Emotion Variation Detection in Discrete English Speech: A Wavelet Transform Use Case in Mental Health Monitoring

Adebanji Adeleye
Department of Computer Science and Software Engineering, Auckland University of Technology
Auckland, New Zealand
adebanji.adeleye@aut.ac.nz

Samaneh Madanian
Department of Computer Science and Software Engineering, Auckland University of Technology
Auckland, New Zealand
sam.madanian@aut.ac.nz

Olayinka Adeleye
Department of Computer Science and Software Engineering, Auckland University of Technology
Auckland, New Zealand
olayinka.adeleye@aut.ac.nz

ABSTRACT

The increasing complexity in modern society has been leading to a series of emotional shifts and mental pressures for individuals. Emotion detection can assist people in managing stress and monitoring mental health. Consequently, recent works are leveraging advancements in vocal/acoustic signal processing and machine learning models to improve emotion detection from speech signals. A challenge in detecting variations in emotion from speech involves the identification of appropriate features that can accurately represent the underlying phenomenon. This paper proposes a set of features derived from energy content and entropy measures extracted through the decomposition signals of the discrete wavelet transform. These features aim to characterize various negative emotions, encompassing fear, sadness, anger, anxiety, and disgust, within speech signals in non-controlled noise conditions. We employ CNN-based architectures to classify the speech signals to detect the embedded emotions. The results of our experiments on publicly available datasets show that the proposed method performs better than the state-of-the-art methods, which use other time-frequency representations. We achieved an unweighted accuracy (UA) of 83.7 ± 2.5 and a weighted accuracy (WA) of 81.7 ± 5.

CCS CONCEPTS

• **Applied computing** → **Health informatics**; *Health informatics*;
• **Information systems** → *Expert systems.*

KEYWORDS

Speech Emotion, Wavelet, Mental health, Data Analytics, CNN, Digital Health, Healthcare Digital Transformation

ACM Reference Format:
Adebanji Adeleye, Samaneh Madanian, and Olayinka Adeleye. 2024. Emotion Variation Detection in Discrete English Speech: A Wavelet Transform Use Case in Mental Health Monitoring. In *2024 Australasian Computer Science Week (ACSW 2024), January 29–February 02, 2024, Sydney, NSW, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3641142.3641167

ACSW 2024, January 29–February 02, 2024, Sydney, NSW, Australia

ACM ISBN 979-8-4007-1730-7/24/01
https://doi.org/10.1145/3641142.3641167

1 INTRODUCTION

Human emotions play a crucial role in everyday life, influencing various domains such as mental health support, criminal investigation, call center operations, and biomedical engineering [2, 16].

Emotion as a mental experience is often characterized by increased physiological responses, including accelerated heartbeat, breathing, perspiration, vocal disorder, and facial expressions. Emotions can be detected through speech, facial expressions, gestures, electroencephalography, and autonomic nervous system signals, among which emotion recognition using speech is more popular [13]. Despite mental health being a subjective experience, changes in emotional states and stress can be objectively measured through physiological signals including speech signals. Studies have demonstrated the intrinsic connection between human psychological activities and the physiological functions of the human body [20]. Maintaining an excellent and stable emotional state contributes to optimal physical functioning. Conversely, disruptions in emotional well-being can adversely affect physical functions and potentially lead to various diseases.

Individuals suffering from different mental disorders often manifest specific emotions as part of their condition. For instance, anxiety and fear are often associated with individuals undergoing stress [4] and seasonal affective disorder [15]. Major depressive disorder (MDD) [14] and mood disorders [5] are commonly associated with feelings of sadness. Generally, individuals diagnosed with Borderline Personality Disorder (BPD) exhibit regular mood swings, display emotional instability, and undergo intense negative emotions, a condition commonly termed affective dysregulation [7]. Therefore, emotions such as anger, fear, sadness, and emotional neutrality can serve as indicators not only of mental disorders but also of other medical conditions. The ability to predict these emotions could significantly assist mental healthcare professionals during the initial stages of investigation, facilitating more accurate diagnosis.

2 RESEARCH BACKGROUND

To detect and classify emotions in speech, three crucial components are [9, 12, 16]: (i) a classifier, (ii) features, and (iii) a database of speech for processing and validation purposes. Identifying appropriate features for an efficient emotion detection process is challenging, prompting the need to use different feature extraction techniques and explore different features. Some studies utilized manually crafted features using statistics derived from time-domain features [17]. Traditional approaches often employ time-frequency

representations, such as spectrograms [12, 19], mel-frequency cepstrum coefficients (MFCCs) [13, 21], and mel-spectrograms [8], utilizing the Fast Fourier Transform (FFT) to convert signals from the time domain to the frequency domain. However, the FFT faces a fundamental limitation imposed by the uncertainty principle, preventing simultaneous high resolution in time and frequency domains. To address this issue, wavelets can provide effective localization in time and frequency domains. In this paper, we propose features derived from the energy content of wavelet-based time-frequency representations for emotion detection from discrete speech (short speech phrases). We use various forms of discrete wavelets to characterize negative emotions that are commonly associated with mental disorders. These include fear, sadness, anger, anxiety, and disgust emotions within speech signals in non-controlled noise conditions. We evaluate the effectiveness of these features using a CNN-based classifier to discriminate between the embedded emotions in the speech signals. Our experiment focuses on three key emotion classification tasks relevant to mental health monitoring: (i) the detection of positive emotions vs negative emotions (ii) the effectiveness of different discrete wavelets (iii) the recognition of negative emotions across two common public short speech dataset, RAVEDESS [11] and TESS [6]. The results obtained with the proposed approach are compared to those obtained from the traditional FFT-based features. Our research's main contributions are:

(1) Investigating the efficiency of discrete wavelet transform-based features in discriminating negative emotions in discrete speech.
(2) Comparing the effectiveness of FFT-based time-frequency-based speech features with discrete Wavelet Transform-based features.
(3) Analysing the effectiveness of different forms of the discrete wavelet transform in classifying emotions using CNN.

3 MATERIAL AND METHODS

Figure 1 shows the proposed emotion classification architecture. The methodology followed in this study comprises three phases (i) Data acquisition and preprocessing (ii) Time-frequency representation of the speech signal and Discrete Wavelet Computation (iii) Implementation of 1D CNN Model as the classifier.

3.1 Discrete Wavelet Transform

DWT enables a time-frequency multi-resolution analysis by decomposing the signal into variable length frames over time. It offers simultaneous localization in the time and frequency domain while FFT loses the time information of a signal when converting it into the frequency domain [18]. Previous works have shown the effectiveness of DWT in the analysis of physiological signals used in medical diagnostics Electromyography (EMG) and Electrocardiography (ECG) [22]. In this work, we utilize the DWT coefficients as features for detecting emotions from short speech signals. These coefficients carry useful temporal information about the transient activity of the analyzed signal and, thus, could serve as indicators of the correlation level between the analyzed signal and the wavelet function at various time instances [1]. Unlike traditional approaches, our proposed feature takes full advantage of the concurrent time-frequency analysis provided by the DWT.

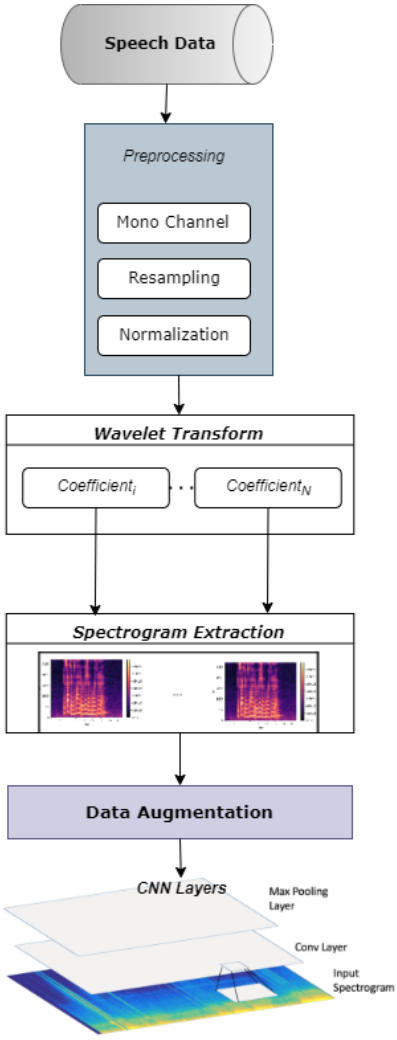

Figure 1: Proposed Emotion Detection Model

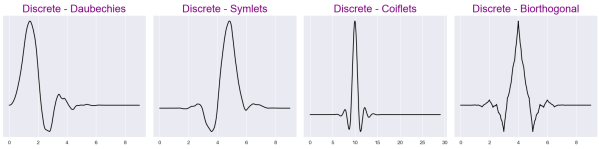

Figure 2: DWT families

The DWT for the normalized speech signal $x[t]$ can be represented in terms of shifted versions of a scaling function $\phi_{j,k}$ and a shifted and dilated version of a so-called mother wavelet function $\psi_{j,k}$. Formally, we define the DWT of a signal $x[t]$ in Equation 1:

$$x[t] = \sum_{k} u_{j_0,k}\phi_{j_0,k}[t] + \sum_{j=-\infty}^{j_0}\sum_{k\in Z} w_{j,k}\psi_{j,k}[t] \quad (1)$$

where $\psi_{j,k}$ represents the wavelet coefficients and $u_{j,k}$ ($j < k$) are the scaling coefficients. Signal decomposition is typically done using scales $a = 2, 4, 8, \ldots, 2^L$ with successive approximations being decomposed in turn, so that it is broken [18]. We analyse our model using different DWT levels ranging from 3-5. Coefficients suitable for different speech emotions are chosen for further analysis since each coefficient represents a particular sub-band of the signal.

3.2 Data Balancing and Augmentation

The emotion classification task provides a good illustration of imbalanced label distribution, given that certain emotion classes, such as disgust, are relatively less compared to more prevalent labels like sadness. As many public speech datasets, such as RAVDESS, have a limited number of samples insufficient for effectively training conventional Neural Network models like CNN that require a larger training set, we implemented a data augmentation technique similar to [23]. An additive white gaussian noise (AWGN) approach [3] was employed to generate additional samples, enhancing the robustness of the proposed model against noise. This allows us to tackle the issues of imbalance and overfitting.

3.3 Wavelet Feature Extraction

Following the formalization of the Discrete Wavelet Transform (DWT) described in Section 3.1, we extract DWT from raw speech signals and spectrograms. To accomplish this, we use the *PyWavelets* library (`pywt`) in Python, employing its functionality for wavelet transform and spectrogram generation. Since the optimal waveform is unknown, we explore various types of wavelet functions. Figure 2 shows the DWT families considered, including Daubechies (`'db4'`) with orders ranging from 1 to 20, Symlets (`'sym'`) with orders 1 to 20, and Coiflets (`'coif'`) with orders 1 to 5. The performance of the different types of wavelet functions would indicate their suitability to detect specific emotions in the speech signals.

3.4 Classification

We construct a CNN model similar to [10] to effectively select emotional features. The model is created using a 1D CNN architecture with fully connected layers. Following each 1D convolutional layer, a batch normalization layer is added to standardize the output and prevent variations in feature distributions between training and test data. The Rectified Linear Unit (ReLU) activation function is applied to induce sparsity in the network, reducing parameter interdependence and mitigating overfitting. The model comprises six 1D convolutional layers, and after the 6th layer, the feature maps are flattened into a single-column matrix to fit into the fully connected layers. Subsequently, the flattened output is passed through two dense layers and two dropout layers. The final layer utilizes the softmax function for classification. The feature map includes informative features and an emotional class label for utterance and loss calculation. In summary, the entire network consists of six 1D convolutional layers and two dense layers with two dropout layers.

4 EXPERIMENTS AND RESULTS

Our experiment focuses on three key emotion classification tasks relevant to mental health monitoring: (i) negative emotion detection in short speeches (ii) evaluating the effectiveness of discrete wavelets in characterizing emotional states(iii) negative emotion recognition in two public speech datasets, RAVEDESS and TESS

4.1 Datasets

We use RAVEDESS and TESS databases for this work as these two datasets share similar emotional classifications and exhibit a diverse representation in terms of age and gender. We also have good representation of negative emotions in each dataset. RAVEDESS [11] is a corpus of actor-based speech, that features recordings from 12 female, 12 male professional actors. RAVDESS comprises 1440 samples that encompass expressions of calm, happiness, sadness, anger, fear, surprise, and disgust. The distribution of emotions includes 192 audio files each for anger, calmness, sadness, happiness, disgust, and surprise, while 96 audio files represent neutral expressions. TESS [6] is a collection featuring 2800 audio files of speech utterances. These recordings are categorized into seven distinct emotions: sad, happy, surprised, angry, disgusted, fearful, and neutral. The actors involved in the dataset belong to both old (64 years old) and young (26 years old) age groups, enhancing the dataset's richness and applicability for research in emotional speech analysis and related fields.

At the beginning of each experiment, data is randomly split into training and testing sets. Specifically, 80% of the samples from each class are allocated for training, while the remaining 20% of samples from each class are designated for testing.

4.2 Evaluation Metrics

To evaluate the effectiveness of the features and performance of the model used, we employ weighted accuracy (WA) and unweighted accuracy (UA). Since WA accounts for the unique characteristics of each emotion class, it provides a more nuanced evaluation basis. We computed WA using Equation 2, as the average accuracy across all samples:

$$WA = \frac{\sum_{k=1}^{K} n_k}{\sum_{k=1}^{K} N_k} \tag{2}$$

where K denotes the number of classes, n_k represents the count of correctly classified samples in class k, and N_k is the total number of samples in class k. UA, expressed through Equation 3, corresponds to the mean accuracy per class:

$$UA = \frac{\sum_{k=1}^{K} \frac{n_k}{N_k}}{K} \tag{3}$$

4.3 Results and Analysis

To validate the effectiveness of the DWT-based features in detecting emotions in short phrase speech, we plugged the feature into the CNN model and compared the accuracy with traditional FFT-based features including MFCC and Mel-Spectrogram. Table 2 shows the comparison of effectiveness (unweighted accuracy in %) of the coefficients of each 4 families of DWT considered in the study. Table 1 shows the mean weighted accuracy (%) performances for each feature. The discrete daubechies at levels 4 and 5 show better results with 81.85% and 79.7% accuracy respectively. Table 3 shows the performance comparison of DWT Coefficients and FFT-based features in the negative emotion detection task. DWT demonstrated

Table 1: Performance Comparison of DWT Coefficients and FFT-based Features in Negative Emotion Detection Task

Features	sad		disgust		fear		anger		neutral	
	WA (%)	**UA (%)**	**WA (%)**	**UA (%)**	**WA (%)**	**UA (%)**	**WA (%)**	**UA (%)**	**WA (%)**	**UA (%)**
MFCC	77.5	77.0	73.7	73.2	72.8	72.6	73.5	71.0	60.2	59.4
FFT-Mel Spectrogram	69.1	68.7	68.7	66.5	69.7	65.5	67.5	66.0	61.3	62.0
DWT Coefficients	81.8	80.5	82	81	83.6	83.5	83.4	83.2	83.7	82.9

The performance comparison of the negative emotion detection solution is reported here using the mean weighted accuracy (WA) and unweighted accuracy (UA)

Table 2: Model Performance for Detecting Negative Emotions Using DWT-based features

Features	RAVDESS	TESS
db4	81.85 ± 6	85.83 ± 3.5
sym4	72 ± 9.0	72 ± 8.0
coif4	65.7 ± 5.6	62. ± 5
db5	79.7 ± 6	77.3 ± 5
sym5	77 ± 4.0	76 ± 5.6
coif5	65.7 ± 5.6	62. ± 5

better performance in recognising negative emotions, particularly, sadness, disgust, fear and anger, which are associated with mental health compared to the state-of-the-art FFT-based features.

Table 3: Discrete Wavelet-Based vs FFT-based Features

Features	RAVDESS	TESS
MFCC	77 ± 5.2	74 ± 5.0
FFT-Mel Spectorgram	68.85 ± 6	69.43 ± 5
DWT Coefficients	81.85 ± 6	85.83 ± 3.5

5 CONCLUSION AND FUTURE WORK

This work evaluates the effectiveness of discrete wavelet-based features in detecting negative emotions associated with mental health disorders using 1D CNN architecture for classification. The experiments are conducted on two different short speech datasets, RAVDESS and TESS, revealing the ability of discrete wavelet transform features to distinguish between various negative emotions. Experimental results show that the specific DWT family (db4 and Symlets) efficiently discriminate between positive and negative emotions across the two databases. Compared to the Traditional FFT features such as Mel-Spectrogram and MFCC, DWT feature demonstrated significant performance improvement. One of the challenges of this work is the curse of dimensionality of wavelet features. In future work, we aim to extend our model by incorporating a layer for dimensionality reductions and experiment with more robust and extensive models.

REFERENCES

[1] Paul S Addison. 2017. *The illustrated wavelet transform handbook: introductory theory and applications in science, engineering, medicine and finance.* CRC press.

[2] Mehmet Berkehan Akçay and Kaya Oğuz. 2020. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication* 116 (2020), 56–76.

[3] Xu Dong An and Zhou Ruan. 2021. Speech Emotion Recognition algorithm based on deep learning algorithm fusion of temporal and spatial features. *Journal of Physics: Conference Series* 1861, 1 (mar 2021), 012064. https://doi.org/10.1088/1742-6596/1861/1/012064

[4] Surekha Reddy Bandela and T Kishore Kumar. 2017. Stressed speech emotion recognition using feature fusion of teager energy operator and MFCC. In *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. IEEE, 1–5.

[5] Fani Deligianni, Yao Guo, and Guang-Zhong Yang. 2019. From emotions to mood disorders: A survey on gait analysis methodology. *IEEE journal of biomedical and health informatics* 23, 6 (2019), 2302–2316.

[6] Kate Dupuis and M Kathleen Pichora-Fuller. 2010. Toronto emotional speech set (tess)-younger talker_happy. (2010).

[7] S Lalitha, Deepa Gupta, Mohammed Zakariah, and Yousef Ajami Alotaibi. 2021. Mental Illness Disorder Diagnosis Using Emotion Variation Detection from Continuous English Speech. *Computers, Materials & Continua* 69, 3 (2021).

[8] Margaret Lech and Ling He. 2013. Stress and emotion recognition using acoustic speech analysis. In *Mental Health Informatics*. Springer, 163–184.

[9] Kapang Legoh, T Tuithung, and U Bhattacharjee. 2015. Features and model adaptation techniques for robust speech recognition: a review. *Communications on Applied Electronics (CAE). New York, USA: Foundation of Computer Science FCS* (2015), 18–31.

[10] Yulan Li, Charlesetta Baidoo, Ting Cai, and Goodlet A Kusi. 2019. Speech emotion recognition using 1d cnn with no attention. In *2019 23rd international computer science and engineering conference (ICSEC)*. IEEE, 351–356.

[11] Steven R Livingstone and Frank A Russo. 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PloS one* 13, 5 (2018), e0196391.

[12] Samaneh Madanian, Talen Chen, Olayinka Adeleye, John Michael Templeton, Christian Poellabauer, Dave Parry, and Sandra L. Schneider. 2023. Speech emotion recognition using machine learning — A systematic review. *Intelligent Systems with Applications* 20 (2023), 200266. https://doi.org/10.1016/j.iswa.2023.200266

[13] Samaneh Madanian, David Parry, Olayinka Adeleye, Christian Poellabauer, Farhaan Mirza, Shilpa Mathew, and Sandy Schneider. 2022. Automatic Speech Emotion Recognition Using Machine Learning: Digital Transformation of Mental Health. (2022).

[14] Nivedhitha Mahendran, PM Durai Raj Vincent, Kathiravan Srinivasan, Vishal Sharma, and DushanthaNalin K Jayakody. 2020. Realizing a stacking generalization model to improve the prediction accuracy of major depressive disorder in adults. *IEEE Access* 8 (2020), 49509–49522.

[15] Sherri Melrose et al. 2015. Seasonal affective disorder: an overview of assessment and treatment approaches. *Depression research and treatment* 2015 (2015).

[16] Hemanta Kumar Palo and Mihir Narayan Mohanty. 2018. Wavelet based feature combination for recognition of emotions. *Ain shams engineering journal* 9, 4 (2018), 1799–1806.

[17] A Pramod Reddy and V Vijayarajan. 2017. Extraction of emotions from speech-a survey. *International Journal of Applied Engineering Research* 12, 16 (2017), 5760–5767.

[18] Celia Shahnaz, SM Shafiul Hasan, et al. 2016. Emotion recognition based on wavelet analysis of Empirical Mode Decomposed EEG signals responsive to music videos. In *2016 IEEE Region 10 Conference (TENCON)*. IEEE, 424–427.

[19] Stanley Smith Stevens, John Volkmann, and Edwin Broomell Newman. 1937. A scale for the measurement of the psychological magnitude pitch. *The journal of the acoustical society of america* 8, 3 (1937), 185–190.

[20] Madhukar H Trivedi. 2004. The link between depression and physical symptoms. *Primary care companion to the Journal of clinical psychiatry* 6, suppl 1 (2004), 12.

[21] Yue Xie, Ruiyu Liang, Zhenlin Liang, Chengwei Huang, Cairong Zou, and Björn Schuller. 2019. Speech emotion classification using attention-based LSTM. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27, 11 (2019), 1675–1685.

[22] Rendi EJ Yohanes, Wee Ser, and Guang-bin Huang. 2012. Discrete wavelet transform coefficients for emotion recognition from EEG signals. In *2012 annual*

international conference of the IEEE engineering in medicine and biology society. IEEE, 2251–2254.

[23] X Zhu, Y Liu, Z Qin, and J Li. [n. d.]. Data augmentation in emotion classification using generative adversarial networks. arXiv 2017. *arXiv preprint arXiv:1711.00648* ([n. d.]).