Perceptual Evaluation of Penetrating Voices through a Semantic Differential Method

Tatsuya Kitamura[1], *Naoki Kunimoto*[1], *Hideki Kawahara*[2], *Shigeaki Amano*[3]

[1]Konan University, Japan
[2]Wakayama University, Japan
[3]Aichi Shukutoku University, Japan

t-kitamu@konan-u.ac.jp

Abstract

Some speakers have penetrating voices that can be popped out and heard clearly, even in loud noise or from a long distance. This study investigated the voice quality of the penetrating voices using factor analysis. Eleven participants scored how the voices of 124 speakers popped out from the babble noise. By assuming the score as an index of penetration, ten each of high- and low-scored speakers were selected for a rating experiment with a semantic differential method. Forty undergraduate students rated a Japanese sentence produced by these speakers using 14 bipolar 7-point scales concerning voice quality. A factor analysis was conducted using the data of 13 scales (i.e., excluding one scale of penetrating from 14 scales). Three main factors were obtained: (1) powerful and metallic, (2) feminine, and (3) esthetic. The first factor (powerful and metallic) highly correlated with the ratings of penetrating. These results suggest that penetrating voices have multi-dimensional voice quality and that the characteristics of penetrating voice related to powerful and metallic aspects of voices.

Index Terms: voice quality, semantic differential method, factor analysis, popping out from noise

1. Introduction

Voice quality varies widely from speaker to speaker, depending on their age, gender, and the physical properties of the speakers' speech organs. Speakers who have penetrating voices are known empirically, with voices that can pop out even from loud noises and be heard in distant places. Such penetrating voices are suitable for effective public address systems, especially on emergency occasions, and exploring their psychoacoustic and acoustic properties might be useful for developing speech synthesis or voice conversion techniques that can produce penetrating voices. To date, however, only a few studies have focused on the "degree of penetration" of voices, thus the present study investigated the perceptual properties of penetrating voices through speech rating experiments.

Yokoyama and Inoue [1] conducted a rating experiment for the Japanese vowel /e/ in 13 speakers using bipolar rating scales for voice quality. They reported that the scores for "penetrating–not penetrating" and "clear–dirty" were highly correlated and that these two rating scales correlated closely with the likability of the voices. They also showed that the frequencies and amplitudes of the third and fourth formants of likable voices were closer and greater, respectively, than those of the others.

Yokoyama and Inoue [1] suggested that the frequency region of the third and fourth formant frequencies is a potential source of penetration. The frequency region corresponds to prominent spectral resonances in the voices of professional male announcers [2] and the singing voices of male singers [3][4]. The spectral resonance of the singing voice is known as the singer's formant or singing formant.

A similar spectral resonance is known as the speaker's formant. Bele [5] reported that better normal voice quality was related to the speaker's formant and the spectral peaks in the frequency region from 3 to 4 kHz of the long-term averaged spectra (LTAS). Other studies have also shown a correlation between the amplitude of the speaker's formant and professional voices [6] or good voice quality [7]. Krause and Braida [8] showed that the amplitude of the LTAS in a lower-frequency region ranging from 1 to 3 kHz increased in clear speech compared with normal speech. The fact that these similar voice qualities ("penetrating," "good," and "clear") have all been pointed out in relation to the spectral resonances around the third and fourth formant frequencies indicates that there is a common principle regarding voice qualities.

This study aimed to explore the perceptual properties of penetrating voices. We first conducted a preliminary experiment to find speakers whose voices pop out the most and least from background noise and then performed a rating experiment using a semantic differential method with 14 bipolar scales.

2. Procedure

2.1. Stimuli

Eleven Japanese participants first scored how the voices of 124 Japanese speakers popped out from babble noise to select speakers used in the main rating session. The voices were collected from the ATR speech database sets B and C [9][10], the ASJ continuous speech corpus for research[1], the RWCP news speech corpus[2], and our original recordings of professional narrators and voice actors. The sentence was "*Gaijin san wa kanpeki shugi de aru*" (the foreign person is a perfectionist) and set at a sampling frequency of 16 kHz with a 16-bit resolution. All the speech data were recorded under quiet conditions. Babble noise was synthesized by overlapping sentences of five male and five female speakers selected randomly from a Japanese speech dataset [11] and added to the voices at a signal-to-noise (SN) ratio of 3 dB.

The participants listened to the voices of the 124 speakers randomly through an audio interface (Roland, Duo-Capture EX) and circumaural headphones (Sony, MDR-CD900ST) and scored their pop-out level on a scale of 1–5, where 1 and 5 corresponded to "never pops out" and "pops out very much," respectively. The scores were averaged among the listeners (hereafter,

[1] https://doi.org/10.32130/src.ASJ-JIPDEC

[2] https://doi.org/10.32130/src.RWCP-SP99

this score is referred to as the ‘pop-out score’), and the speakers were sorted by it.

In the rating sessions using the semantic differential method, the voices of the top five male and top five female speakers (the mean pop-out score of the speakers was 4.3) and the bottom five male and bottom five female speakers (the mean pop-out score of the speakers was 1.8) in ATR speech database set C were rated. The speaker IDs with high pop-out scores were M202, M205, M217, M311, and M508 (male speakers) and F310, F314, F602, F606, and F708 (female speakers), and those with low pop-out scores were M103, M105, M306, M617 and M620 (male speakers) and F105, F201, F218, F504, and F715 (female speakers) of the dataset.

2.2. Scales

Fourteen of the scales used in the rating experiment were bipolar rating scales ranging from 1 to 7. The scales consist of three groups. The first six were basic adjective pairs for the factors of beauty, metallic aspect, and powerful aspect of timbre [12]. The next seven were taken from eight adjective pairs for expressing voice quality proposed by Kido and Kasuya [13] excluding a duplicated one in the first six pairs (“vigorous–frail”). The last, “penetrating–not penetrating,” was added to assess the correlation between the pop-out score and degree of penetration of the voices. The 14 rating scales used in this study are as follows:

1. clear—dirty;
2. beautiful—ugly;
3. powerful—weak;
4. sharp—dull;
5. vigorous—frail;
6. hard—soft;
7. high pitched—low pitched;
8. masculine—feminine;
9. clear—hoarse;
10. calm—excited;
11. youthful—elderly;
12. thick—thin;
13. tense—lax;
14. penetrating—not penetrating.

2.3. Raters

Forty undergraduate students (11 females and 29 males) from the Faculty of Intelligence and Informatics of Konan University, aged 20–22 years, participated in this experiment. The participants were Japanese speakers with no known hearing impairment. They were paid for their participation.

2.4. Rating procedure

The raters were instructed to listen to the stimuli and score their voice quality on an online form (Google Forms). The online forms consisted of 20 pages, and each page was randomly assigned to one of the 20 speakers. On each page, the order of the rating scales and side positions of the two adjectives on each scale were randomized. The rating scales were presented in Japanese. We made four sets of online forms with different orders of speakers and rating scales and randomly assigned one of the sets to each rater.

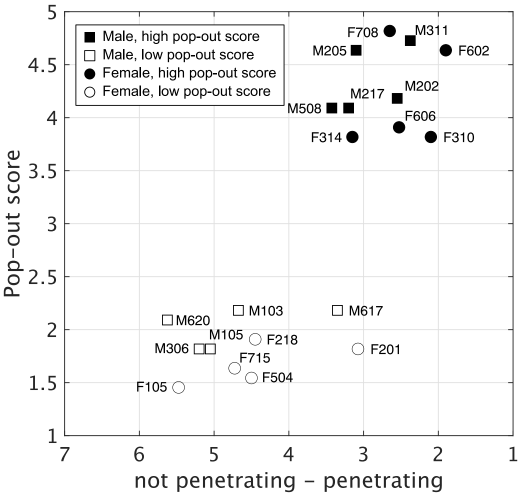


Figure 1: *Scattergram of speakers on a coordinate plane of the pop-out and not penetrating scores. The x-axis is flipped horizontally. The square and circle markers represent the male and female speakers, respectively, and the filled and open makers indicate the speakers with high and low pop-out scores, respectively. Markers are labeled with speaker IDs.*

Stimuli were normalized with their A-weighted amplitudes and presented without background noise. The raters listened to the stimuli through an audio interface (Roland, Rubix 22) and circumaural headphones (Sony, MDR-CD900ST) at an equivalent sound level of 72 dB measured for a 1 kHz pure tone and then scored the rating scales by clicking one of the radio buttons corresponding to the seven points. The raters were allowed to listen to the stimuli any number of times. The experiments were carried out in a quiet room ventilated with fresh air to reduce the risk of infection from coronavirus. Written informed consent was obtained from each participant before the experiment.

3. Results

3.1. Scale reliability

The Cronbach’s coefficient alpha [14] was 0.797 for the rating experiment, indicating the acceptable reliability of the scales.

3.2. Consistency between pop-out score and rating results

A scattergram of speakers on a coordinate plane of the pop-out score obtained in the preliminary experiment and the “not penetrating” score for the speakers is shown in Figure 1. Note that 1 and 7 on the x-axis correspond to “penetrating” and “not penetrating,” respectively, and the axis is flipped horizontally. The voices of speakers with high pop-out scores (filled squares and circles in the plot) showed low ratings (more penetrating) on the scale and vice versa. The correlation coefficient was $r = 0.829$ ($p < 0.05$), indicating that the degree of penetration rated for voices without background noise was highly compatible with the pop-out score measured for voices with background noise.

3.3. Mean scores

Table 1 lists the mean scores for speaker groups with high and low pop-out scores. The left and right adjectives or timbre semantics of the semantic differential scales correspond to scores

Table 1: *Mean rating score of the 14 scales in high and low pop-out score groups. The p-values of t-tests between the mean scores are also shown.*

Scale	Pop-out score		p-value
	High	Low	
clear–dirty	3.54	3.69	0.059
beautiful—ugly	2.28	2.47	0.111
powerful—weak	3.34	5.20	<0.001
sharp—dull	3.38	4.62	<0.001
vigorous—frail	1.86	3.04	<0.001
hard—soft	2.22	3.80	0.001
high pitched–low pitched	3.69	3.96	0.041
masculine—feminine	4.01	4.06	0.743
clear—hoarse	3.19	3.76	<0.001
calm—excited	3.08	2.78	0.021
youthful—elderly	4.06	3.57	<0.001
thick—thin	2.66	3.37	<0.001
tense—lax	2.62	4.91	<0.001
penetrating–not penetrating	1.70	3.61	<0.001

of 1 and 7, respectively. Thus, for example, the results for "powerful–weak" in the table indicate that the raters felt the voices of the high pop-out scores were more powerful (less weak) than the voices of the low pop-out scores.

Paired two-tailed Welch's t-test comparisons were performed on the raters' perceptual scores on the semantic differential scales between the two groups of speakers ($df = 797$). As listed in Table 1, the raters heard significant differences between the groups across all the scales except "clear–dirty," "beautiful–ugly," and "masculine–feminine," showing that the two types of voices gave dedicatedly different impressions to the raters. The mean scores differed largely between the speaker groups for the scales "tense–lax" (the difference was 2.29), "penetrating–not penetrating" (the difference was 1.91), "powerful–weak" (the difference was 1.86), and "hard–soft" (the difference was 1.58).

3.4. Factor loadings

A factor analysis was performed for the scores of the semantic difference scales excluding "penetrating–not penetrating." Factors were extracted using the maximum likelihood method, and missing values were handled using list-wise deletion. Table 2 lists the loadings after the Promax rotation of each of the 13 scales on the 3 main factors.

The percentage of variance accounted for by these three factors was 67.6%. The eigenvalues of the three factors after Promax rotation were as follows: factor 1, 4.397; factor 2, 3.575; factor 3, 1.499. The alpha coefficient of the first three factors were greater than or close to 0.7, indicating that the results were reliable.

In the first factor, high loadings were found for the scales "powerful–weak" and "vigorous–frail," being construed as the powerful factor and the scales "tense–lax," "hard–soft," and "sharp–dull," being construed as the metallic factor; we, therefore, chose the powerful and metallic factor. The second factor was a factor relating the gender, "high pitched–low pitched" and "masculine–feminine" being the scales with the high loadings; we thus chose the feminine factor with consideration for the sign of the loadings. The third factor was labeled esthetic because of the high loadings of the scales "beautiful–ugly" and "clear–dirty."

Figures 2(a) and (b) show plots of the sign-inverted mean scores of each speaker on the Factor 1–Factor 2 and Factor 1–Factor 3 planes. All the speakers with high pop-out scores (filled markers) were plotted in the positive half-plane on the first factor, and the speakers with low pop-out scores (open markers), except for one speaker labeled F201, were plotted in the negative half-plane on the first factor. On the other hand, all the female (circles) and male (squares) speakers were plotted in the positive and negative half-planes, respectively, on the second factor, demonstrating that the factor divided the male and female voices. Factor 3 showed a poor correlation with gender. The results indicated that Factor 1 separated the penetrating voices from the not penetrating voices, suggesting that the penetrating voices can be considered powerful and metallic.

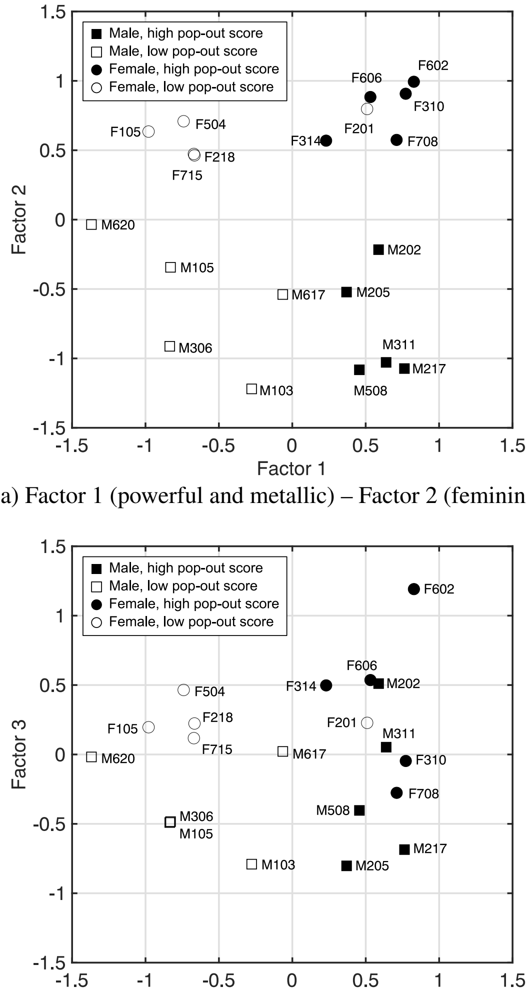
Table 2: *Factor loadings of the voices of the 40 speakers on the three main factors.*

Scale	Factor loadings		
	Factor 1	Factor 2	Factor 3
1. Powerful and metallic			
powerful–weak	0.897	−0.130	0.043
vigorous-–frail	0.892	−0.123	0.110
tense–lax	0.809	0.021	0.112
hard–soft	0.622	0.055	−0.357
sharp–dull	0.429	0.359	0.107
2. Feminine			
high pitched–low pitched	0.109	0.838	0.018
masculine-–feminine	0.008	−0.738	−0.027
thick-–thin	0.372	−0.639	−0.027
youthful-–elderly	−0.176	0.406	0.256
3. Esthetic			
beautiful-–ugly	−0.004	−0.056	0.819
clear-–dirty	0.033	0.171	0.750
clear-–hoarse	0.191	0.117	0.627
calm-–excited	−0.168	−0.483	0.512
Factor correlations			
Factor 1	1.000	0.0002	0.135
Factor 2	0.002	1.000	0.497
Factor 3	0.135	0.497	1.000
Alpha coefficient	0.843	0.779	0.713

Figure 3 shows the correlation between the scores of Factor 1 and the scale "penetrating–not penetrating" for the 20 speakers. Although the factor analysis was performed apart from the scale "penetrating–not penetrating," the correlation coefficient was considerably high ($r = 0.957$, $p < 0.05$), showing a linear relationship between the scores.

4. Discussion

The strong correlation between the pop-out score and the not penetrating score shown in Figure 1 demonstrates that the penetrating voices can truly penetrate or pass through the loud noise. This suggests that there is a way to improve the perceptual SN ratio of speech communication in noisy places, other than simply amplifying the level of the target speech. Although the stimuli were presented at the same level in the rating sessions, the scores on most of the semantic differential scales were significantly different between the two voice quality groups, as listed in Table 1. The difference of the scores was particularly significant for the scales "tense–lax," "powerful–weak," and "hard–soft" and these scales contributed to Factor 1 labeled as powerful and metallic of the results of the factor analysis. Factor 1 separated the speakers into those with a high pop-out score and



(a) Factor 1 (powerful and metallic) – Factor 2 (feminine)

(b) Factor 1 (powerful and metallic) – Factor 3 (esthetic)

Figure 2: *Score plot of the speakers on (a) the Factor 1–Factor 2 plane and (b) the Factor 1–Factor 3 plane. The square and circle markers represent the male and female speakers, respectively, and the filled and open makers indicate the speakers with high and low pop-out scores, respectively.*

those with low pop-out scores (Figure 2(a)) and demonstrated a notable correlation with the not penetrating score (Figure 3). The results indicate that the present study succeeded in extracting the axis corresponding to the degree of penetration of the voices as Factor 1.

The metallic aspect of the timbre related to the spectral structure of sound and sound with the metallic property had a high amplitude in the higher-frequency regions [12]. Thus, penetrating voices may have the same spectral structure. This view supports previous studies on penetrating voice quality [1], the voices of male announcers [2], the singer's formant [3][4], clear speech [8], and the speaker's formant [5], and implies the presence of a fundamental common theory on clear and high-tolerance voice qualities.

Hanayama *et al.* [15] analyzed "metallic voice" for the vowel /e/. They reported that the amplitudes of the second, third, and fourth formants of the voices increased, and the voices were perceived more loudly than others. The results of the present study that penetrating voices had a powerful aspect, aside from metallic voices, supported their results.

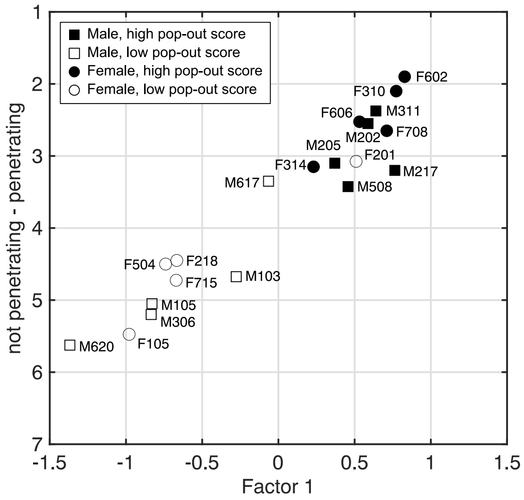


Figure 3: *Correlation between Factor 1 (powerful and metallic) and the not penetrating score.*

5. Conclusions

This study has focused on psychological factors to examine human perceptions of penetrating voices using the semantic differential method. A rating experiment was carried out to evaluate the properties of the voices that pop out from the loud noises, and the data of the experiment were analyzed by factor analysis. As a result: (1) the more penetrating the voice, the more it popped out from the noise; (2) the degree of penetration of the voices corresponded with the powerful and metallic aspect of the voices. Future work will explore the mechanisms of production of penetrating voices and clarify the relationship between penetrating voices and the speaker's and singer's formants.

6. Acknowledgments

This study was supported by MEXT/JSPS KAKENHI (Grant Number JP20H00291). The ATR speech database and sound pressure calibration device for headphones "ChomeJig" used in this study were provided by the Advanced Language Information Forum and Prof. Toshio Irino of Wakayama University, respectively.

7. References

[1] M. Yokoyama and K. Inoue, "An extraction of informations of personal perception based on the sense of voice quality," *Japanese Journal of Ergonomics*, vol. 20, no. 1, pp. 41–48, 1984.

[2] H. Kuwabara and K. Ohgushi, "Acoustic characteristics of professional male announcers' speech sounds," *Acta Acustica united with Acustica*, vol. 55, no. 4, pp. 233–240, 1984.

[3] J. Sundberg, "Articulatory interpretation of the "singing formant"," *J. Acoust. Soc. Am.*, vol. 55, no. 4, pp. 838–844, 1974.

[4] ——, *The Science of the Singing Voice*. DeKalb: Northern Illinois University Press, 1987.

[5] I. V. Bele, "The speaker's formant," *J. Voice*, vol. 20, no. 4, pp. 555–578, 2006.

[6] T. Nawka, L. C. Anders, M. Cebulla, and D. Zurakowski, “The speaker’s formant in male voices,” *J. Voice*, vol. 11, no. 4, pp. 555–578, 1997.

[7] T. Leino, “Long-term average spectrum in screening of voice quality in speech: Untrained male,” *J. Voice*, vol. 23, no. 6, pp. 671–676, 2009.

[8] J. C. Krause and L. D. Braida, “Acoustic properties of naturally produced clear speech at normal speaking rates,” *J. Acoust. Soc. Am.*, vol. 115, no. 1, pp. 555–578, 2004.

[9] Y. Sagisaka and N. Uratani, “ATR spoken language database,” *J. Acoust. Soc. Jpn.*, vol. 48, no. 12, pp. 878–882, 1992.

[10] T. Takezawa, A. Nakamura, and E. Sumita, “Databases for conversational speech translation research at ATR,” *J. Phon. Soc. Jpn.*, vol. 4, no. 2, pp. 16–23, 2000.

[11] Y. Atake, T. Irino, H. Kawahara, J. Lu, S. Nakamura, and K. Shikano, “Robust estimation of fundamental frequency using instantaneous frequencies of harmonic components,” *IEICE Transactions (Japanese Edition)*, vol. J83-D-II, no. 11, pp. 2077–2086, 2000.

[12] S. Iwamiya, N. Osaka, K. Ozawa, M. Takada, N. Fujisawa, and K. Yamauchi, *Science of sound color: Evaluation and creation of timbre and sound quality*. Tokyo: Corona Publishing, 2010.

[13] H. Kido and H. Kasuya, “Representation of voice quality features associated with talker individuality,” *Proc. ICSLP1998*, no. 1005, 1998.

[14] L. J. Cronbach, “Coefficient alpha and the internal structure of tests,” *Psychometrika*, vol. 20, no. 16, pp. 297—-334, 1951.

[15] E. M. Hanayama, Z. A. Camargo, D. H. Tsuji, and S. M. R. Pinho, “Metalic voice: Physiological and acoustic features,” *J. Voice*, vol. 23, no. 1, pp. 62–70, 2009.