SPEAKER INDIVIDUALITY IN THE DURATIONAL CHARACTERISTICS OF VOICED INTERVALS: THE CASE OF CHINESE BI-DIALECTAL SPEAKERS

Yu Zhang [1]; Lei He [1,2]; Volker Dellwo [1]

Institute of Computational Linguistics, University of Zurich
Department of Linguistics, University of Tübingen
{yu.zhang | lei.he | volker.dellwo}@uzh.ch

ABSTRACT

Temporal organizations of the speech signal are highly individual among speakers of the same language. In the present study, we looked at speech production of bi-dialectal speakers using two varieties of the same language. We aimed at testing whether speaker-specific temporal features present in one dialect remain in another dialect of the same speaker. 20 sentences and one passage in both Mandarin and Danyang Dialect of 14 bi-dialectal speakers were recorded. We measured between-speaker variability of the percentage of voiced interval duration (percentVO) in both dialect conditions using linear mixed effect models. Results revealed that speakers exhibited distinct between-speaker variability when dialect variability and style variability were introduced. However, within-speaker variability was also present and the magnitude of the variability differences varied among different speakers and in different speaking styles. Findings of the current study are particularly relevant for forensic voice comparison tasks when a mismatch in speaking languages in trace and suspect materials is present.

Keywords: temporal characteristics, speaker individuality, articulation, bi-dialectal speakers.

1. INTRODUCTION

People differ profoundly in their voices. A number of studies have reported strong between-speaker variability in the temporal organization of the speech signal of the same language (durational- and intensity- based rhythmic measures, formant dynamics, intensity dynamics). Studies were carried out on both segmental and supra-segmental levels. On the segmental level, marked individual differences were observed in speakers' formant frequency dynamics ([15], [16]) and within-syllable durational characteristics ([19]). Supra-segmental or rhythmic characteristics of speech also vary between speakers. Durational variabilities of voiced, voiceless, vocalic and intervocalic intervals were found to be speaker-idiosyncratic ([5], [14]). Dellwo, Leemann and Kolly ([7]) further reported robust between-speaker effects on the durational variability between syllabic amplitude peak points when within-speaker prosodic and linguistic variability were introduced. In addition, syllable intensity characteristics also play an important role in between-speaker rhythmic differences. He and Dellwo ([11]) examined the dispersion of syllable intensity levels and the mean differences between consecutive syllable intensity levels, and obtained stronger speaker specific information in these intensity-based measures compared to duration-based measures of phonetic interval variability. They further expanded the study by looking at the dynamic process of intensity change within a syllable. Strong speaker effect was obtained in this dynamic process as well. Moreover, the parts of speech signal where intensity decreases from the peak point to the trough point were found to contain more speaker idiosyncrasy ([12]).

Voice individuality as manifested in the above-mentioned speaker-specific production of prosodic variabilities is believed to be largely due to idiosyncratic articulation in speech production ([6], [10], [11], [12] [15], [16]). Speakers may differ in their anatomical make-ups of the articulatory apparatus (e.g. jaw, tongue, lips, velum, etc.), which requires them to coordinate the articulators in speaker-specific ways during speech. Articulatory movements may be constrained by one's intentions and expectations, which vary strongly between different speakers ([9], [17]). While there has been a large body of evidence showing that temporal features of speech vary between speakers of the same language ([1], [4], [6], [14], [21], [22]), speaker individuality for those capable of speaking in different languages or language varieties remains a largely neglected scenario, although it is crucial for the understanding of the source of between-speaker rhythmic variability. First evidence that speakers vary systematically in terms of suprasegmental temporal characteristics across languages is present from Italian-German bilinguals ([8]). The researchers reported speaker-specific effects for measures under observation, such as articulation rate, %VO, %V and deltaV. In particular speakers showed systematic higher %VO when speaking in Italian than in German, which might be due to a relatively high number of voiced consonants in the phoneme

inventory and a preference towards open syllables in Italian language compared to German language. Since the acoustic dimensions of speech carry not only the signals about the speaker itself, but also signals about the particular phonological form being produced (i.e. the language), variations arising from distinct phonological systems of different languages are expected to impact how speaker information is conveyed through the acoustic signal.

In the present study, we did a comparable analysis by looking at the speech of bi-dialectal speakers who use two varieties of the same language – local and supralocal – in their linguistic repertoire. Speech of 14 Mandarin – Danyang dialect speakers were measured in terms of the durational variability of voices and voiceless intervals in the speech signal (percentVO) across both varieties. The particular choice of this temporal measure was motivated by recent evidences in the literature pointing to percentVO exhibiting strong between-speaker variability ([5], [6], [8], [14],[18]).

We chose two distinct language varieties in China in consideration of the rich language diversity there, especially in the southeast area. Danyang dialect is a variety of Wu Chinese spoken in the central districts of the city of Danyang. Like other Wu variants, it is mutually unintelligible with other varieties of Chinese, such as Mandarin. Meanwhile, the particular choice of language varieties was motivated by previous research showing that Wu varieties are distinguished by their retention of voiced or murmured obstruent initials, such as stops, affricates and fricatives ([20]), meaning that it reveals strong differences in the parameter under observation compared to Mandarin speech. By choosing these two particular varieties, we aim at testing whether and to what extent bi-dialectal speakers accommodate to two distinct phonological systems, and how dialect variability might influence their productions of voiced interval durations. Speaker idiosyncrasy in the speech of bi-dialectal speakers is influenced by the rhythmic differences between two varieties of the same language. If it is the case that speaker-specific ways of operating the articulators are the driving force behind idiosyncratic speech production, then we would well expect that a bi-dialectal speakers reveal similar voicing patterns when speaking both dialects.

2. METHOD

2.1. The database

14 bi-dialectal speakers (9 female and / 5 male speakers; average age=38, min age=24, max age=48) of Danyang dialect and Mandarin were chosen for the present study. They were all born and living in the city of Danyang where Danyang dialect and Mandarin are the two major languages in use, and thus considered native speakers of both dialects and Mandarin. Speakers were recorded reading the following passage and 20 sentences in both Dialect and Mandarin using Zoom Handy Recorder H2 in a quite room (sampling rate: 44.1 kHz; quantisation depth: 16 bits; format: wav).

2.2. Data processing and measurement

Voiced and unvoiced intervals in the speech signal were automatically processed using the Praat voicing detection function (To TextGrid (vuv)). A voiced interval (v) is the stretch of the signal where periodic laryngeal activity could be detected in the speech signal. All remaining parts of the signal are unvoiced (u).

2.3. Statistical analysis

To test the significance of between-speaker variability captured by the temporal measure percentVO, mixed-effects models were employed using the R package lme4([3]). Language, gender and style were modeled as fixed factors; speaker and sentence were modeled as random intercepts (rationale: speakers were a sample of the bi-dialectal speaking population of Mandarin and Danyang dialect, and sentences were a sample of an infinitely large population of possible Mandarin or Danyang dialect sentences; [2]).

We first carried out interaction tests among all three fixed factors (language, gender and style) and only observed significant inter-dependence between gender and style. So main effect tests were only carried out on the fixed factor language and the random variable speaker, which are the two parameters of interest in the present study

Effects were tested by model comparison between a full model in which the factor in question is included as either a fixed or a random effect and a reduced model in which the factor in question is excluded. The significance of an effect was tested by comparing the results from the two models using standard ANOVAs. For the assessment of the relative goodness of fit we indicate AIC (Akaike Information Criterion) values, which decrease with goodness of fit ([13]). Details of fitted models are presented in Table 1.

3. RESULTS

The results of the mixed-effects models (fitted by maximum likelihood) for between-speaker variability of the percentVO measure with the bi-dialectal speakers were presented in Table 2.

Table1 Description of all fitted mixed-effects models with percentVO as dependent variable

Model ID	Model descriptions			Remarks
	Dependent variable	Fixed effect	Random intercept(s)	
MDL 1	percentVO	language; style; gender	speaker; sentence	full model
MDL 2	percentVO	language; style; gender	sentence	speaker-reduced model
MDL 3	percentVO	style; gender	speaker; sentence	language-reduced model

Table 2 Result of mixed-effects model comparisons using likelihood ratio tests. Alkaike Information criterion values in boldface indicate better fit

Model Comparison	Akaike Information Criterion (AIC)	χ^2[df]	p
MDL1, MDL2	**-2195.6** $_{\mathbf{MDL1}}$ -1560.7$_{MDL2}$	636.87	<2.2e-16
MDL1, MDL3	**-2196.6**$_{\mathbf{MDL}}$ -2186.8 $_{MDL3}$	10.789	0.001021

Figure 1: Box plot showing percentVO values for sentences in two dialects (Danyang dialect [dia] and Mandarin [man]) for 14 speakers. Each speaker uttered 20 sentences in each dialect.

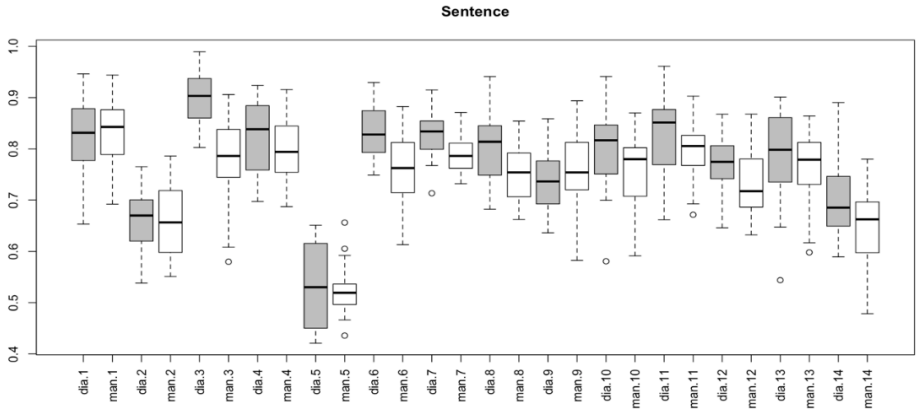
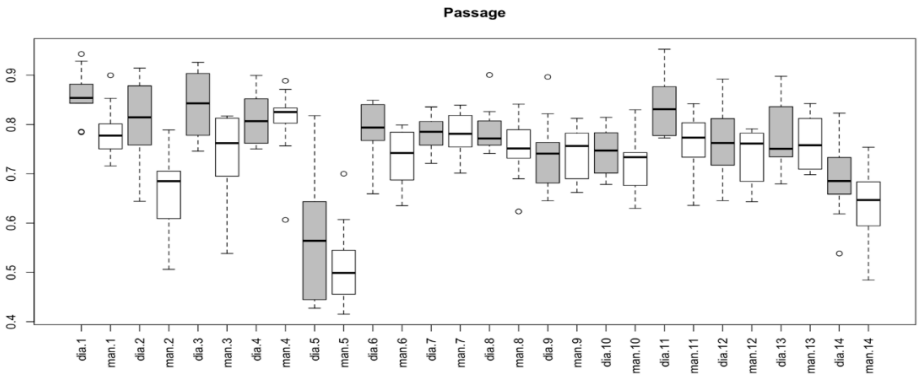


Figure 2: Box plot showing percentVO values for the passage in two dialects (Danyang dialect [dia] and Mandarin [man]) for 14 speakers.



As is shown in Table 2, full models are significantly different from speaker-reduced models with increased goodness of fit (smaller AIC values), indicating that between-speaker variation was highly significant for percentVO. Likewise, the language effect was tested to be very significant for percentVO.

We further examined speaker and language effects by inspecting the boxplots, which showed the percentVO values for sentences and the passage separately. Speaker effect is clearly visible in both styles of reading materials. Speaker specificity is particularly distinct in the case of speaker 5 who showed dramatically lower percentage of voiced intervals in both dialects across both speaking styles, compared to all other speakers. An opposite situation can be observed for speaker 3 who is consistently higher in percentVO than others in both sentences and passage, especially in Danyang dialect. In general, between-speaker durational variabilities of voiced intervals are quite consistent in the two dialect conditions.

Nevertheless, within-speaker variability was also present and the magnitude of the variability differences varied among different speakers and in different speaking styles. There are distinctive cases in which speakers behaved rather similar when producing the sentences in two dialects in terms of the durational variability of voiced intervals, for example, speaker 1, 2 and 11 (see Figure 1). However, the situation completely reversed during the event of the passage where there existed massive between-dialect differences for both speakers (see Figure 2). For speaker 3 and 7, it's completely the other way around. The between-dialect difference within a speaker narrowed when they produced the linguistically more coherent passage rather than single sentences. Other speakers varied in the level of the conformity between two dialect conditions in terms of percentVO, though they showed little within-speaker variation when producing sentences and passage. Clearly, both the dialect and speaking style have profound impacts on the internal stability of a speaker's production of voiced interval duration in speech.

4. DISCUSSION

In the present study, we analyzed durational variabilities of voiced intervals for a group of bi-dialectal speakers in both their native dialects. For the dependent variable under observation (percentVO), both main effects (speaker and language) were significant. Systematic variation existed among all the speakers. An extreme case was speaker 5, who showed a dramatic low percentage of voiced segments in both dialects compared to all other speakers. Though it has been proposed that speakers with distinct creakiness in word or phrase-final positions might have measurable lower percentage of voiced segments in speech, auditory inspection revealed no significant creakiness in this particular speaker. Currently we do not have adequate understanding as to what causes percentVO to vary among these speakers, but it is true that percentVO has been identified as one of the measures revealing the greatest effects of speaker in numerous studies ([6], [8], [14]), even in cases when style variability and channel variability were introduced ([14]). It is quite plausible that speaker-speaker anatomic make-up of the speech organs and individual ways of operating them result in varying overall percentage of voicing time between speakers.

However, durational characteristics of voiced intervals are not always stable within a speaker when two dialects are involved. Within-speaker variability of percentVO was observable across dialect conditions. For some speakers, the durational characteristics of voiced intervals in one dialect were largely obtainable in the other dialect (for example, speaker 4 and 13), though most of the speakers behaved differently to varying degrees when speaking in different dialects, which indicates that percentVO is not robust against dialect variability. The within-speaker dialect variability might be largely due to the phonological differences between Danyang dialect and Mandarin. It was interesting to see that the percentVO values of Danyang dialect is consistently higher than of Mandarin (see Figure 1 and Figure 2). This is in line with previous findings that Wu varieties are distinguished by their retention of voiced or murmured obstruent initials, such as stops, affricates and fricatives ([20]), which can boost the duration of voiced segments in the speech, resulting in a higher percentVO compared to Mandarin speech. This means that when speaking the two dialects, speakers have to accommodate to two distinct phonological systems, resulting in dissimilar production of voiced interval durations across dialects. This is not the first report of language-specific effects for percentVO, Dellwo and Fourcin ([5]) also observed that languages vary in the way their voiced intervals are organized.

Though percentVO performs fairly well in revealing between-speaker temporal variability and we argued that speaker-specific anatomical configurations and ways of moving the articulators are the driving force behind it, this study clearly shows that this supra-segmental temporal feature of speech is not stable within a speaker when two distinct varieties of the same language are involved. This finding bears important implications for forensic phonetic practices, especially those carried out in linguistically diverse areas, when comparisons have to be made on trace and suspect materials with a mismatch in speaking languages or language varieties. It is vital to check the robustness of speaker individuality measures in use against different language variabilities.

So far we only looked at read sentences without any pragmatic context. This is a type of speech that typically does not play a large role in forensic context. Another admittedly drawback of the present study is that the amount of data analyzed is not very large. It would be interesting to expand the study to spontaneous speech and more speakers so that more meaningful interpretations could be obtained.

6. REFERENCES

[1] Arvaniti, A. 2012. The usefulness of metrics in the quantification of speech rhythm. *Journal of Phonetics*, 40, 351–373.

[2] Baayen, R. H. 2008. *Analyzing Linguistics Data: A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.

[3] Bates, D., Maechler, M., Bolker, B. and Walkers, S. 2014. *lmer4: Linear mixed effects models using Eigen and S4* (R package version 1.1-7). http://CRAN.R-project.org/package=lmer4 Accessed 17 November 2018.

[4] Dellwo, V. 2010. Influences of speech rate on the acoustic correlates of speech rhythm: An experimental phonetic study based on acoustic and perceptual evidence. PhD-Dissertation, Universität Bonn.

[5] Dellwo, V. and Fourcin, A. 2013. Rhythmic characteristics of voice between and within languages. *Travaux Neuchâtelois de Linguistique* 59: 87–107.

[6] Dellwo, V, Leemann, A., & Kolly, M.-J. 2012. Speaker idiosyncratic rhythmic features in the speech signal. *Proceedings of Interspeech*, 2012, Portland (OR), USA.

[7] Dellwo, V., Leemann, A. and Kolly, M.-J. 2015. Rhythmic variability between speakers: Articulatory, prosodic, and linguistic factors. *Journal of Acoustical Society of America* 137:1513--1528.

[8] Dellwo, V. and Schmid, S. 2015. Speaker-individual rhythmic characteristics in read speech of German-Italian bilinguals. In: A. Leemann, M.-J. Kolly, S. Schmid and V. Dellwo (eds), *Trends in Phonetics and Phonology: Studies from German speaking Europe*. Bern: Peter Lang, 349-362.

[9] Eriksson, A. 2012. Aural/acoustic vs. automatic methods in forensic phonetic case work. In: A. Neustein and H. A. Patil (eds), *Forensic Speaker Recognition: Law Enforcement and Counter-Terrorism*. NewYork: Springer, 41-69.

[10] He, L. 2018. Development of speech rhythm in first language: The role of syllable intensity variability. *Journal of Acoustical Society of America* 143: EL463--467.

[11] He, L. and Dellwo, V. 2016. The role of syllable intensity in between-speaker rhythmic variability. *International Journal of Speech, Language and the Law* 23: 243--273.

[12] He, L. and Dellwo, V. 2017. Between-speaker variability in temporal organizations of intensity contours. *Journal of Acoustical Society of America* 141: EL488--EL494.

[13] Kliegl, R., Wei, P., Dambacher, M., Yan, M. and Zhou, X. 2011. Experimental effects and individual differences in liner mixed models: estimating the relationship between spatial, object and attraction effects in visual attention. *Frontiers in Psychology*, 1(238): 1-12.

[14] Leemann, A., Kolly, M.-J., and Dellwo, V. 2014. Speaker-individuality in suprasegmental temporal features: Implications for forensic voice comparison. *Forensic Science International* 238: 59-67

[15] McDougall, K. 2004. Speaker-Specific Formant Dynamics: An Experiment on Australian English /aI/. *International Journal of Speech, Language and the Law* 11(1): 103--130.

[16] McDougall, K. 2006. Dynamic Features of Speech and the Characterization of Speakers: Towards a New Approach Using Formant Frequencies. *International Journal of Speech, Language and the Law* 13(1): 89--126.

[17] Runeson, S. 1985. Perceiving people through their movements. In: B. D. Kirkcaldy (ed.), *Individual Differences in Movement*. Lancaster: MTP Press, 43-66.

[18] Schmid, S. & Dellwo, V. 2013. Sprachrhythmus bei bilingualen Sprechern. In: S. Schwab & A. Leemann (Eds.), *L'étude de la prosodie en Suisse: Travaux neuchâtelois de linguistique*, 59, 109–126.

[19] Shriberg, E., Ferrer, L., Kajarekar, S., Venkataraman, A., and Stolcke, A. 2005. Modelling prosodic feature sequences for speaker recognition. *Speech Communication* 46: 455--472.

[20] Wang, Li. 1936. *Chinese Phonology*. Shanghai: The Commercial Press.

[21] Wiget, L., White, L., Schuppler, B., Grenon, I., Rauch, O. and Mattys, S. L. 2010. How stable are acoustic metrics of contrastive speech rhythm? *Journal of the Acoustical Society of America*, 127, 1559–1569.

[22] Yoon, T. J. 2010. Capturing inter-speaker invariance using statistical measures of speech rhythm. Proceedings of Speech Prosody, 2010, Chicago, USA.