SMM19, Workshop on Speech, Music and Mind 2019
14 September 2019, Vienna, Austria

A Successive Difference Feature for Detecting Emotional Valence from Speech

Gauri Deshpande, Venkata Subramanian Viraraghavan, Rahul Gavas

TCS Research and Innovation, Tata Consultancy Services Limited

gauri1.d@tcs.com, venkatasubramanian.v@tcs.com, rahul.gavas@tcs.com

Abstract

Many features have been proposed for detecting emotions from speech. Their detection performance is influenced by the change in contextual parameters such as background noise, speaker variability, expressions, demographics and so on. In this paper, we use a recent, time-domain feature extraction technique for detecting emotional-valence. We report the performance of the time-domain features on data in three different contexts: the Fearless-Steps challenge data for Sentiment Detection with spontaneous emotions, an Indian-demography corpus with induced emotions, and the Berlin database of acted emotions (EmoDB). Data is pre-processed according to the environment they are captured in, but the feature extraction that follows is identical. With these features, a RandomForest Classifier yields an accuracy of 71% on the development set of the challenge data, and 74% on the evaluation set, a significant improvement on the published baseline result of 49%. The same features provide an accuracy of 75% on the Indian-demography corpus and 100% accuracy in classifying happy, sad and neutral emotions from EmoDB, again with RandomForest Classifier. These results are better than those obtained with other prevalent techniques such as Long Short Term Memory (LSTM) with spectrograms, and the RandomForest classifier with the widely accepted features, OpenSMILE and Mel-Frequency Cepstral Coefficients (MFCCs).

Index Terms: Emotion Detection, Time-domain features, Fearless Steps challenge, Sentiment Detection

1. Introduction

A broad spectrum of features derived from audio samples including pitch, energy, Mel Frequency Cepstral Coefficients (MFCCs), Linear predictive cepstral coefficients, log frequency power coefficients, spectrograms [1], raw waveform data [2] and so on have been used to improve the accuracy of speech-to-emotion applications. As mentioned in [3] and [4], valence-scale is known to be the most challenging continuous dimension for detecting emotions from speech acoustics. Despite the array of features and classifiers, detecting valence from speech alone remains a challenge. Although, other modalities such as text and physiology are also being explored along with speech for this task, they have other limitations such as obtrusiveness, need for an extra device, difficulty in capturing data and so on. As detailed by the authors of [5], the role of speech as a key modality for evaluating emotional expression, particularly valence is also established. Apart from being relatively less intrusive, it enables capturing and listening to get emotion-annotation at fine granularity [6].

1.1. Previous Work

At Interspeech 2018 [7], the research on emotion detection from speech was dominated by the analysis using deep neural networks, specifically LSTMs [8], [9] and Bi-directional LSTMs [10], [11], and the dominant feature sets were MFCCs, OpenSMILE and raw data. With both LSTMs and Bi-LSTMs, a banal observation is, arousal is detected better than valence. The authors of [12] have compared hand-crafted features with the features learned by recurrent neural networks and stated that there is no clear winner between the two; which one performs better depends on the data-set. Thus, the need for better features representing emotional speech, and for the most suitable classification schemes, is still unfulfilled. Also, an in-depth analysis of the nature of data is essential to understand and thus model the emotions in it.

In order to understand the data better, one of the significant stages is the process of labeling emotions to an utterance of speech, since the later processing stages heavily depend on it. In our earlier work [13], we found an improvement of 7% in the classification accuracy with utterance-level annotation, than on sentence-level annotation. This is the motivation to analyze the duration of each speech segment in the challenge data-set.

The authors of [14], have studied the effect of acoustic properties on valence and have found fundamental frequency (F0) to be one of the most discriminating features in predicting valence. Another study conducted by [15], found that the mean of the second formant was higher for sentences with positive valence than for sentences with negative valence. Hence, the features proposed in this paper are inspired by the importance of F0 and formants of the speech signal.

1.2. Our Contribution

In this paper, we use a recent time-domain feature, termed as Successive Difference feature, on the Fearless steps challenge data-set for sentiment detection. While the feature was introduced in our earlier work [16], its analysis is presented in Section 2.1. The steps followed in the data pre-processing stage to analyze the challenge data are explained in Section 3. The comparative study of our features with MFCCs and OpenSMILE [16] is augmented by a comparison with LSTMs in Section 4.

2. Successive Difference feature analysis

For each 20 ms of a speech segment, 14 features are extracted, comprising normalized Root Mean Square Energy (RMSE), time-domain auto-correlation, 10 successive differences and the mean and range of the 10 differences. The details of the time-domain auto-correlation feature are given in [13] and the other features are explained next.

2.1. Spectral Analysis of Successive Differences

Consider a time-domain (speech) signal sampled at 8 kHz, or the signal re-sampled to 8 kHz. Let a frame of this signal, after being multiplied by the Hamming window $h[\cdot]$, be denoted by $s[n]$ at $n/8000$ seconds, where n is the sample number, and $0 \leq n < 2N$. The value of N depends on the window size in the application. For the emotion detection task, we chose 160

10.21437/SMM.2019-8

(20 ms) according to previous work [17]. The even signal of this frame is calculated as:

$$x_0[n] = s[2n], 0 \leq n < N \quad (1)$$

where the subscript 0 is used for consistency with (2). Retaining even samples is equivalent to working with the time-domain signal of the even part of the spectrum. To avoid losing information, the odd samples can also be considered, but we found little difference in classification performance when using even, odd or an interleaved combination of the difference feature. In this paper, we work with the even samples $x_0[n]$ only. We define for iteration $m \geq 1$:

$$x_m[n] = \frac{x_{m-1}[n+1] - x_{m-1}[n]}{2}, 0 \leq n < N - m \quad (2)$$

Note that (2) is a linear operation and does not involve cyclic extension of the frame. We first note that, from among $x_m[n], m = \{60, 70, 80\}$, $x_{70}[n]$ is empirically found to be the most useful feature vector for emotional-valence detection.

Let $X_m(e^{j\omega})$ be the Discrete-Time Fourier Transform (DTFT) of $x_m[n]$. We consider the result of applying (2) M times. In the ideal case, when $N - M$ is large, [1] the usual relations hold:

$$\begin{aligned} X_m(e^{j\omega}) &= (e^{j\omega} - 1)X_{m-1}(e^{j\omega})/2 \\ &= je^{-j\omega/2}\sin(\omega/2)X_{m-1}(e^{j\omega}) \quad (3) \\ |X_M(e^{j\omega})| &= \sin^M(\omega/2)|X_0(e^{j\omega})| \quad (4) \end{aligned}$$

The effect of (4) is to enhance the component at half the sampling rate of $x_0[n]$, i.e. at 2 kHz, and attenuate other components. It would seem that the content at $f_a = 2$ kHz is important, but when a sharp filter with a notch at f_a was applied on $s[n]$, the classification results were unchanged. Thus, it is not the content at 2 kHz in the signal that matters; instead it is the content that *manifests* at 2 kHz due to taking successive differences. Note that from (4), $x_{70}[n]$ cannot have a DC component. Thus, it is not surprising that the energy of the frame is needed in addition as a feature for classification.

Another empirical observation is that if (2) is applied on $x_0[n]$ cyclically extended to length $2N$, and $x_M[n], N - M \leq 2N - M$ is chosen as the feature vector, the classification performance reduces by nearly 10%. That is, (4) performs worse than its approximation. This approximation is due to the finite length of the frame (i.e. $N - M$ is not large) and is shown for the first step of the iteration below.

$$\begin{aligned} X_1(e^{j\omega}) &= \sum_{n=0}^{N-2} x_1[n]e^{-j\omega n} \quad (5) \\ &= \sum_{n=0}^{N-2} \frac{x_0[n+1] - x_0[n]}{2} e^{-j\omega n} \quad (6) \\ &= je^{-j\omega/2}\sin(\omega/2)X_0(e^{-j\omega}) - R_1(e^{j\omega}) \quad (7) \end{aligned}$$

where the residual term $R_1(\omega)$ is defined as:

$$R_1(e^{j\omega}) = \frac{x_0[0]e^{j\omega} + x_0[N-1]e^{-j\omega(N-1)}}{2} \quad (8)$$

[1] Equation 3 is similar to the effect of cascaded feedback and accumulation stages on the power spectrum of noise in over-sampled analog-to-digital conversion with noise shaping [18], where $M = 2p$ is even.

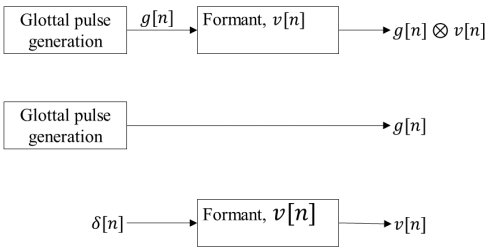

Figure 1: *Simplified diagram of speech production, with the glottal pulse and the formant.*

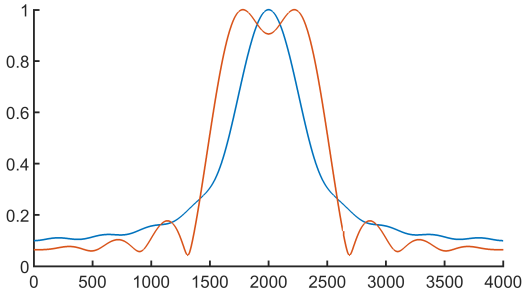

Figure 2: *Spectra of $x_{70}[n]$ for a pitch of 160 Hz and the vowel 'IH'. The x-axis is in Hz and the y-axis is the normalized magnitude. The blue curve corresponds to the glottal pulse alone and the red curve, to the formant alone.*

The residual term appears to play an important part in the classification accuracy. Unfortunately, it becomes more unwieldy with each iteration. It is intractable, and perhaps not even useful, to enumerate the terms for 70 iterations. We therefore analyze artificially generated speech signals and understand the effect of (2) on them. Figure 1 shows a simplified diagram of the essentials of voiced speech production (excluding nasals). The main points of interest in this analysis are the glottal pulse (GP) and the vocal tract, which introduces the formants. We investigate the effect of the GP and formants separately; these paths are also shown in the same figure.

Let the pitch (equated to fundamental frequency in this paper) of a train of GPs be $f_0 = \frac{\omega_0}{2}$. Their spectra may be approximated as delta functions as $G(e^{j\omega}) = \sum_k c_k\delta(\omega - k\omega_0)$, where c_k is the complex coefficient of the k^{th} harmonic of ω_0. The convolution of $G(e^{j\omega})$ with the spectrum of the window function, $H(e^{j\omega})$ results in the spectrum:

$$X_g(e^{j\omega}) = \sum_k c_k H(e^{j(\omega - k\omega_0)}) \quad (9)$$

At $\omega = \frac{2\pi}{2000}$ radians, the effect of the k^{th} harmonic is seen in the *sidelobes* of $H(e^{j\omega})$. When successive differences are found, the effect is magnified exponentially. The formants that convolve with $X_g(e^{j\omega})$ appear to also peak around 2 kHz, but have a different shape.

The precise mechanism for the increased performance is not yet clear, but we suggest an explanation that is corroborated by the observations given above. The pitch of the voice is one of the most important features of emotion recognition [14]. Typical pitch tracking algorithms show several types of errors, especially octave errors. It is possible that such errors make the detected pitch a relatively less reliable feature. Note

that they all work on the low-frequency components of a frame. Similarly, another important cue for valence is the text content [19], which is partly captured by the formants. By finding successive differences, the contribution of the pitch and its components to the information at 2 kHz is enhanced. The contribution of the formant is also enhanced at 2 kHz. However, the two enhancements are slightly different from each other in terms of the shape of the spectrum. Typically, we expect broader spectra for the formant after convolution with $H(e^{j\omega})$ than for the GP. For typical pitch values in human speech, the magnitude of the side-lobes at $f_{\max} \approx 2$ kHz seems to help the amplification in (4). Due to the (aliased) downsampling in (1), the sampling rate needs to be $4f_{\max}$. The closest standard sampling rate for speech signals is 8 kHz.

Although we don't estimate the pitch and the formants from the difference signal ($x_{70}[n]$), machine learning algorithms are able to exploit this information. Interestingly, a previous approach based on wavelets [20] also found that successive high pass components contain valence information. However, interpretability was not discussed in that work. Thus, in summary, it does appear that the use of high-pass filters improves emotional valence detection.

3. Data Pre-processing

In this paper we test the performance of the difference feature using the three data-sets, namely, Fearless Steps challenge data for sentiment detection [21]. For completeness, we also compare the performance on the Indian-demography corpus [13] and EmoDB [22]. The three data-sets are described next.

3.1. Fearless Steps Challenge - Sentiment Detection Dataset

The data-set comprises 116 clips in a train-set, 39 clips in a development (dev) set, and 40 clips in an evaluation (eval) set. Using the transcripts given with the challenge data, the train-set is segmented into 1327 positive, 685 negative, and 1724 neutral segments. There are 1912 positive, 492 negative and 4616 neutral segments in the dev-set. The transcripts are not available for the eval-set, hence they are segmented as explained in Section 3.1.3. Clearly, the three classes are not balanced, with the negative class having the least representation. We thus select a subset of positive and neutral speech segments to form a balanced training set (Section 3.1.2). Also, the model for detecting emotions in the eval-set is formed using the speech segments in the balanced train- and dev-sets. Starting with the analysis of the context in which the data has been captured, we follow the below steps on the challenge data.

3.1.1. Noise Reduction

On examining a handful of the speech files, it was found that the noise consists of a 60 Hz tone and its harmonics. This hum is removed by using a cascade of Infinite Impulse Response notch filters, each with the notch at a harmonic of 60 Hz. All filters have a 3-dB bandwidth of 2 Hz. Twenty five harmonics were removed in this manner.

3.1.2. Duration analysis

In a previous experiment [13], we found that the short clips of 3 seconds and shorter, carry more of emotion specific information. Thus, with a threshold of three seconds, we divide the train-set into two sets with short (SH) and long (LG) segments. The count for each class is given in Table 1. As the table shows, the negative class has the lowest number of segments, 15.8% and 20.3% for Train SH and Train LG respectively. It is therefore intuitive to select a subset of positive and neutral classes, to ensure class level balance. The short train segments have been used to train the models and long train segments are discarded to balance the three class instances. Accuracy is calculated for both short and long dev and eval segments.

3.1.3. Detecting voice segments

The evaluation set provided did not have associated speech segmentation information. The RMSE is normalized by the maximum across all frames, and only the frames whose normalized RMSE value exceeds $0.025(2.5\%)$ are considered for predicting the emotion. The same condition is applied for dev-set as well for getting the prediction of each frame of the defined segments. Further, for the eval-set, a silence duration of 0.5 seconds or longer is considered as a separator of two speech segments. Any speech segment with a duration exceeding 3 seconds, is divided into sub-segments of 3 seconds or shorter. The emotion-valence values of the sub-segments are predicted.

Table 1: *Segment count for each class*

Class	SH Train	LG Train	SH Dev	LG Dev
Positive	650	677	1753	159
Negative	227	458	332	160
Neutral	557	1167	3887	759
Total	1434	2252	5972	1078

3.2. Non-acted Indian Database

The Non-acted Indian (NaI) database consists of emotional utterances from 33 speakers. It is collected by inducing the valence scale emotions, positive affect, negative affect and neutral into the participants. They were shown an audio-visual clip following which they had to verbally answer the automated questions asked by the computer. The design of this experiment is explained in detail in [17]. As a second stage of this experiment, the sentences captured were manually segmented into utterances. The participants themselves were asked to label their own utterance to capture self-annotation and also, a researcher had annotated all the utterances to capture perceptual-annotation. In this experiment, the labels were given with respect to a continuous valence scale so that the intensity of emotions also gets captured. Further details of this experiment are explained in [13]. To report the accuracy in this paper, we have considered the speaker-independent set of this database where, the train set consists of 17 speakers and the test set consists of 10 other speakers.

3.3. EmoDB

It is an acted German database [22], consisting of 10 speakers. To analyze the speaker independent accuracy, we partitioned the train and test-sets such that 4 speakers speech is present in train-set and the rest of the 6 speakers speech are in test-set. Since the focus is on classifying valence scale based emotions only, we considered only happy, sad and neutral classes of EmoDB.

4. Methods and Results

4.1. Classification by Regression

We use the RandomForest Regressor to generate a regression model of the train-set. The model is tested on the train-set itself to get a range of values on the continuous valence scale. The labels for positive, negative and neutral classes are given as 1, -1 and 0 respectively. Hence, the predicted values lie between -1 to +1. The threshold values are derived from these predictions, to separate out 1 (positive) from 0 (neutral) called as positive threshold (PTHR) and -1 (negative) from 0 (neutral) called as negative threshold (NTHR). The same threshold values are then used to classify the regression predictions of dev-set. The thresholds for both EmoDB and NaI database are +0.2 (PTHR) and -0.2 (NTHR). For the challenge data, they are +0.05 (PTHR) and -0.2 (NTHR). These threshold values are learned from dev-set and are applied on the eval-set.

4.2. Analysis of Frame- and Segment-level Predictions

Unlike frame-level labels, the frame-level predictions are not the same for a given speech segment. We devised two methods of interpreting the frame-level predictions to make a segment-level prediction. In Method 1, the average value of all frame-level predictions is used. In Method 2, number of frames above PTHR ($cntP$), number of frames below NTHR ($cntNg$), and for rest of the frames ($cntNu$) are counted. The segment-level prediction is obtained according to Algorithm 1, where $FrCnt$ is the number of 20 ms frames in the speech segment.

Algorithm 1 Frame level predictions for each speech segment

Input: Count values: $cntP$, $cntNg$, $cntNu$, $FrCnt$
Output: Prediction
Procedure:
if (($cntP > cntNg$) and ($cntNu < 0.5 \times FrCnt$)) **then**
    Prediction = Positive
else if ($cntNg >$cntNu) **then**
    Prediction = Negative
else
    Prediction = Neutral

4.3. Classification Results using Difference Features

The performance of the proposed features on all the three data-sets obtained with Method 2 is shown in Figure 3. The proposed features can classify negative class with 100% accuracy, but there is significant confusion in distinguishing positive from neutral. However, two intuitive patterns can be observed. In all the negative utterances, the number of negative frames are more than that of neutral. Also, over 80% of neutral utterances have atleast 50% of the total frames as neutral.

5. Discussion

While comparing the two methods explained in Section 4.2, Method 2 performed the same as that of Method 1 on NaI and EmoDB data-sets, giving an accuracy of 75% and 100% respectively. However, it gave an improvement of almost 25% on the dev-set of the challenge data, giving an accuracy of 71%. Further, it yielded an improvement of 24% on the eval-set of challenge data, giving a score of 74% [2]. This suggests that a short segment of 3 sec along with the long segments carries a combination of positive-neutral and negative-neutral frames. This also reinforces the regression based approach where positive, neutral and negative emotions are thought of lying on the valence scale, and neutral being at the center of positive and negative emotions. With both the methods, there is a significant confusion between positive and neutral classes. The accuracy of LSTMs with spectrograms, and Random Forest (RF) classifier with OpenSMILE and MFCCs features for all the three data-sets are shown in Figure 4. The proposed features perform better with respect to these state of the art methods, giving highest accuracy of 100% on EmoDB. Though LSTMs with the spectrograms also result in 100% accuracy for acted emotions (EmoDB), they did not perform well for spontaneous emotions.

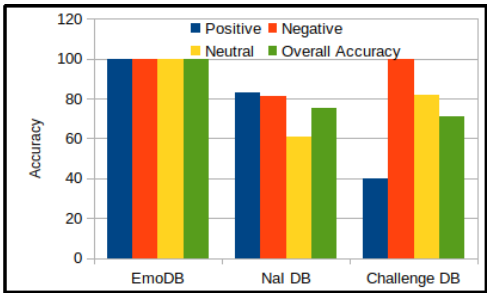

Figure 3: *Proposed feature results on the three data sets*

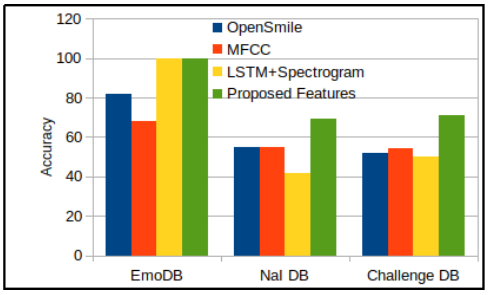

Figure 4: *Comparison of proposed features with other methods*

6. Conclusion and Future Scope

With the advent of multiple conversational agents, there is an increasing need of detecting emotions from spontaneous real time speech. As a step in this direction, the present work aims at classifying speech emotions using a recent time-domain difference feature. The time-frequency domain related study of the proposed feature suggests a possible mechanism. A comparative analysis of the proposed feature against existing standard methods on three different datasets shows their effectiveness in classifying three valence scale emotions. These results also demonstrate that the feature can handle contextual variability across the three datasets. Further improvement can be expected if the discrimination between positive and neutral classes is increased. The study can also be extended to other emotions and fusion with more sensing modalities, such as image and text.

[2] The challenge organizers report this accuracy and the baseline accuracy (we call it 'score' to avoid confusion) by taking the segment lengths into account, while we calculate accuracy irrespective of segment lengths. See `https://exploreapollo-audiodata.s3.amazonaws.com/fearless_steps_challenge_2019/v1.0/Fearless_Step_Evaluation_Plan_v1.2.pdf` for details of how the score is evaluated.

7. References

[1] Z. Yang and J. Hirschberg, "Predicting arousal and valence from waveforms and spectrograms using deep neural networks," *Proc. Interspeech 2018*, pp. 3092–3096, 2018.

[2] R. Lotfidereshgi and P. Gournay, "Biologically inspired speech emotion recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5135–5139.

[3] S. Wu, T. H. Falk, and W.-Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech communication*, vol. 53, no. 5, pp. 768–785, 2011.

[4] F. Eyben, M. Wöllmer, A. Graves, B. Schuller, E. Douglas-Cowie, and R. Cowie, "On-line emotion recognition in a 3-d activation-valence-time continuum using acoustic and linguistic cues," *Journal on Multimodal User Interfaces*, vol. 3, no. 1-2, pp. 7–19, 2010.

[5] S. S. Narayanan, "12 speech in affective computing," *The Oxford Handbook of Affective Computing*, p. 170, 2015.

[6] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, and M. Pantic, "Avec 2017: Real-life depression, and affect recognition workshop and challenge," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 2017, pp. 3–9.

[7] *INTERSPEECH 2018 – 19th Annual Conference of the International Speech Communication Association, September 2-6, Hyderabad, India, Proceedings*, 2018.

[8] C. Gorrostieta, R. Brutti, K. Taylor, A. Shapiro, J. Moran, A. Azarbayejani, and J. Kane, "Attention-based sequence classification for affect detection," *Proc. Interspeech 2018*, pp. 506–510, 2018.

[9] H. Kaya, D. Fedotov, A. Yeşilkanat, O. Verkholyak, Y. Zhang, and A. Karpov, "LSTM based cross-corpus and cross-task acoustic emotion recognition," *Proc. Interspeech 2018*, pp. 521–525, 2018.

[10] S. Rallabandi, B. Karki, C. Viegas, E. Nyberg, and A. W. Black, "Investigating utterance level representations for detecting intent from acoustics," *Proc. Interspeech 2018*, pp. 516–520, 2018.

[11] E. Tzinis, G. Paraskevopoulos, C. Baziotis, and A. Potamianos, "Integrating recurrence dynamics for speech emotion recognition," *arXiv preprint arXiv:1811.04133*, 2018.

[12] J. Wagner, D. Schiller, A. Seiderer, and E. André, "Deep learning in paralinguistic recognition tasks: Are hand-crafted features still relevant?" *Proc. Interspeech 2018*, pp. 147–151, 2018.

[13] G. Deshpande, V. S. Viraraghavan, M. Duggirala, V. R. Reddy, and S. Patel, "Comparing manual and machine annotations of emotions in non-acted speech," *Engineering in Medicine and Biology*, 2018.

[14] C. Busso and T. Rahman, "Unveiling the acoustic properties that describe the valence dimension," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[15] M. Goudbeek, J. P. Goldman, and K. R. Scherer, "Emotion dimensions and formant position," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[16] G. Deshpande, V. S. Viraraghavan, M. Duggirala, and S. Patel, "Detecting emotional valence using time-domain analysis of speech signals," *Engineering in Medicine and Biology*, 2019.

[17] G. Deshpande, V. S. Viraraghavan, M. Duggirala, V. R. Reddy, and S. Patel, "Empirical evaluation of emotion classification accuracy for non-acted speech," in *2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2017, pp. 1–6.

[18] A. V. Oppenheim, J. R. Buck, and R. W. Schafer, *Discrete-time signal processing. Vol. 2*. Upper Saddle River, NJ: Prentice Hall, 2001.

[19] C. Montacié and M.-J. Caraty, "Vocalic, lexical and prosodic cues for the interspeech 2018 self-assessed affect challenge," *Proc. Interspeech 2018*, pp. 541–545, 2018.

[20] S. Deb and S. Dandapat, "Multiscale amplitude feature and significance of enhanced vocal tract information for emotion classification," *IEEE transactions on cybernetics*, no. 99, pp. 1–14, 2018.

[21] J. H. Hansen, A. Sangwan, A. Joglekar, A. E. Bulut, L. Kaushik, and C. Yu, "Fearless steps: Apollo-11 corpus advancements for speech technologies from earth to the moon." in *Interspeech*, 2018, pp. 2758–2762.

[22] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech." in *Interspeech*, vol. 5, 2005, pp. 1517–1520.