Cross cultural differences in arousal and valence perceptions of voice quality

Donna Erickson[1,8] ・*Shigeto Kawahara*[2] ・*Albert Rilliard*[3,8] ・*Ryoko Hayashi*[4] ・*Toshiyuki Sadanobu*[5] ・*Yongwei Li*[6] ・*Hayato Daikuhara*[7] ・*João de Moraes*[8] ・ *Kerrie Obert*[9]

[1]Haskins Laboratories, U.S.A., Kanazawa Medical University, Japan, [2]Keio University, Japan, [3]LIMSI, CNRS, Université France, [4]Kobe University, Japan, [5]Kyoto University, Japan, [6]Institute of Automation Chinese Academy of Sciences, China, [7]Renmin University of China, China, [8]Federal University of Rio de Janeiro, Brazil, [9]The Ohio State University, U.S.A.

ericksondonna2000@gmail.com, kawahara@ici.keio.ac.jp, albert.rilliard@limsi.fr, rhayashi@kobe-u.ac.jp, sadanobu.toshiyuki.3x@kyoto-u.ac.jp, liyongwei2000@126.com, daikuhayato@yahoo.co.jp,jamoraes3@gmail.com, kerriebobert@tmail.com

Abstract

Voice quality differences [1] can convey different attitudes and emotions [2], with speakers of different languages showing different sensitivities to voice qualities, e.g., [3, 4, 5]. It remains to be explored, however, precisely which acoustic properties are perceptually associated with what emotional meanings, and whether such perceptual mappings hold universally or differ across languages. This paper offers a first step addressing these issues. Building upon the previous findings that speakers of different languages demonstrate different sensitivities to voice quality differences, the study examines particularly how the perceptions of arousal and valence are affected by different voice qualities. The current experiment reveals that speakers of the three language groups share similar ratings of arousal in association with breathy voices. Yet the valence ratings vary among the groups: Japanese and Mandarin listeners rate voices with high F0 and small OQ with positivity, whereas Brazilian Portuguese rate voices with low F0 and larger OQ with positivity. The findings of this study have applications for second language teaching, and carry over to the worlds of business, politics, and advertisement; in general, this type of research may have a potential to be useful for improving communication in cross-cultural inter-personal relationships.

Index Terms: voice quality, cross-cultural perception, valence, arousal, acoustics

1. Introduction

Laver [1] described voice quality as long-term settings that change the general sounding of one's voice, without necessarily affecting its phonemic performance; such voice qualities have different settings in terms of source (vocal fold configuration as well as its tension) and filter (supra-laryngeal) configurations. With regard to how attitudes and emotions are conveyed via different voice quality, recent work reports speakers of different languages show different sensitivities to different voice qualities, e.g., [2, 3, 4, 5]. It has also been shown that ethnophonetics—particular cultural/social contexts in which the utterances are made—plays a role in cultural preferences for certain voice qualities (e.g. [6, 7]). For example, a cross-cultural study on Japanese "cake-seller" voices suggests that Japanese listeners prefer the voice with a slight twang (pharyngeal narrowing), while listeners from India prefer a voice without any twang [7]. As for a "seductive, flirtatious" voice, Japanese listeners prefer a non-breathy voice with high F0, while Americans, French, and Brazilian Portuguese prefer a lower, more breathy voice [5, 8]. See also https://www.youtube.com/watch?v=IcT29r33yB0 at 4'59''- for a famous sensual voice of a well-known Brazilian female airport announcer while making airport announcements which shows these acoustic characteristics (the same is observed at French airports [23]). In the advertising world, when doing commercials, voice actors/actresses are requested to change their voice qualities in such a way as to better sell products in different regions. For instance, for selling a yogurt product in London, the voice needs to be perkier and higher pitched; on the other hand, for selling the same yogurt in the U.S., the voice should be lower and more breathy to sound sexy, and in France, it needed to be faster; for selling electronics in the Midwest, the voice needed to be louder with more twang (pharyngeal, epipharyngeal constriction) (p.c. with professional voice actress Kelly Fosdahl Berge). Cross-cultural study of voice qualities can thus be an interesting research topic for phoneticians, with possible interdisciplinary applications. With this general research question in mind, this study examines how the percepts of arousal (excited vs. calm) and valence (positive vs. negative) are affected by different voice qualities. Previous studies have shown that perceptions of arousal increase as F0 increases [9, 10]. Also, valence judgements increase with increased F0, but changes in vocal tract configurations, such as larynx position (high, mid or low) [11] interact with F0 in such a way that for the vowel /i/, low larynx/low F0 is perceived as negative, and high larynx/low F0, as positive.

The current study examines the percepts of arousal and valence by speakers of Japanese, Mandarin Chinese, and Brazilian Portuguese. This study is a part of an on-going, larger project, for which we hope to expand our target languages. The three languages chosen for the current paper are partly due to accessibility; also because they represent different language types (pitch-accent, tonal, intonational) as well as different cultural groupings (Asian vs. Western).

The voice quality material was produced by a female native speaker of English, manipulating the contributions from both the source (the manner of vocal fold approximation) and the filter (the vocal tract area), to produce nine variations of sustained vowels (seven /i/ and two /æ/). The articulatory

manipulations were based on the Estill Voice Training Method of singing [12], described in more detail in the methods section.

The acoustic recordings used for the perception tests were part of a larger MRI study to examine how the vocal fold and vocal tract contribute to voice quality differences. This paper reports primarily on the acoustic analyses of these sounds, along with the cross-cultural perceptual results.

2. Methods

Nine instances of sustained vowels uttered by a female subject (first author) were recorded at ATR, Inc. Kyoto, Japan, using the BAIC MRI recording facilities. The speaker produced the vowels, based on the Estill Training Method, and was coached by the 9th author, a certified course instructor in the Estill Method. In this method, the singer/speaker is taught how to isolate control of several structures involved in speech production in the vocal folds and vocal tract. In these recordings, the speaker changed the mode of vocal fold vibration in three different ways: stiff (breathy), thin (soft, non-breathy) and thick (loud/non-breathy) produced on high and low F0, about an octave apart. (These modes are also often referred to in the literature as breathy, falsetto and modal, respectively). In addition, because we were interested in exploring the twang characteristic of the cake seller voice [6, 7], the speaker narrowed her pharynx in order to produce nasal and oral vowels on a high F0 with thin folds. An additional sound was produced in which the speaker raised the tongue dorsum, keeping F0 low and the vocal folds thin. "Thin" vocal folds were produced with vibrating only the "cover" part of the vocal folds, while thick folds, with both the cover and body of the folds ([13, 14]). Roughly speaking, "thin folds" correspond to falsetto-type phonation, and "thick" folds, to modal-type phonation. Due to time limitations of the MRI recording session, it was not possible to do a complete set of configurations of vowels and voice settings.

For the acoustic analyses of the vowels, the ARX-LF model [15] was used to estimate F0, spectral characteristics of the vocal tract (spectral tilt, i.e., slope of amplitudes of harmonics up to 10kHz) and opening characteristics of the vocal folds (i.e., Open Quotient, OQ).

For the evaluations of arousal and valence, the listeners were 40 Japanese college students in Kobe, 24 Mandarin Chinese college students in Beijing, and 20 Brazilian Portuguese from the Federal University of Rio de Janeiro. The tests were given in their respective languages.

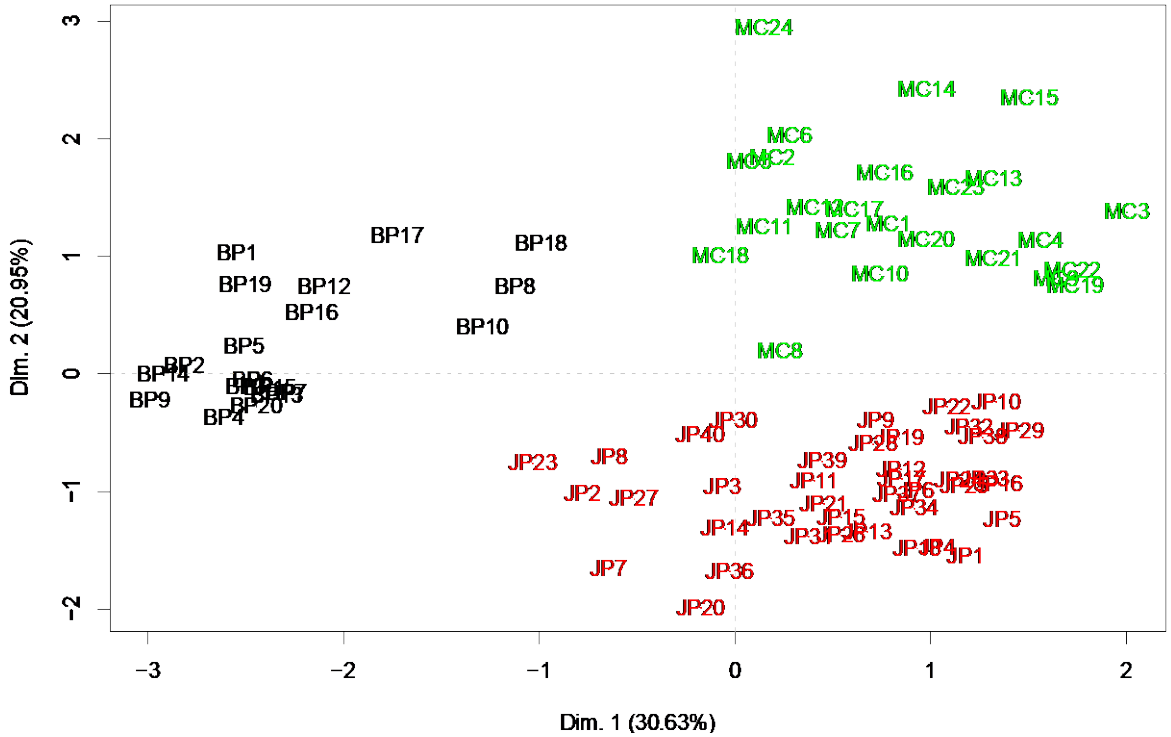

Figure 1: *Factor map obtained by MFA, showing clear separation of the three language groups by their use of the valence and arousal scales: BP grouping is on the left middle, JP, toward the bottom right, and MC, the upper right.*

The Japanese and Mandarin Chinese listeners were presented the 9 sounds from a PowerPoint display one at a time, and asked to evaluate on an ordinal scale of 1 to 5 how calm vs. excited the sound was. Each sound was repeated 4 times in a randomized order, resulting in a total of 36 trials. They were then presented with a different set of 4 randomizations of the same 9 sounds and then asked to evaluate on a scale of 1 to 5 how negative vs. positive the sound was; specifically, if they liked the voice, then it was "positive". The Brazilian Portuguese listeners were presented with the same 9 sounds, using a LiveCode interface and listened with Sennheiser headphones. These listeners were presented with 4 randomizations of each sound and asked to evaluate both scales at the same time. (The reason for this different approach was that the tests were administered individually for the Brazilian Portuguese listeners but as a group for the Mandarin Chinese and Japanese listeners.)

3. Results

Valence and Arousal scores attributed to each stimulus were organized in a table with answer by each listener on the rows, with stimuli as columns, with two parts so to have scores for the two scales. A supplementary column indicating the language group was added. The matrix was subjected to a Multiple Factorial Analysis (MFA; see [16, 17]) so as to compare how the listeners rated the stimuli on these two scales. The first 3 dimensions of the MFA explain more than 64% of the variance, with the first explaining more than 30% and the second 20%.

The first dimension of the MFA mostly separates Brazilian listeners from the others: Brazilian listeners select different stimuli for attributing high scores on both scales, but primarily on the Valence scale in the case of the first dimension. The second dimension separates Japanese listeners from the Mandarin Chinese listeners: The second axis is linked to rating on the Arousal scale—what separates Chinese from Japanese in their perceptual evaluation of these stimuli is mostly linked to this scale. The third dimension is linked negatively to Chinese listeners; it is also linked positively to a set of stimuli which were not attributed high scores by these speakers. In order to have a clearer view of the similarities and differences between listeners, the spread of listeners along the first three dimensions of the MFA was used to run a hierarchical clustering algorithm [16, 17]. A solution with three clusters meets a criterion of inertia gain. The spread of listener on the first two axes and their attributed clusters obtained through this method is presented in Figure 1. The three clusters exactly match the three groups of language listeners: Cluster #1 regroups Brazilian speakers (BP), cluster #2 regroups Japanese listeners (JP), while cluster #3 regroups the Mandarin Chinese ones (MC). The listeners inside each cluster are characterized by the high or low scores they attributed to specific stimuli: there is thus a strong relationship between the perceptual judgments of the two scales of Arousal and Valence, and cultural representations of these voice's acoustic characteristics.

Given this cultural sensitivity, we next investigated what acoustic characteristics the listeners groups may be attending to. Table 1 gives a listing of the 9 vowel sounds, along with the acoustic measurements (F0, spectral tilt and OQ).

To assess how cultural judgments of valence and arousal perceptions were affected by differences in these acoustic measures, a mixed effects model was fit for each language group. The fixed variables were tilt, f0 and OQ (as in Table 1). A random slope for each listener was included in each model. All fixed variables were centered. For the Mandarin Chinese and Japanese, since the dependent variable was based on an ordinal scale, an ordinal logistic regression was fit. For the Brazilian data, since the scale was continuous, a standard linear mixed model was fit. All the statistical analyses were implemented using R [18].

Table 2 shows the results for the valence ratings by the three language groups. Mandarin Chinese listeners preferred vowels that are higher in F0, and with a lower OQ (i.e. less breathy). Japanese listeners preferred those vowels that have a low OQ, with a sharper spectral tilt, and to a lesser extent, with higher F0. Brazilian Portuguese listeners' preferences were largely influenced by spectral tilt: they rate as high valence those with a sharp spectral tilt, and to a lesser extent those with a lower F0.

Table 1: *Nine vowels produced with 3 phonation modes, 4 vocal tract configurations, at low and high F0 values. The last three columns show the acoustic measurements, f0, spectral tilt and open quotient (OQ).*

ID	V	Phonation mode	Vocal tract	F0	Tilt	OQ
1	/i/	stiff (breathy)	normal/ oral	240 Low	-16.9	0.63
2	/i/	thin (falsetto)	normal/ oral	312 Low	-13.4	0.43
3	/i/	thick (modal)	normal/ oral	230 Low	-11.9	0.49
4	/i/	stiff (breathy)	normal/ oral	480 High	-11.4	0.41
5	/i/	thin (falsetto)	normal/ oral	500 High	-16.1	0.47
6	/i/	thick (modal)	normal/ oral	520 High	-10.5	0.4
7	/æ/	thin (twang)	pharyngeal narrowing /oral	520 High	-13.4	0.34
8	/i/	thin; tongue dorsum high	normal/ oral	240 Low	-16.1	0.58
9	/æ/	thin (twang)	pharyngeal narrowing /nasal	520 Low	-4.8	0.35

Table 2: *Summary of mixed-effects models for Valence ratings (negative-positive) for 3 languages (Japanese, Mandarin Chinese, and Brazilian Portuguese). Z-values are reported for ordinal logistic mixed effects models (J and MC); t-tests are reported for linear mixed effects model (BP).*

Lang.		β	St. Err.	*z/t*	*p*
J	tilt	-0.064	0.020	-3.29	<.01
	f0	0.0012	0.0006	1.98	<.05
	OQ	-4.358	1.118	-3.90	<.001
MC	tilt	0.025	0.027	0.94	.35
	f0	0.008	0.001	9.49	<.001
	OQ	-3.252	1.462	-2.22	<.05
BP	tilt	-0.279	0.076	-5.46	<.001
	f0	-0.156	0.064	-2.44	<.05
	OQ	0.110	0.078	1.41	.16

Table 3: *Summary of mixed-effects models for Arousal ratings (calm-excited) for 3 languages (Japanese, Mandarin Chinese, and Brazilian Portuguese).*

Lang.		β	St. Err.	*z/t*	*p*
J	tilt	4.664e-02	2.111e-02	2.21	<.05
	f0	5.698e-03	6.902e-04	8.26	<.001
	OQ	-1.602e+01	1.252e+00	-12.79	<.001
MC	tilt	-0.081	0.026	-3.17	<.01
	f0	-0.0002	0.0009	-0.24	.81
	OQ	-8.561	1.475	-5.81	<.001
BP	tilt	0.017	0.047	0.36	.72
	f0	0.088	0.057	1.55	.12
	OQ	-0.581	0.070	-8.31	<.001

Table 3 shows the results for the arousal ratings by the three language groups. Chinese listeners' arousal ratings were most clearly affected by OQ. The vowels they found to be more excited have a lower OQ (i.e. are less breathy) and also have a shallower spectral tilt. Japanese listeners' arousal ratings were primarily affected by OQ and F0: the vowels that were judged to be exciting are high in F0 and have a low OQ (less breathy), and to a lesser extent have a shallower spectral tilt. Brazilian Portuguese rated with higher arousal those vowels that have a lower OQ (less breathy).

All three language groups rated non-breathy sounds as being more excited (more aroused). This consistency across languages arises maybe because activation is a global property of the whole utterance, based on physiology. However, valence ratings vary more across the three language groups, with Japanese and Chinese listeners preferring high F0, less breathy sounds, while Brazilian Portuguese prefer lower, more breathy sounds. The Japanese group differs also from the two others for its judgments of arousal from acoustic characteristics, with Japanese paying attention to F0 (high F0 being judged highly aroused), while this cue was not used by the two other groups for this scale.

Figure 2 shows spectral slices of the non-breathy high F0 /i/ sound preferred by Japanese and Mandarin Chinese listeners compared with the breathy low F0 one preferred by the Brazilian Portuguese listeners. Notice also the difference in amplitudes of F3 at around 2.5 kHz for each of these voices: the preferred voice by Japanese and Mandarin Chinese listeners has higher energy (54 dB) compare to that preferred by Brazilian Portuguese with 10 dB lower energy (44 dB). This increase in amplitude around 2.5 kHz is similar to that seen for pharyngeal narrowing for producing a twang-like sound; the increase in 10 dB substantially increases the perception of loudness of the sound (see, e.g., [24]).

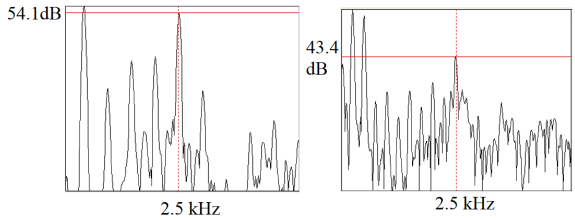

Figure 2: *Spectra of /i/ vowels. Ut. 6 produced at high F0 thick folds on left, ut 8 produced at low F0, thin folds, tongue dorsum raised on right. The horizontal lines indicate the amplitude of F3 of each of the /i/: ut 6 with thick folds has an amplitude of 54dB, ut 8 with thin folds and tongue dorsum raised has an amplitude of 44dB.*

4. Discussion & Conclusions

Ratings of arousal (calm-excited) are similar for the three language groups. OQ, a measure of breathiness, plays a key role in whether a voice is heard as excited or calm, i.e., a non-breathy voice is rated as more excited (aroused) than a breathy voice. Japanese listeners also pay attention to F0, and judge voices with higher F0 to be excited than to do the other language listeners. However, valence ratings vary among the groups: Japanese and Mandarin Chinese listeners give more positive ratings to voices with high F0 and small OQ (less breathy); Brazilian Portuguese give more positive ratings to voices with low F0 and larger OQ (more breathy). Combining the two scales of arousal and valence, the languages separate into three groups.

The finding that Japanese listeners prefer high-pitched, non-breathy voices, while Brazilian Portuguese listeners prefer low pitched breathy voices is compatible with findings from previous studies by [5] about which voices are considered to be seductive in each of these languages. Similarly, these findings are compatible with those reported in [7] about the cake-seller voice preferred by Japanese listeners compared with the voice preferred by Indian Portuguese listeners in Goa, India: the Japanese preferred the high-pitched voice with pharyngeal narrowing (twang), while the Indian Portuguese listeners preferred the lower-pitched voice with no twang.

Future work will examine more detailed aspects of glottal source contributions from the perspective of strength of glottal excitation, based on EGG data, as well as peak to peak amplitude of the glottal-flow waveform divided by the amplitude of the minimum of the flow derivative (AQ) and also pitch-Normalized Amplitude Quotient (NAQ) (AQ multiplied by the fundamental frequency of voicing [19, 20]). [21] reports that interlocutor, speaking-style, and speech-act all have significant interactions with NAQ; [22] reports that Japanese listeners "tune into" F0 to indicate speaking style (mood, brightness, and interest), while Americans "tune into" AQ to indicate speaking style (softness).

In addition, vocal tract areas measured from the MRI images will be compared with those estimated from the ARX-LF modelling, in order to fine-tune the ARX-LF model to deal with various types of laryngeal settings.

Some final comments are offered about what "valence" and "arousal" might have meant to the listeners. There may be large differences in the interpretation of the images of active ones like "cute," "cheerful," "energetic" vs. passive ones like "calm," "gentle," "warm." In order to better understand listeners' ratings, future experiments will elicit listeners short adjectival impressions of each of the voices.

The findings of this study have applications for second language teaching, and carry over to the worlds of business, politics, and advertisement; in general, the type of research may have a potential to be useful for improving communication in cross-cultural inter-personal relationships.

5. Acknowledgments

This study was partially supported by the Ministry of Education, Culture, Sports, Science and Technology, Grant-in-Aid for Scientific Research (A), 15H02605 and 23242023 to Toshiyuki Sadanobu, and by research money granted by the Keio University to Shigeto Kawahara.

6. References

[1] J. Laver, "The Phonetic Description of Voice Quality," Cambridge University Press. ISBN 0-521-23176-0, 1980.

[2] I. Yanushevskaya, C. Gobl, A. Ni Chasaide "Cross-language differences in how voice quality and f_0 contours map to affect,"The Journal of the Acoustical Society of America, vol. 144, pp. 2730-2750, https://doi.org/10.1121/1.5066448, 2018.

[3] Erickson, D., Rilliard, A., Shochi, T., Han, J., Kawahara, H., and Sakakibara, K. (2008) A cross-linguistic comparison of perception to formant frequency cues in emotional speech. *COCOSDA, Kyoto, Japan,* 163-167.

[4] Erickson, D. (2006). Some gender and cultural differences in perception of vocally-expressed affect. *Speech Prosody 2006*, Dresden, Germany, May, 2006, PS6-2-29.

[5] Rilliard, A., Erickson, D., Shochi, T., de Moraes, J.A.(2013) Social face to face communication – American English attitudinal prosody, *Interspeech 2013. Proceedings of Interspeech 2013. Lyon, 2013,* 1648-1652.

[6] Sadanobu, T., Zhu, C., Erickson, D., Obert, K. (2016) Japanese "street seller's voice". *Proc. Mtgs. Acoust*. **29**, 060003, doi: 10.1121/2.0000404.

[7] Erickson, D., Sadanobu, T., Zhu, C., Obert, K., Daikuhara, H. (2018) Exploratory study in ethnophonetics: Comparison of cross-cultural perceptions of Japanese cake seller voices among Japanese, Chinese and American English listeners. *Speech Prosody 2018.*

[8] Rilliard and J. A. de Moraes, "Social affective variations in Brazilian Portuguese: a perceptual and acoustic analysis," Revista de Estudos da Linguagem, Belo Horizonte, vol. 25.3, pp. 1043-1074, DOI: 10.17851/2237-2083.25.3.1043-1074, 2017.

[9] Juslin, P. N., and Laukka, P. "Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion," Emotion 1(4), 381–412 (2001).

[10] Scherer, K. R. "Vocal communication of emotion: A review of research paradigms,"Speech Commun. **40(1)**, 227–256 (2003).

[11] Erickson, D., Li, Yongwei., Obert, K. and Masaato Akagi, M. 2019. Estimation of glottal source and vocal tract components of perceptions of valence and arousal due to changes in larynx height. *Acoustical Society of Japan, Spring Meeting*.

[12] Estill, J., Steinhauer, K., McDonald, M. "The Estill Voice Model: Theory and Translation". Estill Voice International, LLC: Pittsburgh PA (2017).

[13] M. Hirano, "Structure and vibratory behavior of the vocal folds," Dynamic Aspect of Speech Production, University of Tokyo Press, Tokyo, Japan, pp. 13–27, 1977.

[14] M. Hirano, S. Kurita and T. Nakashima, "The structure of the vocal folds," Vocal Fold Physiology, edited by K. Stevens and M. Hirano, University of Tokyo, Tokyo, pp. 33–41, 1981.

[15] Y. Li, J. Li, M. Akagi, "Contributions of the glottal source and vocal tract cues to emotional vowel perception in the valence-arousal space", J. Acoust. Soc. Am., vol.144, doi: 10.1121/1.5051323, 2018.

[16] Husson, F., Lê, S., Pagès, J. "Exploratory multivariate analysis by example using R." Chapman and Hall/CRC, 2017.

[17] Lê, S., Josse, J., Husson, F. (2008). FactoMineR: An R Package for Multivariate Analysis. Journal of Statistical Software, 25(1), 1-18. 10.18637/jss.v025.i01

[18] R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[19] P. Alku and E. Vilkman, ''Amplitude domain quotient for characterization of the glottal volume velocity waveform estimated by inverse filtering,'' Speech Commun., 18, 131– 138 (1996).

[20] P. Alku, T. Bäcksröm and E. Vilkman, ''Normalized amplitude quotient for parameterization of the glottal flow,'' J. Acoust. Soc. Am., 112, 701–710 (2002).

[21] N. Campbell and P. Mokhtari, ''Voice quality: The 4th prosodic parameter,'' Proc. 15th Int. Congr. Phonetic Sciences, pp. 2417–2420, 2003.

[22] Erickson, D., Ohashi, S., Makita, Y., Kajimoto, N., Mokhtari, P. (2003). Perception of naturally-spoken expressive speech by American English and Japanese listeners. *Proceedings of The 1st JST/CREST International Workshop on Expressive Speech Processing, Kobe,* Feb. 21,22, 2003, pp. 31-36.

[23] P. Léon, "Précis de phonostylistique," Paris, Nathan, 1993.

[24] R. Titze and A. Palaparthia, "Vocal loudness variation with spectral slope," J. Speech, Language, and Hearing Research , 63, pp. 74–82 , 2020.