UNSUPERVISED EMOTIONAL PATTERN RECOGNITION USING RHYTHMIC AND VOCAL FEATURES

Neda Mousavi[1], *Seyyed Saeed Sarfjoo*[2], *Sven Grawunder*[1],[3]

[1]*Martin Luther University Halle-Wittenberg, Germany,* [2]*Dialpad Canada Inc,* [3]*Max Planck Institute for evolutionary Anthropology, Leipzig, Germany*
neda.mousavi@sprechwiss.uni-halle.de, saeed.sarfjoo@dialpad.com,
sven.grawunder@sprechwiss.uni-halle.de

Abstract: In this study, we address the complex dynamics of emotional speech and comprehensively examine the integration of rhythmic and vocal features to recognize emotional patterns. Our exploration is conducted using two German emotional corpora: VMEmo and EmoDB. Employing a combination of supervised methods (here linear discriminant analysis, LDA) and unsupervised techniques (here k-means clustering), we aim to uncover nuanced patterns within the emotional speech in these corpora. The application of LDA highlights salient patterns across different feature sets and focuses on the classification of speakers and prosodic characteristics. In addition, k-means clustering uncovers latent structures that reveal subtle mapping between emotions and speech behavior. Our results suggest that it is possible to cluster data based on prosodic behaviors that are influenced by emotional changes. Although precise mapping to the actual clusters derived from emotional labels could not be fully achieved, the results nonetheless reveal a moderate level of success in this investigation.

1 Introduction

The study of communication, which encompasses various aspects from linguistic analysis to the recognition of emotions, is a diverse and important field of research. Within this broad field, the study of emotional speech occupies an important place that encourages in-depth investigation and research. To this end, numerous corpora have been compiled to explore emotional speech and shed light on the intricate interplay between vocal features, linguistic cues, and underlying emotional states. However, the methods used to elicit and represent these emotional nuances differ considerably between the various emotional corpora, not only in terms of how different emotional categories are handled but also in terms of the fundamental question of how emotions are conceptualized in speech.

According to the literature in the field, individuals typically can experience different emotional states at the same time. Psychologists have sought to understand the structures and subtleties associated with mixed emotions [14]. Within the extensive spectrum of human emotions, around 34,000 different variants have been identified [19]. However, many of the available corpora contain emotion labels that only represent prototypes of such mixed patterns. Datasets such as the Danish Emotional Speech Corpus [6], the Berlin Emotional Speech Database (EmoDB) [4], and the FAU Aibo Emotion Corpus [1] can be mentioned in this context, which apply limited labels to emotional speech emulated by humans. Other corpora, such as the VMEmo corpus [23], follow a different methodology, according to which the speaker's emotional fluctuations are not directly indicated by emotional labels. Instead, these changes are derived from linguistic peculiarities that comprise the speaker's expression to convey his or her feelings.

DOI: 10.35096/othr/pub-7102, URN: urn:nbn:de:bvb:898-opus4-71026

This serves as the starting point for our investigation, which aims to determine the necessity of explicit emotional labeling for the classification of a speaker's emotional state. Specifically, we investigate the possibility of using unsupervised methods to classify emotions in speech that are subject to emotional fluctuations and are only reflected in labeled linguistic behavior. To achieve this objective, we performed a focused analysis using the VMEmo corpus and then used the EmoDB corpus as a reference to evaluate and validate the results obtained. VMEmo is a subset of the broader German Verbmobil project (VM), a project conducted between 1993 and 2000 aiming to develop an automatic speech-to-speech translation system for German, American English, and Japanese [23]. The corpus includes speech signals derived from the interaction between humans and a machine involving an agreement on a specific topic. Within these dialogues, deliberate attempts were made to elicit specific emotions from participants through the system's responses. As a result, speakers' speech patterns are subject to emotional changes that influence their linguistic behavior and are characterized by lexical, conversational, or prosodic peculiarities. Although there is no direct labeling of emotional mode, variations in emotional mode are thought to trigger the use of different peculiarities, thus revealing latent emotional dimensions of speakers. In addition, the EmoDB corpus, a collection of acted emotional speech in German, serves as a reference. This corpus consists of ten sentences performed by ten actors, each representing six different emotions. In this way, EmoDB relies on actors evoking emotions through e.g. the Stanislavski method, drawing on past intense emotional experiences to express them authentically. The emotional labels include, notably (German terms in brackets) neutral (*Neutral*), anger (*Ärger*), fear (*Angst*), joy (*Freude*), sadness (*Trauer*), disgust (*Ekel*) and boredom (*Langeweile*).

Investigating consistent cues and acoustic measures for analyzing emotional speech is the next aspect addressed in this study, with previous research dating back to [7]. Several studies [17], [21], [15] indicate differences in acoustic measures such as pitch, pitch range, rate of speech, voice quality, and articulation accuracy across different emotions. In addition, research [9], [8], [12] has investigated the use of rhythmic features to identify the emotional state of the speaker. Therefore, the aim could be to find out which attributes - whether vocal, rhythm-based, or a combination of both - provide better results in such analysis. Consequently, this study addresses two key aspects: firstly, the clustering of emotion patterns without explicit emotion labels, and secondly, exploring the role of rhythmic and vocal features in the representation of these patterns.

2 Method

Focusing on two datasets, VMEmo and EmoDB, this study evaluates three distinct sets of features – rhythm-related, voice-related, and a combination of both – to assess their performance in recognizing emotional patterns. The VMEmo-derived dataset contains the human-generated audio files of 33 individual speakers, the textual content of each spoken phrase, and the particular prosodic tags that characterize the speakers' different linguistic strategies during the engagement. To ensure the accuracy of the data for the rhythm analysis, phrases shorter than 4 seconds were excluded according to [22]. The audio files were automatically segmented using the WebMaus service [13], whereby the corresponding orthographic version was used for each signal. Additional tiers were integrated into the *Praat*[3] TextGrids to enhance the annotations, including the phrase number, the peculiarity tags, and the intervals regarding consonants and vowels. The EmoDB corpus was subjected to the same procedure in preparation for this analysis.

The analysis of acoustic characteristics included an evaluation of various indices associated with both rhythmic patterns and vocal attributes. For rhythm analysis, this included an assessment of the duration of each utterance, as well as metrics such as V% (the percentage of vocalic intervals) and the standard deviation of consonant intervals [20], pairwise variability measures (nPVIv, rPVIc) [10], varcoC and CV rate [5]. In addition, standard deviation (SD) and nPVI measures for peak and mean values within the intensity and frequency domains [11], [16] were included in the analysis to capture domains beyond time. Finally, a total of fifteen different rhythmic features were included, all of which were normalized using the StandardScaler class in *scikit-learn* library [18]. The *Praat* based voice analysis included the assessment of metrics for pitch stability (known as jitter), amplitude fluctuations (shimmer), and parameters of sound quality (HNR) and fundamental frequency. Thirteen different features were generated for the voice analysis, comprising metrics such as jitter (local, rap, ppq5), absolute jitter (local absolute), and shimmer (local, local dB, apq3, apq5, apq11 and dda), HNR, and mean and standard deviation of the fundamental frequency.

Employing *scikit-learn* library [18] to implement linear discriminant analysis (LDA) as a supervised model, two different models were used in this study - one for speaker recognition and one for prosodic peculiarity recognition. Both models were trained based on explicit labels corresponding to speakers and prosodic peculiarity tags in the corpus. The study is then extended to an unsupervised clustering based on the same package, using the k-means algorithm to identify and classify emotional states in the absence of direct emotional tags.

3 Results

The preliminary results consist of the performance metrics derived from linear discriminant analysis (LDA) applied to speaker recognition and prosodic peculiarity identification within the VMEmo corpus. Different metrics, including accuracy, precision, recall, and F1 score, provide insights into the classification performance achieved with different feature sets. Using *scikit-learn* library, a function was implemented to obtain the metrics using the predictions generated by the model on the test data and comparing them to the true labels. In this context, the effects of using different feature sets – rhythm-based, voice-based, and a combination of both – can be compared.

subcaption In the context of speaker recognition, the performance metrics for different feature sets reveal varying levels of effectiveness (s. Fig 1). Rhythm features exhibit an accuracy of 37%, precision of 39%, recall of 37%, and an F1-score of 36%. Voice features, on the other hand, demonstrate lower performance with an accuracy of 22%, precision of 19%, recall of 22%, and an F1-score of 19%. The combined features, encompassing both rhythm and voice aspects, achieve better results with an accuracy of 46%, precision of 48%, recall of 46%, and an F1-score of 46%. For the recognition of prosodic peculiarity, distinct patterns emerge (s. Fig 2). Rhythm features exhibit an accuracy of 44%, precision of 33%, recall of 44%, and an F1-score of 34%. Voice features demonstrate an accuracy of 38%, precision of 41%, recall of 38%, and an F1-score of 23%. The combined features, integrating both rhythm and voice characteristics, yield an accuracy of 45%, precision of 37%, recall of 45%, and an F1-score of 37%. These results underscore the significance of considering combined features for enhanced prosodic peculiarity recognition in comparison to individual rhythm or vocal feature sets. Rhythm and combined features also show relatively comparable performance patterns in the areas of speaker and prosodic feature recognition. In contrast, vocal features differ significantly from each other, showing significant differences in results between speaker recognition and prosodic peculiarity recognition. However, a more detailed analysis is required to clarify the factors contributing to this observed trend.

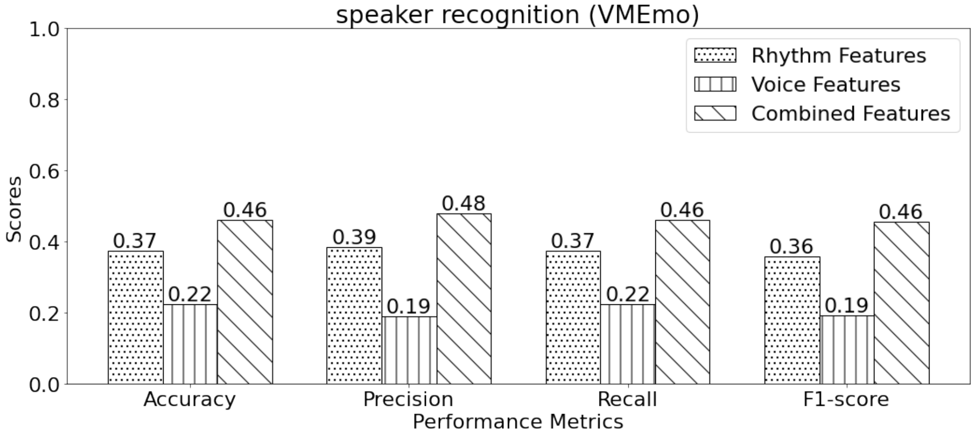

Figure 1 – Performance metrics for the classification of speakers across different feature sets

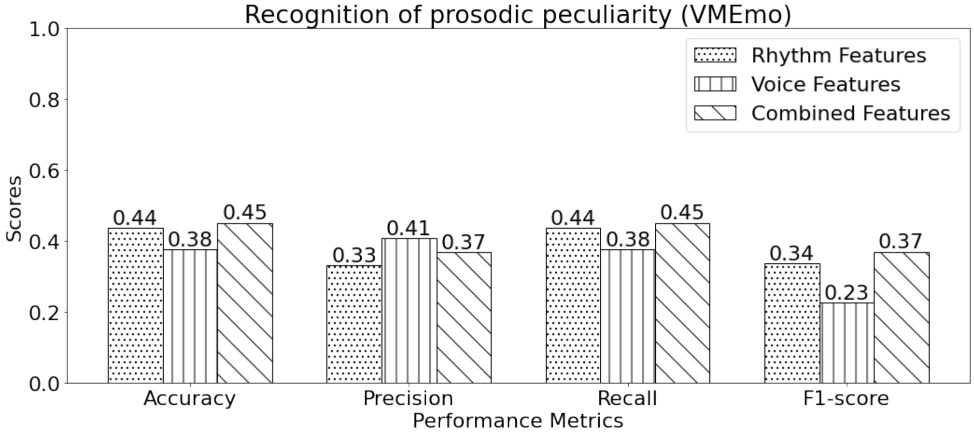

Figure 2 – Performance metrics for the classification of prosodic peculiarity tags across different feature sets

3.1 Unsupervised clustering using k-means algorithm

Concerning the main focus of this study, the k-means clustering algorithm as an unsupervised model is used to recognize emotional states during Human-Machine interaction. The primary challenge is to identify emotional patterns in speech when no direct emotional labels are available, relying on the assumption that variations in linguistic behavior reflect underlying emotional fluctuations. This framework emphasizes prosodic features that act as mediators to capture the nuanced interplay between linguistic cues and emotional states. Through the application of the k-means clustering method to the prosodic tags of the datasets, we aim to reveal distinctive clusters that signify emotional expressions embedded within the speech data.

The optimal number of clusters (k) in the k-means clustering analysis was determined by the *sum of squares within clusters* (WCSS) metric [2]. The WCSS calculations were performed over a spectrum of potential cluster counts, and the elbow point, which denotes the optimal clustering point, was determined. Using this metric, a consistent cluster selection of 5 was achieved for all feature sets of VMEmo corpus. Then, based on this selected number of clusters, the distribution of observations was visualized using the reduced-dimensional space of principal component analysis (PCA). The centroids of each cluster are indicated and provide a visual representation of the effectiveness of the clustering algorithm in separating observations (s. Fig 3). Clusters that overlap may indicate similarities or mixed characteristics between classes, while clearly separated clusters indicate clear and distinct categories.

To quantitatively assess the discriminative power of our model and identify potential areas for improvement in feature selection or clustering algorithms, we use the silhouette metric.

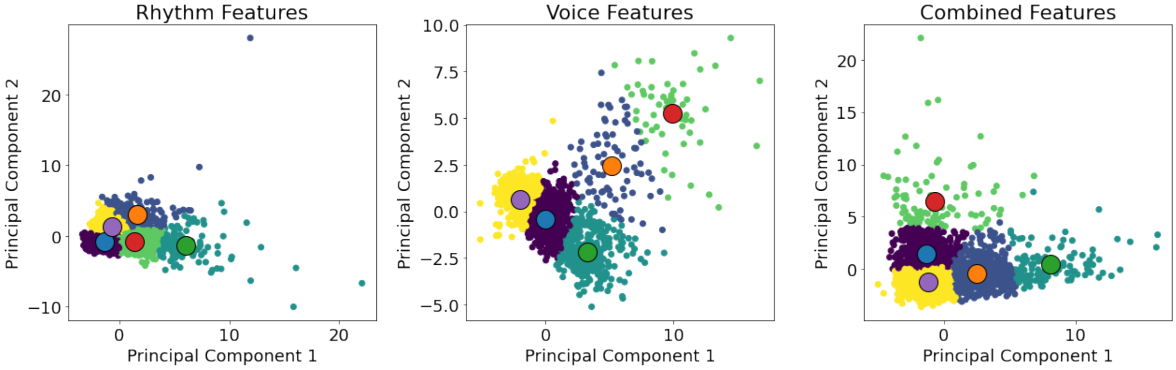

Figure 3 – Unsupervised clustering of VMEmo across different feature sets

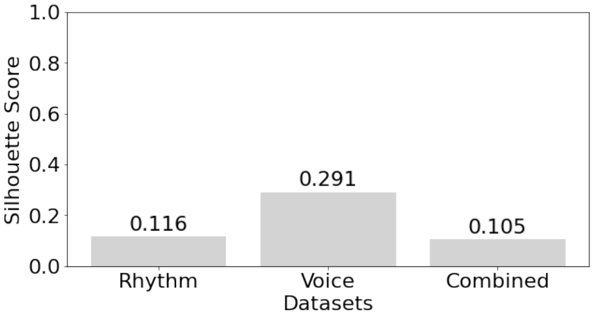

Figure 4 – Quantification of the degree of separation between the clusters based on the silhouette value

This metric, which ranges from -1 to 1, provides information about the degree of overlap or separation between clusters (s. Fig 4). A value close to 1 indicates well-separated clusters, suggesting robust clustering performance. Conversely, values close to 0 indicate clusters with some degree of overlap, indicating areas where either feature selection or clustering methods need to be improved. Negative values indicate possible misclassification of data points.

In Figure 4, the value of 0.116 for the rhythm features indicates a medium level of consistency within the clusters. The vocal features have a stable silhouette score of 0.291, indicating better-defined clusters. However, the combined features have a lower silhouette value of 0.105, indicating weaker separation and less consistency within the clusters.

3.2 Analyzing the EmoDB corpus as a reference for the clustering analysis

In the following, we use the EmoDB corpus as a reference to evaluate the unsupervised clustering results. Here, the clustering results are compared with the known emotional categories in the EmoDB dataset, which provides insights into the effectiveness and accuracy of the clustering algorithm in capturing different emotional patterns in the speech data. Similar to the previous series of analyses, three different data sets were created, emphasizing rhythmic features, vocal features, and their combination. Applying WCSS method to these data sets consistently resulted in an optimal cluster number of 4 for each set (s. Fig 5).

To evaluate and compare the clustering outcomes with explicit emotion labels in the three datasets, two widely used metrics were used - the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI). These metrics range from 0 to 1 and quantify the similarity between the clusters generated by the algorithm and the true emotion labels. Consequently, they serve as tools to assess the accuracy of the clustering results. Considering the 7 emotional labels of the main corpus, we performed an evaluation using two clustering approaches. The first one used the optimal cluster number of 4, which was determined by the WCSS criterion (s. Fig 6). The second approach used the number of 7 clusters to match the actual number of emotional labels

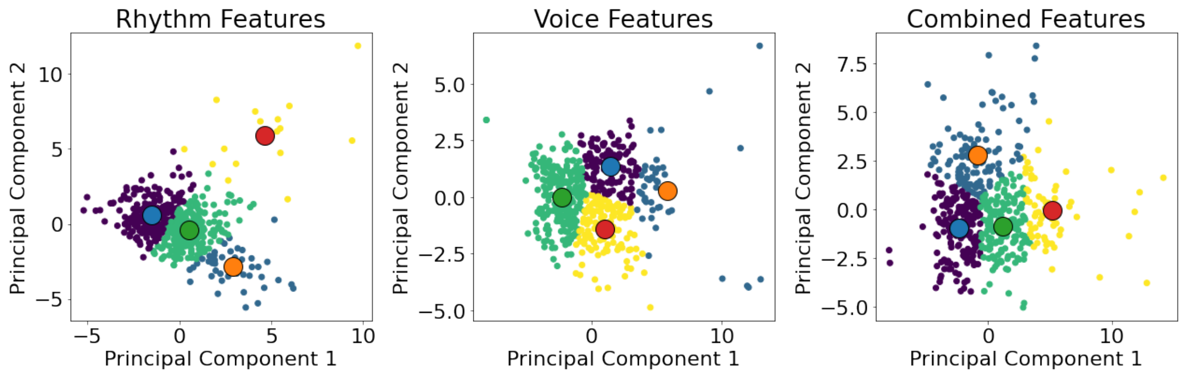

Figure 5 – Unsupervised clustering of EmoDB across different feature sets

present in the corpus (s. Fig 7).

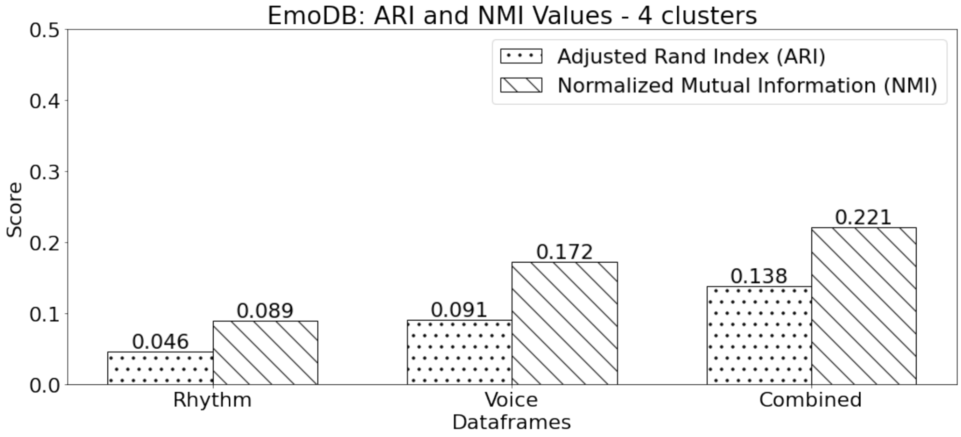

Figure 6 – ARI and NMI values for 4 clusters (EmoDB)

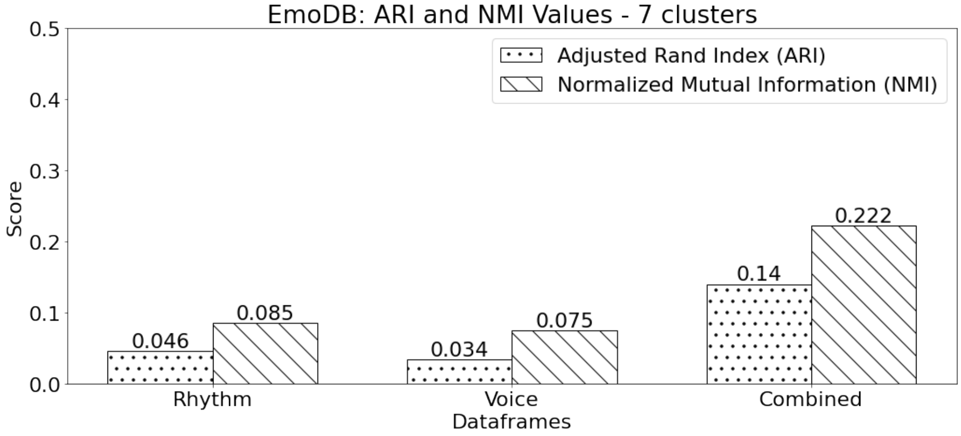

Figure 7 – ARI and NMI values for 7 clusters (EmoDB)

A comparison of these results reveals that the combined dataset shows stronger agreement between the clustering and emotion labels, with higher ARI (0.1401) and NMI values (0.2237) compared to the rhythm dataset (ARI: 0.0468, NMI: 0.0839) and the voice dataset (ARI: 0.0924, NMI: 0.1718). This suggests that merging rhythm and voice features can improve the correspondence of clusters with reported emotion labels. However, all datasets show moderate agreement between the clustering and the explicit emotion labels, suggesting that the clusters may not fully represent the labeled emotions. In the second graph, similar to the choice of 4 clusters, the voice features outperformed the rhythm features in cluster performance, and a combination of rhythm and voice features improved performance compared to the individual features. In addition, considering the minimal difference between the results of the two clustering with 4 and 7 clusters, it can be concluded that by selecting the number of clusters from WCSS analysis, the algorithm focuses on the most well-shaped clusters of the actual data that reach ARI and NMI scores similar to the condition when we use the actual number of clusters.

4 Discussion

This study explored the potential of clustering observations by utilizing prosodic information derived from the speaker's emotional fluctuations and assigning them to respective emotional classes. Using the unsupervised K-means algorithm, the analysis focused on the VMEmo corpus, which consists of German emotional speech. In addition, the influence of feature selection on the clustering results was investigated. The three different groups - rhythmic features, vocal features, and their combination - were compared. The results show that although the feature groups considered ultimately produce a relatively equal number of clusters, the separability of the resulting clusters varies between the different features of the VMEmo corpus. Vocal features show clusters with higher differentiation, while rhythmic features and the combined sentence represent a moderate degree of differentiation. Moreover, the potential mapping of the resulting clusters to the actual emotional clusters in the corpus was assessed using the EmoDB corpus, which contains explicit emotional labels. The results show a moderate mapping between the clusters of the unsupervised algorithm and the actual emotional labels in the corpus. Even when the number of clusters of the unsupervised algorithm is intentionally set equal to the emotional labels in the corpus, there is only a minimal difference in this mapping. Although these results have limitations, future research could investigate other feature sets, such as spectral features, which may yield clusters with greater correspondence to emotional clusters.

References

[1] Anton Batliner, Stefan Steidl, and Elmar Nöth. Releasing a thoroughly annotated and processed spontaneous emotional database: the fau aibo emotion corpus. 2008.

[2] C Bishop. Pattern recognition and machine learning. *Springer google schola*, 2:531–537, 2006.

[3] Paul Boersma and David Weenink. Praat: doing phonetics by computer [Computer program], 2022.

[4] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, Benjamin Weiss, et al. A database of german emotional speech. In *Interspeech*, volume 5, pages 1517–1520, 2005.

[5] Volker Dellwo, P Karnowski, and I Szigeti. Rhythm and speech rate: A variation coefficient for deltac. 2006.

[6] Inger Samso Engberg and Anya Varnich Hansen. Documentation of the danish emotional speech database des. *Internal AAU report, Center for Person Kommunikation, Denmark*, 22, 1996.

[7] Grant Fairbanks and Wilbert Pronovost. An experimental study of the pitch characteristics of the voice during the expression of emotions. *Speech monographs*, 1939.

[8] John Gideon, Emily Mower Provost, and Melvin McInnis. Mood state prediction from speech of varying acoustic quality for individuals with bipolar disorder. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 2359–2363. IEEE, 2016.

[9] Frederick K Goodwin and Kay Redfield Jamison. *Manic-depressive illness: bipolar disorders and recurrent depression*, volume 2. Oxford university press, 2007.

[10] Esther Grabe and Ee Ling Low. Acoustic correlates of rhythm class. *Laboratory phonology*, 7(515-546), 2002.

[11] Lei He and Volker Dellwo. The role of syllable intensity in between-speaker rhythmic variability. *International Journal of Speech, Language & the Law*, 23(2), 2016.

[12] Soheil Khorram, John Gideon, Melvin G McInnis, and Emily Mower Provost. Recognition of depression in bipolar disorder: Leveraging cohort and person-specific knowledge. In *INTERSPEECH*, pages 1215–1219, 2016.

[13] Thomas Kisler, Uwe Reichel, and Florian Schiel. Multilingual processing of speech via web services. *Computer Speech & Language*, 45:326–347, 2017.

[14] Sylvia D Kreibig and James J Gross. Understanding mixed emotions: paradigms and measures. *Current opinion in behavioral sciences*, 15:62–71, 2017.

[15] Donn Morrison, Ruili Wang, and Liyanage C De Silva. Ensemble methods for spoken emotion recognition in call-centres. *Speech communication*, 49(2):98–112, 2007.

[16] Neda Mousavi and Sven Grawunder. The classification of speaker and prosodic peculiarities in emotional speech based on rhythmic patterns. In *14th International Conference of Experimental Linguistics (ExLing 2023)*, Athens, Greece, 2023.

[17] Iain R Murray and John L Arnott. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America*, 93(2):1097–1108, 1993.

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[19] Robert Plutchik. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350, 2001.

[20] Franck Ramus, Marina Nespor, and Jacques Mehler. Correlates of linguistic rhythm in the speech signal. *Cognition*, 73(3):265–292, 1999.

[21] Klaus R Scherer. Vocal communication of emotion: A review of research paradigms. *Speech communication*, 40(1-2):227–256, 2003.

[22] Sam Tilsen and Amalia Arvaniti. Speech rhythm analysis with decomposition of the amplitude envelope: Characterizing rhythmic patterns within and across languages. *The Journal of the Acoustical Society of America*, 134(1):628–639, 2013.

[23] Wolfgang Wahlster. *Verbmobil: foundations of speech-to-speech translation*. Springer Science & Business Media, 2013.