PERSIAN SPEAKER CLASSIFICATION USING RHYTHMIC FEATURES

Neda Mousavi[1], *Sven Grawunder*[1,2]
[1]*Martin-Luther-Universität Halle-Wittenberg,* [2]*Max-Planck-Institut für evolutionäre Anthropologie, Leipzig*
neda.mousavi@student.uni-halle.de, grawunder@sprechwiss.uni-halle.de

Abstract: We applied three common supervised classification models, including Random Forest (RF), Naive Bayes (NB), and Support Vector Machine (SVM), to classify speakers based on rhythm features. The dataset consisted of a set of read speech by 8 Persian speakers. Following previous studies, rhythm parameters in domains other than time, namely intensity and frequency, had been considered in the selection of rhythmic features and subsequently used for classifying speakers. Whereas PVIs and rate features contribute most to accuracy and Gini in particular for RF, there seems to be no aggravating difference between model schemes (OvR and OvO) with respect to performance.

1 Introduction

In recent years, rhythm research has evolved from a debate about the controversial dichotomy of syllable timing and stress timing classes to an examination of the concept of variation within language and even within speakers. Several studies indicate that the method of data collection [1], the different complexity of phonotactics of sentences [1, 2, 3], the patterns of stress and intonation [2, 4], speaker age [5, 6], and speaker health [7] can lead to rhythmic variations. Such views in rhythm analysis go beyond the main assumption of typological approaches [8], which assume small differences between and within speakers. On the contrary, the general idea is that specific characteristics based on the effects of individual anatomical configurations, such as laryngeal size, duration of vocal tract activity, or size of tongue movement area, lead to specific temporal organization and influence rhythmic measures such as $\%V$, $\%VO$, ΔV, and (pairwise variability indeces) PVI, which vary considerably between speakers [3, 9, 10, 11].

The approaches applied in this paper are primarily concerned with determining models for evaluating and representing such rhythmical variations. The use of machine learning (ML) methods for speaker recognition based on verbal features has yielded relatively good results so far, but the limitations in terms of computational complexity and effort cannot be ignored [12]. Recently, there have been some attempts to use rhythmic metrics to improve speech recognition models such as speaker identification (SID) and language identification (LID) or even speech intelligibility [13]. However, the instability of rhythmical indicators in different speech contexts seems to be often ignored. These approaches attempt to model rhythmic units (e.g., "pseudo-syllables", see [14, 15]) to extract features for training ML models. Others involve statistical studies of rhythm-related features in the extracted signal periodicities [16, 17]. In addition, other studies have applied ML models in order to recognize speakers based on emotion, dialect, or social class. As an example, Alotaibi et al. [18] investigated the relationship between rhythm metrics and the ability to classify speakers of a Levantine Arabic dialect by gender and social class using rhythm metrics and artificial neural networks (ANN) as a classification model. Their results show that the classification accuracy is higher when interval measures and PVI rhythm metrics are used, higher than when only interval measures are used as rhythm indicators.

Furthermore, the useful effect of adding rhythmic features for classification of a speaker's emotions has been confirmed [19, 20]. Another study [21] on speech identification with rhythmic features, used an automatic rhythm extraction method from Music Information Retrieval to describe the periodicity of speech and applied the Support Vector Machine approach [cf. 22]. They developed a novel method for extracting rhythm-related speech features by using several novel features such as spectral fluency, spectral flatness, spectral centroid, RMS amplitude, and fundamental frequency in speaker identification (see [23] for more information on these features), and sub-features were extracted from the above features using a beat histogram. Their results only partially confirmed the original form of the rhythm class hypothesis. Lykartsis et. al [12] applied the above novel features to speaker identification using the TEVOID corpus [24] to determine whether the proposed rhythmic features could be as successful for SID as for LID. They achieved an average accuracy of 26.95%, although the values varied for different speakers. Using a set of spectral features, the performance results of their model changed significantly, reaching an accuracy of 87.6% without much difference between speakers. Unlike the results they showed in speech recognition, or the effect of the specific nature of the data used, the features used may not be able to capture the variability between speakers. However, the change in the accuracy of the model in speaker recognition when using rhythmic features and adding spectral features may be a matter of debate. In this study, we attempt to follow this ML approach by using supervised speaker classification models and rhythm indicators and applying them to a Persian corpus. Our goal is *inter alia* to compare the performance of different models in classifying Persian speakers and to analyze the effects of model schemes (one-to-one versus one-to-rest classification, see below) on model performance. In addition, we intend to evaluate the importance of different features in classifying Persian speakers and to gain more insight into the variability in Persian speech rhythm as such.

2 Methods

Our goal in this research is to compare different ML algorithms in order to increase the accuracy of speaker identification. Here, we apply three common supervised classification methods, including Random Forests (RF), Naive Bayes (NB), and Support Vector Machines (SVM), to classify speakers based on rhythm features.

2.1 Feature selection and model evaluation

Following the computational approaches in rhythm analysis and the use of rhythm indicators in speech recognition, the question of selecting and extracting rhythm-related features arises again. These features will form the basis for data training in ML models. The question arises how the selection of rhythm metrics from different acoustic domains (time, intensity and frequency) affects the accuracy of the classification models. Is it better to consider only one domain to achieve more accurate performance, or will classification accuracy benefit from adding features from other domains? From this point of view and based on the common rhythmic metrics introduced in phonetic studies (see above), we define two categories of features. First, features defined exclusively in the time domain, and second, features that have applied metrics to quantify rhythm in other domains. Previous studies have already demonstrated the application of these metrics in other domains (e.g., [25]). Here, the features in the time domain include *nPVIv*, *rPVIc*, *%V*, *ΔC*, *varcoC*, $syllable\ rate$, $CV\ rate$, whereas features in the frequency domain include $nPVI_{f0}$, $rPVI_{f0}$ of the syllables, and features in the intensity domain include $rPVI_{intensity}$, $nPVI_{intensity}$ of the syllables. The goal is to decide on the features that lead to the optimal performance of the model and discard the redundant or irrelevant ones. In addition, the performance of classification models is evaluated using a set of data mining metrics such as accuracy, sensitivity, specificity, precision, detection, and F-Measure, which indicate the classification capability of the model and are used to determine how well the algorithm performs on a given data set (for more details see [26]).

2.2 Classification scheme

Types of classification are usually divided into two main groups depending on the number of classes. One side there are binary classifications, in which the data are divided into two classes, and the other side there are multi-class classifications, dealing with so-called multi-class problems, which is not limited to a certain number of classes. In the speaker classification in the present study, we are naturally confronted with a multi-class situation. In language classification, the situation is also inherently multi-class, but here we use binary classification due to the limited amount of data. For ML models, it is not the case that all classification prediction models support multi-class classification. Algorithms such as logistic regression and SVMs are natively suitable for binary classification and do not support classification tasks with more than two classes. However, decomposition strategies have been proposed [cf. 22] to use binarization techniques to deal with multi-class problems and to partition the original problem into binary classifiers. The most common strategies are "one-vs-rest" (OvR for short, also referred to as One-vs-All or OvA) and "one-vs-one" (OvO). In the OvO approach, the problem is divided into so many binary problems that all possible combinations between pairs of classes are covered. In the OvR approach, a classifier is trained for each class such that this class can be distinguished from all other classes [27]. In a study to classify speakers based on rhythm features, [12] used a OvO multiclass supervised classification setting and explained that this setting was chosen to determine how well the algorithm can distinguish one speaker compared to another, rather than all the others combined.

2.3 Dataset

The data of the present study includes Persian read speech. Previous studies have considered Persian rhythm as syllable-timed, and the simple syllabic structure of this language (i.e., CV(C)(C)) and the absence of vowel reduction patterns have been taken as evidence for this assumption (e.g., [28, 29, 30, 31]). The hypothesis of a more syllable-timed rhythm of formal Persian using $rPVI$ and $nPVI$ indices [8] and by considering vowel intervals as the basic units had been confirmed [32]. Further, the rhythm of Persian based on the length of vowel intervals and consonance as well as the intensity of syllables was investigated by [33]. Assuming that Persian is a syllable-timing language they demonstrated the differences between speakers based on the indicators $\%V$, ΔV, ΔC and *nPVI-V*. Following this research approach, we use Persian read speech as our research data. However, our goal is not to identify the rhythmic class of Persian, but to evaluate the possibility of speaker classification based on rhythmic features. The data consists of audio recordings of eight speakers (4 males and 4 females) reading part of a short story at their normal speaking rate in a quiet room. The signal extracted of each speaker is also divided into three sections according to the text paragraphs. It is assumed that the speaker is likely to have a different speaking rate at the beginning, middle and end of the text, which may affect the rhythm measurement. The speech signals were segmented and divided into consonantal, vocalic, and syllabic intervals in the *Praat* environment ([34] v. 6.2.14), and then the aforementioned rhythm metrics were calculated in the *R*-environment [35]. Subsequently, the obtained data frame was divided into training and test data in a ratio of 80:20. Previously, due to the different values of the features, especially in the different acoustic domains, a standardization of the data was performed using the scale function in *R*.

3 Results

3.1 Unsupervised classification

As a first step, we visually represent the distribution of speakers based on rhythmic features to gain an understanding of the data. In order to consider the effect of all features on speaker dispersion, we used the score plot (Fig. 1 left) obtained from a principal component analysis (PCA) and grouped the available features into two dimensions. For each speaker, there are three points in the plot representing the signal extracted at the beginning, middle and end of reading the text by the speaker and the points of each speaker are clustered together by the color. Also, the loading (bi-)plot (Fig. 1 right) obtained in *factoextra* package [36] shows the relationship between PCs and rhythm variables. As can be seen, the PVI metrics go in the same direction for both intensity and frequency domains. The speech rate metrics also go in the same direction. Other metrics in the time domain, with the exception of the vocalic percent, follow the same direction to varying degrees.

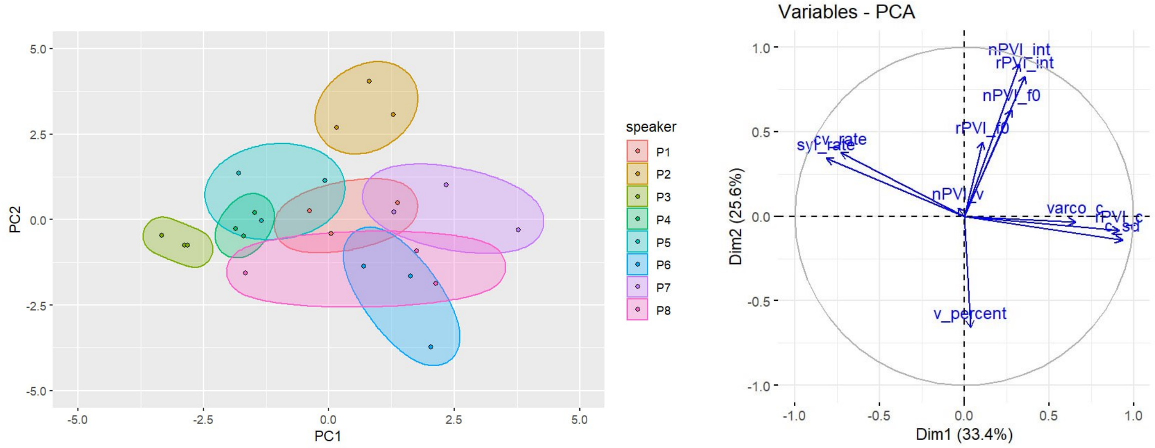

Figure 1 – PCA (left) component distribution per speaker; (right) biplot for individual parameters

3.2 Supervised Speaker classification: Comparing the performance of the models

The question of choosing the appropriate learning model in classification problems has also arisen in the field of speaker classification. The question is which learning model has higher performance in recognizing speakers in a given language, and the reasons for this have been discussed in terms of phonetic content, phonotactic constraints, speech rhythm, and so on (see e.g., [37]). Here, we compared RF, SVM, and NB models in two different schemes (OvO and OvR) based on accuracy, precision, recall, and f1 score.

The plots (s. Fig. 2) show no clear difference between the results of the two schemes in

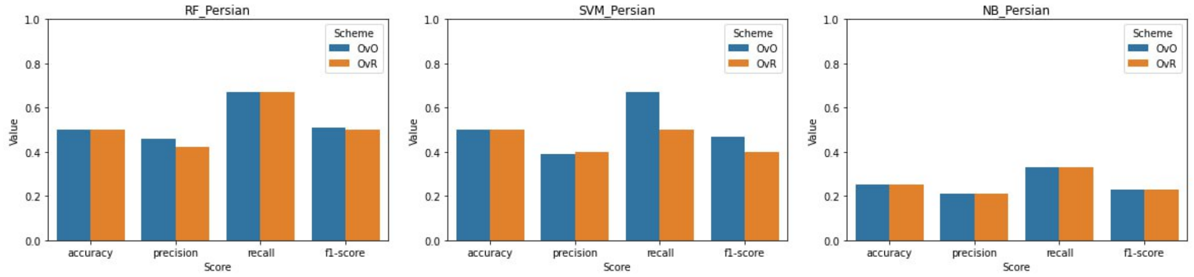

Figure 2 – Comparison of classification models

each classification model. Only in the SVM model do the results of the two schemes differ slightly for the recall score. In general, the figure (2) shows that the scheme has no effect on the speaker classification results. However, comparing the results obtained between different classification models is worth considering that the NB model has significantly low performance in classifying Persian speakers. The scores obtained in the other two models are quite similar. These results can be seen as confirming the effect of the choice of a model on the accuracy of speaker classification.

3.3 Feature importance

Evaluating the impact of features in the classification model is another interesting topic in model development. In classification models, the importance of features can be ranked based on their usefulness in predicting the target variable. There are several algorithms for determining the importance of features. Among them, we consider the mean decrease accuracy and mean decrease Gini coefficients as the basis for analysis (s. Fig. 3). Mean decrease accuracy coefficient shows how much removing each feature affects the accuracy of the model. Eliminating more important features of the model further affects its accuracy. Mean decrease Gini coefficient indicates how each feature contributes to the homogeneity of nodes and leaves in the resulting Random Forest. The higher the value of mean decrease accuracy or mean decrease Gini, the more important the feature is to the model under study.

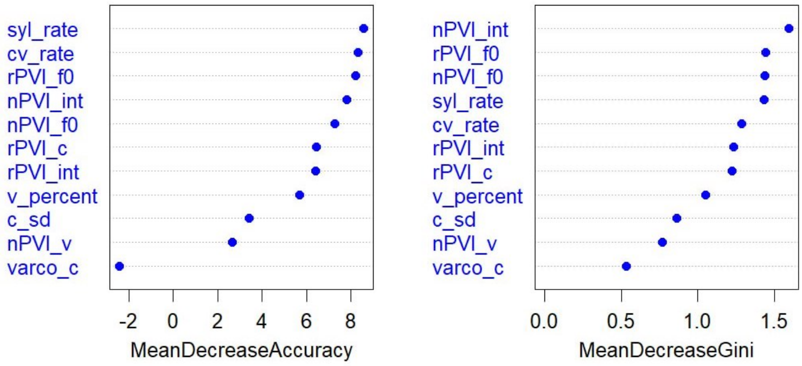

Figure 3 – Feature Importance in RF Model

4 Discussion & Conclusion

Studies in the field of speaker identification and speaker classification based on phonetic features have increased, and various research works attempt to shed new light on the studies in this field, taking into account different aspects. The present study was an attempt to evaluate the effects of model selection, schema selection and feature selection in speaker classification based on rhythmic features. Although the selected aspects cannot be considered completely innovative compared to previous studies, their application to a Persian corpus provides new results in the field of studies on this language. The first impression of the obtained results shows the impact of "model selection" on the performance of the classification model. The reason for this effect requires more extensive studies on the structure of the language and the effects of the selected features. Although, due to the limitations of this study, such as the small size of the corpus in terms of the number of speakers and the consideration of only one type of speech, caution should be advised in generalizing the results of this study to the Persian language as a whole. Moreover, the question of the selection of appropriate rhythmic features or even, from a broader perspective, the suitability of rhythmic features engineering for speaker classification algorithms remains an open and controversial issue, especially since the performance of these models is relatively low compared to models based on spectral features. In our studies, the accuracy of the RF model was about 50%, while the studies based on spectral features gave an accuracy of over 80% (see, e.g., [12]). However, a review of the performance of the features in the present analysis indicates that improvement in the performance of classification models based on rhythmic features should be sought in the selection of indicators that reflect the greatest differentiation among speakers. It should be noted that the interval measures and PVIs used as features here were initially introduced as features against the isochrony hypothesis, according to which sequences of syllables (syllable timing) or feet (stress timing) occur more or less simultaneously [38]. These indices were originally responsible for differentiating between languages in terms of rhythmic class, but on the basis of other characteristics of speech patterns such as syllabic complexity, vowel reduction, and stress-based lengthening. In addition, other measures (e.g., *VarcoC*, *VarcoV*) were first introduced to show rhythmic class differences at other levels such as the coefficient of variation of C and V interval durations and rate variability. However, there is still evidence that rhythmic measures such as $\%V$, $\%VO$, ΔV and PVI vary significantly between speakers [3, 9, 11]. Moreover, using the measures of $\%V$, [39] show that timing patterns in speech are proportional to the size of tongue movement area (TMA). In other words, the larger the TMA, the higher the average $\%V$ values. Others [40] showed that this measure reveals more between-speaker differences than other rhythm measures such as *Varco* or PVI. In the ranking of the features in the current study, this feature was placed in the middle position after the features based on frequency and intensity domains and those based on speech rate. However, in order to arrive at individual characteristics, it is necessary to consider other aspects in future research. In the present study, considering the values of the PVI index in other domains such as frequency and intensity seems to be a good step, especially since their influence on the ranking of the importance of the features became obvious. Thus, in setting the overall perspective for determining features, attention must be paid to approaches that take into account the effects of individual anatomical configurations that lead to different articulatory gestures and ultimately cause idiosyncratic pathways in the occurrence of temporal patterns. The culmination of this approach may be seen in frameworks that search for rhythmic features in the domain of intensity and based on patterns based on the amplitude envelope. Dellwo et al. [24] applied the technique of measuring the variability of peak-to-peak intervals in the amplitude envelope. He & Dellwo [41] stated that such a measure assumes that the anatomic characteristics of the speech organs and their movement leads to a certain temporal organization of in the specific characteristics of the amplitude envelope. On this basis, the characteristic of intensity variability in inter-syllable time intervals is used as another criterion for examining individual characteristics, which of course, reflect a significant speaker effect.

References

[1] Arvaniti, A.: *The usefulness of metrics in the quantification of speech rhythm. Journal of Phonetics*, 40(3), pp. 351–373, 2012.

[2] Prieto, P., M. del Mar Vanrell, L. Astruc, E. Payne, and B. Post: *Phonotactic and phrasal properties of speech rhythm. evidence from catalan, english, and spanish. Speech Communication*, 54(6), pp. 681–702, 2012.

[3] Wiget, L., L. White, B. Schuppler, I. Grenon, O. Rauch, and S. L. Mattys: *How stable are acoustic metrics of contrastive speech rhythm? JASA*, 127(3), pp. 1559–1569, 2010.

[4] White, L. and S. L. Mattys: *Calibrating rhythm: First language and second language studies*. *Journal of Phonetics*, 35(4), pp. 501–522, 2007.

[5] Payne, E., B. Post, L. Astruc, P. Prieto, and M. Vanrell: *Rhythmic modification in child directed speech*, pp. 147–183. Supplementi alla biblioteca di linguistica. Aracne, 2015.

[6] Pellegrino, E., L. He, and V. Dellwo: *The effect of ageing on speech rhythm: A study on Zurich German*. *University of Zurich*, 2018.

[7] Liss, J. M., S. LeGendre, and A. J. Lotto: *Discriminating dysarthria type from envelope modulation spectra*. *ASHA*, 2010.

[8] Grabe, E. and E. L. Low: *Durational variability in speech and the rhythm class hypothesis*. *Papers in laboratory phonology*, 7(1982), pp. 515–546, 2002.

[9] Loukina, A., G. Kochanski, B. Rosner, E. Keane, and C. Shih: *Rhythm measures and dimensions of durational variation in speech*. *JASA*, 129(5), pp. 3258–3270, 2011.

[10] Dellwo, V. and A. Fourcin: *Rhythmic characteristics of voice between and within languages*. *Revue Tranel (Travaux neuchâtelois de linguistique)*, 59, pp. 87–107, 2013.

[11] Leemann, A., M.-J. Kolly, and V. Dellwo: *Speaker-individuality in suprasegmental temporal features: Implications for forensic voice comparison*. *Forensic science international*, 238, pp. 59–67, 2014.

[12] Lykartsis, A., S. Weinzierl, and V. Dellwo: *Speaker identification for swiss german with spectral and rhythm features*. In *Audio Engineering Society Conference: 2017 AES International Conference on Semantic Audio*. Audio Engineering Society, 2017.

[13] Lykartsis, A.: *On the analysis of speech rhythm for language and speaker identification*. Phd dissertation, Technische Universität Berlin, 2020.

[14] Rouas, J.-L., J. Farinas, F. Pellegrino, and R. André-Obrecht: *Rhythmic unit extraction and modelling for automatic language identification*. *Speech Communication*, 47(4), pp. 436–456, 2005.

[15] Rouas, J.-L.: *Automatic prosodic variations modeling for language and dialect discrimination*. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(6), pp. 1904–1911, 2007.

[16] Tilsen, S. and K. Johnson: *Low-frequency fourier analysis of speech rhythm*. *JASA*, 124(2), pp. EL34–EL39, 2008.

[17] Tilsen, S. and A. Arvaniti: *Speech rhythm analysis with decomposition of the amplitude envelope: characterizing rhythmic patterns within and across languages*. *JASA*, 134(1), pp. 628–639, 2013.

[18] Alotaibi, Y. A., A. H. Meftah, S.-A. Selouani, and Y. M. Seddiq: *Speaker environment classification using rhythm metrics in levantine arabic dialect*. In *2014 9th International Symposium on Communication Systems, Networks & Digital Sign (CSNDSP)*, pp. 706–709. IEEE, 2014.

[19] Mefiah, A., Y. A. Alotaibi, and S.-A. Selouani: *Arabic speaker emotion classification using rhythm metrics and neural networks*. In *2015 23rd European Signal Processing Conference (EUSIPCO)*, pp. 1426–1430. IEEE, 2015.

[20] Ringeval, F., M. Chetouani, and B. Schuller: *Novel metrics of speech rhythm for the assessment of emotion*. In *INTERSPEECH*. 2012.

[21] Lykartsis, A. and S. Weinzierl: *Using the beat histogram for rhythm description and language identification*. In *16th INTERSPEECH*. Dresden, 2015.

[22] Bishop, C. M. and N. M. Nasrabadi: *Pattern recognition and machine learning*, vol. 4 of *Information Science and Statistics*. Springer, 2006.

[23] LERCH, A.: *An introduction to audio content analysis: Applications in signal processing and music informatics*. Wiley-IEEE Press, 2012.

[24] DELLWO, V., A. LEEMANN, and M.-J. KOLLY: *Speaker idiosyncratic rhythmic features in the speech signal*. Interspeech Conference Proceedings, 2012.

[25] HE, L.: *Development of speech rhythm in first language: The role of syllable intensity variability*. *JASA*, 143(6), pp. EL463–EL467, 2018.

[26] SANGAIAH, I. and A. VINCENT ANTONY KUMAR: *Improving medical diagnosis performance using hybrid feature selection via relieff and entropy based genetic search (rf-ega) approach: application to breast cancer prediction*. *Cluster Computing*, 22(3), pp. 6899–6906, 2019.

[27] GALAR, M., A. FERNÁNDEZ, E. BARRENECHEA, H. BUSTINCE, and F. HERRERA: *An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes*. *Pattern Recognition*, 44(8), pp. 1761–1776, 2011.

[28] WINDFUHR, G. L.: *Persian grammar*. In *Persian Grammar*. De Gruyter Mouton, 1979.

[29] LAZARD, G.: *A grammar of contemporary Persian*. Mazda Publishers, 1992.

[30] SADEGHI, V.: *A phonetic study of vowel reduction in persian*. *Language Related Research*, 6(3), pp. 165–187, 2015.

[31] SHEYKH, S. S. and M. BIJANKHAN: *The study of vowel reduction in persian spontaneous speech*. *Journal Of Researches In Linguistics*, 2(1), pp. 35–48, 2010.

[32] BOUBAN, N.: *Quantitative rhythm indices in Persian, in comparison with English and French*. In *7th Iranian Conference on Linguistics*. Tehran, Allame Tabatabaie University, 2007.

[33] ASADI, H., M. NOURBAKHSH, L. HE, E. PELLEGRINO, and V. DELLWO: *Between-speaker rhythmic variability is not dependent on language rhythm, as evidence from persian reveals*. *Intern. Journal of Speech, Language and the Law*, 25(2), pp. 151–174, 2018.

[34] BOERSMA, P.: *Praat: doing phonetics by computer*. *http://www. praat. org/*, 2022.

[35] TEAM, R. C.: *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2012*. 2021.

[36] KASSAMBARA, A., F. MUNDT ET AL.: *Factoextra: extract and visualize the results of multivariate data analyses*. *R package version*, 1(5), pp. 337–354, 2020.

[37] KLEYNHANS, N. T. and E. BARNARD: *Language dependence in multilingual speaker verification*. *PRASA*, 2005.

[38] ABERCROMBIE, D.: *Elements of General Phonetics*. Edinburgh University Press, 1967.

[39] TOMASCHEK, F. and A. LEEMANN: *The size of the tongue movement area affects the temporal coordination of consonants and vowels—a proof of concept on investigating speech rhythm*. *JASA*, 144(5), pp. EL410–EL416, 2018.

[40] DELLWO, V., A. LEEMANN, and M.-J. KOLLY: *Rhythmic variability between speakers: Articulatory, prosodic, and linguistic factors*. *JASA*, 137(3), pp. 1513–1528, 2015.

[41] HE, L. and V. DELLWO: *The role of syllable intensity in between-speaker rhythmic variability*. *Intern. Journal of Speech, Language & the Law*, 23(2), 2016.