Cross Cultural perception of Valence and Arousal

Erickson, D.[1], Rilliard, A.[2], Li, Y.[3], Menezes, C.[4], Kawahara, S.[5], Sadanobu, T.[6], Hayashi, R.[7], Shochi, T.[8] Moraes, J.[9], Obert, K.[10]

Haskins, CT. USA[1], Université Paris Saclay, CNRS, LISN, Universidade Federal do Rio de Janeiro, CNPq [2], Institute of Automation, Chinese Ac. Sciences, Beijing, Ch [3], Toledo University, Ohio, USA[4], Keio University, Jn[5], Kyoto University, Jn[6], Kobe Univeristy, Jn[7], Université Bordeaux Montaigne [8], Universidade Federal do Rio de Janeiro, CNPq [9], Voice Study Centre, University of Wales Trinity Saint David, Lampeter, UK [10]
EricksonDonna2000@gmail.com[1], Albert.Rilliard@limsi.fr[2], yongwei.li@nlpr.ia.ac.cn[3], Caroline.Menezes@utoledo.edu[4], kawahara.research@gmail.com[5], sadanobu.toshiyuki.3x@kyoto-u.ac.jp[6], rhayashi@kobe-u.ac.jp[7], takaaki.shochi@labri.fr[8] , jamoraes3@gmail.com[9],kerriebobert@gmail.com[10]

ABSTRACT

This paper reports on the perception of Valence and Arousal by listeners of different cultural backgrounds, trying to reproduce previous results. Listeners (105) from five language groups were presented with vowel sounds produced by one speaker with varying voice qualities; they had to rate each on a "Calm-Excited" (Arousal) and a "Positive Negative" (Valence) five-points scales. The sounds acoustic characteristics were analysed in terms of fundamental frequency, spectral tilt and open quotient. The results show variation across language groups in their use of these acoustic cues to produce their judgements. These differences also apply to results from previous experiments, that differ for the Valence scale, but not for the Arousal one. A multidimensional analysis also showed some tendency of in-group similarity in the way to answer on these two scales.

Keywords: Cross-cultural, Valence, Arousal, Acoustic analysis, Voice quality.

1. INTRODUCTION

The aim of this pilot study is to better understand how the perception of voice quality is affected by a speaker's language and culture. It is well-known that segmental characteristics of speech, i.e., consonants and vowels, are perceived differently depending on the language of the speaker. However, there is a growing body of research reporting that voice quality characteristics of speech may also be perceived differently, depending on the language/culture of the speaker, see e.g., [1]. One approach to examining these language/cultural differences is by using a Valence-Arousal-Dominance 3-Dimensional framework of emotions, e.g., [2]–[4]. In this paper, we look at Valence and Arousal percepts of vowel sounds produced with different voice qualities. We conducted a linear effect mixed model as well as a multifactorial analysis to ascertain which acoustic cues may be contributing to the Valence and Arousal percepts.

According to some of the previous research, voices that are loud, and high-pitched tend to be perceived as aroused/excited (e.g., [1], [4]–[8]). Voices that are breathy [9] or have a steep spectral slope [4] are perceived as having positive valence. Mandarin Chinese and Japanese listeners find high-pitched, less breathy voices pleasant [1] yet German listeners find them unpleasant [8]; Brazilian Portuguese listeners prefer lower pitched, breathy voices [1].

The languages we examine for this study are five: American English (AE), Japanese (JP), Mandarin Chinese (MC), Brazilian Portuguese (BP), and Goa Portuguese (GP), as spoken in Goa, India. The study includes results from [1] re BP listeners, but reports on new perception data collected using headphones from Japanese and Mandarin Chinese listeners, since the previous study collected the perception data in a classroom with loudspeakers.

2. METHODS

2.1. Corpus

Nine vowels (either /i/ or /æ/) were produced during an MRI study by an American English female speaker, with varying voice quality: the controlled changes were (1) the phonation mode (thick, thin, and stiff vocal folds, as described within the Estill model of voice production [10], which roughly correspond to modal, falsetto, and breathy, respectively), (2) vocal tract changes produced by (nasal or not) pharyngeal narrowing (resulting in a twang voice quality (see, e.g., [11]) or a lowered/backed bunching of the tongue, and (3) F0 changes (high or low F0) (see table 1 for details). Since the recordings were collected in an MRI session, which has a limited time requirement, it was not possible to do recordings of a complete set of configurations of voice settings.

2.2. Acoustic analysis

The model does not use inverse filtering, rather it fits the residual signal to the LF model to simultaneously estimate the glottal source waveform and the vocal tract shape parameters using an analysis-by-synthesis strategy (e.g., [12], [13]). The measurements are shown in Table 1.

2.3. Listeners

Five groups of listeners with different cultural backgrounds were recruited: (1) 18 Japanese students from Kobe University; (2) 21 Mandarin Chinese University students from Beijing; (3) 20 Brazilian Portuguese students from Federal University of Rio de Janeiro; (4) 25 American Students from Toledo University (Ohio); (5) 19 listeners from Goa (India), speakers of Goa Portuguese (mean age=66 years). The five language groups were selected due to both their representing diverse languages, and also the location of the authors.

Table 1: *Articulatory and acoustics characteristics of the stimuli: vowel (V), phonation type (Phonation), pharynx narrowed (PhN), presence of nasality (N), fundamental frequency (F0), spectral tilt, and open quotient (OQ).*

	V	Phonation	PhN	N	F0 (Hz)	Tilt	OQ
1	i	stiff (breathy)	-	-	Low 240	-16.9	0.63
2	i	thin (falsetto)	-	-	Low 312	-13.4	0.43
3	i	Thick (modal)	-	-	Low 230	-11.9	0.49
4	i	Stiff (breathy)	-	-	High 480	-11.4	0.43
5	i	Thin (falsetto)	-	-	High 500	-16.1	0.4
6	i	Thick (modal)	-	-	High 520	-10.5	0.4
7	ae	Thin (twang)	+	-	High 520	-13.4	0.34
8	i	Thin (tongue lowered/backed)	+	-	Low 240	-16.1	0.58
9	ae	Thin (twang)	+	+	High 520	-4.8	0.35

2.4. Paradigm

These nine vowels were presented to the five groups of listeners, individually, through headphones, and they had to judge on a scale of 1 to 5 how "Negative" or "Positive" the sound was (if they liked the sounds, it was positive), and how "Calm" or "Excited" the sound were (5 corresponding to "Excited"). There were four randomizations of each of the nine sounds, presented through a LiveCode computer interface.

2.5. Statistics

The perceptual ratings of each of the two scales were fitted using linear mixed-effects models (using R's lme4 library [14], [15]) – one for each language group and each scale. The dependent variable was the answer on each scale (Arousal and Valence), and the fixed effects were the acoustics characteristics of the stimuli, as indicated in Table 1; random intercepts for each listener were included in the model.

Then, the four answers given by each listener to each stimulus, for each of the two scales, were summed up and arranged into a table with one row for each subject, and one column for each stimulus and scale, thus having 105 rows x 18 columns (twice 9 stimuli) table. This table was submitted to a Multiple Factor Analysis (using R's FactoMineR library [16]) so to analyze the main dimensions linked to the way listeners answer to these two scales, according to the acoustic specificities of each stimulus. The spread of the listeners along the 6 first axes of the MFA was used as an input for hierarchical clustering. An inertia gain indicated a 3-cluster solution.

Table 2: *Summary of Linear mixed-effects model for Arousal (calm-excited) for 5 language groups (AE, BP, GP, JP, MC)*

Grp	Factor	Estimate	Std. Error	t value	P
AE	F0	0.275	0.044	6.2	0.000
	tilt	0.063	0.037	1.7	0.089
	OQ	-0.496	0.055	-9.1	0.000
PB	F0	0.094	0.060	1.6	0.121
	tilt	0.018	0.050	0.4	0.719
	OQ	-0.616	0.074	-8.3	0.000
GP	F0	0.317	0.063	5.0	0.000
	tilt	-0.004	0.052	-0.1	0.940
	OQ	-0.236	0.078	-3.0	0.000
JP	F0	0.461	0.048	9.7	0.000
	tilt	0.080	0.039	2.0	0.042
	OQ	-0.281	0.059	-4.8	0.000
MC	F0	0.141	0.051	2.8	0.006
	tilt	0.053	0.043	1.3	0.211
	OQ	-0.441	0.063	-7.0	0.000

Table 3: *Summary of Linear mixed-effects model for Valence (negative-positive) for 5 language groups (AE, BP, GP, JP, MC)*

Grp	Factor	Estimate	Std. Error	t value	P
AE	F0	0.341	0.052	6.5	0.000
	tilt	-0.109	0.044	-2.5	0.012
	OQ	-0.212	0.065	-3.3	0.001

PB	F0	-0.165	0.067	-2.4	0.015
	tilt	-0.305	0.056	-5.5	0.000
	OQ	0.117	0.083	1.4	0.159
GP	F0	-0.082	0.076	-1.1	0.283
	tilt	-0.036	0.063	-0.6	0.573
	OQ	0.007	0.093	0.1	0.940
JP	F0	0.384	0.058	6.6	0.000
	tilt	-0.068	0.048	-1.4	0.154
	OQ	-0.004	0.071	-0.1	0.953
MC	F0	0.101	0.063	1.6	0.109
	tilt	0.028	0.052	0.5	0.588
	OQ	-0.207	0.078	-2.7	0.008

3. RESULTS

3.1. Regression analysis

Tables 2 and 3 present the coefficients of the linear models fitted, for each language group, to the Arousal scale (table 2), and Valence scale (table 3). AE, GP, and MC listeners gave high Arousal ratings to stimuli with high F0 and small OQ, while PB listeners relied on OQ cues only, and JP listeners on high F0, low OQ and higher spectral tilt.

AE listeners tend to rate with a positive valence the stimuli with high pitch and low tilt and OQ; conversely, BP listeners rated as positive stimuli with low pitch and low tilt. GP do not show any systematic relation with acoustic characteristics for their valence judgments, while JP relied on high pitch only, and MC on low OQ only for attributing positive valence judgments.

Table 4: *Main association with the three clusters obtained from the spread of the listeners along the main dimensions of the MFA: language groups more frequent within each cluster; stimuli whose rating by listeners of the cluster are significantly higher (+) or lower (–) than the mean ratings. The numbers in the cells represent the utterance number.*

Cluster	#1	#2	#3
Group	AE, JP	BP	GP, MC
Val +	6, 7, 9, 15	1, 8, 2, 3	3
Val –	1, 8, 3, 2	6, 9, 7, 5	–
Aro +	7, 9	6, 9, 13	1, 8, 3
Aro –	3, 1	1, 4	9, 7, 6

3.2. Multiple Factor Analysis

The hierarchical agglomeration of individual listeners is shown on the dendrogram of Figure 2, from which was derived the three clusters. The characteristics of the stimuli (high positive or negative value given to a specific stimulus on a specific scale) significantly associated with each of these clusters are given in Table 4, with the language groups that are the most representative of each cluster (i.e., groups whose frequency within the cluster is significantly higher than the global frequency).

Figure 1 presents the spread of listeners along the first two main axes of the MFA: one can show that language groups show large overlaps, but that some tendency of listeners from the same cultural background to fit on the same portion of the plane may hint for some amount of culture-specific behavior. Thus, by agglomerating all listeners on the first 6 axes, we can observe the results on a more general level representing about 80% of the variance.

Figure 2 shows how the language groups cluster, based on these first six dimensions of the MFA. Cluster #1 is predominantly composed of AE and JP listeners, cluster #2 with BP listeners, and cluster #3 with GP and MC ones.

In terms of Valence, listeners from clusters #1 and #2 tend to give opposite judgments, with the most typical stimuli for this scale being opposed systematically: (Ut5,6,7,9) were judged positive by #1 and negative by #2; (Ut1,2,3,8) were judged the other way. Listeners in cluster #3 have few associations of stimuli with the valence scale: only Ut3 was judged as positive, as was also seen for cluster #2.

For Arousal, clusters #1 and #3 tend to be opposed in their judgments, with (Ut1,3) receiving high arousal by cluster #3 and low by cluster #1, and (Ut7,9) the other way. Listeners from cluster #2 present a mixed judgment, with some oppositions and some similarities with both of the other clusters.

An interpretation of the data in terms of acoustic cues triggering the perceptions is as follows: Re Valence, AE and JP indicate a positive feeling for high F0 vowel sounds (Ut 6, 7, 9 & 5), and negative feelings for low F0 vowel sounds (Ut 1, 8, 3, & 2). BP, however, show the opposite pattern, in that they prefer low F0 vowel sounds (Ut 1, 8, 2, & 3), not high F0 vowel sounds (Ut 6, 9, 7, and 5). GP and MC show a preference for low F0 vowel sounds produced with a modal somewhat breathy voice (Ut 3).

Re Arousal, all listeners rate low F0 breathy or somewhat breathy vowels as not aroused. And, all groups (AE, JP, BP, MC and GP) rate high F0 not breathy vowels produced with twang as aroused. In addition, BP rate vowels with sustained energy in the upper frequencies are rated as aroused. They rate low breathy vowels (Ut 1) as well as high F0 somewhat breathy vowels (Ut 4) as not aroused. It is interesting that BP rate Ut 3 as aroused, but the other language groups rate it as not aroused.

As for the MC and GP groups, according to the MFA, only moderate ratings are given for arousal by these language speakers. A possible interpretation of this is that for these speakers, these specific acoustic cues are not the salient ones for perceiving Arousal for these vowel sounds.

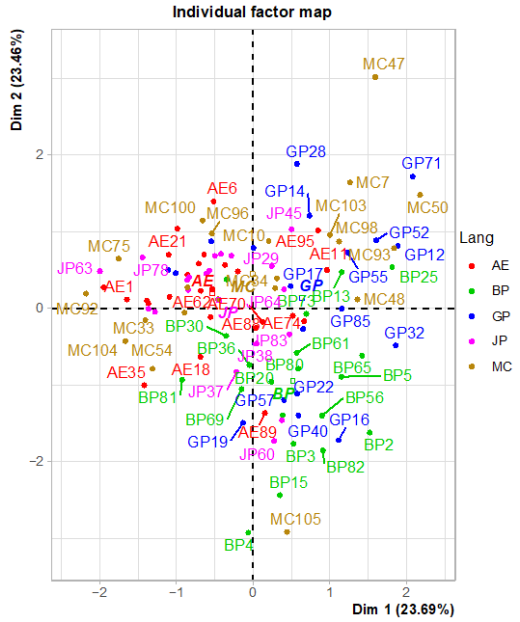

Figure 1: *Spread of the listeners on the first two dimensions of the MFA (about 45% of the variance), coloured by language group*

4. DISCUSSION & CONCLUSIONS

The preliminary results of this study confirm that listeners from different language groups perceive voice quality changes differently. We report an in-group similarity in the way to answer on these two scales, with listeners from each group being consistently clustered together on the basis of their answers.

A comment on OQ and spectral tilt: generally, when OQ increases, spectral tilt becomes steeper. Falsetto voice, often thought of a voice with high F0, generally has a large OQ and steep spectral tilt. However, adding pharyngeal narrowing to a falsetto voice changes the Speed Quotient (rate of increased contact of the vocal folds to decreased contact of the vocal folds per glottal cycle) [17], [18], as occurs when a singer/speaker produces a twang phonation. The abruptness of the vocal fold closure can affect the spectral tilt [19].

Future work is needed to further explore these cross-cultural perceptions of Valence and Arousal characteristics of voice quality in speech; specifically, we would like to include more acoustic measurements of voice quality, as well as more groups of language listeners. The five languages studied here represent a diverse group, and we would like to add more diversity.

Voice quality differences due to language/culture has of yet not been well-investigated; however, we feel this research has valuable applications for a number of professional areas, such as teaching in general, teaching second languages, clinical work, interpersonal relationships, professional speaking, advertisement, etc., basically any situation where we use our voices.

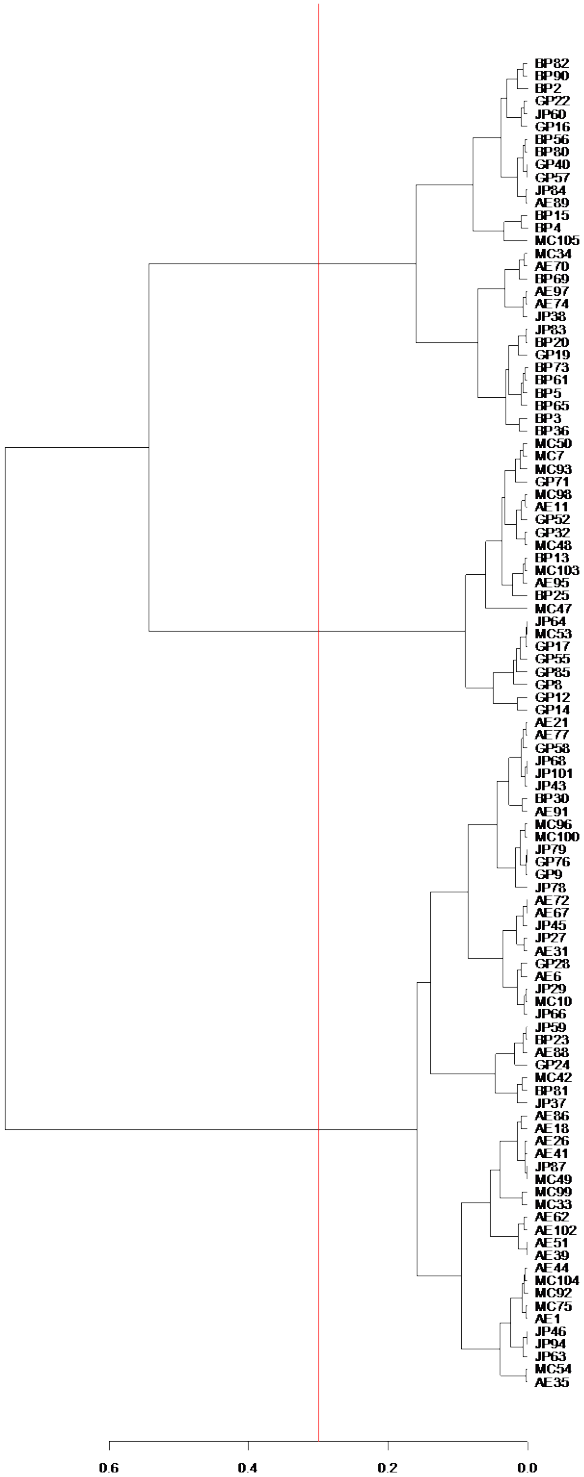

Figure 2: *Dendrogram obtained from the spread of listeners along the first six dimensions of the MFA; the red lines indicate the place where the tree has been cut, according to an inertia gain criterion.*

5. ACKNOWLEDGMENTS

We thank the subjects who participated in the perception tests. This study was partially supported by the Ministry of Education, Culture, Sports, Science and Technology, Grant-in-Aid for Scientific Research (S), 20H05630

6. REFERENCES

[1] D. Erickson *et al.*, 'Cross cultural differences in arousal and valence perceptions of voice quality', in *Speech Prosody 2020*, ISCA, May 2020, pp. 720–724. doi: 10.21437/SpeechProsody.2020-147.

[2] C. E. Osgood, W. H. May, and M. S. Miron, *Cross-cultural universals of affective meaning*. University of Illinois Press, 1975.

[3] J. R. J. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth, 'The World of Emotions is not Two-Dimensional', *Psychol. Sci.*, vol. 18, no. 12, pp. 1050–1057, Dec. 2007, doi: 10.1111/j.1467-9280.2007.02024.x.

[4] M. Goudbeek and K. Scherer, 'Beyond arousal: Valence and potency/control cues in the vocal expression of emotion', *J. Acoust. Soc. Am.*, vol. 128, no. 3, pp. 1322–1336, 2010, doi: 10.1121/1.3466853.

[5] P. N. Juslin and P. Laukka, 'Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion.', *Emotion*, vol. 1, no. 4, pp. 381–412, 2001, doi: 10.1037/1528-3542.1.4.381.

[6] K. Scherer, 'Vocal communication of emotion: A review of research paradigms', *Speech Commun.*, vol. 40, no. 1–2, pp. 227–256, Apr. 2003, doi: 10.1016/S0167-6393(02)00084-5.

[7] T. Bänziger and K. R. Scherer, 'The role of intonation in emotional expressions', *Speech Commun.*, vol. 46, no. 3–4, pp. 252–267, Jul. 2005, doi: 10.1016/j.specom.2005.02.016.

[8] J. Schmidt, E. Janse, and O. Scharenborg, 'Perception of Emotion in Conversational Speech by Younger and Older Listeners', *Front. Psychol.*, vol. 7, May 2016, doi: 10.3389/fpsyg.2016.00781.

[9] A. Anikin, 'A Moan of Pleasure Should Be Breathy: The Effect of Voice Quality on the Meaning of Human Nonverbal Vocalizations', *Phonetica*, vol. 77, no. 5, pp. 327–349, Sep. 2020, doi: 10.1159/000504855.

[10] K. Steinhauer, M. McDonald Klimek, and J. Estill, *The Estill voice model: theory & translation*. Pittsburgh, Pennsylvania: Estill Voice International, 2017.

[11] K. Perta, Y. Bae, and K. Obert, 'A pilot investigation of twang quality using magnetic resonance imaging', *Logoped. Phoniatr. Vocol.*, vol. 46, no. 2, pp. 77–85, Apr. 2021, doi: 10.1080/14015439.2020.1757147.

[12] Y. Li, J. Li, and 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, no. 2, pp. 908–916, Aug. 2018, doi: 10.1121/1.5051323.

[13] Y. Li, J. Tao, D. Erickson, B. Liu, and M. Akagi, 'F_0-Noise-Robust Glottal Source and Vocal Tract Analysis Based on ARX-LF Model', *IEEEACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 3375–3383, 2021, doi: 10.1109/TASLP.2021.3120585.

[14] D. Bates, M. Mächler, B. Bolker, and S. Walker, 'Fitting Linear Mixed-Effects Models Using **lme4**', *J. Stat. Softw.*, vol. 67, no. 1, 2015, doi: 10.18637/jss.v067.i01.

[15] R Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2022. [Online]. Available: https://www.R-project.org/

[16] F. Husson, S. Lê, and J. Pagès, *Exploratory multivariate analysis by example using R*, Second edition. Boca Raton: CRC Press, 2017.

[17] D. Erickson, J. Yun, J. Gao, and K. Obert, 'Interaction between phonation mode and pharyngeal narrowing: A pilot EGG study', in *Proceedings of the 12th International Seminar on Speech Production*, M. Tiede, D. H. Whalen, and V. Gracco, Eds., New Haven, CT, USA: Haskins Press, 2020, pp. 190–193.

[18] D. Erickson, J. Yun, K. Obert, M. Reeve, H. Rowson, and K. Møller, 'Voice quality: Interactions among F0, vowel quality, phonation mode and phayngeal narrowing.', in *Nordic Prosody 13*, 2022.

[19] K. Stevens, 'Prosodic influences on glottal waveform: preliminary data', in *international symposium on Prosody, Yokohama, Japan*, 1994, pp. 53–64.