Bilingual acoustic voice variation is similarly structured across languages

Khia A. Johnson, Molly Babel, Robert A. Fuhrman

Department of Linguistics, University of British Columbia, Canada

`khia.johnson@ubc.ca, molly.babel@ubc.ca, robert.fuhrman@ubc.ca`

Abstract

When a bilingual switches languages, do they switch their "voice"? Using a new conversational corpus of speech from early Cantonese-English bilinguals (N = 34), this paper examines the talker-specific acoustic signature of bilingual voices. Following prior work in voice quality variation, 24 filter and source-based acoustic measurements are estimated. The analysis summarizes mean differences for these dimensions, in addition to identifying the underlying structure of each talker's voice across languages with principal components analyses. Canonical redundancy analyses demonstrate that while talkers vary in the degree to which they have the same "voice" across languages, all talkers show strong similarity with themselves.

Index Terms: Bilingual speech production, Corpus phonetics, Voice quality, Voice variation, Principal components analysis

1. Introduction

Voices can tell you a lot about the person who is talking, and have been discussed as an "auditory face" [1]. They simultaneously provide information about the talker's current physical and emotional state, as well as cues to who they are [1]. However, voices are highly variable, and while different voices share some dimensions of acoustic variability, the way that voices vary is argued to be largely idiosyncratic [2, 3]. Despite the high degree of variability and idiosyncrasies, listeners nonetheless use talker-specific information to recognize and discriminate voices. Listeners are good at identifying familiar voices, but perform poorly on the same tasks with unfamiliar voices [4]. It was suggested by [2] that familiarity with a voice largely comes from learning how that voice varies across time and space, whether within an utterance or across environments, physical states, and emotions.

Bilingualism brings an additional dimension of variability into the picture. Prior research in both perception and production suggests that while some aspects of voice variability differ for linguistic reasons, other talker-indexical features remain constant across languages, and still others can be influenced by both linguistic and non-linguistic factors. That bilingual listeners are sensitive to this information signals its importance [5, 6].

In this study, we examine how voice varies across a bilingual's two languages. Some differences are expected. Languages differ in terms of their consonant and vowel inventories, which affect the spectral properties of a language. While all languages have consonants and vowels, they differ with respect to distribution, articulation, and acoustics (e.g., [7]). Suprasegmental and prosodic properties also vary across languages. Languages can differ in terms of whether a suprasegmental dimension is exploited at all, in addition to how a language might carve up a potential suprasegmental space (e.g., number of tonal contrasts). Within an individual bilingual, the acoustic variability within each language can also be related to the social identities a talker adopts within each language (see discussion in [8]).

In an effort to understand what aspects of an individual's voice vary across languages and what are more or less fixed talker-specific attributes, researchers have compared spectral properties of bilingual speech. Results have been decidedly mixed [8, 9, 10]. For example, a small group of English-Cantonese bilinguals (n = 9) in did not differ in mean fundamental frequency (F0), but exhibited greater variability in F0 [9]. This was not the case in [11], which examined voice differences with Cantonese-English bilinguals reading passages (n = 40). Based on Long-Term Average Spectral measures, females exhibited higher F0 in English than Cantonese, but males did not [11]. In the same study, all participants had greater mean spectral energy values (mean amplitude of energy between 0–8 kHz) and lower spectral tilt (ratio of energy between 0–1 kHz and 1–5 kHz) in Cantonese [11]. Respectively, these findings suggest a greater degree of laryngeal tension and breathier voice quality in Cantonese compared to English.

Together, these bodies of literature invite us to consider whether bilingual talkers have the "same" voice in each of their languages. Using a new corpus of conversational Cantonese-English bilingual speech—SpiCE [12]—we look at spectral properties [8, 9, 10, 11], and also examine how acoustic variation is structured, following the work of [2, 3, 13].

2. Methods

2.1. Data

The 34 talkers (17 female, 17 male) studied here come from the SpiCE corpus of bilingual speech in Cantonese and English [12]. The corpus comprises two conversational interviews with early bilinguals—one in each language. While these interviews were conducted in either Cantonese or English, code-switching occurred regularly within each interview. Talkers were of a similar age (*Mdn* = 21, *M* = 22.4, *SD* = 4), learned both languages from an early age (Cantonese: *Mdn* = 0, *M* = 0.03, *SD* = 0.2; English: *Mdn* = 0, *M* = 1.32, *SD* = 1.8), and considered themselves to be proficient speakers (Cantonese: Good/Excellent = 29, Fair = 5; English: Good/Excellent = 33; Fair = 1). Talkers, however, had varied language background profiles (e.g., locations lived, dialects, knowledge of other languages), and should not be characterized as a homogeneous bilingual group. This heterogeneity is an accurate representation of the Cantonese speech community [14] and it is considered a boon in this study, as analyses are conducted on a within-talker basis.

The audio recordings are high quality, with a 44.1 kHz sampling rate, 24-bit resolution, and minimal background noise. Both the participant and interviewer wore head-mounted microphones connected to separate channels, and levels were adjusted to minimize speech from the other talker. The participant channel was extracted, including any code-switches they made during the interview, as this process was done without orthographic transcripts. We then identified all voiced segments with

the *Point Process (periodic, cc)*[1] and *To TextGrid (vuv)* Praat algorithms [15], implemented with the Parselmouth Python package [16]. While speech from the interviewer can occasionally be heard in the participant channel, it is quiet enough to have been ignored by the Praat algorithms. This method captures vowels and approximants, as well as some voiced obstruents, differing slightly from [2].

2.2. Acoustic measurements

All voiced segments were subjected to the same set of acoustic measurements of voice quality made by [2][2] in VoiceSauce [17]. The suite of selected measurements are drawn from a psychoacoustic voice quality model [13]. Measurements were made every 5 ms during voiced segments, as in [2].

F0 is a correlate of pitch, associated with linguistic (e.g., lexical tone), prosodic, and talker characteristics. **F1**, **F2**, and **F3** are the three formant frequencies typically discussed in relation to linguistic contrasts (e.g., for vowels and sonorant consonants). The fourth formant frequency, **F4**, is not typically discussed in linguistic contexts, and is instead associated with talker characteristics.

The corrected amplitude difference between the first two harmonics, **H1*–H2***, is a measure of harmonic spectral slope associated with phonation type. The asterisk indicates that the value has been corrected [18]. The corrected amplitude difference between the second and fourth harmonics, **H2*–H4***, is a measure of harmonic spectral slope in a slightly higher frequency band. **H4*–H2kHz*** is the corrected amplitude difference between the fourth harmonic and the harmonic closest to 2000 Hz; it is a measure of harmonic spectral slope in a higher frequency band. The corrected amplitude difference between the harmonics closest to 2000 Hz and 5000 Hz, **H2kHz*–H5kHz***, is a measure of harmonic spectral slope that does not depend on F0. Together, these harmonic-based measures characterize source spectral shape [13].

Cepstral Peak Prominence (**CPP**) is a measure of the ratio between harmonic energy and spectral noise, and is associated with non-modal phonation types. Root Mean Square (RMS) **Energy** is a measure of overall amplitude. The subharmonics-harmonics amplitude ratio (**SHR**) is a measure of spectral noise associated with period doubling or irregularities in phonation. Together, these three measures characterize key aspects of spectral noise.

2.3. Exclusionary criteria and post-processing

Observations were removed if they included impossible or erroneous values, including instances where F0, F1–F4, CPP, or H5kHz was equal to zero. Filtering was done with just one of the uncorrected harmonic amplitude measures, as erroneous values tended to co-occur on the same observation, and the distribution of H5kHz did not span zero, with the exception of (erroneous) values equal to zero. This minimizes the removal of correctly measured zero values.

Moving standard deviations were calculated for each of the 12 measures using a centered 50 ms window ($\approx$ 10 observations). This captures dynamic changes for each of the voice quality measures, which is important as they may better reflect

what listeners attend to in talker identification and discrimination tasks [2].[3] Observations missing a moving standard deviation value (i.e., those near a voicing boundary) were removed. Including both the values measured in VoiceSauce and their moving standard deviations, a total of 24 measures were used in the analysis described in the next section. Across the 34 talkers, there were 3,126,267 observations after winnowing the data.

2.4. Principal components analysis

Principal components analysis (PCA) is a dimensionality reduction technique appropriate for data that include a large number of (potentially) correlated variables. The distillation into components helps identify and facilitate describing the internal structure, in this case, of a voice. We adapt methods from work on voices [2] and faces [19, 20]. The goal is to the capture similarities or differences for each talker's voice across languages. As such, we conducted PCAs separately for each talker-language pair, and compared the results of each talker's English and Cantonese PCAs. All 24 measures were normalized (z-scored) on by-PCA basis for the analysis. PCAs were implemented with the *parameters* package [21] in R [22], using an oblique *promax* rotation to simplify the factor structure, as the measurements reported in the previous section were expected to be somewhat correlated [2].

Each PCA included the number of components for which all resulting eigenvalues were greater than 0.7 times the mean eigenvalue, following Jolliffe's [23] recommended adjustment to the Kaiser-Guttman rule. We used this rule, rather than a more sophisticated test (e.g., broken sticks), as it is not detrimental to our exploratory analysis to err on the side of including marginal components. Additionally, across each of the components, only loadings with an absolute value of 0.32 or higher were interpreted [24, 2]; however, all loadings were retained for the canonical redundancy analysis described in the next section.

2.5. Canonical redundancy analysis

In order to assess whether variation in a talker's voice is structurally similar across both languages, we compare the PCA output from English and Cantonese by calculating redundancy indices from a canonical correlation analysis (CCA) [23, 25]. CCA is a statistical method used to explore how groups of variables are related to one another. The two sets of variables are transformed such that the correlation between the rotated versions is maximized. This is useful here, as a talker may have similar components in their English PCA and Cantonese PCA, but these components might not necessarily be in the same order, even if they account for comparable amounts of variance.

Redundancy is a relatively simple way to characterize the relationship between the loadings matrices of two PCAs—the two sets of variables under consideration here. The two indices represent the amount of variation in a talker's Cantonese PCA output that can be accounted for via canonical variates by their English PCA output, and vice versa. Notably, the two redundancy indices are not symmetrical [25].

We computed redundancy indices for all pairwise combinations, including cases where similar values were expected (same talker, different language), and cases where we expected dissimilarity (different talker and language). Considering that the PCA analyses retain the lower-dimensional structure within each language, these redundancy indices effectively reflect the degree to

---

[1]The pitch range settings differed by gender (female: 100–500, male: 75–300).

[2]The exception was formant dispersion, which was excluded because it was almost perfectly correlated with the measured value of F4 and led to problems with the principal components analyses.

[3]We used *SD*, as opposed to the coefficients of variation used by [2]. Regardless, all variables were scaled prior to inclusion the PCAs.

which the lower-dimensional structure of the voice variability is retained across a talker's two languages.

3. Results

3.1. Crosslinguistic comparison of acoustic measurements

For each acoustic measurement and talker, we conducted a Student's t-test and calculated Cohen's d, in order to give a high-level assessment of whether variable means differed across the two languages. These comparisons have no bearing on how a given variable *varies*. Table 1 reports counts of talkers by effect size. Notably, across all talkers and variables, only 21.1% yielded non-trivial Cohen's d values. Most talkers (32/34) had at least one non-trivial comparison. The distribution of these counts is depicted in Figure 1.

Table 1: *This table reports counts of Cohen's d for crosslinguistic comparisons of each of the acoustic measurements by talker. Degrees of freedom ranged between 49,274–136,644 across t-tests. For most talkers and variables, the difference in means was trivial, which is reflected in that column's high counts.*

| | Cohen's d | | |
Variable	Trivial 0.0–0.2	Small 0.2–0.5	Medium 0.5–0.8
F0	21	10	3
F0 s.d.	34	0	0
F1	24	9	1
F1 s.d.	29	5	0
F2	26	8	0
F2 s.d.	32	2	0
F3	24	9	1
F3 s.d.	29	5	0
F4	30	3	1
F4 s.d.	28	6	0
H1*–H2*	18	15	1
H1*–H2* s.d.	32	2	0
H2*–H4*	25	9	0
H2*–H4* s.d.	31	3	0
H4*–2kHz*	25	8	1
H4*–2kHz* s.d.	34	0	0
H2kHz*–5kHz*	23	10	1
H2kHz*–5kHz* s.d.	31	3	0
CPP	21	10	3
CPP s.d.	32	2	0
Energy	17	14	3
Energy s.d.	18	16	0
SHR	31	3	0
SHR s.d.	29	5	0

For the non-trivial comparisons, there were consistent patterns across languages for H1*–H2* and F0. For the remaining variables, while some talkers exhibited a difference in mean values, the direction of the difference varied, or relatively few talkers exhibited the difference.

H1*–H2* was significantly higher in Cantonese for a relatively large subset of the talkers (13/34), lower for a small number (3/34), but trivial for most (18/34). While based on a different measure than [11], this is consistent with the finding that Cantonese tends to be breathier, or English creakier—the current analysis does not distinguish between these interpretations.

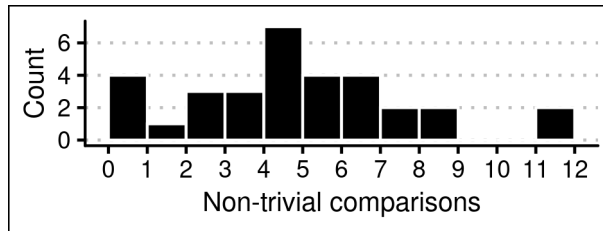If there was a non-trivial difference in F0 across languages,

Figure 1: *A summary of the number of non-trivial comparisons from Table 1 across the 34 talkers.*

then Cantonese had a lower mean F0 than English (13/34; Female = 7), though most talkers did not exhibit a difference (21/34). This is consistent with prior findings that when a difference between English and Cantonese was found, Cantonese had a lower mean F0 for females [11, 9]. We also observe this difference for a small number of males.

3.2. PCA results

The PCAs across both languages for all 34 talkers resulted in 10–15 components and accounted for 74.6–85.8% of the total variation. To assess whether talkers exhibit the same structure in voice variability across their languages, we first consider the patterns present across the different PCAs, as this provides context for understating what unique structural characteristics in talkers' voices looks like. To this end, we briefly summarize common patterns across PCA components, regardless of how much variance they account for, as the difference is often quite small. Figure 2 shows the first four components of a single talker's Cantonese and English PCAs, illustrating some examples of how components can vary (or not) across languages.
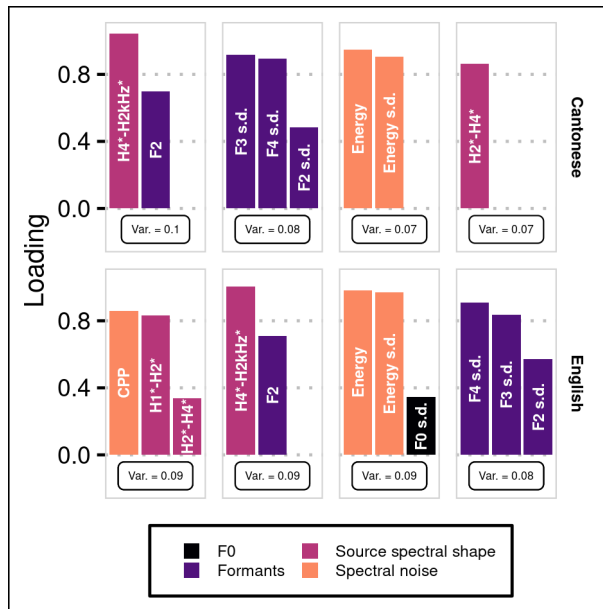

Figure 2: *In the first four components of a talker's Cantonese and English PCAs, loadings are represented by bar height and are labelled with the variable name; color represents conceptual groupings; and, the component's variance is superimposed.*

Broadly speaking, there were a lot of similarities in component composition across both talkers and languages, with the eight most commonly occurring components summarized in Table 2. For context, recall that PCAs had anywhere from 10–15 components total. These eight components consisted of source spectral shape, spectral noise, as well as formant variables. On the other hand, F0 co-occurred with a wide variety of variables (often Energy), but in a manner that was less consistent across talkers. There were additional components (not reported here) that were shared by less than half of talkers. In summary, despite the greater amount of shared structure across PCAs than found in [2], there is still ample room for idiosyncratic variation, both in terms of which variables co-occur, as well as in how much variance different components account for.

Table 2: *A summary of the most commonly occurring components across all PCAs. Variables are only included if |Loading| > 0.32. Italics indicate additional variables that were present on a component for a subset of talkers (i.e., an alternative but related configuration). N indicates the number of times a component occurred (out of 34), and Var. % gives the range of percent variance accounted for by the component.*

| | Cantonese | | English | |
Variables	N	Var. %	N	Var. %
H4*–H2kHz*, H2kHz*–H5kHz*, F2, *F3, F4*	34	9.3–15.5	32	9.2–16.7
H4*–H2kHz* s.d., H2kHz*–H5kHz* s.d.	32	6.3–8.3	34	4.1–5.0
Energy, Energy s.d, *F0*	31	5.8–9.4	33	6.3–9.1
CPP s.d.	29	4.1–5.0	31	4.1–4.9
SHR, SHR s.d.	30	3.8–7.5	29	5.4–7.3
F3, F4, *F2*	26	6.0–8.5	29	5.8–8.5
F3 s.d., F4 s.d., *F2 s.d.*	26	5.3–8.6	29	4.7–8.6
H2*–H4* s.d., H1*–H2* s.d.	26	4.2–6.5	28	4.2–6.8

3.3. Within-talker analysis

A slight majority of talkers had the same number of components for each of their languages (18/34). Of the remainder, most talkers had a difference of one in the number components (14/34), and far fewer differed by two (2/34). Redundancy indices for within-talker comparisons ranged from 0.82 to 0.99, (*Mdn* = 0.93, *M* = 0.92, *SD* = 0.04), and are displayed in Figure 3, with the two redundancy indices for a given pair plotted against one another. Comparisons across talkers within-language (range: 0.63–0.98, *Mdn* = 0.84, *M* = 0.84, *SD* = 0.6) and across-language (range: 0.66–0.98, *Mdn* = 0.83, *M* = 0.84, *SD* = 0.6) are generally lower, but still relatively high. Within-talker values were confirmed to be higher than across-talker comparisons [*Welch's t*(71.36) = −17.83, *p* < 0.001, d = 1.76].

The high values are not unexpected. As PCA is a dimensionality reduction technique, the discarded components almost certainly contain idiosyncratic variation. Moreover, and following from Section 3.2, there were a substantial number of commonly occurring patterns across talkers and languages.
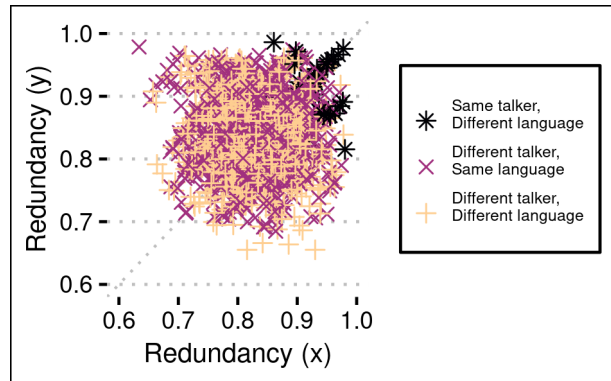

Figure 3: *The relationship between the two redundancy indices for three different types of comparisons. Within-talker comparisons are clustered at the top right.*

4. Discussion and conclusion

This study examines spectral properties and structural similarities in an individual's voice in two languages. A clear result is that most of the bilinguals studied here exhibit similar spectral properties, and similar lower-dimensional structure in voice variation, despite substantial segmental and suprasegmental differences across English and Cantonese [26]. In this sense, a majority appear to have the same "voice" across languages, which renders voice-as-an-auditory-face an apt comparison.

The comparison of these 34 Cantonese-English bilinguals' voices across languages suggest more similarity for an individual across languages than found within a more tightly controlled group of monolingual English speakers [2, 3]—several analysis decisions may have contributed to this. We compared similar components independent of order, which ignores the fact that similar components may account for different amounts of variance, but ensures that any comparisons made are among like items. Any downside to this methodological decision is mitigated by the fact that most components made relatively small contributions, accounting for 4.2–10.3% (95% highest density interval) of the PCA's total variance.

While statistical choices may have affected these results, the data differences between the current and previous studies are also important to note. This study uses substantially longer passages than the short samples in [2] and [3]. The larger speech sample may allow for a more stable underlying structure to showcase itself, as opposed to the potential for ephemeral variation in a shorter sample. This possibility is easily testable by manipulating the length of the speech sample in the analysis.

Ultimately, the goal is to understand how the acoustic variability and structure of talkers' voices maps onto listeners' organization of a voice space for use in talker recognition and discrimination. Turning to listener and behavioural data will help in deciphering what is meaningful variation within a voice from low level noise that cannot be attributed to a particular vocal signature. Verification from listener performance will help adjudicate which statistical choices present an acoustic voice space that matches listener organization.

5. References

[1] P. Belin, S. Fecteau, and C. Bédard, "Thinking the voice: Neural correlates of voice perception," *Trends in Cognitive Sciences*,

vol. 8, no. 3, pp. 129–135, 2004.

[2] Y. Lee, P. Keating, and J. Kreiman, "Acoustic voice variation within and between speakers," *The Journal of the Acoustical Society of America*, vol. 146, no. 3, pp. 1568–1579, 2019.

[3] Y. Lee and J. Kreiman, "Within- and between-speaker acoustic variability: Spontaneous versus read speech," Poster presented at at the 178th Meeting of the Acoustical Society of America, San Diego, USA, December 2019.

[4] L. C. Nygaard and D. B. Pisoni, "Talker-specific learning in speech perception," *Perception & Psychophysics*, vol. 60, no. 3, pp. 355–376, 1998.

[5] A. J. Orena, L. Polka, and R. M. Theodore, "Identifying bilingual talkers after a language switch: Language experience matters," *The Journal of the Acoustical Society of America*, vol. 145, no. 4, pp. EL303–EL309, 2019.

[6] M. Fricke, J. F. Kroll, and P. E. Dussias, "Phonetic variation in bilingual speech: A lens for studying the production-comprehension link," *Journal of Memory and Language*, vol. 89, pp. 110–137, 2016.

[7] B. Munson, J. Edwards, S. K. Schellinger, M. E. Beckman, and M. K. Meyer, "Deconstructing phonetic transcription: Covert contrast, perceptual bias, and an extraterrestrial view of vox humana," *Clinical linguistics & phonetics*, vol. 24, no. 4-5, pp. 245–260, 2010.

[8] A. Cheng, "Cross-linguistic F0 differences in bilingual speakers of English and Korean," *The Journal of the Acoustical Society of America*, vol. 147, no. 2, pp. EL67–EL73, 2020.

[9] E. P. Altenberg and C. T. Ferrand, "Fundamental frequency in monolingual English, bilingual English/Russian, and bilingual English/Cantonese young adult women," *Journal of Voice*, vol. 20, no. 1, pp. 89–96, 2006.

[10] R. Ryabov, M. Malakh, M. Trachtenberg, S. Wohl, and G. Oliveira, "Self-perceived and acoustic voice characteristics of Russian-English bilinguals," *Journal of Voice*, vol. 30, no. 6, pp. 772.e1 – 772.e8, 2016.

[11] M. L. Ng, Y. Chen, and E. Y. Chan, "Differences in vocal characteristics between Cantonese and English produced by proficient Cantonese-English bilingual speakers—a long-term average spectral analysis," *Journal of Voice*, vol. 26, no. 4, pp. e171–e176, 2012.

[12] K. A. Johnson, M. Babel, I. Fong, and N. Yiu, "SpiCE: A new open-access corpus of conversational bilingual speech in Cantonese and English," in *Proceedings of the 12th Language Resources and Evaluation Conference*, May 2020.

[13] J. Kreiman, B. R. Gerratt, M. Garellek, R. Samlan, and Z. Zhang, "Toward a unified theory of voice production and perception," *Loquens*, vol. 1, no. 1, p. e009, 2014.

[14] S. Liang, *Language Attitudes and Identities in Multilingual China: A Linguistic Ethnography*. Springer International Publishing, 2015.

[15] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer [computer program]," 2020, version 6.1.09. [Online]. Available: http://www.praat.org/

[16] Y. Jadoul, B. Thompson, and B. de Boer, "Introducing Parselmouth: A Python interface to Praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.

[17] Y.-L. Shue, P. Keating, C. Vicenik, and K. Yu, "VoiceSauce: A program for voice analysis," in *Proceedings of the 17th International Congress of Phonetic Sciences*, vol. 3, Hong Kong, 2011, pp. 1846–1849. [Online]. Available: https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2011

[18] M. Iseli, Y.-L. Shue, and A. Alwan, "Age, sex, and vowel dependencies of acoustic measures related to the voice source," *The Journal of the Acoustical Society of America*, vol. 121, no. 4, pp. 2283–2295, 2007.

[19] A. M. Burton, R. S. S. Kramer, K. L. Ritchie, and R. Jenkins, "Identity from variation: Representations of faces derived from multiple instances," *Cognitive Science*, vol. 40, no. 1, pp. 202–223, 2016.

[20] M. Turk and A. Pentland, "Face recognition using eigenfaces," in *Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE Comput. Sco. Press, 1991.

[21] D. Makowski, M. S. Ben-Shachar, and D. Lüdecke, "Describe and understand your model's parameters," *CRAN*, 2019, r package. [Online]. Available: https://github.com/easystats/parameters

[22] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2020. [Online]. Available: http://www.R-project.org/

[23] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. New York: Springer-Verlag, 2002.

[24] B. G. Tabachnick and L. S. Fidell, *Using Multivariate Statistics*, 6th ed. Pearson Education, Inc., 2013.

[25] D. Stewart and W. Love, "A general canonical correlation index," *Psychological Bulletin*, vol. 70, no. 3, pt.1, pp. 160–163, 1968.

[26] S. Matthews, V. Yip, and V. Yip, *Cantonese: A Comprehensive Grammar*. Routledge, 2013.