Analysing the effect of language on speaker-specific speech rhythm in Cantonese-English bilinguals

Adas Li[1,2], Peter French[1,3], Volker Dellwo[4] and Eleanor Chodroff[1]
[1]University of York, [2]University of Hong Kong, JP French Associates, York,
[4]University of Zürich

Background. Undertaking forensic speaker comparison (FSC) cases in which the questioned sample is in a different language from the known sample has caused concern within the forensic phonetics community (see clause 3.10 of [1]). A major reason for such reservations is the lack of research knowledge concerning which aspects of speech are robust to language switching. To develop an empirically-justified basis for conducting cross-language analysis, it is important to widen the bilingual focus of FSC research. Within a single language, rhythm is known to be a speaker-specific parameter in FSC. However, little is known about its potential as a speaker discriminant for bilinguals in forensic practice. Few studies have examined how rhythm varies between bilingual speakers and between their L1 and L2 languages; and even fewer have examined this between two typologically distinct languages such as Cantonese and English [2][3].

This research extends previous literature on bilingual rhythm variability and L2 rhythm studies to assess the robustness of rhythm characteristics against language-specific effects and to identify speaker-specific parameters within FSC research: 1) Is there significant between-speaker rhythmic variability among Cantonese-English bilinguals?, 2) To what extent do other factors (language and speaker effects) contribute to such variability?, and 3) To what extent can we predict a speaker's rhythm in their L2 English speech based on their L1 rhythmic patterns, and vice versa?

Methods. The speech data used in this study was retrieved from the ALLSSTAR (Archive of L1 & L2 Scripted and Spontaneous Transcripts and Recordings) corpus [4][5]. The speech materials were produced by 14 Cantonese-English bilinguals (8 females and 6 males, mean age = 22, range = 19–27), whose English language proficiency ranged from sufficiently high for US undergraduate studies to native proficiency. All speakers indicated Cantonese as their L1. Speakers produced two 'Hearing in the Noise Test' sentence sets and the 'North Wind and the Sun' passage in both Cantonese and English.

The audio files (44.1kHz sampling rate, 16 bit depth, mono) were segmented by utterance, and then aligned at the phone level with Montreal Forced Aligner [6]. Speech disfluencies were removed and manual correction was carried out to ensure the precision of the alignment. Following [7], the five rhythm measurements (RMs) of rateCV, V%, ΔV(ln), ΔC(ln), and ΔPeak(ln) were taken using the *durationAnalyzer* Praat script [8][9].

Statistical analysis. General observations were made using descriptive analysis and visual plotting in *R*. To investigate the degree to which rhythm varied by language and speaker, a series of linear regression models was performed and compared: a) a model with a single fixed effect of language, b) a Linear Mixed-Effects (LMMs) Model with a fixed effect of language and a random intercept of speaker and, lastly, c) a LMMs Model with a fixed effect of language, a random intercept for each speaker, as well as a random slope of language for each speaker. The three models' goodness of fits were then compared using the Akaike Information Criterion (AIC). To assess the potential utility of speaker-specific rhythm for speaker identification, a closed-set speaker classification task was implemented using a multinomial logistic regression that predicted the speaker's identity from the 5 RMs and their interactions for each utterance. 4 models were trained and tested using mutually-exclusive subsets within and across the 2 languages.

Results. AIC strongly favoured the model with by-speaker intercepts and slopes for language for each of the five RMs. The considerable improvement in model fit indicated sizable speaker-specific influences on overall and language-specific rhythm variability.

The potential of using speaker-specific rhythm for classification was assessed with a multinomial logistic regression. Average accuracy in speaker classification was the highest when the model was trained and tested on the same language: 25% for Cantonese and 19% for English but it decreased in cross-language prediction (in which the model was trained on the rhythm features of one language and tested on those of the other): 12% for the Cantonese-trained-English-tested model and 11% for the English-trained-Cantonese-tested-model. Despite the low accuracy of the overall training models, the accuracy was significantly above chance for all four models, as estimated by bootstrapping with 10,000 samples. This finding reveals some contribution of speaker-specific rhythm for speaker classification both within a language, and critically across the speaker's two languages. Further research is, however, necessary to determine the extent to which classification accuracy is influenced by other intervening variables such as the speakers' level of competence or intelligibility in their L2 English, the type of L1 dialects spoken, as well as the stimuli used in the experiment.

Conclusion. Overall, this paper examines the effect of language and speakers on rhythm variability and finds it of some – but limited – use in speaker comparisons. Consistent with previous findings, we identified significant effects of language and speaker on rhythm variability [2][3], thus leaving reservations about undertaking cross-linguistic analysis within the field of FSC casework relatively unallayed. Nevertheless, a small and significant degree of speaker-specificity is present in the realisation of rhythm. This individuality and cross-language predictability of rhythm could still be useful in FSC cases, at least in combination with more canonical speaker-specific features. These findings highlight the need for all aspects of speech to be researched as potential candidates for inclusion in cross-language comparisons.

References.

[1] International Association for Forensic Phonetics and Acoustics, Code of Practice (2020). http://www.iafpa.net/the-association/code-of-practice/

[2] Dellwo, V., Schmid, S., Leemann, A., Kolly, M.-J., & Müller, M. (2012, July). Speaker identification based on speech rhythm: the case of bilinguals. *Perspectives on Rhythm and Timing (PoRT)*. https://doi.org/10.5167/uzh-111811

[3] White, D., & Mok, P. (2019). L2 Speech Rhythm and Language Experience in New Immigrants. *The 19th International Congress of Phonetic Sciences (ICPhS 2019)*, 1–5.

[4] Ackerman, L., Burchfield, L. A., Hesterberg, L., Bradlow, A. R., Luque, J. S., & Mok, K. (2010). *ALLSSTAR Project Manual*, https://groups.linguistics.northwestern.edu/speech_comm_group/allsstar2/#!/manual

[5] Bradlow, A. R. (n.d.) ALLSSTAR: Archive of L1 and L2 Scripted and Spontaneous Transcripts And Recordings. Retrieved from https://oscaar3.ling.northwestern.edu/ALLSSTARcentral/#!/recordings.

[6] McAuliffe, Michael, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger (2017). Montreal Forced Aligner [Computer program]. Version 0.9.0, retrieved 17 January 2017 from http://montrealcorpustools.github.io/Montreal-Forced-Aligner/

[7] Dellwo, V., Leemann, A., & Kolly, M.-J. (2015). Rhythmic variability between speakers: Articulatory, prosodic, and linguistic factors. *Journal of the Acoustical Society of America*, *137*(3), 1513–1528. https://doi.org/10.1121/1.4906837

[8] Dellwo, V. (2019). *Praat script: Duration Analyzer (version 0.03)*. University of Zurich. https://www.pholab.uzh.ch/static/volker/software/plugin_duratio%0AnAnalyzer.zip

[9] Boersma, Paul & Weenink, David (2020). Praat: doing phonetics by computer [Computer program]. Version 6.1.22, retrieved 24 September 2020 from http://www.praat.org/.