, and O. JOKISCH: *Prosodic correlates of voice preference in Mandarin Chinese and German: A cross-linguistic comparison*. In J. TROUVAIN, I. STEINER, and B. MÖBIUS (eds.), *Elektronische Sprachsignalverarbeitung 2017*, pp. 83–90. Saarbrücken, Germany, 2017.