BILINGUAL ACOUSTIC VOICE VARIATION IS SIMILARLY STRUCTURED ACROSS LANGUAGES

UBC DL-NLP | December 16, 2020

KHIA A. JOHNSON
DEPARTMENT OF LINGUISTICS, UBC

@khia_johnson
@UBCSpeeCon

…THE PAPER YOU READ

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.

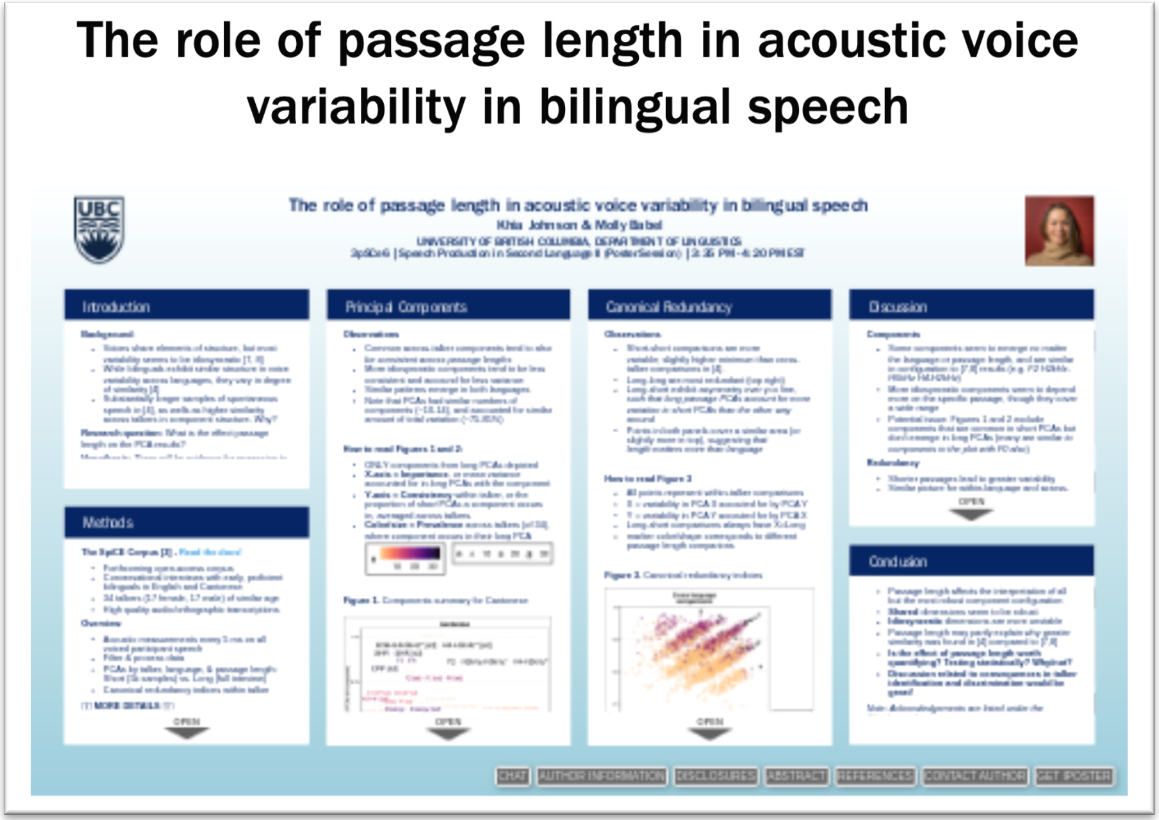
Index Terms: Bilingual speech production, Corpus phonetics, Voice quality, Voice variation, Principal components analysis

1. Introduction

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

… AND A RECENT POSTER BUILDING ON IT



THE BIG PICTURE: UNDERSTANDING HOW BILINGUAL VOICES VARY (ACOUSTICALLY), AND HOW LISTENERS LEVERAGE STRUCTURED VARIABILITY FOR SPEECH PERCEPTION

THE BIG PICTURE:

UNDERSTANDING HOW BILINGUAL VOICES VARY (ACOUSTICALLY), AND HOW LISTENERS LEVERAGE STRUCTURED VARIABILITY IN SPEECH PERCEPTION

THIS IS PART OF MY DISSERTATION

SOME ACOUSTIC & VOICE QUALITY BACKGROUND INFO

SOURCE-FILTER THEORY

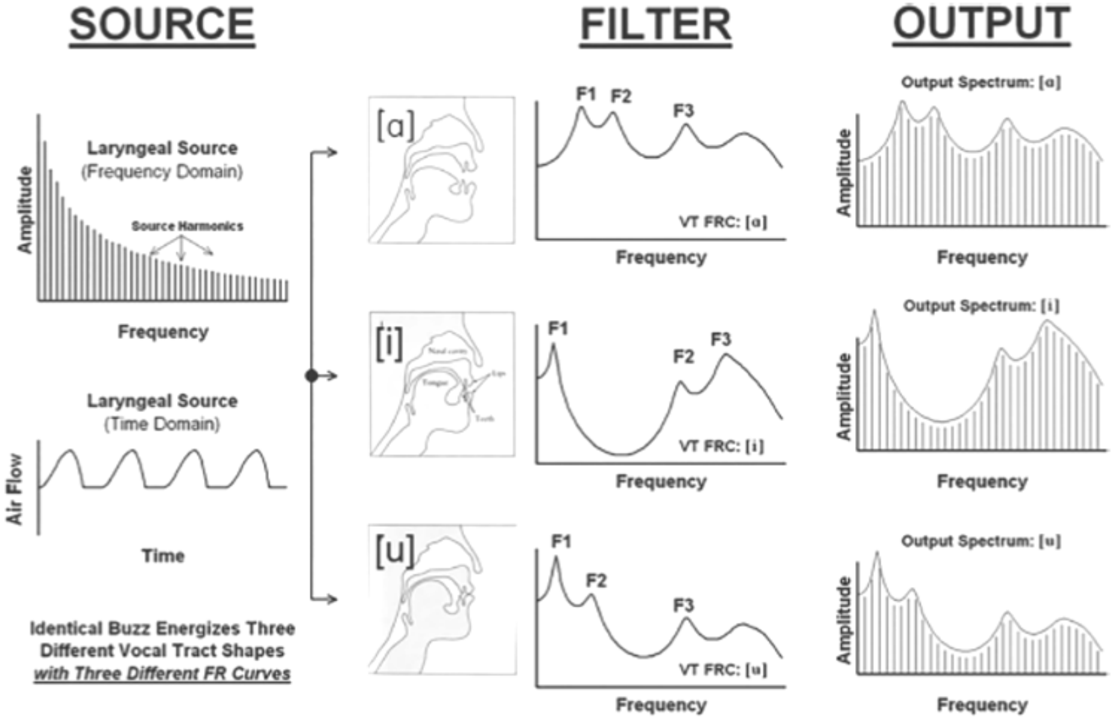


Figure 9.1 Slightly simplified version of source-filter theory for three static, phonated vowels. A single periodic glottal source signal serves as the input to three different vocal tract filters whose frequency response curves are controlled by the positions of the tongue, jaw, and lips. Amplitudes in the output spectrum are derived by multiplying the amplitudes at each frequency by the gain of the filter at those frequencies.

Hillenbrand (2019)

SOURCE-FILTER THEORY

An overly simplified representation of the source…

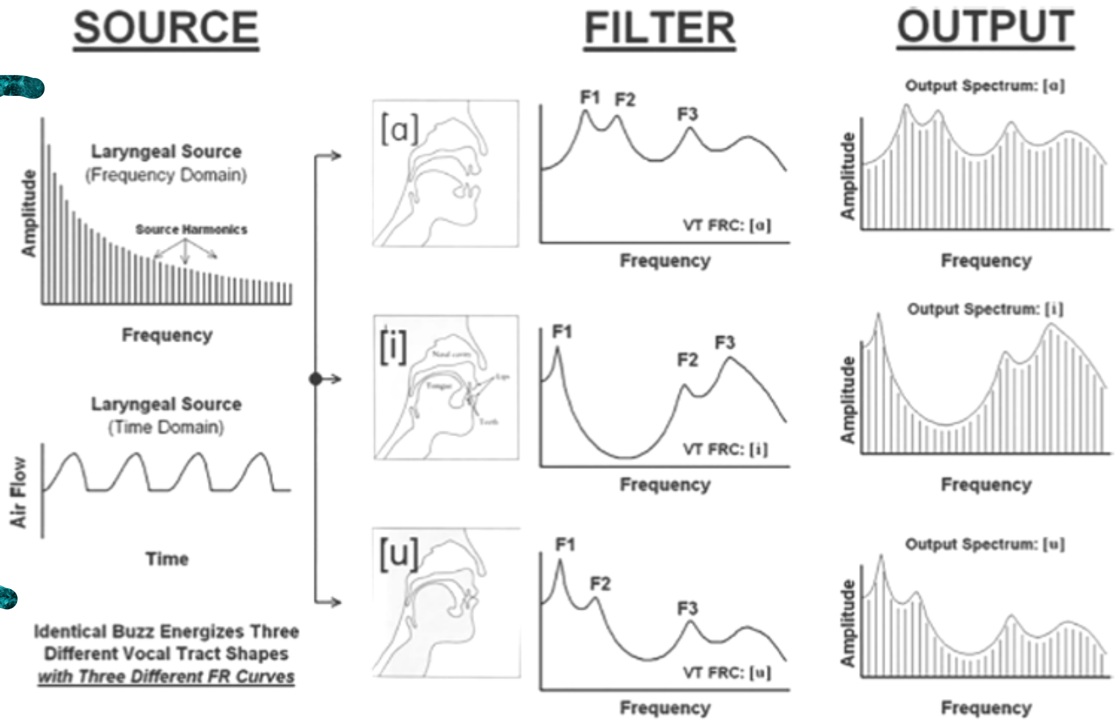


Figure 9.1 Slightly simplified version of source-filter theory for three static, phonated vowels. A single periodic glottal source signal serves as the input to three different vocal tract filters whose frequency response curves are controlled by the positions of the tongue, jaw, and lips. Amplitudes in the output spectrum are derived by multiplying the amplitudes at each frequency by the gain of the filter at those frequencies.

Hillenbrand (2019)

SO, WHAT'S THE SOURCE?

- Airflow + vocal fold configuration
- Varies on talker & linguistic dimensions

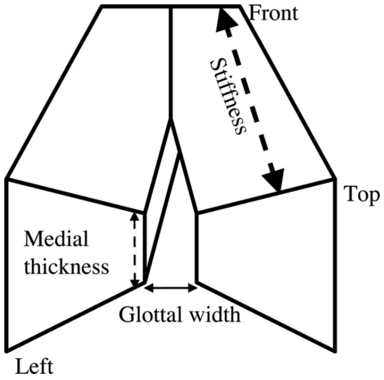
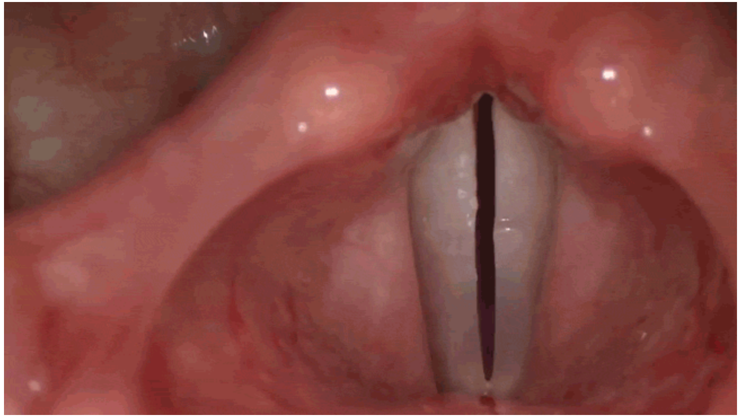


Figure 4.9 Simplified model of the vocal folds, after (Zhang, 2015, 2016a). The primary voice dimensions are influenced mainly by glottal width (the angle between the folds, thin unbroken arrow), their medial vertical thickness (thin dashed arrow), their stiffness from front to back (thick dashed arrow), and the interactions of these parameters with the subglottal pressure. Transverse stiffness is also included in later models (Zhang, 2017), but is less relevant for the primary voice dimensions in language (as discussed in text).

Garellek (2019)



GIF of https://youtu.be/9Tlpkdq8a8c

THE VOCAL FOLDS DO A LOT

Table 4.1 Primary vocal fold movements and their use in sounds of the world's languages.

Dimension	*Articulatory description*	*Relevant sounds*
Approximation	How far apart the vocal folds are from each other	All voiced sounds All voiceless sounds, e.g., aspirated sounds, glottalized sounds, fricatives, trills, and ejectives
Voicing	Whether the vocal folds are vibrating	All voiced sounds, e.g., sonorant consonants (voiced) vowels
Rate	Rate of vibration	Tone Intonation Stress
Quality	Constriction of vibration Irregularity/noise	Register Contrastive voice quality ("phonation type")

Garellek (2019)

THE BACKGROUND IN THE PAPER (& POSTER)

DIMENSIONS OF FACE VARIABILITY ARE MORE INTUITIVE

- The original inspiration for this methodology is Burton et al. (2016)
- Some example dimensions include…
  - Commonalities: direction looking, lighting conditions
  - Idiosyncrasies: facial expression, hairstyle

204 *M. Burton et al./Cognitive Science 40 (2016)*

Fig. 1. Different images of the same face, all identifiable to a familiar viewer (see Acknowledgments for attributions).

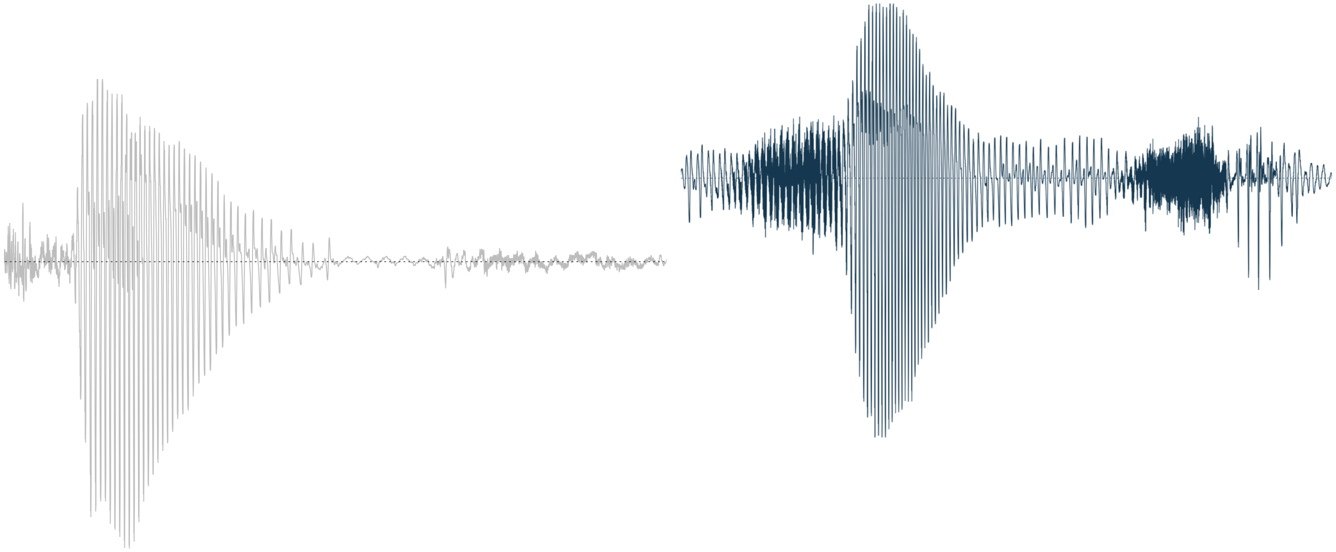
VOICES ARE ARE *ALSO* HIGHLY VARIABLE

- Apart from a small number of key commonalities, voice variability seems to be largely idiosyncratic (Lee, Keating, & Kreiman, 2019)
    - Note: to date, this area of research hasn't addressed subgroupings
- To know a voice is to know how it varies across environments, physical states, and emotions
- Is this variation influenced by language?

THE ROLE OF LANGUAGE IN VOICE VARIABILITY

- Segmental, suprasegmental, & aspects of languages vary
- Few Cantonese-English voice quality comparisons (Ng, Chen, & Chan, 2012):
  - ➔ English tends to be creakier (or less breathy)
  - ➔ Cantonese tends to have lower, more variable pitch
- Perceptual evidence that bilingual talkers can be identified after a language switch, especially by other bilinguals (Orena, Polka, & Theodore, 2019)

DO BILINGUAL TALKERS HAVE THE SAME VOICE IN EACH OF THEIR LANGUAGES?



ON TO METHODS & RESULTS…

DATA

www.spice-corpus.rtfd.io

- SpiCE Corpus (Johnson et al., 2020)
  - 34 high-proficiency, early Cantonese-English bilinguals
  - 30-minute conversational interviews in Cantonese & English
  - High-quality audio
- Pre-processing:
  - Select all voiced participant speech with Praat algorithm (Boersma & Weenink, 2020)
  - Includes vowels, approximants, & some voiced obstruents

ACOUSTIC MEASUREMENTS

- Drawn from psychoacoustic voice quality model (Kreiman et al., 2014), measurements every 5 ms with VoiceSauce (Shue et al., 2011)

Pitch F0	**Source spectral shape** H1*-H2*, H2*-H4*, H4*-2kHz*, H2kHz*-5kHz*
Formants F1, F2, F3, F4	**Spectral noise** CPP, Energy, SHR

- Post-processing
  - Remove impossible values
  - Calculate moving s.d. for each measure

A PSYCHOACOUSTIC MODEL OF VOICE QUALITY

- "…listeners perceive voice quality as an integral pattern, rather than as the sum of a number of separate features."
- "An adequate voice source model should…
    1) include enough parameters that it can model any voice quality…
    2) should only include parameters to which listeners are sensitive"

(Kreiman et al., 2014)

A PSYCHOACOUSTIC MODEL OF VOICE QUALITY

Table 4.5 Summary of psychoacoustic voice model's parameters according to primary phonological dimensions of voice.

Dimension	*Relevant model parameters*
Vocal fold approximation	Absence of f_0 track
	Aspiration noise (if vocal folds are spread)
	Voice quality changes on adjacent voiced sounds
Voicing	Presence of f_0 track
Rate of vibration	Frequency of f_0 track
Voice quality	**Breathy voice:**
(compared with modal)	Higher H1–H2, H2–H4, H4–H2 kHz, H2 kHz–H5 kHz
	Lower HNR
	Unconstricted creaky voice:
	Higher H1–H2 H2–H4, H4–H2 kHz, H2 kHz–H5 kHz
	Lower HNR Lower f_0
	Constricted creaky voice qualities
	(Prototypical creaky, tense voice, and vocal fry):
	Lower H1–H2 H2–H4, H4–H2 kHz, H2 kHz–H5 kHz
	Lower HNR (prototypical creaky voice)
	Lower f_0 (prototypical creaky voice and vocal fry)

Garrellek (2019)

Don't worry if this doesn't make sense… it's mostly to give a rough idea of what these acoustic parameters map onto

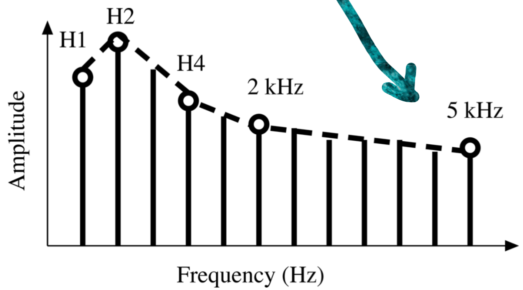


Figure 4.4 The four-parameter harmonic source spectrum model.
Source: Based on Kreiman et al., 2014.

Garrellek (2019)

YOU MAY HAVE SEEN…

190 IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, VOL. 7, NO. 2, APRIL-JUNE 2016

The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing

Florian Eyben, Klaus R. Scherer, Björn W. Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong

Abstract—Work on voice sciences over recent decades has led to a proliferation of acoustic parameters that are used quite selectively and are not always extracted in a similar fashion. With many independent teams working in different research areas, shared standards become an essential safeguard to ensure compliance with state-of-the-art methods allowing appropriate comparison of results across studies and potential integration and combination of extraction and recognition systems. In this paper we propose a basic standard acoustic parameter set for various areas of automatic voice analysis, such as paralinguistic or clinical speech analysis. In contrast to a large brute-force parameter set, we present a minimalistic set of voice parameters here. These were selected based on a) their potential to index affective physiological changes in voice production, b) their proven value in former studies as well as their automatic extractability, and c) their theoretical significance. The set is intended to provide a common baseline for evaluation of future research and eliminate differences caused by varying parameter sets or even different implementations of the same parameters. Our implementation is publicly available with the openSMILE toolkit. Comparative evaluations of the proposed feature set and large baseline feature sets of INTERSPEECH challenges show a high performance of the proposed set in relation to its size.

Index Terms—Affective computing, acoustic features, standard, emotion recognition, speech analysis, geneva minimalistic parameter set

1 Introduction

Interest in the vocal expression of different affect states has a long history with researchers working in various fields of research ranging from psychiatry to engineering. Psychiatrists have been attempting to diagnose affective states. Psychologists and communication researchers have been exploring the capacity of the voice to carry signals of emotion. Linguists and phoneticians have been discovering the role of affective pragmatic information in language pro-

There's overlap with the Kreiman et al. (2014) model, but some of the parameters in GeMAPS have no real perceptual grounds, *even if they're useful for engineering*

ACOUSTIC MEASUREMENTS

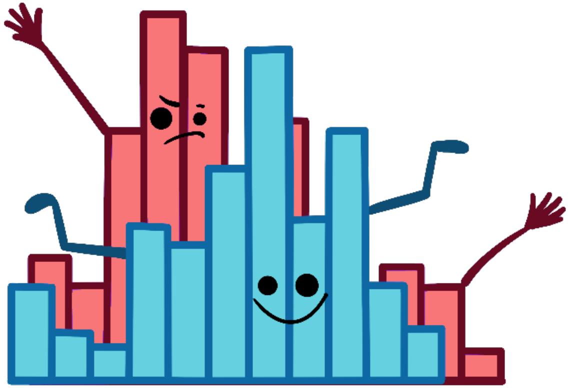
- Drawn from psychoacoustic voice quality model (Kreiman et al., 2014), measurements every 5 ms with VoiceSauce (Shue et al., 2011)

Pitch F0	**Source spectral shape** H1*-H2*, H2*-H4*, H4*-2kHz*, H2kHz*-5kHz*
Formants F1, F2, F3, F4	**Spectral noise** CPP, Energy, SHR

- Post-processing
  - Remove impossible values
  - Calculate moving s.d. for each measure

METHODS 1/3

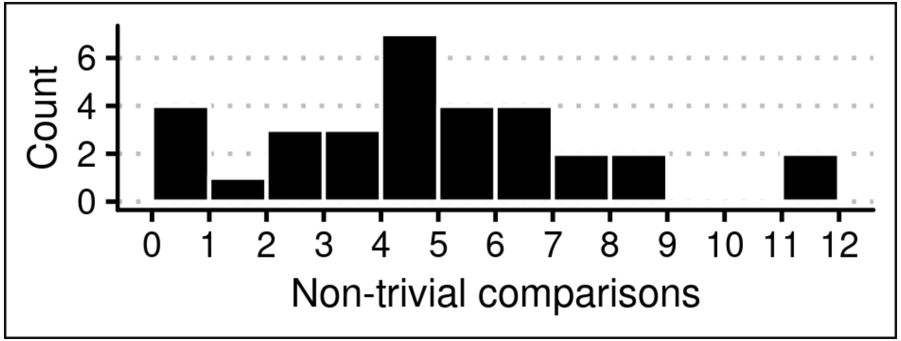
- Crosslinguistic comparison of acoustic measurements
  - → Do bilingual talkers have the same mean values for each measure?
- Within- talker Principal components analyses (PCAs)
  - → How is voice variability structured? How much of it is Idiosyncratic?
- Canonical redundancy analysis
  - → How similar are talkers across languages?



artwork by @allison_horst

COMPARISON OF ACOUSTIC MEASUREMENTS

- Cohen's *d* for t-tests within-talker, across language
- Most talkers have relatively few non-trivial comparisons

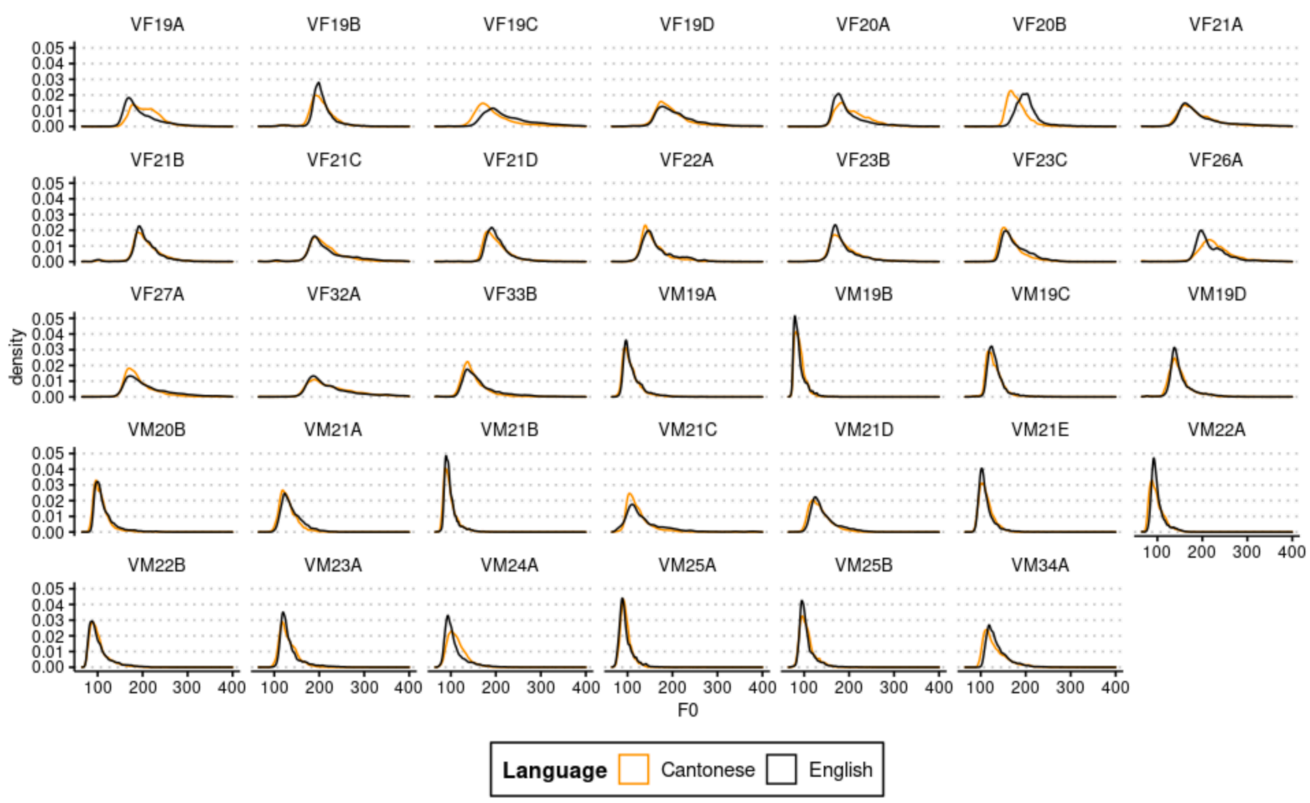


COMPARISON OF ACOUSTIC MEASUREMENTS

- Non-trivial differences tend to…
  - ➞ be small
  - ➞ lack a consistent direction
- When there is a consistent direction, it mirrors prior work
  - ➞ F0 tends to be lower in Cantonese
  - ➞ H1*-H2* consistently puts English on creakier end of spectrum

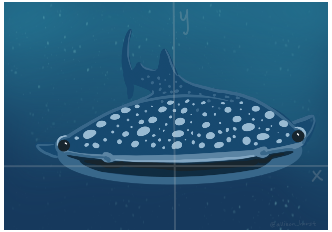
	Cohen's *d*		
Variable	**Trivial** *0.0–0.2*	**Small** *0.2–0.5*	**Medium** *0.5–0.8*
F0	21	10	3
F0 s.d.	34	0	0
F1	24	9	1
F1 s.d.	29	5	0
F2	26	8	0
F2 s.d.	32	2	0
F3	24	9	1
F3 s.d.	29	5	0
F4	30	3	1
F4 s.d.	28	6	0
H1*–H2*	18	15	1
H1*–H2* s.d.	32	2	0
H2*–H4*	25	9	0
H2*–H4* s.d.	31	3	0
H4*–2kHz*	25	8	1
H4*–2kHz*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

A CLOSER LOOK AT… FUNDAMENTAL FREQUENCY (~PITCH)



METHODS 2/3

- Crosslinguistic comparison of acoustic measurements
  - ➞ Do bilingual talkers have the same mean values for each measure?
- Principal components analyses (PCAs)
  - ➞ How is voice variability structured? How much of it is idiosyncratic?
- Canonical redundancy analysis
  - ➞ How similar are talkers across languages?



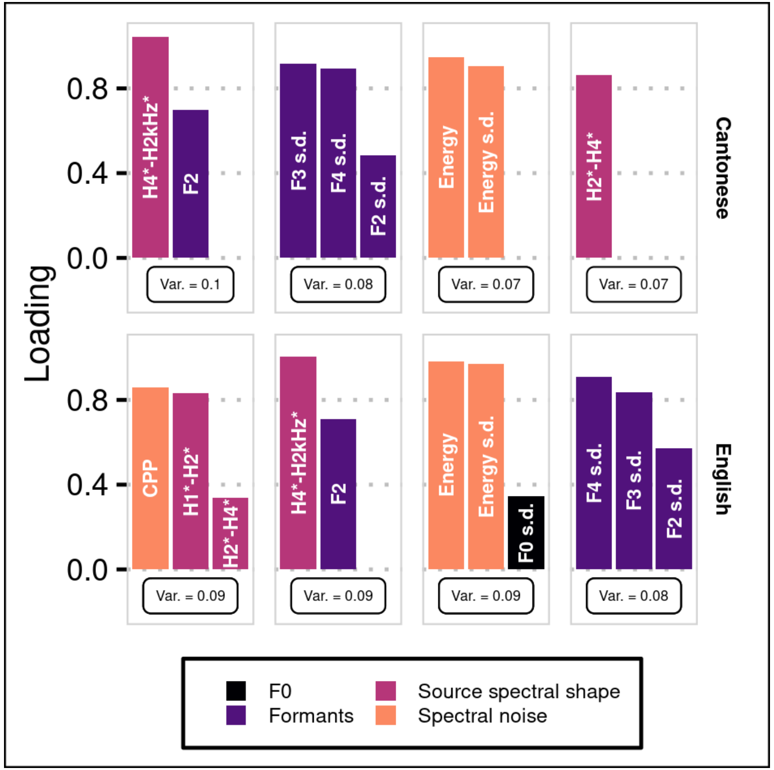
artwork by @allison_horst

PCA DETAILS (LARGELY BASED ON LEE ET AL., 2019)

- PCAs by talker and language
- All 24 measures (standardized)
- Oblique promax rotation (as measures expected to be correlated)
- Components retained if eigenvalues were > 0.7x the mean eigenvalue, a conservative choice (Joliffe, 2002)
- Only |loadings| > 0.32 were interpreted

COMPONENT VARIABILITY

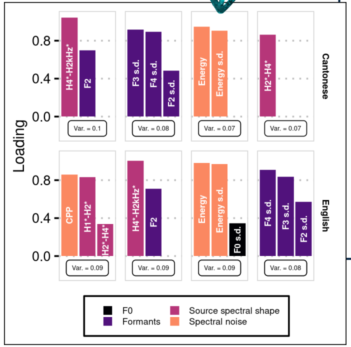
- 10–15 components accounted for 74.6–85.8% of the total variation
- Similar component structure across languages, but variable order



COMPONENT STRUCTURE

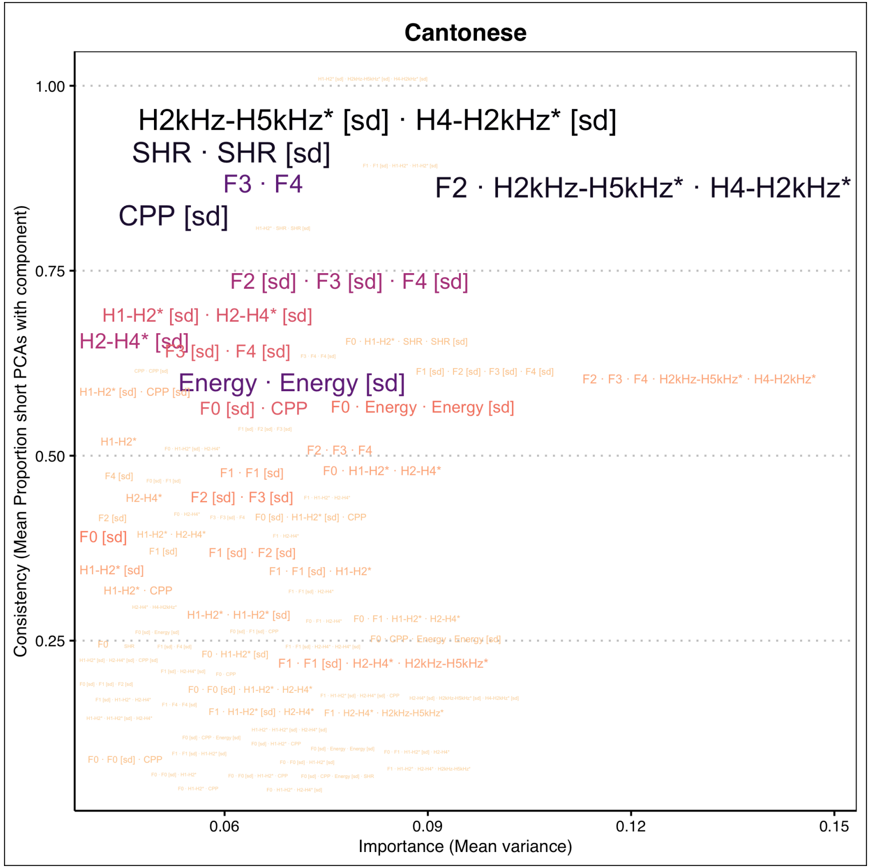
- Similar component composition across talkers and languages
- F0 is a less consistent variable
- Plenty of idiosyncratic variation

Variables	Cantonese N	Cantonese Var. %	English N	English 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



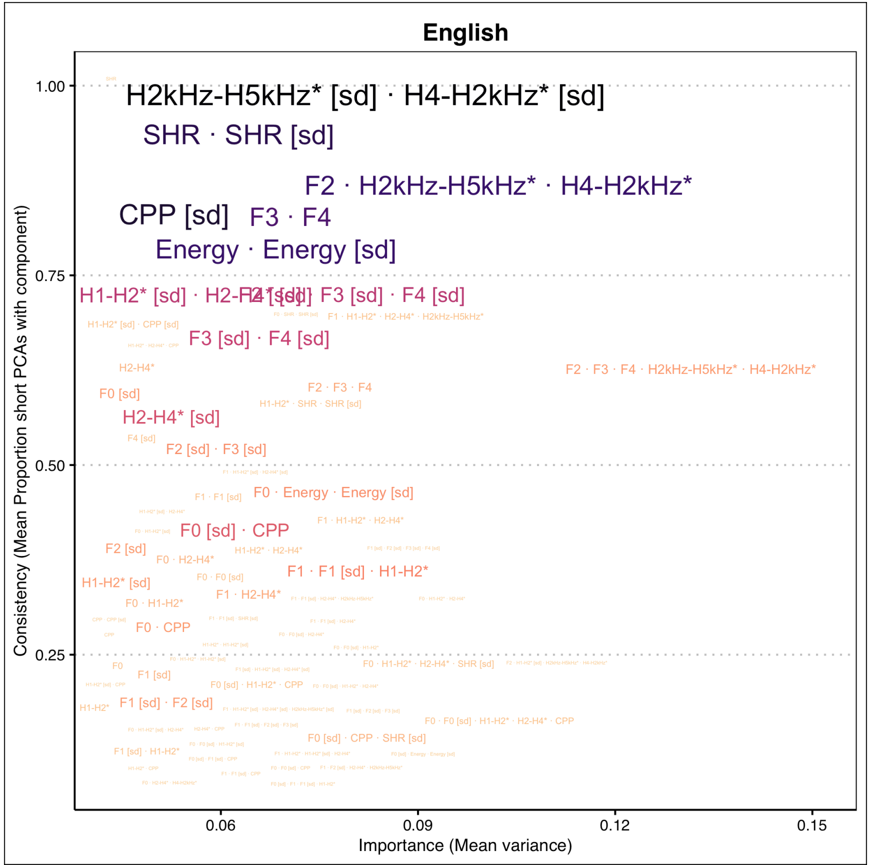
HOW COMPONENT STRUCTURE HOLDS UP ACROSS PASSAGE LENGTHS (POSTER)

- Short (~25 sec) vs. Long (~4 min) of contiguous voiced speech
- X: Importance (~ variance accounted for in long PCA)
- Y: Consistency (~how many of short PCAs have component)
- Color/size: Prevalence (~how many talkers have component)



HOW COMPONENT STRUCTURE HOLDS UP ACROSS PASSAGE LENGTHS (POSTER)

- However, there are some robust component structures
- Seemingly strong relationship between consistency and prevalence
- Passage length matters



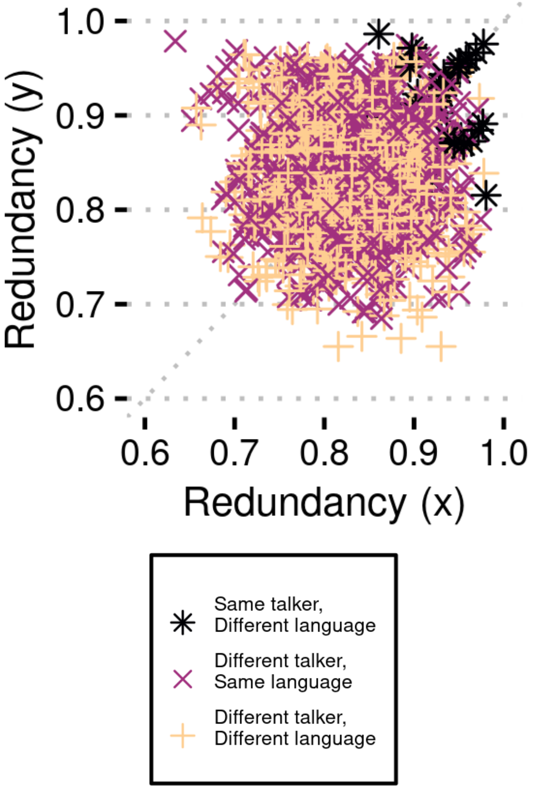
METHODS 3/3

- Crosslinguistic comparison of acoustic measurements
  - ➞ Do bilingual talkers have the same mean values for each measure?
- Within- talker Principal components analyses (PCAs)
  - ➞ How is voice variability structured? How much of it is Idiosyncratic?
- Canonical redundancy analysis
  - ➞ How similar are talkers across languages?

artwork by @allison_horst

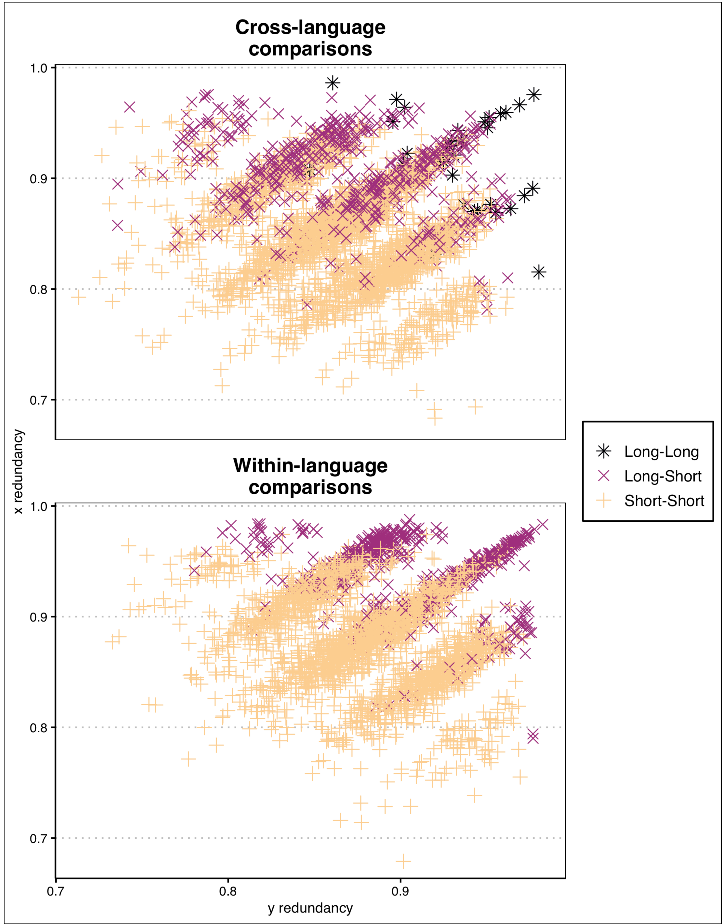
CANONICAL REDUNDANCY ANALYSIS

- Allows for comparison of two PCAs, and accounts for different component orders
- Asymmetrical, so variation in A accounted for by B and vice versa
- All loadings retained
- Within-talker comparisons are significantly more redundant: *Welch's t*(71.36) = –17.83, $p < 0.001$, d = 1.76



HOW REDUNDANCY HOLDS UP ACROSS PASSAGE LENGTHS (POSTER)

- Short comparisons are more variable
- Within-language might have higher redundancy, but not immediately clear
- Takeaway: passage length matters
- More to do here!



DISCUSSION

- Methodological differences from work by Lee & colleagues (2019, 2020)
  Initial follow up with passage length…
  - Seems to be important for all but the most robust components
  - Might be more important than language differences
- Robust components seem to show up no matter
- Despite substantial segmental & suprasegmental differences across English & Cantonese, bilinguals exhibit similar spectral properties and structure in voice variability ➞ voices are like "auditory faces"

MOVING FORWARD

- Refine the analysis to better account for passage length, etc.
- This work generates predictions related to bilingualism and cognitive organization of voices in speech perception
    - currently testing perception
    - interest in comparing differences in human and machine identification/discrimination with voice (but no black boxes!)

THANK YOU!

SpiCE was developed with support from Nancy Yiu, Ivan Fong, Ariana Zattera, Christina Sen, Kristy Chan, Katherine Lee, Rachel Wong, Rachel Soo, and members of the Speech-in-Context Lab: **www.speechincontext.arts.ubc.ca**

The SpiCE corpus will be available soon! Follow for updates: **@khia_johnson**

www.spice-corpus.rtfd.io

Natural Sciences and Engineering Research Council of Canada
Conseil de recherches en sciences naturelles et en génie du Canada
Canada

Social Sciences and Humanities Research Council of Canada
Conseil de recherches en sciences humaines du Canada
Canada

a place of mind
THE UNIVERSITY OF BRITISH COLUMBIA

Graduate and Postdoctoral Studies
WWW.GRAD.UBC.CA

REFERENCES

Boersma, P., & Weenink, D. (2020). Praat: Doing phonetics by computer [Computer software]. http://www.praat.org/

Burton, A. M., Kramer, R. S. S., Ritchie, K. L., & Jenkins, R. (2016). Identity From Variation: Representations of Faces Derived From Multiple Instances. Cognitive Science, 40(1), 202–223. https://doi.org/10/f778gx

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., Andre, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., & Truong, K. P. (2016). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Transactions on Affective Computing, 7(02), 190–202. https://doi.org/10/f3sfxq

Garellek, M. (2019). The phonetics of voice. In The Routledge Handbook of Phonetics. https://doi.org/10.4324/9780429056253-5

Hillenbrand, J. M. (2019). The acoustics and perception of North American English vowels. In The Routledge Handbook of Phonetics. https://doi.org/10.4324/9780429056253-10

Jolliffe, I. T. (2002). Principal Component Analysis (2nd ed.). Springer-Verlag. https://doi.org/10.1007/b98835

Johnson, K. A., & Babel, M. (2020). The role of passage length on acoustic voice variability in bilingual speech [abstract]. The Journal of the Acoustical Society of America, 148, 2659–2659. https://doi.org/10.1121/1.5147411

Johnson, K. A., Babel, M., Fong, I., & Yiu, N. (2020). SpiCE: A New Open-Access Corpus of Conversational Bilingual Speech in Cantonese and English. Proceedings of The 12th Language Resources and Evaluation Conference, 4089–4095. https://www.aclweb.org/anthology/2020.lrec-1.503

Johnson, K. A., Babel, M., & Fuhrman, R. A. (2020). Bilingual Acoustic Voice Variation is Similarly Structured Across Languages. Proc. Interspeech 2020, 2387–2391. https://doi.org/10.21437/Interspeech.2020-3095

Lee, Y., & Kreiman, J. E. (2019). Within- and between-speaker acoustic variability: Spontaneous versus read speech [abstract]. The Journal of the Acoustical Society of America, 146, 3011–3011. https://doi.org/10.1121/1.5137431

Lee, Y., & Kreiman, J. (2020). Language effects on acoustic voice variation within and between talkers [abstract]. The Journal of the Acoustical Society of America, 148(4), 2473–2473. https://doi.org/10.1121/1.5146847

Lee, Y., Keating, P., & Kreiman, J. (2019). Acoustic voice variation within and between speakers. The Journal of the Acoustical Society of America, 146(3), 1568–1579. https://doi.org/10.1121/1.5125134

Kreiman, J., Gerratt, B. R., Garellek, M., Samlan, R., & Zhang, Z. (2014). Toward a unified theory of voice production and perception. Loquens, 1(1), 009. https://doi.org/10.3989/loquens.2014.009

Ng, M. L., Chen, Y., & Chan, E. Y. K. (2012). Differences in Vocal Characteristics Between Cantonese and English Produced by Proficient Cantonese-English Bilingual Speakers—A Long-Term Average Spectral Analysis. Journal of Voice, 26(4), e171–e176. https://doi.org/10.1016/j.jvoice.2011.07.013

Orena, A. J., Polka, L., & Theodore, R. M. (2019). Identifying bilingual talkers after a language switch: Language experience matters. The Journal of the Acoustical Society of America, 145(4), EL303–EL309. https://doi.org/10.1121/1.5097735

Orena, A. J., Polka, L., & Theodore, R. M. (2019). Identifying bilingual talkers after a language switch: Language experience matters. The Journal of the Acoustical Society of America, 145(4), EL303–EL309. https://doi.org/10.1121/1.5097735