Speaker and prosodic peculiarity classification in emotional speech

Neda Mousavi[1], Sven Grawunder[1,2]

[1]Martin Luther University Halle-Wittenberg, Germany
[2]Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

https://doi.org/10.36505/ExLing-2024/15/0023/000648

Abstract

In this study, the relationship between rhythmic metrics, emotion recognition, and speaker variability is investigated using the German emotional speech corpus (VMEmo). Using principal component analysis and linear discriminant, the results show accuracies close to 0.40 when rhythmic features from different acoustic domains of time, intensity, and frequency are merged to identify linguistic behavior. However, the fluctuating accuracies of 0.44 to 0.17 in classifying speakers based on specific rhythmic feature categories emphasize the significant differences within these feature subgroups. These variations suggest possible nuances or complexities that require deeper exploration and thorough investigation to better understand the differences between these features and their impact on speaker classification accuracy.

Keywords: rhythm, speaker classification, between-speaker variation, prosodic peculiarity, emotional speech

Introduction

In the literature, rhythmic patterns in speech have been introduced not only as potential markers for individual features but also as cues for emotion recognition (Lykartsis, 2020; Mefiah et al. 2015). In this study, we investigate the interaction between rhythmic measures, emotion recognition, and speaker variability, focusing on the German emotional speech corpus (VMEmo, Batliner et al., 2000). The emotional data in this corpus were elicited through an experimental method in which participants interact with machines in tasks such as making an appointment. The experiment was purposefully designed to elicit various emotional fluctuations in the speaker as a result of the machine's responses. However, emotional fluctuations were not labelled directly, rather the corpus tags mark the speaker's linguistic behavior during emotional expressions. With our focus on rhythmic patterns, we examined the prosodic dimensions of this linguistic behavior and considered the prosodic tags of each utterance, including peculiarities such as pauses between words, strong contrastive stress, pauses between syllables, syllable lengthening, etc. The research focused on answering these two questions: how accurately can we identify different prosodic strategies for expressing emotion based on rhythmic

indicators? and, how effectively can speakers be distinguished based on these rhythmic indicators?

Method

Data preparation involved extracting key information from the VMEmo corpus, including start and end times of human-generated segments, speaker phrases, and prosodic peculiarity tags. Machine-generated segments were excluded, and waveforms from 33 speakers were selected. Phrases under 4 seconds were omitted following Tilsen and Arvaniti (2013). Automatic segmentation was conducted using WebMAUS (Kisler et al., 2017), and annotations were enriched in Praat TextGrids with details on phrase numbers, peculiarity tags, and consonant/vowel intervals.

In feature selection, rhythmic indices such as %V, ΔC (Ramus et al., 1999), nPVIv, rPVIc (Grabe and Low, 2002), varcoC and CV rate (Dellwo, 2006) were examined. Following He and Dellwo (2016), intensity-based rhythmic indicators, including the SD and nPVI of peak and mean intensity values, were analyzed. Also referring to Mousavi and Grawunder (2023), metrics from the frequency domain reflecting intensity indicators were also included. In total, 14 rhythmic features were categorized into four subgroups: duration-based, intensity-based, frequency-based and all metrics .

Results

The study employed principal component analysis (PCA) and linear discriminant analysis (LDA) as the models of analysis. Figure (1) illustrates the distribution of the values of the first component across different speakers, sorted by the variance of values. The distribution of PC1 values within each violin plot shows the range and distribution of these values across the different phrases uttered by each speaker.

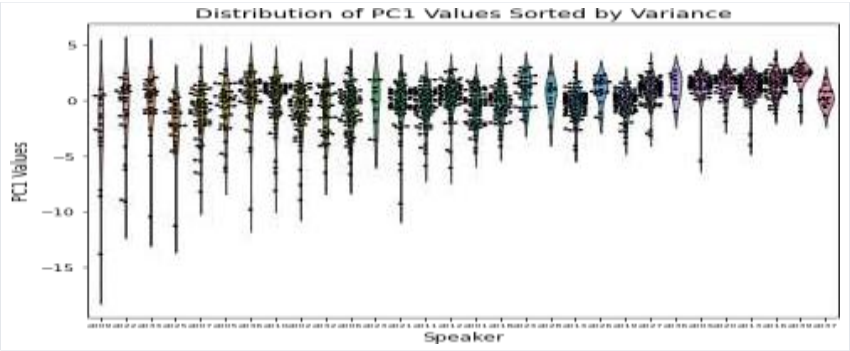

Figure 1. Visualizing within-speaker variability through PCA Analysis.

Figure (2) depicts the variance among speakers based on their PC1 and PC2 values. Closely clustered points indicate speakers with similar rhythmic patterns, while greater distances between points suggest greater variability in the rhythmic measures.

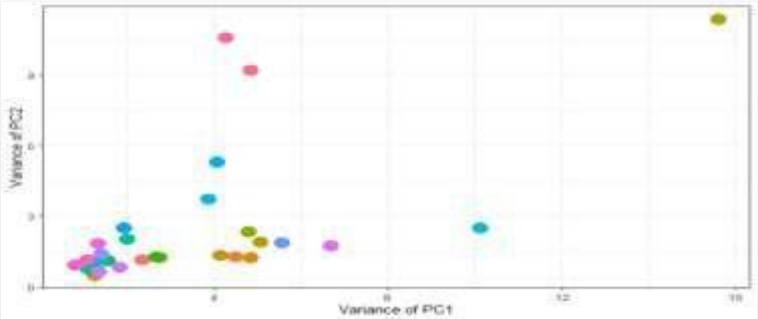

Figure 2. The variance among speakers based on their PC1 and PC2 values

Following this, we employed linear discriminant analysis for two main tasks: classification of phrases into prosodic peculiarities, regardless of speaker identity and evaluate the discriminative power of rhythmic features in distinguishing prosodic patterns, and speaker recognition, which focuses on speaker identification under the abstraction of emotional context. In addition, rhythmic groupings were compared across acoustic domains for their discrimination potential.

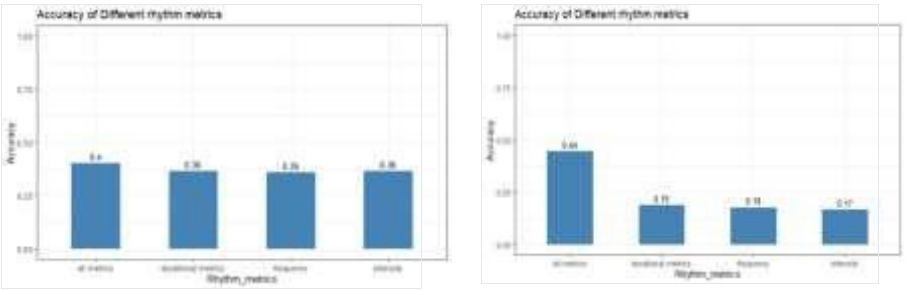

Figure 3. The accuracy of the recognition of prosodic peculiarities (left) and speaker (right)

While the differences in accuracy between these rhythmic metric categories suggest slight discrepancies in their effectiveness, the significance of these discrepancies requires further investigation. However, the notable differences in these accuracy values suggest possible differences in the discriminative abilities of these metrics for speaker discrimination.

Discussion

In this study, we investigated the effectiveness of rhythmic features extracted from vowel and consonant intervals in three different acoustic domains. The results show an approximate accuracy of 40% in distinguishing linguistic behaviors using rhythmic features in different acoustic domains. Furthermore, speaker identification achieved a similar level of accuracy when merging features from all acoustic domains. These results are consistent with previous research using rhythmic indicators extracted from different approaches to rhythm measurement, e.g. from Music Information Retrieval (Lykartsis, 2020). However, the interpretation of the accuracy levels achieved depends on their context of use and must be evaluated in accordance with the intended application and study objectives. Further research could also include a detailed investigation of the robustness of these rhythmic features in different emotional states, different linguistic contexts, or even in specific communication environments to allow a more nuanced understanding of their applicability.

References

Batliner, A., Huber, R., Niemann, H., Nöth, E., Spilker, J., Fischer, K. 2000. The Recognition of Emotion. In: Wahlster, W. (Ed.), Verbmobil: Foundations of Speech-to-Speech Translation. Artificial Intelligence. Springer, Berlin, Heidelberg.

Dellwo, V. 2006. Rhythm and speech rate: a variation coefficient for deltaC. In Karnowski P. & Szigeti, I. (ed.) Language and language processing. Frankfurt am Main: Peter Lang, 231-241.

Grabe, E., Low, E.L. 2002. Acoustic correlates of rhythm class. In: Gussenhoven, Warner (Eds.), Laboratory Phonology, vol. 7. Berlin: Mouton de Gruyter: 515–546.

He, L., Dellwo, V. 2016. The role of syllable intensity in between-speaker rhythmic variability. International Journal of Speech, Language & the Law, 23(2), 243-273.

Kisler, T., Reichel, U.D., Schiel, F. 2017. Multilingual processing of speech via web services. Computer Speech & Language, 45, 326–347.

Lykartsis, A. 2020. On the analysis of speech rhythm for language and speaker identification. PhD dissertation, Technische Universität Berlin.

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

Mousavi, N., Grawunder, S. 2023. Persian speaker classification using rhythmic features. In Draxler, C., editor, Studientexte zur Sprachkommunikation: Elektronische Sprachsig-nalverarbeitung 2023, pages 194–201. TUDpress, Dresden.

Ramus, Fr., Nespor, M., Mehler, J. 1999. Correlates of Linguistic Rhythm in the Speech Signal. Cognition, 73, 265-292.

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