Investigating the Forensic Applications of Global and Local Temporal Representations of Speech for Dialect Discrimination

Leah Bradshaw, Vincent Hughes, Eleanor Chodroff

University of York, Department of Language and Linguistic Science, UK

lkb523@york.ac.uk, vincent.hughes@york.ac.uk, eleanor.chodroff@york.ac.uk

Abstract

Languages, dialects, and speakers can differ substantially in the temporal structure of speech. With the exception of only a handful of studies, the application of this information has been fairly limited in forensic speech research and casework. This may in part be due to existing differences in how to quantify temporal differences, as well as the limited research on the efficacy of such operationalisations for speaker discrimination. Standard operationalisations of temporal information have included measures reflecting *global* aspects of vowel or consonant duration alternations (e.g., Rhythm Metrics: RMs), as well as *local* measures of the change and acceleration of the cepstrum (e.g., delta and delta-delta coefficients in Automatic Speech or Speaker Recognition). This paper investigates the utility of these temporal measures for discriminating among four dialects of British English that contrast in region and language contact: Cambridge, Multicultural London, Leicester, and Punjabi-Leicester English. Using linear regression, log-likelihood model comparison and k-means clustering, we identified significant differences between dialects in all investigated RMs and substantially better performance of RMs in comparison to delta and delta-delta coefficients in dialect clustering. These findings suggest that temporal information in speech, and particularly global temporal information, is highly useful for dialect and speaker discrimination.

Index Terms: rhythm, timing, dialectal variation, forensic phonetics, automatic speaker recognition

1. Introduction

Languages, dialects, and individual speakers can differ substantially in the temporal structure of speech [e.g., 1–10]. As suggested in [9], temporal variation may be highly useful in forensic speech analysis, but has been limited in its application in both research and casework. This may in part be due to existing differences in how to quantify temporal differences between speakers, dialects, and languages, as well as the limited research on the efficacy of such operationalisations for speaker discrimination. The present study focuses on temporal variation within and between four dialects of British English, and examines the extent to which global and local representations of temporal structure discriminate between these dialects. To the extent that temporal representations can discriminate between dialects, they may also prove useful for speaker discrimination tasks.

Temporal representations of speech can be constructed at varying levels of granularity. Global temporal representations of speech can include long-term alternations in vocalic and consonantal intervals which may approximate the *rhythmic* pattern of speech [1]. (We note, though, that the acoustic approximation of rhythm is more complex than can be adequately addressed here [11].) Local temporal representations of speech can include the change between adjacent spectral properties, which may also be diagnostic of speaker or dialect identity. Existing research points towards the usefulness of global temporal properties for describing and discriminating speakers and dialects [e.g., 2, 4–9], and also for automatic speech and speaker recognition [12–15]. Previous research in the former area has largely employed Rhythm Metrics (RMs) as a global representation of temporal structure, which we adopt as well. Local temporal information is known to improve the performance of automatic speaker and language recognition systems [16–20]. This information is standardly represented by delta (Δ) and delta-delta (ΔΔ) features, which reflect the change in spectral properties between adjacent temporal frames and the acceleration of that change. We thus ask two primary questions in our research: how well do global temporal properties (RMs) discriminate among four varieties of British English that differ in region and language influence? Second, how do RMs compare to delta and delta-delta features in dialect discrimination?

Rhythm metrics (RMs) are considered here to represent global temporal properties of speech that may relate to speech timing and/or rhythm. These metrics were devised as a response to early studies suggesting that languages may be classified as either "stress-timed" or "syllable-timed" languages [21–22]. A stress-timed language is marked by regular intervals between stressed syllables whereas a syllable-timed language has syllables of roughly equal length. Findings from [23] suggested a much more nuanced range of temporal and rhythmic patterns in languages related to vowel reduction and complexities in consonantal clusters. RMs aim to capture these variances by quantifying patterns in the duration and temporal alternation of vocalic and/or consonantal intervals. Vocalic measures capture differences in vowel reduction and variation in tense vowels and diphthongs, while consonantal measures correlate with the complexity of consonantal structures [1, 3, 24–25].

These measures have been applied in a number of analyses considering rhythmic patterning across a range of languages, dialects, and speakers [e.g., 1–5, 7–10, 24–26]. Among British English dialects, [27] suggest that L1 varieties show little variation in speech rhythm, arguing any differences result from differences in speaking style. The rhythmic patterns of British English varieties, however, may be influenced by dialect and language contact. Using various RMs [3], [28] identified a range of rhythmic patterns among young speakers in the London area, depending on ethnic influence and dialect contact: non-Anglo Hackney speakers (speakers of Multicultural London English) were more syllable-timed than Anglo Hackney speakers, who were in turn more syllable-timed than Anglo speakers in Havering, a relatively Anglo-dominant area. Further work investigated rhythmic differences between a so-called "stress-timed" variety, Leeds English, and a "syllable-

timed" variety spoken by Punjabi-English bilinguals from neighbouring Bradford [29]. Though the results showed numerical differences in several RMs, these differences were not statistically supported, suggesting that these two varieties were unlikely to be two extremes of a continuum.

With respect to their forensic application, durational ratio measures capturing the percentage over which speech is voiced or vocalic (%V, %VO) have been shown to be successful discriminators for speakers of German and Swiss German [9] and Persian speakers [26]. Additionally, [10] showed durational variability measures to be capable of speaker discrimination. A growing body of research has also considered the possibility of these rhythmic differences to discriminate between dialects of a language [2, 4–8]. The present study evaluated the utility of global RMs for discriminating among four varieties of British English: Cambridge (CE), Multicultural London English (MLE), Leicester (LE) and Punjabi-Leicester (PLE). These varieties are relatively balanced along dimensions of region and language contact: CE and MLE are geographically Southern whereas LE and PLE are spoken in the Midlands; CE and PE are Anglo varieties whereas MLE and PLE are contact varieties. In addition, we investigated the relative performance between global temporal representations (RMs) and local temporal representations (Δs, ΔΔs) in discriminating these dialects.

2. Experiment 1: Discriminability of Rhythm Metrics

An empirical study was conducted using six RMs to capture variation between these dialects. Our first question was to examine whether between-dialect variability was greater than within-dialect variability. The combination of these measures was then submitted to a k-means cluster analysis to evaluate their efficacy in clustering speakers of the same variety.

2.1. Methods

2.1.1. Materials

Recordings of the CE and MLE speakers came from the International Varieties of English (IViE) corpus [30]. These recordings consisted of 24 speakers (12 CE, 12 MLE), reading a short passage (The Cinderella Passage). Speakers were split roughly equally between male and female and were aged 16 at the time of recording. The MLE participants were monolingual speakers of Caribbean descent. All of the speakers had a moderate to good reading ability.

The recordings of the LE and PLE speakers were obtained from [31] and consisted of 30 speakers (8 LE, 22 PLE) reading a short passage (Fern's Star Turn). Speakers were split roughly equally between male and female; ages ranged from 20 to 53 years. The LE ('Anglo-Leicester') speakers were all British-born with no heritage language other than English, and both parents and grandparents were born in the United Kingdom. The PLE ('Punjabi-Leicester') speakers were second-generation (British-born) speakers with 'Punjabi language heritage' and characterised as having at least one parent who is a native Punjabi speaker and first-generation immigrant from the Indian Punjab. The heritage language for these speakers is close to Modern Standard Punjabi.

2.1.2. Measurement

The rhythm metrics presented in Table 1 were calculated by the 'Duration Analyzer (version 0.03)' Praat script [32]. Only normalised measures were considered within the analysis to account for slight differences in the length of the passages read by the speakers. This minimised influences due to the overall number of vocalic/consonantal segments in each corpus. Critically, all silent intervals were removed prior to the analysis.

For the CE and MLE speakers, phone- and utterance-level alignments were obtained using the English acoustic models in the Praat EasyAlign software extension (English models based on British English) [33]. For the LE and PLE speakers, phone- and utterance-level alignments accompanied the recordings. All phone alignments were manually adjusted. Consonantal and vocalic segments and intervals were derived from these alignments. Following [3], vocalic intervals were defined using the vowel onset and offset, while intervocalic (consonantal) intervals were the stretches from vowel offset to onset. Glides and liquids were treated as consonants, aside from occurrences of /l/-vocalisation.

Table 1: *Rhythm Metrics used for analysis. All durations were log-transformed.*

Metric	Description
stdevV	Standard deviation of vocalic interval duration
stdevC	Standard deviation of consonantal interval duration
VarcoV	Coefficient of variation for the vocalic interval duration
nPVI-V	Pairwise Variability Index for vocalic interval durations. Mean of the differences between successive vocalic interval durations, divided by their sum
nPVI-C	Pairwise Variability Index for consonant interval durations. Mean of the differences between successive consonantal interval durations, divided by their sum
nPVI-CV	Normalised pairwise variability index for summed vocalic and consonantal interval durations. Mean of the differences between successive vocalic and consonantal interval durations, divided by their sum

2.2. Results

We conducted a series of analyses to investigate the utility of rhythm metrics (RMs) for dialect discrimination and classification. First, we analysed the effect of dialect on each RM using linear regression. In addition to examining major differences in RM realisation among dialects, we also examined whether dialect significantly improved model fit for each RM through model comparison [8]. Finally, we assessed the utility of RMs for dialect discrimination in a k-means cluster analysis.

2.2.1. Descriptive statistics

As shown in Figure 1, variation was observed between dialects in the z-scored values obtained for each RM. Based on visual inspection, no one measure discriminated all four dialects, though individual dialects or small groups of dialects certainly differed from others for several of the measures.

2.2.2. Linear regression models

Linear regressions models were implemented for each of the six RMs using the lme4 package in R [34], with the RM values as a continuous independent variable and dialect and gender as

predictors. As there was only one RM value per recording, by-speaker random effects were not included. Factors were sum-coded, such that the interpretation of the following comparisons are relative to the average production across all four dialects (LE held out). Alpha levels were adjusted to 0.008 using a Bonferroni correction for multiple hypothesis testing. Cambridge speakers exhibited significantly higher values for *stdevV*, *VarcoV*, and *nPVI-CV*, and significantly lower *stdevC* values (*stdevV*: β = 0.047, *stdevC*: β = -0.016, *VarcoV*: β = 0.032, *nPVI-CV*: β = 3.87, each $p < 0.008$). MLE speakers had significantly lower values for *stdevC*, *nPVI-V*, and *nPVI-C* (*stdevC*: β = -0.031, *nPVI-V*: β = -2.88, *nPVI-C*: β = -2.60, each $p < 0.008$). Punjabi-Leicester speakers had significantly higher *stdevC* values, but significantly lower *stdevV*, *VarcoV*, and *nPVI-CV* values (*stdevV*: β = -0.041, *stdevC*: β = 0.016, *VarcoV*: β = -0.037, *nPVI-CV*: β = -2.20, each $p < 0.008$). Gender was not significant in any model.

Log-likelihood model comparisons were conducted to compare the model fit of models in which dialect was included as a predictor and those in which it was not [8]. For all RMs, the model fit was significantly improved with dialect as a predictor (for all six comparisons, $p < 0.008$).

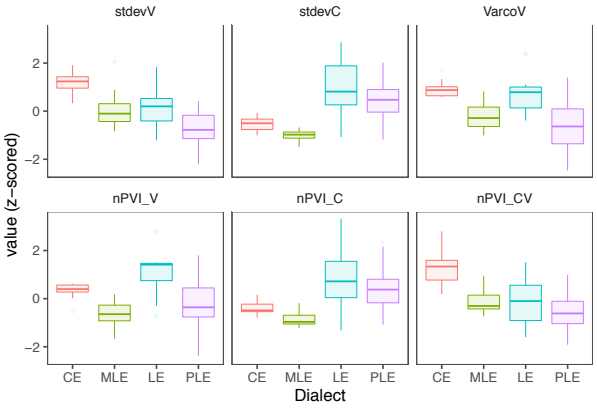

Figure 1: *Scaled results for each rhythm metric (RM) plotted by dialect. Boxplots reflect variation across speakers within each dialect*

2.2.3. K-means cluster analysis

A k-means cluster analysis was implemented using the 'factoextra' package in R to examine the extent to which these six measures properly cluster these varieties of British English [35]. For this analysis, each speaker was represented by the six RMs (their rhythmic profile), and the number of clusters was set to four to reflect the number of expected varieties.

Overall, classification was respectable: assuming that each cluster corresponded to the dialect most represented in that cluster, the overall purity was 0.64 (where perfect classification corresponds to a purity score of 1) [36]. The slightly diminished purity may have been driven by the PLE speakers who were notably diverse in background and occupied several clusters. The two primary dimensions of variation were largely interpretable: Dimension 1 appeared to reflect the divide between Anglo and Ethnic varieties, and Dimension 2, the divide between Southern (Cambridge/London) and Midlands (Leicester) speakers (or alternatively, the spoken passage). All CE speakers were successfully grouped together, and despite previous studies suggesting that Anglo–British English varieties are not distinctive on the basis of their rhythmic patterning [27], none of the LE speakers were grouped with the CE speakers on the basis of the full rhythm profile. Further investigation into the optimal set of RMs for dialect discrimination would be beneficial.

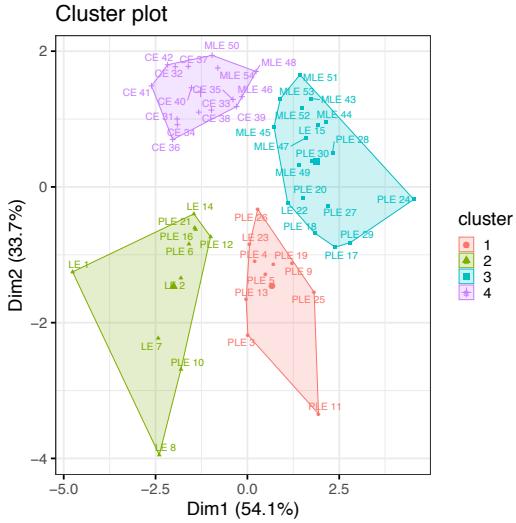

Figure 2: *Results of the k-means cluster analysis using RMs. Clusters are visualised along the two primary dimensions of variation*

2.3. Discussion

Overall, these findings reveal greater between-dialect variability than within-dialect variability for each of the tested RMs. Significant differences were observed among dialects for several RMs, and dialect significantly improved model fit for all six RMs. Anglo varieties tended to have more variability in vocalic durations and longer durations between vocalic intervals (vocalic RMs) than contact varieties, and Leicester varieties tended to have more variability in consonantal durations and longer durations between consonantal intervals (consonantal RMs) than Southern varieties. (We note that region is confounded with reading passage here, which should be disentangled in future research.) Relative to average, CE was marked by more variable vowel duration, less variable consonantal duration, and overall more variability in the time between utterances; MLE was marked by less variable consonantal duration, less time between successive vocalic intervals, and less time between successive consonantal intervals (the latter two may suggest a fast speech rate); LE was marked by more variable consonantal durations, more time between successive vocalic intervals, and more time between successive consonantal intervals (the latter two may suggest a slow speech rate); finally, PLE was marked by more variable consonantal durations, less variable vocalic durations, and less variability in the time between utterances. The complex temporal characterisation of these dialects and the interaction between vocalic and consonantal RMs between region and language influence suggest that temporal aspects of speech may be more fruitfully described by a rhythmic profile than a binary distinction or even a (uni-dimensional) continuum [3]. Moreover, the clustering of these dialects using the rhythmic profile was respectable, with a purity of 0.64.

3. Experiment 2: Discriminability of Delta Metrics

Temporal information is known to improve the performance of automatic speaker and language recognition systems [19]. These systems use delta (Δ) and delta- delta (ΔΔ) coefficients to capture these temporal aspects of the speech signal. Δs, also known as differential coefficients, measure the degree of spectral change across adjacent frames; ΔΔs, or acceleration coefficients, measure the degree of change across the Δs. However, it is not clear whether the information being captured by the Δs and ΔΔs within automatic systems performs as well as the linguistically-motivated RMs for speaker or dialect discrimination. While the RMs capture durational differences between syllables (global information), the Δs and ΔΔs capture changes between much smaller frames of speech (local information). In this experiment, we examined the potential of Δ and ΔΔ coefficients for dialect discrimination.

3.1. Methods

Δs and ΔΔs were extracted in MatLab. The recordings from Experiment 1 were first subjected to voice activity detection using the *vadsohn* function in the Voicebox toolkit [37]. Using the speech-active portions of each recordings, Mel-frequency cepstral coefficients (MFCCs) were then extracted within a 0-4000 Hz range from 20 ms frames shifted by 10 ms using the *melfcc* function in the Rastamat toolkit [38]. Cepstral mean and variance normalisation was then applied to the MFCCs in an attempt to reduce the effects of the different equipment and room conditions used in the collection of the original recordings [39]. This was done using the *cmvn* function from the MSR Identity toolkit [40]. Δs and ΔΔs were then appended to the normalisation MFCC feature vector for each frame using the *delta* function in Rastamat.

The Δs and ΔΔs were then averaged for each recording – the MFCCs themselves were not used for the purpose of analysis. Thus, each speaker's recording was described using 12 Δs and 12 ΔΔs. Following the methods from Experiment 1, these values were used as input for k-means clustering, to assess how well they were able to group speakers of the same variety.

3.2. Results

As shown in Figure 3, the assignment of speakers to clusters was skewed towards Cluster 1: the majority of speakers for each variety was grouped into this cluster (9 CE, 12 MLE, 5 LE, 10 PLE). The overall purity of the clusters was 0.44, which was much lower than the value of 0.64 obtained using RMs [36]. This is particularly striking given the fact that the delta analysis employed 24 temporal features, whereas the RM analysis employed only 6. The first two dimensions of variation shown in Figure 3 also accounted for much less of the variation relative to those in the RM analysis.

3.3. Discussion

This experiment investigated the utility of local temporal representations commonly employed in ASR systems for dialect discrimination. No clear clustering of dialects was found using the combination of local Δs and ΔΔs (Figure 3), unlike that found using the global RMs (Figure 2). This suggests that Δs and ΔΔs are not capturing the same information as the linguistic measures of rhythm used in Experiment 1 and may be missing useful information regarding meaningful variation between dialects, and thus speakers. This, in itself, is a positive finding for forensic purposes, since it suggests that speaker and accent/language recognition systems are not sensitive to the types of rhythm information analysed by linguists. Therefore, there is considerable potential for improving the performance of such systems using the measures described in Table 1. The extent to which this is the case, of course, remains an empirical question, and one that deserves further attention.

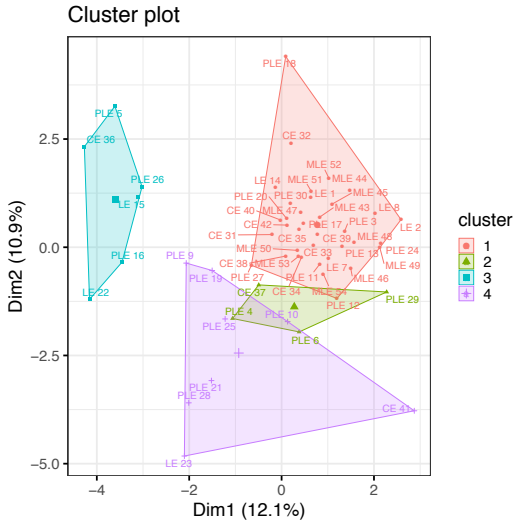

Figure 3: *Results of the k-means cluster analysis using* Δs and ΔΔs*. Clusters are visualised along the two primary dimensions of variation*

4. Conclusion

The purpose of this study was two-fold: to investigate the extent to which four British English varieties differed from one another in their temporal aspects, and to discern how well global and local temporal representations performed in dialect discrimination. The findings from Experiment 1 showed that the realisation of six RMs differed significantly between dialects, and speakers of the same variety were mostly grouped together using a set of these metrics. This corroborates findings from Swiss German [8] and further strengthens the argument for using speech rhythm as a cross-dialectal discriminator. Moreover, significant differences were observed between Southern and Midlands varieties of British English, and between Anglo and contact varieties. In particular, Southern and Leicester varieties differed primarily in consonantal RMs, whereas Anglo and contact varieties differed primarily in vocalic RMs.

Experiment 2 compared local temporal representations of Δ and ΔΔ coefficients with linguistically-motivated RMs for dialect discrimination. Delta features performed relatively worse than RMs, which successfully captured variation capable of discriminating between dialects — variation that may also be valuable for ASR systems. These results suggest that globally-defined temporal measures of speech may prove useful for forensic and ASR applications [see also 12–15], and that rhythmic profiles of speakers, dialects, and languages may be beneficial for linguistic description and analysis.

5. Acknowledgements

We would like to thank Paul Foulkes, Peter French, and Sam Hellmuth for helpful feedback, and Jessica Wormald for generously sharing her data for this study.

6. References

[1] F. Rasmus, M. Nespor, and J. Mehler, "Correlates of linguistic rhythm in the speech signal," Cognition, 73(3): 265-292, 1999.

[2] S. Ghazali, R. Hamdi, and M. Barkat, "Speech rhythm variation in Arabic dialects," in *Speech Prosody 2002, France*, pp. 331-334, 2002.

[3] E. Grabe and E. L. Low, "Durational variability in speech and the rhythm class hypothesis," In N. Warner, & C. Gussenhoven (Eds.), *Papers in Laboratory Phonology* 7 (pp. 515-546). Berlin: Mouton de Gruyter, 2002.

[4] E. Ferragne and F. Pellegrino, "Rhythm in Read British English: Interdialect variability," in *Proceeding of the 8th International Conference on Spoken Language Processing, Jeju, Korea,* pp. 1573-1576, 2004.

[5] E. O'Rourke, "Speech rhythm variation in dialects of Spanish: Applying the Pairwise Variability Index and Variation Coefficients to Peruvian Spanish," in *Speech Prosody 2008, Brazil*, pp. 431-434, 2008.

[6] F. Biadsy and J. Hirschberg, "Using prosody and phonotactics in Arabic dialect identification", *In INTERSPEECH 2009 – 20th Annual Conference of the International Speech Communication Association, Brighton, UK,* pp. 208-211, 2009.

[7] A. Arvaniti, "The usefulness of metrics in the quantification of speech rhythm," *Journal of Phonetics*, vol. 40, no. 3: 351-373, 2012.

[8] A. Leemann, V. Dellwo, M.-J. Kolly, and S. Schmid, "Rhythmic variability in Swiss German Dialects," in *Speech Prosody 2012*, China, pp. 607-610, 2012.

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

[10] V. Dellwo, A. Leemann, and M. J. Kolly, "Rhythmic variability between speakers: Articulatory, prosodic and linguistic factors," Journal of the Acoustic Society of America, 137(3): 1513-1528, 2015.

[11] A. Arvaniti, "Rhythm, timing and the timing of rhythm," Phonetica, 66: 46-63, 2009.

[12] A. G. Adami, R. Mihaescu, D. A. Reynolds, and J. J. Godfrey, "Modeling prosodic dynamics for speaker recognition," In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing,* vol. 4, 2003.

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

[14] N. Dehak, P. Kenny, and P. Dumouchel, "Continuous Prosodic Features and Formant Modelling with Joint Factor Analysis for Speaker Verification," In *INTERSPEECH 2007*, Antwerp, pp. 1234-1237, 2007.

[15] J. Rouas, "Automatic prosodic variations modeling for language and dialect discrimination," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 15, no. 6, pp. 1904-1911, 2007.

[16] C-H. Lee, E. Giachin, L. R. Rabiner, R. Pieraccini, and A. E. Rosenberg, "Improved acoustic modeling for continuous speech recognition," In *Speech and Natural Language: Proceedings of a Workshop*, *Hidden Valley, Pennsylvania, June 24-27*, 1990.

[17] T. Matsui and S. Furui, "Text-independent speaker recognition using vocal tract and pitch information," *First International Conference on Spoken Language Processing*, 1990.

[18] H. Gish and M. Schmidt, "Text-independent speaker identification," *IEEE Signal Processing Magazine*, vol. 11, pp. 18–32, 1994.

[19] J. P. Campbell, "Speaker recognition: A tutorial. *Proceedings of the IEEE*, *85*(9), pp. 1437-1462, 1997.

[20] D. Martínez, E. Lleida, A. Ortega, and M. Antonio, "Prosodic features and formant modelling for an iVector-based language recognition system," *Proceedings of ICASSP 2013*, pp. 6847-6851, 2013.

[21] K. Pike, The intonation of American English. Ann Arbor: University of Michigan Press, 1945.

[22] D. Abercrombie, *Elements of general phonetics*. Edinburgh: Edinburgh University Press, 1967.

[23] R. M. Dauer, "Stress-timing and syllable-timing reanalysed." *Journal of Phonetics*, 11: 51-62, 1983.

[24] V. Dellwo, "Rhythm and speech rate: A variation coefficient for deltaC," *Proceedings of the 38th Linguistic Colloquium, Peter Lang*, *Frankfurt*, pp. 231-241, 2006.

[25] S. Schmid, "Syllable typology and the rhythm class hypothesis: Evidence from Italo-Romance dialects," In J. C. Reina and R. Szczepaniak (Eds.), *Syllable and Word Languages*, (pp. 421-453) Berlin: Walter de Gruyter, 2014.

[26] H. Asadi, M. Nourbakhsh, L. He, E. Pellegrino, and V. Dellwo, "Between-speaker rhythmic variability is not dependent on language rhythm, as evidence from Persian reveals," *International Journal of Speech Language and the Law*, 25(2): 151-174, 2018.

[27] L. White and S. L. Mattys, "Calibrating rhythm: First language and second language studies," *Journal of Phonetics*, 35: 501-522, 2007.

[28] E. N. Torgersen and A. Szakay, "An investigation of speech rhythm in London English," *Lingua*, 122: 822-840, 2012.

[29] T. V. Rathcke and R. H. Smith, R. H., "Speech timing and linguistic rhythm: On the acoustic bases of rhythm typologies," *The Journal of the Acoustical Society of America*, 137: 2834-2845, 2015.

[30] E. Grabe, B. Post, and F. Nolan, *The IViE Corpus*. Department of Linguistics, University of Cambridge. [ESRC grant R000237145.], 2001.

[31] J. Wormald, *Regional Variation in Punjabi-English*, PhD Dissertation, University of York, 2016.

[32] V. Dellwo, "Praat script: Duration Analyzer (version 0.03). [Computer software*]," Retrieved 3 July 2019 from https://www.pholab.uzh.ch/static/volker/software/plugin_durationAnalyzer.zip,* 2019.

[33] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer program]," *Version 6.0.43, retrieved from http://www.praat.org/*, 2019.

[34] D. Bates, M. Maechler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *Journal of Statistical Software*, 67(1), 1-48. doi:10.18637/jss.v067.i01, 2015.

[35] A. Kassambara and F. Mundt, "factoextra: Extract and Visualize the Results of Multivariate Data Analyses," *R package version 1.0.5. https://CRAN.R-project.org/package=factoextra*, 2017.

[36] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Online edition, 2008.

[37] M. Brooks, "VOICEBOX: Speech Processing Toolbox for MATLAB," *http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html,* 1999.

[38] D. Ellis, "PLP and RASTA (and MFCC, and inversion) in Matlab," *http://www. ee. columbia. edu/~ dpwe/resources/matlab/rastamat/*, 2005.

[39] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalisation for noise robust speech recognition," *Speech Communication*, 25(1-3): 133-147, 1998.

[40] S. O. Sadjadi, M. Slaney, and L. Heck, "MSR Identity Toolbox v1.0: A MATLAB Toolbox for Speaker Recognition Research," IEEE Speech and Language Processing Technical Committee Newsletter, 2013.