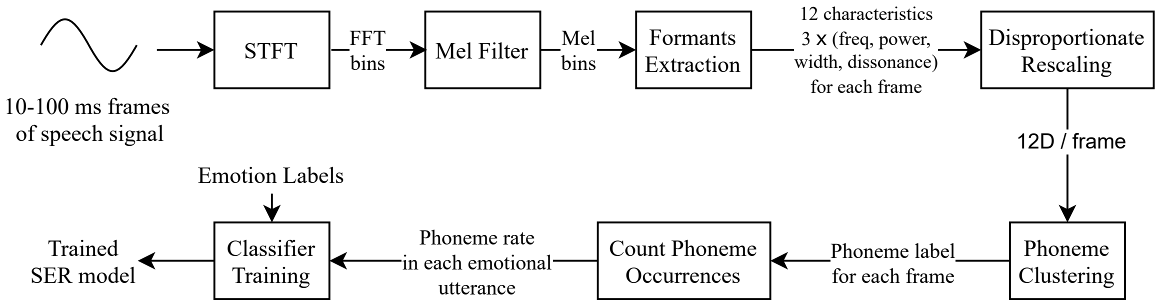
Graphical Abstract

Speech Emotion Recognition Based on Formant Characteristics Feature Extraction and Phoneme Type Convergence

Zhen-Tao Liu,Abdul Rehman,Min Wu,Wei-Hua Cao,Man Hao



Highlights

Speech Emotion Recognition Based on Formant Characteristics Feature Extraction and Phoneme Type Convergence

Zhen-Tao Liu,Abdul Rehman,Min Wu,Wei-Hua Cao,Man Hao

- Emotions are recognized using the occurrences of auto detected phonological units.
- Phonemes are clustered together based on the similarity of formant characteristics.
- Experiment results indicate reduced computational cost and increased robustness.

Speech Emotion Recognition Based on Formant Characteristics Feature Extraction and Phoneme Type Convergence

Zhen-Tao Liu [a,b,c,*], Abdul Rehman[a,b,c], Min Wu [a,b,c], Wei-Hua Cao [a,b,c] and Man Hao [a,b,c]

[a] *School of Automation, China University of Geosciences, Wuhan 430074, China*
[b] *Hubei Key Laboratory of Advanced Control and Intelligent Automation for Complex Systems, Wuhan 430074, China*
[c] *Engineering Research Center of Intelligent Technology for Geo-Exploration, Ministry of Education, Wuhan 430074, China*

ARTICLE INFO

Keywords:
speech
emotion recognition
formants extraction
phonemes
clustering
cross-corpus

ABSTRACT

Speech Emotion Recognition (SER) has its applications in numerous novel domains particularly in human-robot interaction, online gaming, and health care assistance. The deep learning-based approaches achieve considerable precision but come with high computational and time costs. This is because feature learning strategies have to search for important features in a large amount of speech data. With an objective to reduce the time and computational costs, we propose an alternative solution that focuses on major formant characteristics of speech signal while ignoring the insignificant signal nuances. Speech segments with similar formant characteristics are clustered together and labeled as the same phoneme. Then the phoneme occurrence rates in emotional utterances are used as the input features for classifier. This pre-processing step reduces the input dimensions for the classifier and thus reduces the overall training time without any significant loss of accuracy. The proposed technique is evaluated on 6 databases, i.e., EmoDB, RAVDESS, IEMOCAP, ShEMO, DEMoS and MSP-Improv. The experiment results show that the accuracy of our proposed technique is comparable to that of current state-of-the-art methods while significantly reducing the required training time from hours to minutes.

1. Introduction

Speech carries many kinds of cognitive signals that convey important affective information between humans. Although computers are getting better at understanding human communication, yet some affective signals are unintelligible to computers. One of the reasons is that speech can be colored by different emotions that have obscure differences in discrete speech signals. Differentiating one emotion from another is not easy especially when even humans can't validate them with certainty. The adjectives that define emotions can vary by meaning and intensity for different humans due to which there is always uncertainty in emotional annotations as well as their respective speech signals. This makes speech emotion recognition an interesting research area which has recently gained increased attention in Human-Computer Interaction (HCI) field.

An SER system takes speech signal as input (usually few seconds) and predicts the type and/or intensity of emotion being conveyed in that speech signal. This is usually achieved by first extracting some useful features (e.g., pitch) from the speech signal and then mapping these features on to an emotional construct using a machine learning classifier or a neural network. The training of a machine learning classifier or a neural network is performed by using few hundreds or thousands of annotated examples of speech signal. Therefore, the availability of speech emotional corpora is a prerequisite for all SER systems. Emotions could be either annotated on a continuous scale between positive and negative or divided up in 4 to 12 classes. If considering the gender and intensity, unique emotional labels can be up to 24. The higher the number of classes is, the higher the chances of overlapping definitions of emotional labels are, which increases the level of complexity for classifiers.

This work was supported in part by the National Natural Science Foundation of China under Grant 61976197, 61403422 and 61273102, in part by the Hubei Provincial Natural Science Foundation of China under Grant 2018CFB447 and 2015CFA010, in part by the Wuhan Science and Technology Project under Grant 2017010201010133 and 2020010601012175, in part by the 111 Project under Grant B17040, and in part by the Fundamental Research Funds for National University, China University of Geosciences, Wuhan, under Grant 1910491T01.

Preprint submitted to Information Sciences on February 17, 2020. Revised on September 13, 2020.

[*] Corresponding author

liuzhentao@cug.edu.cn (Zhen-Tao Liu); abdulrehman@cug.edu.cn (Abdul Rehman); wumin@cug.edu.cn (Min Wu); weihuacao@cug.edu.cn (Wei-Hua Cao); haoman@cug.edu.cn (Man Hao)

ORCID(s): 0000-0003-4100-9557 (Zhen-Tao Liu); 0000-0003-2345-2256 (Abdul Rehman); 0000-0002-0668-8315 (Min Wu); 0000-0002-9677-9586 (Wei-Hua Cao)

Major challenges reported by the related works are robustness and autonomy, i.e., an SER system needs to adapt to different speakers. But due to constraints in feature definitions, one model that works well for one group of speakers, performs poorly for another group. The training of spectrogram based deep learning models can take 2 to 14 days for few hours of training data. Then if a new training data arrives (in case of active learning or speaker dependent learning), then model needs to be retrained again which can again take few hours. A part of the reason for large size of required input is the importance of modulation along the temporal axis of features. Features such as Mel Frequency Cepstral Coefficients (MFCCs) are difficult to further condense into fewer variables because each contextual frame has its individual importance, therefore MFCCs or Mel-filter energies of all frames need to be considered as individual inputs.

To solve the above problems, we propose a model that can extract relevant features quickly by focusing the attention on formant characteristics of speech signals while ignoring the rest of the input. As a result, training or retraining a SER model by our method takes only few minutes. This is achieved by a phoneme type convergence method that takes 8-48 kHz audio input and converges it into a label of phoneme type for each contextual frame of 10-50 ms by K-means clustering. We aim to narrow down the distinguishable features to only a few components such that a classifier achieves a sensitivity for few specific phonemes. This method helps decrease the computational cost since the recognition model will only have to look for specific components instead of the whole spectra, and it also helps to avoid blindsided over-fitting of the model on unimportant features such as minor formants created by noise or silent regions. The proposed method focuses on the preprocessing of speech signal to extract distinct unlabeled phonemes. All contextual frames of speech data with similar values along the 12 dimensions (i.e., frequency, power, width, and dissonance of three major formants of contextual frame) are clustered together and labeled as the same phoneme. Then occurrence rates of phonemes that occur at different frequencies in different emotional classes are selected as input features for classifier. This whole process takes 1-25 minutes of training and matches the accuracy of state-of-the-art methods.

Our approach to speech emotion recognition (SER) is based on the detection of unlabeled phonemes that occur more frequently in expression of certain emotions than others. The generally accepted theory of linguistic arbitrariness rejects the notion that vocal expression of words is driven from psychobiological mechanisms such as emotions, because if that was the case then most languages would have the similar sounding words for emotional symbols [1, 2]. Contrarily, recent studies have shown that the semantics of words could be influenced by the underlying emotional symbols which suggest that it is possible to predict the emotional valance of a text by analyzing syllabic content [3]. The first phoneme of a word is shown to have a higher importance for predicting emotional valance of word's lexical meaning regardless of the phonetic features (e.g., intonation, tone, stress, and rhythm) [4]. Similarly, particular phonological units are shown to convey basic affective tone of poems to readers [5]. The results of these studies suggest that there exists a general relation between individual phonemes and emotional symbols, however the relation is still not very well understood.

Our technical contributions are two folds. Firstly, our method captures the phoneme sound quality in a compact set of only 12 features including 6 new features, i.e., width and dissonance of 3 formants. Secondly, clustering of phonemes using the disproportionately scaled formant characteristics provides a basis of information convergence to basic units (i.e., phonemes) that helps to speed up the recognition of the distinguishing features in data, thus increasing the computational efficiency.

The main advantage of the proposed method over existing SER techniques is that raw input speech data is converged into a few variables which speeds up the training process. Our method actively refines the speech data by focusing on 12 specific formant characteristics, whereas other methods search for features in the whole spectrum or temporal arrays of acoustic features. Our method abbreviates the whole temporal axis into occurrence rates of a few important unlabeled phonemes hence decreasing the size and cost of the model.

The rest of the paper is organized as follows. Related works of SER are introduced in Section 2. A new method of phoneme type convergence is proposed in Section 3. A phoneme selection method is given in Section 4. Experiments and analysis are given in Section 5.

2. Related Works

An SER system performs emotion recognition decisions by implementing intermediate mappings between the speech signal features and an emotional construct [6]. The types of speech signal features and the method of intermediate mappings vary for different systems. Creating spectrograms or periodograms are the preferred methods for speech input preprocessing of Deep Neural Network (DNN) input [7], whereas spectral features such as MFCC, Mel filters with or without log, and temporal features such as Linear Descriptor (LD), Auto-correlation, and Zero-Crossing Rate (ZCR) are preferred inputs for machine learning classifiers. Formant based features such as Linear Predictive Coding (LPC) coefficients are quite helpful to capture speech information into a few variables therefore they are widely used in the detection of phoneme cues as well as speech emotion detection [8]. MFCCs have been a popular choice of features to detect all kinds of cues in speech. Many studies have shown the importance of MFCCs in recognizing emotions in speech [9]. MFCCs are derived from Mel spectrograms [10], and some studies have suggested that using log-Mel spectrograms as CNN input creates a more efficient architecture [11, 12, 13].

Phoneme based methods are widely used for Automatic Speech Recognition (ASR) [14, 15]. But there are relatively fewer examples of phoneme based methods for emotion recognition. In an earlier work, phoneme-class specific HMM models were used for 5 broad classes of phonemes rather than a generic classifier for all speech segments [16]. More recently, Jing et al. [17] proposed a method of SER that uses a combination of acoustic features and occurrence rates of prominent syllables as input features to classifiers. Similarly, Yenigalla et al. used a set of 47 phonemes labels in combination with spectrograms as an input to CNN for speech emotion recognition [18]. Both of these studies indicated better performance for phonetically aware methods. A phonetically aware acoustic feature set was also shown to make significant improvement in emotional arousal and valence recognition [19].

Activities in certain regions of the brain are linked to formant detection which suggests that the human brain is wired to detect formant characteristics [20]. Understanding of human perception of formants is still limited but there is evidence that suggests that inter-formant distance and ratios are important factors for human perception of sound [21, 22, 23]. Studies show that small changes in prosodic patterns of syllables cause significant changes in the perceived prominence of syllables [24, 25, 26]. Phoneme-level signal manipulations were shown to be effective in modifying the emotional information conveyed in a speech utterance [27].

The research on the topic of emotions has been far from mathematical modeling which offers little help for cross-field research. But there is hope for better, more elaborate and accurate models that can help classify emotions such that there is less confusion for the machines [28, 29]. Meanwhile, emotions are being labeled with little to no validation. Most of the research-purpose databases are collected by acting because it helps to minimize variances from other sources, i.e., same sentences are repeated by different actors at a controlled intensity. The naturally recorded emotional corpora are only a fraction of all publicly available databases [30].

Considering the above mentioned state of affairs, researchers working on SER are doing their best using AI tools such as convolution neural network (CNN) [31, 32, 33, 34], recurrent neural network (RNN) [35, 36, 37, 38] and single or multiple machine classifiers [39, 40]. A few brain-inspired SER models have also been proposed that work in the same manner as the limbic brain of humans [41, 42]. By far, CNN has been the most preferred choice. Most of the recent work has been on the front-end of the classification model. However, for back-end, the speech signal preprocessing methods are borrowed from the decade old speech recognition methods [43].

Some feature selection algorithms select both personalized and non-personalized features in an attempt to create speaker-independent SER models. A two-layer fuzzy multiple random forest algorithm was proposed that considers speaker differences and creates different classifier trees of different depths for difficult to recognize emotions [44]. Similarly, an extreme learning machine was used to classify emotions using a fusion of personalized and non-personalized speech features [45].

There has been a focus on decreasing training time as well as the latency of prediction for emotion or lexical recognition from speech because it usually takes days to train large corpora [46]. Satt et al. [11] showed that latency and accuracy of SER can be improved by cleaning the input signal from noise. Another way is to focus the attention on salient regions of the spectrogram by learning the prominent discriminant regions [47, 48]. Attention-based feature learning has shown improvements in recognition accuracy of LSTM architectures [49, 50, 51].

3. Phoneme Type Convergence Method

In traditional SER methods, spectrograms and/or audio features are used as the inputs for a deep learning classifier. The general framework starts with the segmentation of audio signal into 20-150 millisecond frames, then spectrogram and/or audio features of each frame are passed as the input to a deep neural network. The RNN based neural networks use LSTM or similar technique to learn the temporal order of frames (e.g., [37, 13]). There are few other methods to learn the temporal cues (e.g., pyramid matching by [52] and 3D CNN by [53]). There are thousands of audio features but usually only few of them are used, most noticeably are MFCCs and LPC [8]. The number of audio features is usually

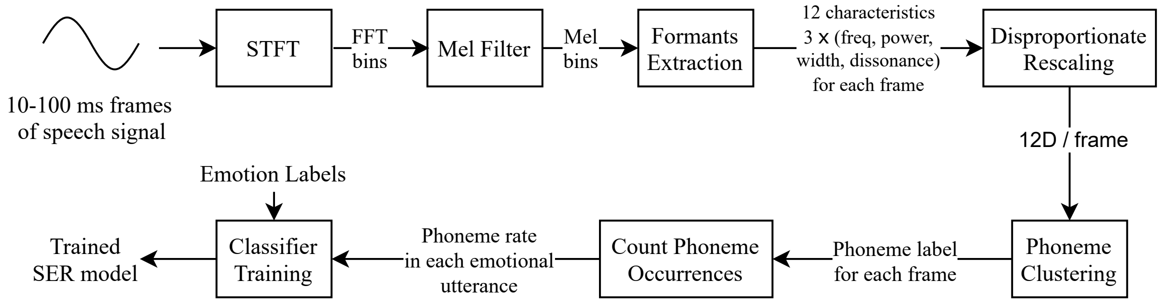

10-100 ms frames of speech signal
STFT
FFT bins
Mel Filter
Mel bins
Formants Extraction
12 characteristics 3 x (freq, power, width, dissonance) for each frame
Disproportionate Rescaling
12D / frame
Phoneme Clustering
Phoneme label for each frame
Count Phoneme Occurrences
Phoneme rate in each emotional utterance
Emotion Labels
Classifier Training
Trained SER model

Figure 1: Overview of the proposed phoneme type convergence method.

within the range of 20 to 100 per frame. The LPC has been used extensively in speech processing as a compression technique because it captures the sound quality in 4 to 20 variables [54]. Spectrogram based methods use varying sizes of 2-65 KB image inputs per frame. Decreasing the input size increases the risk of crucial information loss, whereas increasing the input size creates a risk of over-fitting and increases the computational processing cost.

As an alternative to the traditional approach, a phoneme type convergence method is proposed, which reduces the Mel-frequency cepstrum to only 12 variables per frame. Feature information per frame is further reduced to less than 10 variables by labeling the frames with phoneme labels generated based on the similarity of formant characteristics of frames. This size reduction not only helps process large amount of data but also helps in only focusing attention on the useful information while discarding the rest. An overview of the phoneme type convergence method is shown in Fig. 1.

Phonemes are the smallest phonetic units which are often labeled as the parts of words or syllables in lexical recognition of speech. Those cues that have no meaningful lexical label (e.g., breath sounds, pauses and most prosody level features) or variation in cues (long or short inflections) are considered as noise or different variants of the same phoneme. Such cues carry important affective meanings but most of them have no specific defining labels and vary across speakers and emotional contexts. Therefore, an unsupervised K-means clustering of phonemes is proposed to recognize the unlabeled phonemes and classify them into distinct types. The level of distinction between phoneme types can vary from syllable to suprasegmental level depending on the clustering parameters. All frames in speech corpora are clustered using a metric of 12-dimensional Euclidean distance from each other (each characteristic feature is regarded as a dimension) and numeric phoneme labels are assigned to all frames belonging to the same cluster. Because phonemes come in different shapes and sizes, the phoneme cluster types are mainly divided into two categories, i.e., one for the short uninterrupted segments and another for the differential changes in the syllable level segments. The occurrence rate of these phoneme labels in an utterance is then taken as the input for a classifier.

The scaling factors of formant characteristics are introduced, which create a difference in importance of formant characteristics in phoneme perception, e.g., a small difference in fundamental frequency changes the phoneme label, whereas relatively a big difference in the width changes the phoneme label provided all other characteristics are the same. We have determined the scaling factors for different formant characteristics based on the related work on just noticeable differences in hearing perception [55, 56]. These scaling factors ensure that the similarity level of phoneme cluster members is more dependent on the more important characteristics.

3.1. Formant Characteristic Features

The proposed phoneme type convergence method is designed to account for the most useful variations in the speech signal with minimum number of variables. Speech signal usually has two or more formants which carry most of the energy, whereas the rest of the spectrum has less energy and significantly less information. Frequency formants usually lie at harmonic distances from each other. The fundamental frequency along with the relative power of its harmonics determines the timbre of a sound. Like music, the emotional quality of speech is conveyed partially through the differences in timbre [57]. Since it is a quality, it is usually measured by multi-dimensional variables. Although it is difficult to encapsulate the voice quality in discreet quantities with perfect accuracy, but measuring the most defining components of the sound signal provides enough information to judge timbre quantitatively. Therefore, according to the proposed method, four variables, i.e, frequency, power (amplitude), width, and dissonance of the top three formants with the highest amplitude are taken as the 12 defining characteristics of a signal window.

The process of characteristic features extraction starts from dividing a speech signal into a few millisecond frames such that each frame has minimum variation within its timeframe. A contextual frame window of time duration T_w is iterated through the speech signal with a stride of T_s. Then a Hamming window is applied to each window

$$x_t(n) = (0.54 - 0.46 \cos(\frac{2\pi n}{W-1}))s_t(n) \tag{1}$$

where s_t is the input signal of frame t, x_t is the windowed frame, $0 \leq n \leq W-1$, and W is size of window (T_w times sampling rate). Then power spectrum of each frame is calculated by taking Short-term Fourier Transform ($STFT$) of x_t

$$P_t = \frac{(STFT(x_t))^2}{N_{FFT}} \tag{2}$$

where P_t power of N_{FFT} for frame t. When all P_t frames in an utterance are combined is a 2D matrix (periodogram) columns represent the FFT bins, rows represent the frames. Then Mel-filter is applied to each frame that coverts linear Hertz to a non-linear log scale, which is commonly used for many speech recognition methods due to similarity with the human ear perception [43]. The Mel-filter banks have triangular shape and get wider as the frequency increases. The Mel scale frequency can be converted to Hertz scale by

$$f = 700(10^{m/2595} - 1) \tag{3}$$

where m is Mel frequency and f is the Hertz scale frequency. A high number of Mel-filter banks (128-256) and number of FFT bins (512-2048) is recommended so that the resolution of formant characteristics is preserved.

3.1.1. Frequency and power

As a rule of harmonics, formants of fundamental frequency have the highest magnitude as compared to the rest of frequencies for harmonic sounds (usually vowel sounds). However, there can be non-harmonic sounds due to differences in eloquence and imperfections of the larynx, which usually convey consonants or blatant noise. We consider top three Mel-filter banks with the highest magnitude as the top three formants. Formants (i.e., high power frequency bands) are usually separated from each other by low energy frequency bands. Formants are detected by comparing the local maxima and minima of power of Mel-filter banks with each other as shown in Fig. 2. Then the central frequencies of Mel-filter banks are calculated as

$$f_c(l) = 700(10^{(m_l - m_{l+1})/5190} - 1) \tag{4}$$

where f_c is the central frequency of filter bank l on Hertz scale and m_l is the lower limit of filter bank l on Mel scale of a certain frame t. Figure 2 shows formants of a sample frame on Hertz scale of 30 Hz to 4000 Hz. The ranking order of mel-filter banks are determined in order to assign the formant ranks to mel-filter banks based on their power from highest to lowest. The mel-filter bank indices of peaks (highest peak between two valleys) are ordered by rank as

$$\phi_h \in \{\phi_0, \phi_1, \phi_2, \dots, \phi_{N_h - 1}\} \tag{5}$$

where ϕ_h gives the index l of mel-filter bank that is ranked h based on it's relative peak power within the current frame, and N_h is the maximum number of formants we need to extract. If there are more than one peaks between two valleys, then only the highest one is assigned a formant peak rank while other peaks are used for calculating dissonance. Then the power amplitude of each formant peak is log-scaled with decaying coefficient as

$$p_h = 100k^h \log_{10}(p(\phi_h)) \tag{6}$$

where p_h is the rescaled peak power of formant h of the current frame, $p(\phi_h)$ is the power of mel-filter bank at index $l = \phi_h$, and $k \leq 1$ is the decay constant that decreases the scale of power for each proceeding rank of formant from highest to lowest in power. Similarly, the frequency of formants are rescaled again using

$$f_h = 200k^h \ln(f_c(\phi_h)) \tag{7}$$

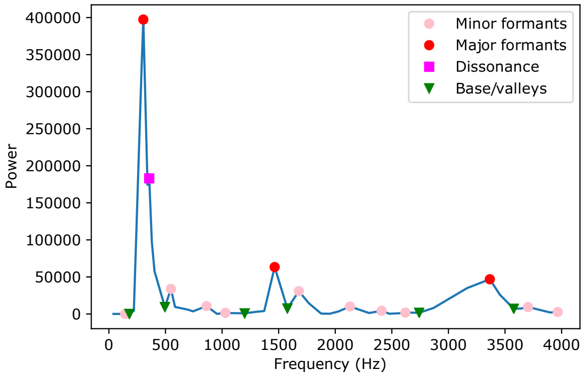


Figure 2: A sample of 25ms window frame showing the power of 256 bins of Mel-filter at their central frequencies (converted from Mel-scale to Hz). Formants (the highest local maxima) are separated by valleys (local minima) in between them. Other peaks within the peak-valley threshold around the highest local peak are considered as dissonance. The width of a formant is the distance (in Hz) in between the two valleys (local minima) on both sides of the formant peak.

where f_h is the rescaled central frequency, and $f_c(\phi_h)$ is Hertz-scale central frequency of formant h of current frame The purpose of this rescaling is to increase the scaling ratio of each predeceasing formant over its succeeding (lower in rank) formant such that the higher rank formant has the bigger component in Euclidean distance while measuring the multi-dimensional distances for creating clusters.

3.1.2. Formant Width and Dissonance

A high number of Mel-filter banks is recommended so that the width and dissonance of narrow formants can be measured. Formant width and dissonance are important in music analysis because the quality of melody is a function of consonance in it. Formant width can be guessed by the shape of the instrument that produced it. For example, small and narrow instruments like a flute or clarinet have narrow formant widths whereas bigger instruments like a horn or bassoon have wider formant width [58]. Similar to formant frequency and power, formant width w_h is calculated and rescaled as

$$w_h = 50k^h \log_{10}(f_{h,v_1} - f_{h,v_0}) \tag{8}$$

where f_{h,v_0} and f_{h,v_1} are the amplitude local minima points (valleys) on the lower side and higher side of the current frame's formant h on Hertz frequency scale, respectively. There is a threshold condition for local minima to be considered as valleys, i.e., their amplitude should be lower than the quarter of the formant's maximum amplitude peak.

Dissonance and consonance are what make a sound unpleasant or pleasant to hear. 'Pleasant' is an adjective of emotional nature but there have been very few related works on measuring dissonance for emotion detection [59, 60]. The biological or cultural nature of dissonance perception is debated among ethnologists [61, 62]. A sound or voice is perceived as pleasant when there is a higher ratio of harmonic components to non-harmonic components. The dissonance does not occur regularly around major formants, but it is recorded as a feature since it carries an important emotional signal if it is not caused by noise. In a noisy signal, this method will not guarantee a valid measurement of unpleasantness. The consonant sounds formed by the shape of the vocal tract are more likely to have a higher component of dissonance in them [63].

To measure the dissonance, non-harmonic formants occurring in the vicinity of major formants are taken as a measure of dissonance. The sum of all dissonance power maxima is divided by the formant's peak power and rescaled as

$$d_h = 100k^h \frac{\sum_{j=1}^{M_h-1} p(\phi_{h,j})}{p(\phi_{h,0})} \tag{9}$$

where d_h is the dissonance for formant h, M_h is the total number of amplitude local maxima (peaks) between the amplitude valleys f_{h,v_0} and f_{h,v_1} on the either side of formant h on frequency axis, k is the formant decay constant, and $p(\phi_{h,j})$ is the power of mel-filter bank where $\phi_{h,j} \in \{\phi_{h,0}, \phi_{h,1}, \phi_{h,2}, \dots, \phi_{h,M_h-1}\}$ represents the mel-filter bank indices of local maxima ranked at j from highest to lowest including the highest peak and the dissonance peaks between f_{h,v_0} and f_{h,v_1}.

3.1.3. Differential Phoneme Features

Another set of characteristic features is created by using a combination of adjacent frames in order to capture the consonant sounds. The gradient in formant characteristics happening within a few adjacent frames are important indicators of temporal cues. These features include the differences in frequency of the first and second formants and power difference of the first formant. The number of adjacent frames can be varied from 6 to 12 depending on the context frame size. The best results of an uninterrupted window of speech sound are usually acquired at a 25ms window. Therefore, a minimum of 60ms (i.e., 6 frames at 10ms stride) of a combined window of adjacent frames is recommended so that the change in consonant sound is captured within that window. Then by measuring the difference between the first half frames and the second half frames, a measure of the transitions that have happened in that window is calculated. Mean μ_f and difference δ_f of frequency of formant h for g adjacent frames with initial frame at t is calculated of as

$$\mu_{fh}(t) = \frac{\sum_{y=t}^{t+g} f_h(y)}{g} \tag{10}$$

$$\delta_{fh}(t) = \frac{2}{g}(\sum_{y=t+\frac{g}{2}}^{t+g} f_h(y) - \sum_{y=t}^{t+\frac{g}{2}} f_h(y)) \tag{11}$$

where $0 \leq t \leq T - g$, g is the number of adjacent subsequent frames covered by transition. Similarly, mean μ_p and difference δ_p of power of formant h with initial frame at t is calculated as

$$\mu_{ph}(t) = \frac{\sum_{y=t}^{t+g} p_h(y)}{g} \tag{12}$$

$$\delta_{ph}(t) = \frac{2}{g}(\sum_{y=t+\frac{g}{2}}^{t+g} p_h(y) - \sum_{y=t}^{t+\frac{g}{2}} p_h(y)) \tag{13}$$

Equations (11) and (13) calculate the differences between the means of first and second halves of segments. Since the number of frames per segment is constant, $\delta_{fh}(t)$ and $\delta_{ph}(t)$ can be taken as a measure slopes of frequency and power of formants at frame t. Total 6 differential features are calculated as a measure of the transitions that have happened in g subsequent frames following the initial frame at t. Although only three characteristic features, i.e., frequencies of first two formants (f_{t0}, f_{t1}) and power of first formant (p_{t0}) are recommended to be used for calculation of differential features, other features can also be included. It should be noted that while each differential feature is calculated using g frames (6 to 12 frames) the stride rate is still t (1 frame), which means that for two adjacent frames there is very less difference in differential feature values. Nonetheless, this difference is important to detect the transitional gradient in phonemes.

3.2. Phoneme Clustering

Different permutations of 12 formant characteristic features can create innumerable varieties of phonemes. Therefore, we limit the scope of phoneme heterogeneity by clustering using the metric of similarity of formant features. But not all formant features are equally important, so we rescale all features such that their variability range is a function of their importance in phoneme classification. The first and second formants carry most of the information needed for vowel detection, and most of the frames are observed to have a drastic drop in power after the second or third formant [64]. Therefore, the use of only the first three formants is recommended for clustering. As explained in Section 3.1.1 and 3.1.2, all features are rescaled such that f_0 has the highest range of variability, then f_1, then p_0 and p_1. The Euclidean distance along the 12-dimensions is used to calculate the distance among frames while creating K-means

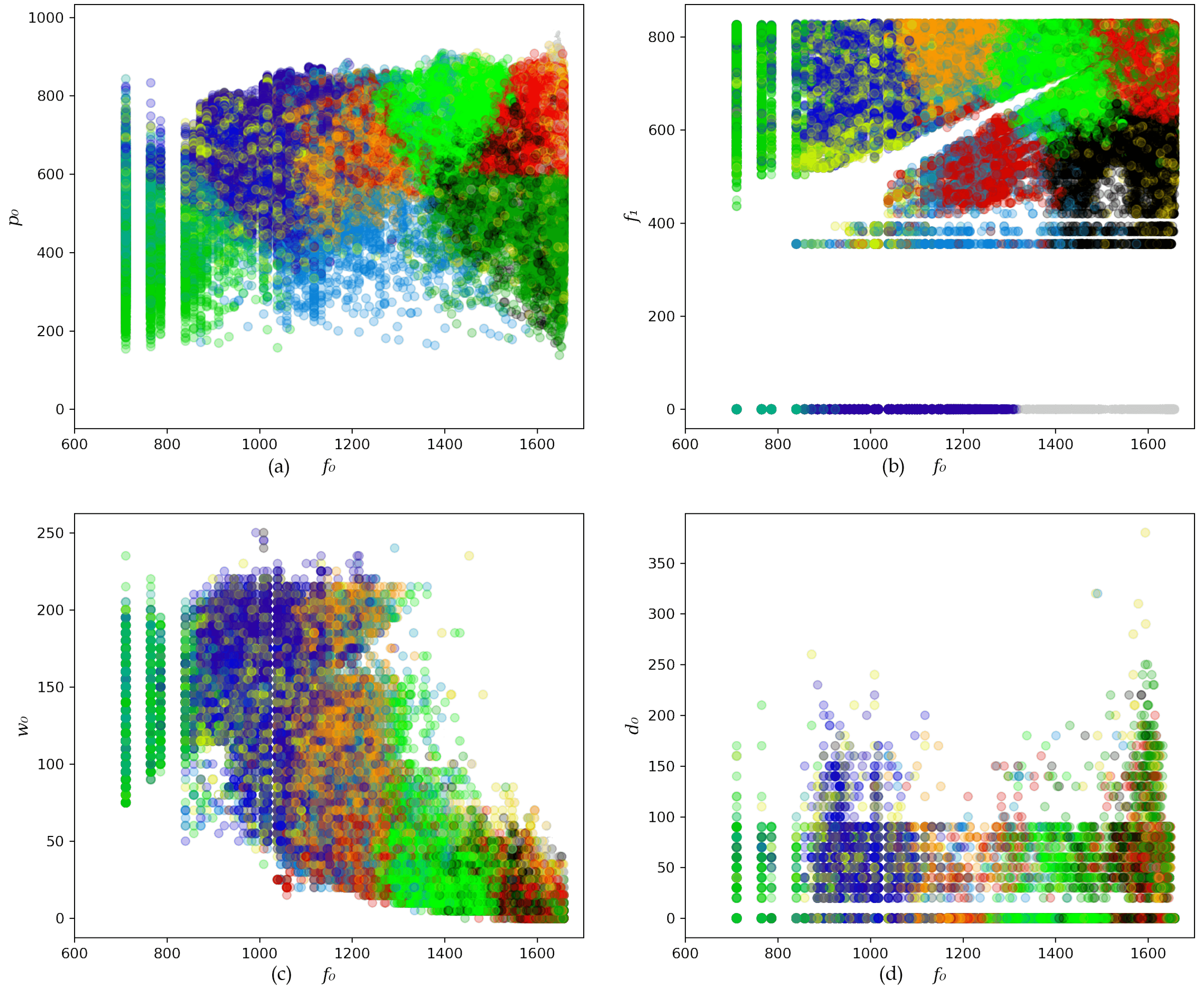


Figure 3: Scatter plots of clustering data points across 5 out of total 12 dimensions extracted from EmoDB. Different colors of points represent the cluster labels of total 16 clusters created by K-Means clustering (parameters same as No. 1 in Table 1). All features are scaled differently according to Eq. (7), (6), (8) and (9) for f_0, p_0,w_0 and d_0 respectively with $h = 0$, $h = 1$ for f_1, and $k = 0.5$ for all features. Cluster labels show more variability along f_0, p_0 and f_1 axes as compared to w_0 and d_0 axes due to the differences in scale of dimensions.

clusters. Figure 3 shows the 16 clusters created from formant characteristic features of more than 67 thousand 25ms frames. It can be observed that cluster labels (colors) vary in both dimensions in the (f_0, f_1) plot, whereas labels remain the same in vertical dimension in (f_0, w_0) plot.

Another group of clusters is created using the differential features between adjacent frames which are referred to as differential phonemes. The number of clusters (N_m) for both groups can be varied from 16 to 128 depending on the size and variability in the dataset or a set of different cluster sizes can be used together.

The size parameters of clustering are a function of phonetic similarity of cluster members. We recommend using a set of clustering models covering cluster sizes from 16 to 128 such that big cluster models group sparingly similar frames together whereas small cluster models differentiate between phonemes at phonetic level. In short, multiple cluster models with different sizes for two types of phonemes (instantaneous and differential) are recommended.

K-means clustering is preferred due to its high processing speed and easy scalability [65, 66]. Other algorithms such as Agglomerative and BIRCH clustering were pretested for method selection which did not show any improvement in accuracy of the whole model but caused an increase in the duration of training. BIRCH clustering did not distribute the clusters members uniformly but rather all members were clustered together in one or two dense groups. Other

Table 1
Computation time, Silhouette coefficient (S_{Sil}) [67], Calinski-Harabasz index (S_{CH}) [68] and Davies-Bouldin index (S_{DB}) [69] as metrics of clustering quality of K-Means and Hierarchical clustering algorithms [70] with different parameters. Clustering is performed using 12 formant features for each frame in 50 utterances of EmoDB.

No.	Algorithm	Time	Clusters	S_{Sil}	S_{CH}	S_{DB}	Parameters
1	K-Means	0.3s	16	0.806	259798	0.944	init='k-means++', batch=2000, n_init=5
2	K-Means	0.3s	16	0.802	256547	0.946	init='k-means++', batch=1000, n_init=5
3	K-Means	0.4s	16	0.804	254595	0.978	init='k-means++', batch=2000, n_init=10
4	K-Means	0.3s	11	0.801	214833	1.087	init='random', batch=1000, n_init=10
5	K-Means	4.1s	16	0.794	258722	0.979	init='k-means++', max_iter=500
6	Hierarchical	121s	16	0.800	237413	0.960	linkage='ward', affinity='euclidean'
7	Hierarchical	101s	16	0.806	182715	0.804	linkage='average', affinity='euclidean'

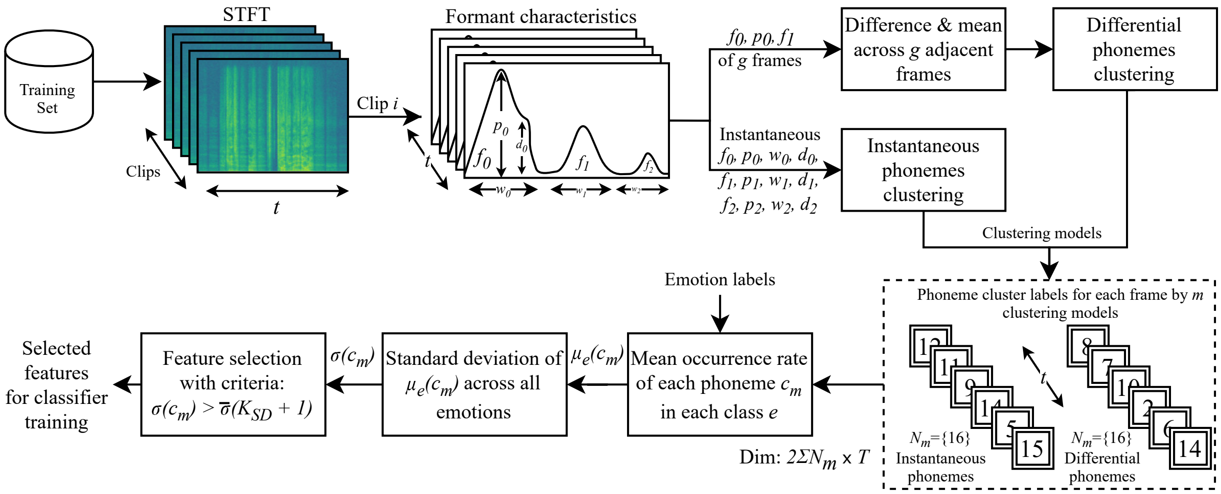

Training Set
STFT
Clips
t
Clip i
Formant characteristics
t
p_0
d_0
f_0
f_1
f_2
w_0
w_1
w_2
f_0, p_0, f_1 of g frames
Difference & mean across g adjacent frames
Differential phonemes clustering
Instantaneous $f_0, p_0, w_0, d_0, f_1, p_1, w_1, d_1, f_2, p_2, w_2, d_2$
Instantaneous phonemes clustering
Clustering models
Phoneme cluster labels for each frame by m clustering models
12
11
9
14
5
15
8
7
10
2
6
14
t
N_m={16} Instantaneous phonemes
N_m={16} Differential phonemes
Emotion labels
Mean occurrence rate of each phoneme c_m in each class e
Dim: $2\Sigma N_m \times T$
$\mu_e(c_m)$
Standard deviation of $\mu_e(c_m)$ across all emotions
$\sigma(c_m)$
Feature selection with criteria: $\sigma(c_m) > \overline{\sigma}(K_{SD} + 1)$
Selected features for classifier training

Figure 4: An overview of the feature extraction and selection process. First, formant based characteristic features are extracted from raw WAV files, then at least 2 types of phonemes are created using K-means clustering. Then cluster labels, regarded here as phoneme labels, are counted to calculate the occurrence rate in utterances of different emotions. The standard deviation of the occurrence rate of a phoneme in different emotions is used as a deciding factor in feature selection.

hierarchical models (Agglomerative Clustering) were not able to converge on a bigger dataset due to memory restraints. Another advantage of K-means is the batch-by-batch processing of dataset, which means it can be used in active applications for real-time updates of the classification model for big databases.

Table 1 gives Silhouette coefficient (S_{Sil}) [67], Calinski-Harabasz index (S_{CH}) [68] and Davies-Bouldin index (S_{DB}) [69] as metrics of clustering quality of K-Means and Agglomerative clustering algorithms with different parameters. Comparing by computation time and Calinski-Harabasz index (S_{CH} is higher when clusters are dense and well separated), K-Means clustering performed better than Hierarchical clustering. Silhouette coefficient (Higher S_{Sil} value relates to a model with better defined clusters) and Davies-Bouldin index (S_{DB} closer to zero indicate a better partition) did not show significant difference to draw a comparative judgement from these two metrics.

4. Phoneme Based Feature Selection

Occurrence rates of a few selected phonemes are proposed as input features of a machine learning classifier. After K-mean clusters are created for instantaneous and differential phonemes, occurrence rates of phonemes in for all utterances in a dataset are calculated. Only a few of these phonemes are selected as input features of a classifier based on the standard deviation of occurrence rate across all emotion classes in sample space. Occurrence rate R_{i,c_m} of phoneme

c_m for utterance clip i is calculated as

$$R_{i,c_m} = \frac{\sum_{t=0}^{T_i} u_{i,t}}{T_i} \tag{14}$$

where

$$u_{i,t} = \begin{cases} 1 & \text{if } c_{m,i,t} = c_m \\ 0 & \text{otherwise} \end{cases}, \tag{15}$$

where $0 \leq c_m < N_m$, N_m is the total number of clusters for clustering model m, $c_{m,i,t}$ is the phoneme label assigned by model m to frame t, and T_i is total number of frames in clip i. Then standard deviation σ_{c_m} of mean occurrence rate of phoneme c_m in different emotions is calculated as

$$\sigma(c_m) = \sqrt{\frac{\sum_{e=0}^{N_e} (\mu_e(c_m) - \overline{\mu(c_m)})^2}{N_e - 1}} \tag{16}$$

where N_e is the total number emotion classes, μ_e is a mean calculation function of occurrence rate of phoneme c_m in all clips belonging to emotion class e, $\overline{\mu}$ is the mean function of μ_e for all emotions e. Equation (16) gives a measure of importance of phoneme c_m for class discrimination. Using the calculated σ for each phoneme, important phonemes are selected by the criterion as

$$\sigma(c_m) > \overline{\sigma}(K_{SD} + 1) \tag{17}$$

where $\overline{\sigma}$ is the mean of σ for all phonemes labels by all models, and K_{SD} is a constant that determines the limit of selection above the mean of σ (i.e., $\overline{\sigma}$). The higher the K_{SD} is, the higher the selection limit for the standard deviation of phoneme occurrence rate in different emotion classes will be, therefore fewer features will be selected. As explained in Section 3.2, two types of clustering models are created, one from instantaneous phoneme features and another from differential phoneme features. The number of clusters in these two models can be different, or a set of different clustering models can be created using these two types of features separately. In this section and in experimentation in Section 5; to avoid any confusion between the number of instantaneous and differential phonemes, the same set of clusters count (N_m) is used for both phoneme types. For example, if $N_m = \{16, 32, 64, 128\}$ for both instantaneous and differential phonemes, the total clustering models will be 8 and the total number of unique phoneme labels (N_p) will be 480 (i.e., 240 for both types of phonemes).

When the dataset is huge and there is a significant imbalance then there are higher chances of all phonemes to be occurring with higher probabilities in a class with the most numerous number of samples [71]. If one class has comparatively more numerous samples, then the phoneme clustering is likely to be biased towards that class which means a high number of all phonemes will be occurring in that class. To avoid this problem, Bayesian probabilities of all phonemes in all classes are estimated to make sure at least $N_e/2$ of the selected features have the highest Bayesian probabilities for belonging to class e. If this condition is not satisfied, then K_{SD} will be decreased further to include more or all features.

$$P(e|c_m) = \frac{P(c_m|e)P(e)}{P(c_m)} \tag{18}$$

where $P(e)$ is the probability of an emotion class e occurrence (it reflects the class imbalance in samples), $P(c_m)$ is the probability of a phoneme c_m occurrence in any class, and $P(c_m|e)$ is the probability of phoneme c_m occurring in class e. Equation (18) can also be used to calculate the weight of an individual phoneme. A single phoneme is not enough to make a decision of utterance classification, but a combination of occurrence probabilities of many phonemes can make predictions. Therefore, a classifier can be trained to recognize emotion classes using the occurrence rates of the selected distinguishing phonemes.

5. Experimentation

We evaluated our SER method on six databases based on accuracy (weighted and unweighted), time cost, and robustness (within corpus and cross-corpus). An overview of the method used for experiments is shown in Fig. 5. All

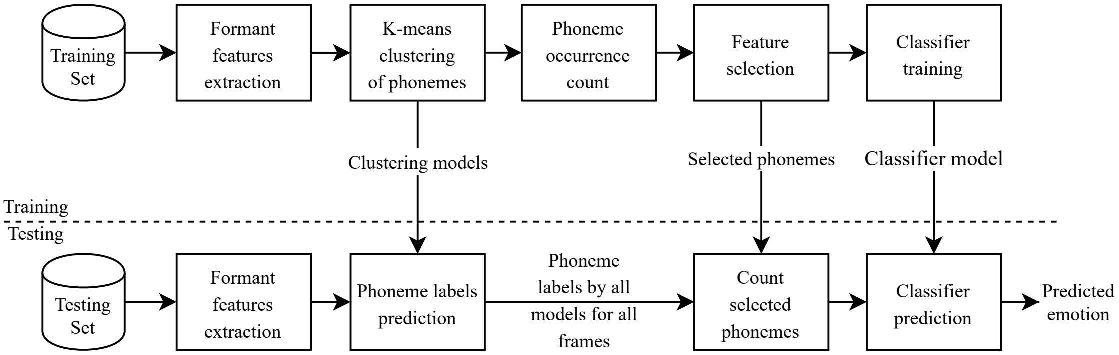


Figure 5: Overview of the experiment pipeline.

experiments are carried out in Python 3.7.4 (64-bit) environment on a computer of the 64-bit Windows 10 system with 16G memory and Intel Core i7-8550U processor. The code used for experiments has been published on GitHub [1] for open source distribution. The details and results are discussed in the following section.

5.1. Databases

Six databases were used, i.e., Berlin EmoDB [72], Ryerson audiovisual database of emotional speech and song (RAVDESS) [73], IEMOCAP (Interactive emotional dyadic motion capture database) [74], Sharif Emotional Speech Database (ShEMO) [75], DEMoS (Database of Elicited Mood in Speech) [76] and MSP-Improv [77]. A short summary of information about databases is given in Table 2. The count and duration of labeled utterances that were used for experimentation are given in Table 3.

Emotions in most of these databases were acted, but the data collectors have tried their best to verify the naturalness ([72]) or genuineness ([73]) of emotions. The validity of annotations of different databases is reported by different metrics by their respective data collectors. According to the evaluation of EmoDB in [72], the mean recognition rate by 20 listeners was more than 80%, with lowest for label 'disgust' at 80% and highest for label 'anger' at 97%.

For RAVDESS database (speech only), an average of Fleiss' kappa (κ) as a measure of inter-rater agreement among 20 annotators is 0.61, with the weakest agreement ($\kappa = 0.53$) for label 'sad' and strongest agreement ($\kappa = 0.67$) for label 'angry'. Similarly, for ShEMO database, [75] reported Cohen's kappa of 0.64 among 12 annotators. According to [78], the value of κ can be interpreted as moderate agreement within range of 0.41–0.60 and as substantial agreement within range of 0.61–0.80.

IEMOCAP and MSP-Improv databases include scripted and improvised utterances, we only used the improvised utterances from both to control the scenario. From MSP-Improv database, we only selected the utterances which have at least 67% agreement among raters. In IEMOCAP, there are only 1405 (out of 2943) utterances that have more than 67% agreement among raters. However, in accordance with the comparative works we used all the 2943 utterance of IEMOCAP so that a comparison can be made based on the same sample data [79, 80].

5.2. Experiment Pipeline

The main objective of the proposed method was to converge the information size of raw speech signal to compact phoneme based features without any significant drop in recognition accuracy. Using the method as explained in Section 3, experiments were carried out to measure the following effects of three parameters: 1) the effect of phoneme cluster size (N_m) on recognition accuracy, 2) effects of change in feature selection parameter K_{SD} on number of selected features (N_F), within corpus recognition accuracy and cross-corpus recognition accuracy, and lastly 3) the effect of change in hold-off ratio (test:total) on recognition accuracy. The SER model training is carried out in four steps:

Step 1: Twelve instantaneous features ($f_0, p_0, w_0, d_0, f_1, p_1, w_1, d_1, f_2, p_2, w_2, d_2$) for each frame t are extracted using 25 ms window, 10 ms stride, 2048 FFT bins, 256 Mel-filter banks, frequency range of 30 Hz to 4 kHz, and

[1] Source code is available at `https://github.com/tabahi/Phoneme-Converge-SER`

[2] Excitement considered as happiness.

Table 2
Information of databases used in experiment

	EmoDB	RAVDESS	IEMOCAP	ShEMO	DEMoS	MSP-Improv
Total clips	535	1440	2943	3000	1896	4835
Language	German	English	English	Persian	Italian	English
Speakers (M:F)	5M:5F	12M:12F	5M:5F	56M:31F	38M:21F	6M:6F
Environment	Script	Script	Improv	Radio drama	Elicited	Improv
Emotions	7	8	4	6	8	4
Total duration	24.8m	1h 16m	3h 28m	3h 25m	1h 25m	5h 7m
Trimmed duration	23.5m	46m	3h 26m	3h 19m	1h 23m	5h 3m
Avg. utterance length	2.6s	1.9s	4.2s	4s	2.6s	3.8s
Sampling rate	16 kHz	48 kHz	16 kHz	44.1 kHz	44.1 kHz	44.1 kHz
WAV files size	47 MB	589 MB	399 MB	1 GB	458 MB	1.6 GB
Formant features size	10 MB	27 MB	54 MB	56 MB	35 MB	98 MB
Trained model size	1.6 MB	4.2 MB	7.8 MB	7.1 MB	8.6 MB	17 MB

Table 3
Utterances count and total duration(minutes) for each label in databases.

Emotion	Label	EmoDB		RAVDESS		IEMOCAP		ShEMO		DEMoS		MSP-Improv	
		count	min	count	min	count	min	count	min	count	min	count	min
anger	A	127	5.1	192	6.4	289	21	1059	62	246	9.6	252	17.4
boredom	B	81	3.6										
calm	C			192	6.4								
disgust	D	46	2.4	192	7.1					140	7.1		
fear	F	69	2.4	192	6.1			38	1.9	177	8.6		
guilt	G									209	9		
happiness	H	71	2.8	192	5.9	[2]947	61	201	12.5	167	7.3	1859	109
neutral	N	79	3	96	2.6	1099	74	1028	81.5	332	15	2284	141
sadness	S	62	4.1	192	6.4	608	51	449	35.2	422	17.4	440	35
surprise	U			192	5.3			225	6.3	203	8.9		

formant decay constant of $k = 0.5$ according to the method explained in Section 3.1. Then using three of these instantaneous features, six differential phonemes features ($\mu_{f0}, \delta_{f0}, \mu_{f1}, \delta_{f1}, \mu_{p0}, \delta_{p0}$) are calculated at each frame ($g = 6$, 60 ms including adjacent windows).

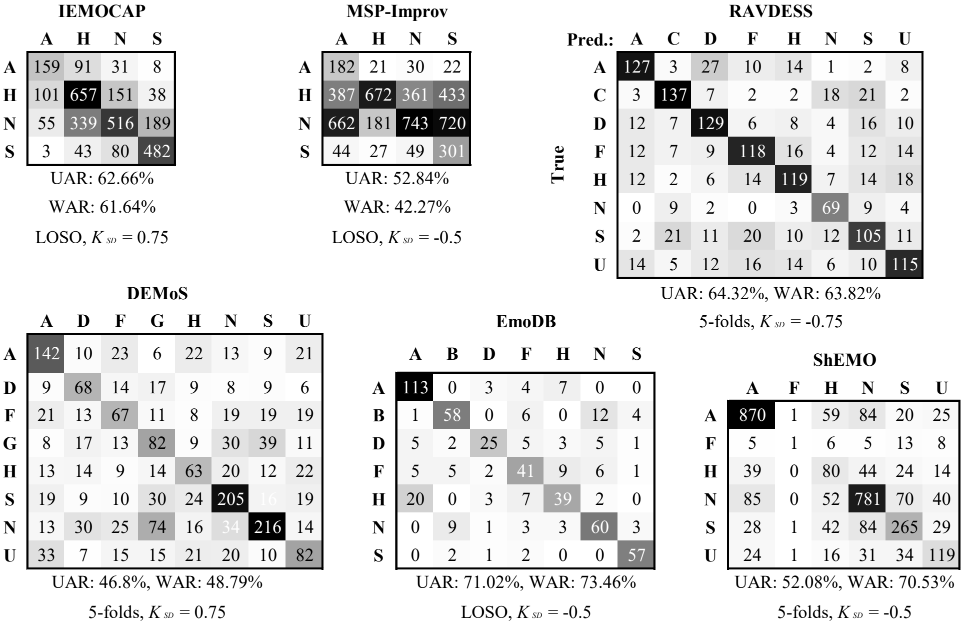
Step 2: K-means clustering (parameters same as No. 1 in Table 1) is performed for both instantaneous and differential phonemes using one or more cluster sizes (N_m) resulting into two or more clustering models and their respective generated phoneme labels for each frame of each clip. Training and testing dataset are split according to the validation scheme before clustering and only training set is used for creating clustering models.

Step 3: Phoneme occurrence rates are calculated in each emotion class of training set. Then highly discriminative phonemes are selected using the criteria described in Section 4.

Step 4: A classifier is trained using the selected phoneme occurrence rates as input features to predict emotional labels.

After training, phoneme cluster labels are predicted for each frame in audio clips of the testing set and then emotion class is predicted for each test clip using the phoneme occurrence rate as an input of the trained classifier. Three different types of validation schemes were used for different experiments. In accordance with the state-of-the-art methods, Leave-one-speaker-out (LOSO) validation was used for EmoDB and IEMOCAP databases, whereas 5-folds cross-validation was used for all other databases. In cluster size models comparison (Fig. 7), classifier comparison (Table 5) and within dataset K_{SD} sweep experiment (Fig. 9) only 5-folds cross-validation scheme was used for all databases. For hold-off ratio sweep experiment (Fig. 10), training and testing sets were split at certain ratios after a random shuffling of the complete set. For cross-corpus experiments (Fig. 9 and Table 6), training and testing sets were composed of the whole databases without intra-dataset splits.

The proposed method is analyzed based on four metrics, i.e., Unweighted Average Recall (UAR), number of classifier input features and training duration. Confusion matrices, UAR and WAR (Weighted Average Recall or accuracy)



IEMOCAP

	A	H	N	S
A	159	91	31	8
H	101	657	151	38
N	55	339	516	189
S	3	43	80	482

UAR: 62.66%
WAR: 61.64%
LOSO, $K_{SD} = 0.75$

MSP-Improv

	A	H	N	S
A	182	21	30	22
H	387	672	361	433
N	662	181	743	720
S	44	27	49	301

UAR: 52.84%
WAR: 42.27%
LOSO, $K_{SD} = -0.5$

RAVDESS

True \ Pred.:	A	C	D	F	H	N	S	U
A	127	3	27	10	14	1	2	8
C	3	137	7	2	2	18	21	2
D	12	7	129	6	8	4	16	10
F	12	7	9	118	16	4	12	14
H	12	2	6	14	119	7	14	18
N	0	9	2	0	3	69	9	4
S	2	21	11	20	10	12	105	11
U	14	5	12	16	14	6	10	115

UAR: 64.32%, WAR: 63.82%
5-folds, $K_{SD} = -0.75$

DEMoS

	A	D	F	G	H	N	S	U
A	142	10	23	6	22	13	9	21
D	9	68	14	17	9	8	9	6
F	21	13	67	11	8	19	19	19
G	8	17	13	82	9	30	39	11
H	13	14	9	14	63	20	12	22
S	19	9	10	30	24	205	16	19
N	13	30	25	74	16	34	216	14
U	33	7	15	15	21	20	10	82

UAR: 46.8%, WAR: 48.79%
5-folds, $K_{SD} = 0.75$

EmoDB

	A	B	D	F	H	N	S
A	113	0	3	4	7	0	0
B	1	58	0	6	0	12	4
D	5	2	25	5	3	5	1
F	5	5	2	41	9	6	1
H	20	0	3	7	39	2	0
N	0	9	1	3	3	60	3
S	0	2	1	2	0	0	57

UAR: 71.02%, WAR: 73.46%
LOSO, $K_{SD} = -0.5$

ShEMO

	A	F	H	N	S	U
A	870	1	59	84	20	25
F	5	1	6	5	13	8
H	39	0	80	44	24	14
N	85	0	52	781	70	40
S	28	1	42	84	265	29
U	24	1	16	31	34	119

UAR: 52.08%, WAR: 70.53%
5-folds, $K_{SD} = -0.5$

Figure 6: Confusion matrices of emotion recognition using phoneme clustering models $N_m = \{16, 32, 64, 128\}$ and SVM classifier. Emotion labels are given in Table 3.

Table 4
Recorded computation time of four steps using an Intel(R) i7-8550U processor and Python 3.7.4 (64-bit) environment.

Step	EmoDB	RAVDESS	IEMOCAP	ShEMO	DEMoS	MSP-Improv
Formants extraction from WAV files	1.1 min	3.7 min	9.1 min	9.4 min	3.5 min	12.2 min
Phoneme clustering ({16,32,64,128})	7 sec	16 sec	52 min	49 sec	15 sec	1.4 min
Occurrence count & feature selection	3 sec	7 sec	50 sec	20 sec	6 sec	59 sec
SVM training	1 sec	2 sec	4 sec	2 sec	2 sec	5 sec

of within corpus validation are shown in Fig. 6. The training duration of each step is shown in Table 4. In the following subsections, experiment results at each step of the proposed method are analyzed.

5.2.1. Formant Features Extraction

The top three formants are preferred because the proposed method did not achieve as high accuracy for one or two formant based features, therefore only three formant based experiment results are reported here for analysis. Four formants based phonemes were not tested as there were not enough audio clips that had a significant ratio of f_3/f_0. Formant feature extraction took the longest duration in the whole training process. However, that includes writing the extracted formant features to an HDF (Hierarchical Data Format) database on disk which is later read again for clustering. The convergence of data size from WAV files to the HDF database can be seen in Table 2.

The convergence ratio for different datasets correlates with the duration of feature extraction (see Table 4). This ratio can be further reduced by limiting the frame numbers per utterance or by limiting the floating-point precision. In our experiment, we used 16-bit unsigned integer to store formant characteristics, but the maximum value (after scaling according to Section 3.1) was within range of 1000 for all formant characteristic features, which means there is more room for compression. Moreover, if a marginal increase in UAR is not as an important factor for method selection, then by decreasing the phoneme cluster sizes to a half we can get almost as good accuracy as the best UAR model (i.e., $N_m = \{16, 32, 64, 128\}$).

5.2.2. Phoneme Clustering

Different phoneme cluster sizes (N_m) were used for experiments ranging from 16 to 128. Fig. 3 shows K-means clusters of 16 types of phonemes. There was a marginal increase in UAR with the increase in the number of clusters

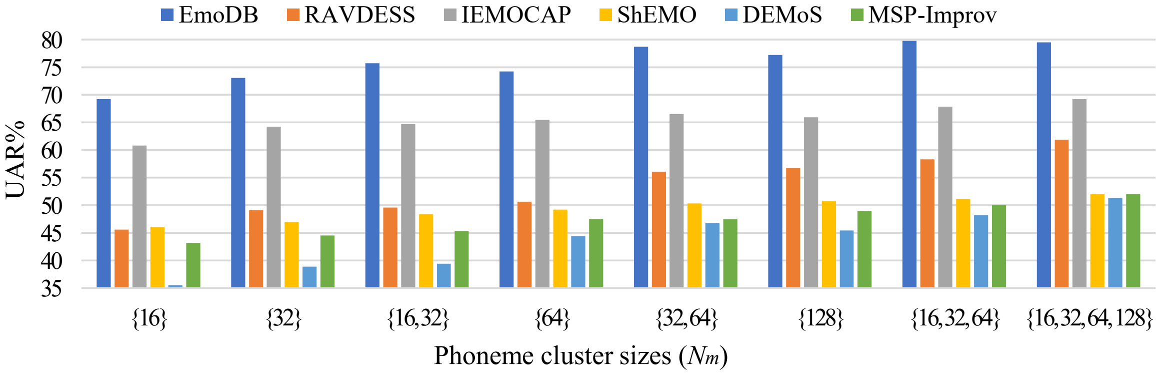


Figure 7: SVM prediction UAR using different phoneme cluster size sets (N_m). Parameter of feature selection is constant at $K_{SD} = -0.5$ for all databases.

as shown in Fig. 7. The best results were achieved by using a set of cluster sizes for both instantaneous and differential phonemes with $N_m = \{16, 32, 64, 128\}$. The difference between lowest ($N_P = 32$) and highest ($N_P = 480$) number of clusters was not as big as it was expected. There was less than 10% increase in UAR with a 15 times increase in the number of phoneme clusters for all databases. For a clean and reliable (high inter-rater agreement) database such as EmoDB, the proposed model works at 69% UAR using only 32 phonemes. UAR for the RAVDESS database increases linearly relative to UAR for the IEMOCAP database, which shows that an increase in explained variance among phonemes increases much more for the RAVDESS database as compared to IEMOCAP database.

5.2.3. *Phoneme Occurrence Rate and Feature Selection*

Mean occurrence rate R_{i,c_m} is calculated by dividing each phoneme count to the total number of frames (T) in a clip according to Eq. (14). In this experiment, the mean occurrence rate is also multiplied by 100 for the purpose of increasing floating-point depth. A visualization by error bar plots of phoneme occurrence rate for different emotions is shown in Fig. 8 for 16 clusters extracted from training set (80%) of EmoDB. It can be observed that almost half of the phonemes have a distinguishable occurrence rate for one or two emotions. By using the feature selection method as proposed in Section 4, only highly discriminative features are selected (11 features are selected in case of Fig. 8). These occurrence rates of selected phonemes are then used as input features for SVM training.

The number of selected features is dependent on parameter K_{SD}, which has different optimum values for different datasets. Plots in Fig. 9 show change in UAR and number of selected features (N_F) as the number K_{SD} sweeps through -1 to +1. The number of selected features (N_F) can be slightly different for each fold (5 training sets are created for 5-folds cross-validation), therefore the average value of N_F is shown in Fig. 9. The UAR for IEMOCAP shows a little change for a wide range of number of selected features (N_F). However, UAR for the EmoDB and RAVDESS shows a linear decrease in accuracy with an increase in K_{SD}.

5.2.4. *Classifier Training*

Four types of classifiers were tested using Scikit-learn library [70] (v0.22.1). SVM-RBF (Support Vector Machine with Radial Basis Function) performed relatively better than others. The UAR and WAR of four classifiers are given in Table 5. The classifier comparison experiment was performed with all other parameters constant as $N_m = \{16, 32, 64, 128\}$, $K_{SD} = -0.5$. SVM (parameters: ‘OVR’, ‘RBF’, C=1, gamma= ‘scale’) performed better in most of the cases as compared to RF (Random Forest, parameters: max_depth=30, estimators=100, max_features=10), KNN (K-Nearest Neighbors, n_neighbors=20), and MLP (Multi-Layer Perceptron, parameters: alpha=1, max_iter=2000, validation=0.3, hidden_layers={400,100,50}).

Due to the superior performance of SVM, all other experiments were performed only with SVM classifier. Figure 6 shows confusion matrices for all databases using SVM and K-means clustering sizes of $N_m = \{16, 32, 64, 128\}$. The number of selected features were different for different datasets as there was different optimum K_{SD} for different speech corpora. UAR, WAR, and the number of features (as input size) of the classifier is compared with the state-of-the-art works in Table 7.

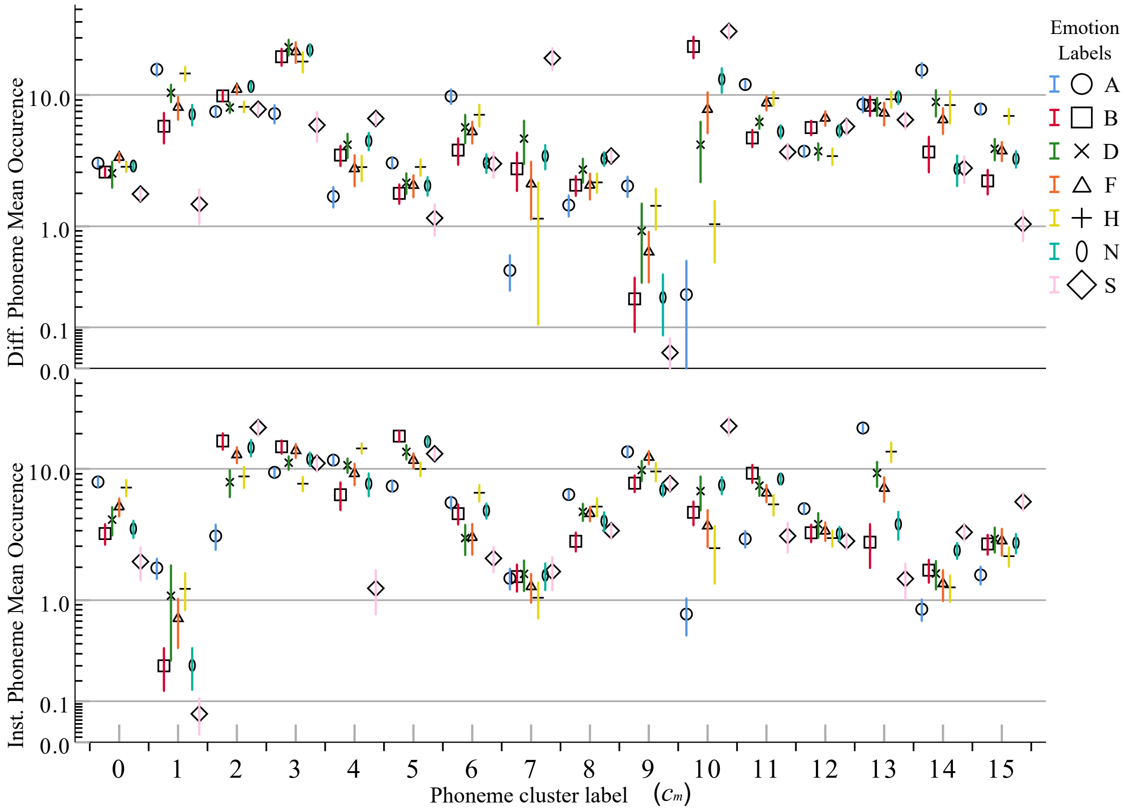


Figure 8: Mean occurrence rate of phonemes per 100 frames in EmoDB. Error bars confidence interval is 95%. $N_m = \{16\}$, i.e., 16 clusters for both instantaneous and differential types. Legends represent emotions as A=anger, B=boredom, D=disgust, F=fear, H=happy, N=neutral, and S=sad. Phonemes with high standard deviation in occurrence rate for different emotions are preferred in feature selection process.

Table 5
UAR% of different classifiers using 5-folds cross validation.

Classifier	EmoDB	RAVDESS	IEMOCAP	ShEMO	DEMoS	MSP-Improv
SVM	78.66	64.32	66.19	52.08	46.06	54.43
RF	71.05	49.2	62.01	44.91	38.33	36.78
KNN	60.5	43.24	59.28	34.32	37.56	30.38
MLP	72.91	61.02	61.91	49.08	44.27	52.84

5.3. Discussion

In terms of accuracy or UAR, an SER model can perform as good as the annotation reliability. Moreover, the imbalance in labels affects the overall UAR. As it can be seen in Figure 6, there are relatively fewer samples for fear ('F') in ShEMO's confusion matrix, which drops the whole UAR. All confusion matrices show higher class-accuracies for anger ('A'), which can be explained by the number of distinctive phonemes by occurrence rate in Fig. 8.

In Fig. 9, the dependence of UAR on the number of features (N_F) indicates that the complexity of the database can be explained by increasing number of features. The complexity of a database can be judged by numerous measures, such as number of speakers, number of emotions, number of recordings, variation of sentences, naturalness, and inter-rater agreement. According to all these measures, the RAVDESS database is far more complex than all other databases, as it shows an inclined slope for UAR in Fig. 10. But the number of samples per label in RAVDESS are not as much as in IEMOCAP which performs better at generalization due to the increased variance within the same labels.

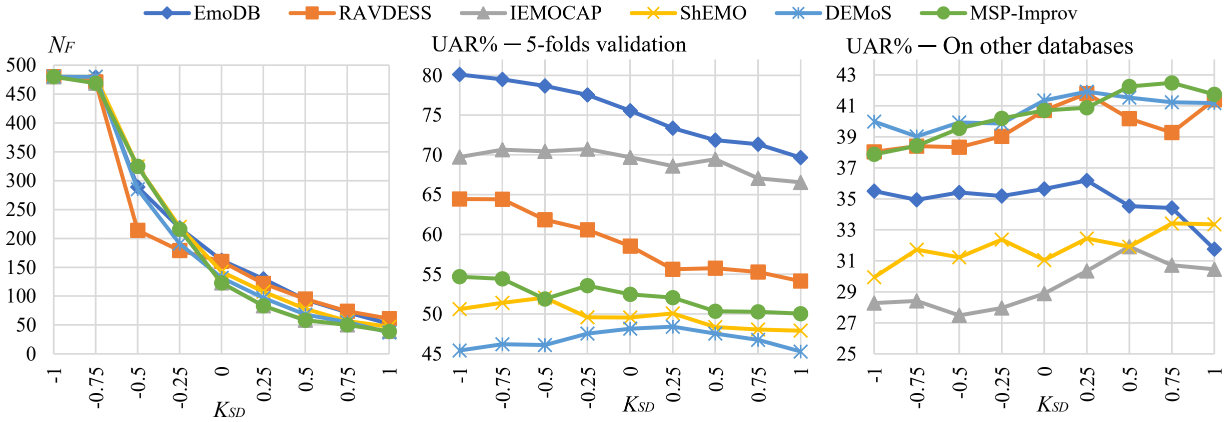


Figure 9: These plots show the effects of parameter K_{SD} on the average number of selected features (N_F out of total 480), validation UAR% and the cross-corpus UAR%. For cross-corpus tests, models are trained on one dataset and tested for robustness on other 6 datasets using only the 4 common emotions.

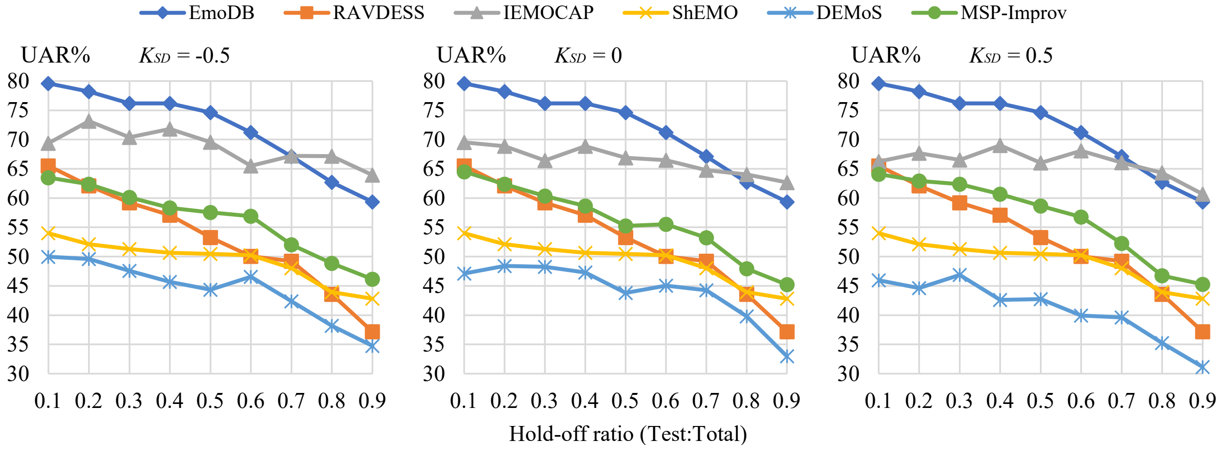


Figure 10: The effect of decreasing the training set size (and increasing the testing set size) on the validation UAR of models with more features ($K_{SD} = -0.5$) verses models with fewer features ($K_{SD} = 0.5$) shows that the number of features has bigger effect on the UAR for databases with more categories, i.e., RAVDESS and DEMoS.

5.3.1. Generalizability Evaluation

The relation between generalization capability and number of features can be observed in Fig. 9 and 10. As the number of features is decreased, within corpus UAR decreases but the cross-corpus UAR increases. The UAR for bigger datasets (e.g., IEMOCAP) show relatively change when the number of features is increased or decreased.

Within corpus, the increase in the hold-off ratio in Fig. 10 shows mixed results for different datasets. IEMOCAP and ShEMO databases show relatively less drop in UAR as the size of the training set is reduced. Ideally, there should be no difference in UAR for the best SER model (zero slope in Fig. 10). For EmoDB and RAVDESS, however, there was a linear drop in accuracy as the hold-off ratio was increased. The relatively less change in UAR is an indicator that the model doesn't learn anything much from a big chunk of data (90%) compared to a small chunk of data (10%) for IEMOCAP database. The UAR for IEMOCAP shows the same insensitivity to the number of features (N_F) and feature selection parameter K_{SD}.

In Table 6, the results of experiments with mutually exclusive sample spaces as well as overlapping sample spaces between training and testing sets are given. As we increase the number of features (N_F), the UAR of tests with mutually exclusive sets decreases, but the UAR of tests with overlapping sets increases. The number of features (N_F) has a generalization optimum (usually at $K_{SD} = 0.75$), which supports our assumption that number of features per utterance is a big factor for regularization. Fewer number of features imply a higher sensitivity to the selected few

Table 6

Cross-corpus and cross-lingual UAR% for 4 basic emotions (anger, happiness, neutral and sadness). Results of mutually exclusive training and testing sets are **bolded** to illustrate the cross-corpus performance. Decreasing the number of features N_F increases the cross-corpus UAR%.

	Testing set						Parameters	
Training set	MS	IE	Sh	DE	RAVDESS	EmoDB	K_{SD}	N_F
MSP-Improv (MS)	59.46	**46.94**	**35.58**	**33.98**	**32.29**	**44.06**	0.75	39
MSP-Improv (MS)	75.32	**45.87**	**35.29**	**31.46**	**35.16**	**48.7**	0	117
MSP-Improv (MS)	76.38	**45.04**	**34.1**	**31.05**	**36.72**	**43.5**	-0.75	480
IEMOCAP (IE)	**37.53**	68.7	**35.99**	**32.89**	**42.32**	**72.72**	0.75	42
IEMOCAP (IE)	**36.57**	77.71	**35.11**	**31.33**	**40.89**	**69.59**	0	106
IEMOCAP (IE)	**35.36**	84.06	**34.88**	**31.7**	**40.36**	**67.56**	-0.75	480
ShEMO (Sh)	**31.86**	**32.01**	62.75	**42.7**	**41.41**	**44.52**	0.75	57
ShEMO (Sh)	**29.59**	**31.31**	83.21	**40.45**	**42.32**	**42.45**	0	140
ShEMO (Sh)	**27.58**	**30.87**	84.41	**40.27**	**41.41**	**46**	-0.75	480
DEMoS (DE)	**33.01**	**41.75**	**43.13**	63.41	**43.36**	**53.73**	0.75	50
DEMoS (DE)	**31.64**	**41.31**	**43.47**	82.44	**44.27**	**48.66**	0	119
DEMoS (DE)	**31.03**	**40.94**	**44.26**	84.53	**45.31**	**50.28**	-0.75	480
MS+IE	50.32	61.33	**36.24**	**40.44**	**40.23**	**61.05**	0.75	46
Sh+DE	**32.83**	**37.94**	67.42	54.99	**39.84**	**56.08**	0.75	51
MS+DE	58.94	**44.9**	**43.29**	60.2	**41.93**	**49.45**	0.75	54
MS+Sh	63.35	**40.97**	61.04	**41.57**	**36.98**	**48.32**	0.75	46
IE+Sh	**33.21**	61.71	61.03	**35.78**	**33.85**	**59.69**	0.75	43
IE+DE	**33.05**	63.99	**35.29**	48.96	**38.28**	**54.09**	0.75	41

features which can be a double edge sword if those few features only exist within the corpus's domain. But when we increase the variety in training set, the selected phonemes become less domain dependent hence perform better on out of domain corpora, e.g., the UAR was higher when we trained the model on a variety of datasets (and languages) and tested on EmoDB. The variety in languages can have the opposite effect within domain, e.g., when we tested the models on IEMOCAP, the UAR was lower when we trained the model with variety of datasets (and languages) as compared to training only on RAVDESS which has the same language.

There is room for optimization to improve the generalizability of the current model. Scaling factors, decay rate, and many other parameters were kept constant to control other factors such as K_{SD}. These parameters can be optimized to fit the right domain at the right time.

5.3.2. Comparative Analysis

A comparison of methods and their UARs are given in Table 7. There is not much difference in the UAR of our method with the state-of-the-art methods. But there is a big difference in the number of features used as input of classifiers. The earlier studies mostly used the various low-level acoustic features (e.g., pitch, RMS Energy, loudness, MFCCs), but the trend has been shifting towards the spectrum based using deep learning. In [79], authors showed a slight improvement of their proposed deep spectrum features learning strategy over the low-level descriptors (LLDs) and SVM based model. Their method, by best of our knowledge, gave the best UAR of 68% for IEMOCAP dataset. A recent work by [80] proposed a quantum-behaved particle swarm optimization (QPSO) for dimension reduction, which increased the accuracy of Gaussian elliptical basis function (GEBF) neural network classifier from 74.55% to 79.94% for EmoDB.

Related works on SER rarely report the training time and computational costs. For some perspective, an evaluation by [83] reported a training duration of 2 to 14 days for a deep neural network using spectrogram inputs extracted from the IEMOCAP dataset. Recently, Daneshfar et al. reported that despite the higher accuracy of their method, high computational cost and low convergence speed are the major drawbacks for their method [80]. Very different from it, our method finishes training within minutes. The prediction latency is also not limited by any computational constraint other than the length of speech itself. The average utterance length for all datasets was less than 5 seconds (see Table 2), which means a latency of 5 seconds should be expected from the proposed method.

Table 7
UAR% and estimated input size per utterance (/u) or per frame (/t) of other methods are compared with our method.

DB	Ref.	Emotions	Method Base	Validation	UAR%	WAR%	/t	/u
EmoDB	[81]	All 7	LLDs, SVM	LOSO	84.6	85.6	N/A	>6K/u
EmoDB	[52]	All 7	CNN	LOSO	**86.3**	87.3	4096/t	N/A
EmoDB	[80]	All 7	LLDs, GEBF	LOSO	76.81	79.94	213/t	N/A
EmoDB	Ours	All 7	$N_P = 480, K_{SD} = -0.5$	LOSO	71.02	73.45	N/A	192/u
EmoDB	[82]	All 7	LLDs, DNN	5-folds	N/A	63.6	N/A	**59/u**
EmoDB	Ours	All 7	$N_P = 480, K_{SD} = -0.5$	5-folds	78.66	80.75	N/A	218/u
IEMOCAP	[83]	A,H,N,S	CNN, LSTM	LOSO	60.89	64.78	1600/t	N/A
IEMOCAP	[79]	A,H,N,S	CNN, LSTM	LOSO	**68.0**	65.0	16K/t	N/A
IEMOCAP	[84]	A,H,N,S	CNN, LSTM	LOSO	60.23	N/A	16K/t	N/A
IEMOCAP	[80]	A,H,N,S	LLDs, GEBF	LOSO	65.73	65.71	213/t	N/A
IEMOCAP	Ours	A,H,N,S	$N_P = 480, K_{SD} = 0.75$	LOSO	62.66	61.64	N/A	**42/u**
IEMOCAP	[85]	A,H,N,S	LLDs, BLSTM	5-folds	63.9	62.8	32/t	N/A
IEMOCAP	[11]	A,H,N,S	CNN, LSTM	5-folds	59.4	68.8	60K/t	N/A
IEMOCAP	Ours	A,H,N,S	$N_P = 480, K_{SD} = 0.75$	5-folds	66.19	66.21	N/A	**47/u**
RAVDESS	[86]	All 8	CNN, ResNets	5-folds	**64.5**	N/A	512/t	N/A
RAVDESS	Ours	All 8	$N_P = 480, K_{SD} = -0.5$	5-folds	64.32	63.82	N/A	**471/u**
ShEMO	[75]	5, ∄F	LLDs, SVM	5-folds	58.2	N/A	N/A	N/A
ShEMO	Ours	5, ∄F	$N_P = 480, K_{SD} = -0.5$	5-folds	**60.87**	72.72	N/A	**335/u**

The results of cross-corpus experiments are difficult to compare with the existing literature due to the differences in emotion labeling structure and differences in train-test splits. Most of the works used the two or three valance classes ([87, 88, 89]). An analysis of LSTM, CNN, and CNN-LSTM was given by [89] in which authors trained the deep learning architectures using IEMOCAP dataset (classes: negative, positive and neutral) and tested on EmoDB (UAR 42% for CNN-LSTM) and RAVDESS (UAR 33.3% for CNN and LSTM). Gideon et al. proposed an adversarial discriminative domain generalization method which was trained on MSP-IMPROV dataset and tested on IEMOCAP (UAR $47 \pm 1\%$) for recognizing three valance classes [87]. Probabilistically, their model made 1.42 times more correct predictions compared to the random chance. Although non-comparable, our model makes 1.84 times more correct predictions to that of random chance for 4 categorical classes in IEMOCAP.

Recently, some studies have proposed adversarial networks that were trained on IEMOCAP and tested on MSP-IMPROV dataset, autoencoding and attentive CNN [90], and CycleGAN (Cycle-consistent generative adversarial networks) [91], both of which gave UAR figures of $45 \pm 1\%$. In [88], the authors used Generative Adversarial Network (GAN) based model for unsupervised domain adaption which achieved 65.3% UAR for binary classification (positive and negative emotions) when trained on Urdu speech database (400 utterances) and tested on EmoDB. Our model achieves up to 61% UAR (4 classes) for EmoDB if it is trained on a combination 2 datasets.

As shown in Figure 10, when the hold-off ratio is 90% (i.e., testing set is 9 times the training set) then adding more features (by decreasing K_{SD}) doesn't affect the accuracy of prediction. Similarly, in Table 6 it can be observed that decreasing the number of features (N_F) increases the cross-corpus UAR%. We present an argument that models that are trained with high number of dependent factors (or features) have a high specificity on the training domain and therefore such models are less likely to make correct predictions in the wild. The feature selection criteria used by deep learning methods are usually designed to avoid underfitting or overfitting but it takes relatively long time for neural networks to find out the optima of thousands of input features. Our approach presents a practical compromise between the generalizability, cost and accuracy. As an alternative to the CNN strategy that uses frame-by-frame Mel spectrograms to find features, our method can be used to cut short the time usually taken by the feature-finding process by directly looking at formants features, grouping them together and then recognizing those groups that occur frequently in certain emotional categories.

6. Conclusion

Speech emotion recognition faces many challenges to improve recognition accuracy as well as to decrease the computational cost of the overall model. Due to these challenges, we proposed a fast salient features extraction mechanism to improve the accuracy and achieve a reduced computational cost of the overall SER model. We used Mel-filter banks to extract each frame's formant characteristics using which phoneme labels were generated by K-means clustering. Then the occurrence count of selected phonemes was used to train a classifier.

The effectiveness of the proposed method was evaluated on six databases, two of which are the standard benchmark databases (EmoDB and IEMOCAP). Our method achieves UAR of 62% on the IEMOCAP database and 71% on EmoDB with a lesser number of features and reduced training time of few minutes, yielding a computational-friendly SER system while achieving the same accuracy as state-of-the-art methods. We also tested the robustness by cross-corpus experiments which gave better performance with fewer selected phonemes, which hints that only the most discriminative phonemes are transferable to other domains.

In the future, we plan to continue to make improvements in the current method to effectively minimize supervision in noisy and unpredictable environments. In the current model, we tried to make sure that it can be used online using small batches of training data, because we plan to add a domain adaption module over the current functionality. One drawback of the current model is that it assumes that the theory of arbitrariness is false, i.e., one model fits all domains. However, in real-world it is likely to come across a highly adverse domains, in which case an adaptive deep learning or neuro-evolution based model would be useful. We hope to improve the applicability of SER further so that it can be employed in more real-world applications such as online gaming, social media, and self-driving cars.

References

[1] E. Finegan, Language: Its structure and use, Cengage Learning, 2014.

[2] R. L. Trask, Language: the basics, Routledge, 2003.

[3] V. Slavova, Towards emotion recognition in texts–a sound-symbolic experiment, International Journal of Cognitive Research in Science, Engineering and Education 7 (2) (2019) 41–51.

[4] J. S. Adelman, Z. Estes, M. Cossu, Emotional sound symbolism: Languages rapidly signal valence via phonemes, Cognition 175 (2018) 122–130.

[5] A. Aryani, M. Kraxenberger, S. Ullrich, A. M. Jacobs, M. Conrad, Measuring the basic affective tone of poems via phonological saliency and iconicity., Psychology of Aesthetics, Creativity, and the Arts 10 (2) (2016) 191.

[6] S. S. Narayanan, Emotion recognition system, uS Patent 8,209,182 (Jun. 26 2012).

[7] M. Neumann, N. T. Vu, Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech, arXiv preprint arXiv:1706.00612.

[8] N. Dave, Feature extraction methods lpc, plp and mfcc in speech recognition, International journal for advance research in engineering and technology 1 (6) (2013) 1–4.

[9] S. Wu, T. H. Falk, W.-Y. Chan, Automatic speech emotion recognition using modulation spectral features, Speech communication 53 (5) (2011) 768–785.

[10] S. K. Kopparapu, M. Laxminarayana, Choice of mel filter bank in computing mfcc of a resampled speech, in: 10th International Conference on Information Science, Signal Processing and their Applications (ISSPA 2010), IEEE, 2010, pp. 121–124.

[11] A. Satt, S. Rozenberg, R. Hoory, Efficient emotion recognition from speech using deep learning on spectrograms, in: INTERSPEECH, 2017, pp. 1089–1093.

[12] T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, O. Vinyals, Learning the speech front-end with raw waveform cldnns, in: Sixteenth Annual Conference of the International Speech Communication Association, 2015.

[13] C. Etienne, G. Fidanza, A. Petrovskii, L. Devillers, B. Schmauch, Cnn+ lstm architecture for speech emotion recognition with data augmentation, arXiv preprint arXiv:1802.05630.

[14] J. Zhou, R. Liang, L. Zhao, L. Tao, C. Zou, Unsupervised learning of phonemes of whispered speech in a noisy environment based on convolutive non-negative matrix factorization, Information Sciences 257 (2014) 115–126.

[15] O. Farooq, S. Datta, Phoneme recognition using wavelet based features, Information Sciences 150 (1-2) (2003) 5–15.

[16] C. M. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, S. Narayanan, Emotion recognition based on phoneme classes, in: Eighth International Conference on Spoken Language Processing, 2004.

[17] S. Jing, X. Mao, L. Chen, Prominence features: Effective emotional features for speech emotion recognition, Digital Signal Processing 72 (2018) 216–231.

[18] P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, J. Vepa, Speech emotion recognition using spectrogram & phoneme embedding., in: Interspeech, 2018, pp. 3688–3692.

[19] Z. Huang, J. Epps, An investigation of partition-based and phonetically-aware acoustic features for continuous emotion prediction from speech, IEEE Transactions on Affective Computingdoi:10.1109/TAFFC.2018.2821135.

[20] A. C. Rampinini, G. Handjaras, A. Leo, L. Cecchetti, M. Betta, E. Ricciardi, G. Marotta, P. Pietrini, Formant space reconstruction from brain activity in frontal and temporal regions coding for heard vowels., Frontiers in human neuroscience 13 (2019) 32.

[21] C. Alain, J. S. Arsenault, L. Garami, G. M. Bidelman, J. S. Snyder, Neural correlates of speech segregation based on formant frequencies of adjacent vowels, Scientific reports 7 (2017) 40790.

[22] A. J. King, S. Teki, B. D. Willmore, Recent advances in understanding the auditory cortex, F1000Research 7.

[23] R. D. Kent, H. K. Vorperian, Static measurements of vowel formant frequencies and bandwidths: A review, Journal of communication disorders 74 (2018) 74–97.

[24] S. Kakouros, O. Räsänen, Statistical learning of prosodic patterns and reversal of perceptual cues for sentence prominence., in: CogSci, 2016.

[25] S. Kakouros, O. Räsänen, 3pro–an unsupervised method for the automatic detection of sentence prominence in speech, Speech Communication 82 (2016) 67–84.

[26] S. Baumann, B. Winter, What makes a word prominent? predicting untrained german listeners' perceptual judgments, Journal of Phonetics 70 (2018) 20–38.

[27] M. Bulut, C. Busso, S. Yildirim, A. Kazemzadeh, C. M. Lee, S. Lee, S. Narayanan, Investigating the role of phoneme-level modifications in emotional speech resynthesis, in: Ninth European Conference on Speech Communication and Technology, 2005.

[28] W. E. Frankenhuis, Modeling the evolution and development of emotions., Developmental psychology 55 (9) (2019) 2002.

[29] P. E. Smaldino, Models are stupid, and we need more of them, in: Computational social psychology, Routledge, 2017, pp. 311–331.

[30] M. Swain, A. Routray, P. Kabisatpathy, Databases, features and classifiers for speech emotion recognition: a review, International Journal of Speech Technology 21 (1) (2018) 93–120.

[31] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, S. Zafeiriou, Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network, in: 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, 2016, pp. 5200–5204.

[32] Z. Aldeneh, E. M. Provost, Using regional saliency for speech emotion recognition, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2017, pp. 2741–2745.

[33] M. Hao, W.-H. Cao, Z.-T. Liu, M. Wu, P. Xiao, Visual-audio emotion recognition based on multi-task and ensemble learning with multiple features, Neurocomputingdoi:10.1016/j.neucom.2020.01.048.

[34] M. S. Hossain, G. Muhammad, Emotion recognition using secure edge and cloud computing, Information Sciences 504 (2019) 589–601.

[35] H. Meng, T. Yan, F. Yuan, H. Wei, Speech emotion recognition from 3d log-mel spectrograms with deep learning network, IEEE Access 7 (2019) 125868–125881.

[36] O. V. Verkholyak, H. Kaya, A. A. Karpov, Modeling short-term and long-term dependencies of the speech signal for paralinguistic emotion classification, Proceedings of SPIIRAS 18 (1) (2019) 30–56.

[37] D. Le, Z. Aldeneh, E. M. Provost, Discretized continuous speech emotion recognition with multi-task deep recurrent neural network., in: INTERSPEECH, 2017, pp. 1108–1112.

[38] S. Ghosh, E. Laksana, L.-P. Morency, S. Scherer, Representation learning for speech emotion recognition., in: Interspeech, 2016, pp. 3603–3607.

[39] D. Kamińska, Emotional speech recognition based on the committee of classifiers, Entropy 21 (10) (2019) 920.

[40] J. Deng, X. Xu, Z. Zhang, S. Frühholz, D. Grandjean, B. Schuller, Fisher kernels on phase-based features for speech emotion recognition, in: Dialogues with social robots, Springer, 2017, pp. 195–203.

[41] Z.-T. Liu, Q. Xie, M. Wu, W.-H. Cao, Y. Mei, J.-W. Mao, Speech emotion recognition based on an improved brain emotion learning model, Neurocomputing 309 (2018) 145–156.

[42] Y. Mei, G. Tan, Z. Liu, An improved brain-inspired emotional learning algorithm for fast classification, Algorithms 10 (2) (2017) 70.

[43] B. J. Shannon, K. K. Paliwal, A comparative study of filter bank spacing for speech recognition, in: Microelectronic engineering research conference, Vol. 41, 2003.

[44] L. Chen, W. Su, Y. Feng, M. Wu, J. She, K. Hirota, Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction, Information Sciences 509 (2020) 150–163.

[45] Z.-T. Liu, M. Wu, W.-H. Cao, J.-W. Mao, J.-P. Xu, G.-Z. Tan, Speech emotion recognition based on feature selection and extreme learning machine decision tree, Neurocomputing 273 (2018) 271–280.

[46] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen, et al., Deep speech 2: End-to-end speech recognition in english and mandarin, in: International conference on machine learning, 2016, pp. 173–182.

[47] Q. Mao, M. Dong, Z. Huang, Y. Zhan, Learning salient features for speech emotion recognition using convolutional neural networks, IEEE transactions on multimedia 16 (8) (2014) 2203–2213.

[48] S. Mirsamadi, E. Barsoum, C. Zhang, Automatic speech emotion recognition using recurrent neural networks with local attention, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2017, pp. 2227–2231.

[49] P. Li, Y. Song, I. V. McLoughlin, W. Guo, L.-R. Dai, An attention pooling based representation learning method for speech emotion recognition.

[50] C.-W. Huang, S. S. Narayanan, Attention assisted discovery of sub-utterance structure in speech emotion recognition., in: INTERSPEECH, 2016, pp. 1387–1391.

[51] Y. Xie, R. Liang, Z. Liang, L. Zhao, Attention-based dense lstm for speech emotion recognition, IEICE TRANSACTIONS on Information and Systems 102 (7) (2019) 1426–1429.

[52] S. Zhang, S. Zhang, T. Huang, W. Gao, Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching, IEEE Transactions on Multimedia 20 (6) (2017) 1576–1590.

[53] N. Hajarolasvadi, H. Demirel, 3d cnn-based speech emotion recognition using k-means clustering and spectrograms, Entropy 21 (5) (2019) 479.

[54] D. O'Shaughnessy, Linear predictive coding, IEEE potentials 7 (1) (1988) 29–32.

[55] A. Jongman, Z. Qin, J. Zhang, J. A. Sereno, Just noticeable differences for pitch direction, height, and slope for mandarin and english listeners, The Journal of the Acoustical Society of America 142 (2) (2017) EL163–EL169.

[56] E. Zwicker, H. Fastl, Just-noticeable sound changes, in: Psychoacoustics, Springer, 1999, pp. 175–201.

[57] G. Lemaitre, P. Susini, Timbre, sound quality, and sound design, in: Timbre: Acoustics, Perception, and Cognition, Springer, 2019, pp. 245–272.

[58] S.-A. Lembke, S. McAdams, The role of spectral-envelope characteristics in perceptual blending of wind-instrument sounds, Acta Acustica united with Acustica 101 (5) (2015) 1039–1051.

[59] A. C. Elkins, Vocalic markers of deception and cognitive dissonance for automated emotion detection systems.

[60] K. M. Liew, Meaningful noise: auditory roughness and dissonance predict emotion recognition and cross-modal perception, Ph.D. thesis (2018).

[61] J. H. McDermott, A. F. Schultz, E. A. Undurraga, R. A. Godoy, Indifference to dissonance in native amazonians reveals cultural variation in music perception, Nature 535 (7613) (2016) 547.

[62] A. J. Oxenham, How we hear: The perception and neural coding of sound, Annual review of psychology 69 (2018) 27–50.

[63] B. H. Story, K. Bunton, Relation of vocal tract shape, formant transitions, and stop consonant identification, Journal of Speech, Language, and Hearing Research.

[64] B. Prica, S. Ilić, Recognition of vowels in continuous speech by using formants, Facta universitatis-series: Electronics and Energetics 23 (3) (2010) 379–393.

[65] D. Sculley, Web-scale k-means clustering, in: Proceedings of the 19th international conference on World wide web, 2010, pp. 1177–1178.

[66] B. Bahmani, B. Moseley, A. Vattani, R. Kumar, S. Vassilvitskii, Scalable k-means++, arXiv preprint arXiv:1203.6402.

[67] P. J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, Journal of computational and applied mathematics 20 (1987) 53–65.

[68] T. Caliński, J. Harabasz, A dendrite method for cluster analysis, Communications in Statistics-theory and Methods 3 (1) (1974) 1–27.

[69] D. L. Davies, D. W. Bouldin, A cluster separation measure, IEEE transactions on pattern analysis and machine intelligence (2) (1979) 224–227.

[70] 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, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[71] Z.-T. Liu, B.-H. Wu, D.-Y. Li, P. Xiao, J.-W. Mao, Speech emotion recognition based on selective interpolation synthetic minority oversampling technique in small sample environment, Sensors 20 (8) (2020) 2297.

[72] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss, A database of german emotional speech, in: Ninth European Conference on Speech Communication and Technology, 2005.

[73] S. R. Livingstone, F. A. Russo, 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.

[74] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, S. S. Narayanan, Iemocap: Interactive emotional dyadic motion capture database, Language resources and evaluation 42 (4) (2008) 335.

[75] O. M. Nezami, P. J. Lou, M. Karami, Shemo: a large-scale validated database for persian speech emotion detection, Language Resources and Evaluation 53 (1) (2019) 1–16.

[76] E. Parada-Cabaleiro, G. Costantini, A. Batliner, M. Schmitt, B. W. Schuller, Demos: an italian emotional speech corpus, Language Resources and Evaluation (2019) 1–43.

[77] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, E. M. Provost, Msp-improv: An acted corpus of dyadic interactions to study emotion perception, IEEE Transactions on Affective Computing 8 (1) (2016) 67–80.

[78] J. R. Landis, G. G. Koch, The measurement of observer agreement for categorical data, biometrics (1977) 159–174.

[79] Z. Zhao, Y. Zhao, Z. Bao, H. Wang, Z. Zhang, C. Li, Deep spectrum feature representations for speech emotion recognition, in: Proceedings of the Joint Workshop of the 4th Workshop on Affective Social Multimedia Computing and first Multi-Modal Affective Computing of Large-Scale Multimedia Data, 2018, pp. 27–33.

[80] F. Daneshfar, S. J. Kabudian, A. Neekabadi, Speech emotion recognition using hybrid spectral-prosodic features of speech signal/glottal waveform, metaheuristic-based dimensionality reduction, and gaussian elliptical basis function network classifier, Applied Acoustics 166 (2020) 107360.

[81] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, A. Wendemuth, Acoustic emotion recognition: A benchmark comparison of performances, in: 2009 IEEE Workshop on Automatic Speech Recognition & Understanding, IEEE, 2009, pp. 552–557.

[82] S. B. Alex, B. P. Babu, L. Mary, Utterance and syllable level prosodic features for automatic emotion recognition, in: 2018 IEEE Recent Advances in Intelligent Computational Systems (RAICS), IEEE, 2018, pp. 31–35.

[83] H. M. Fayek, M. Lech, L. Cavedon, Evaluating deep learning architectures for speech emotion recognition, Neural Networks 92 (2017) 60–68.

[84] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Epps, Direct modelling of speech emotion from raw speech, Proc. Interspeech 2019 (2019) 3920–3924.

[85] J. Lee, I. Tashev, High-level feature representation using recurrent neural network for speech emotion recognition, in: Sixteenth annual conference of the international speech communication association, 2015.

[86] Y. Zeng, H. Mao, D. Peng, Z. Yi, Spectrogram based multi-task audio classification, Multimedia Tools and Applications 78 (3) (2019) 3705–3722.

[87] J. Gideon, M. McInnis, E. Mower Provost, Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (addog), IEEE Transactions on Affective Computing (2019) 1–1doi:10.1109/TAFFC.2019.2916092.

[88] S. Latif, J. Qadir, M. Bilal, Unsupervised adversarial domain adaptation for cross-lingual speech emotion recognition, in: 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE, 2019, pp. 732–737.

[89] J. Parry, D. Palaz, G. Clarke, P. Lecomte, R. Mead, M. Berger, G. Hofer, Analysis of deep learning architectures for cross-corpus speech emotion recognition, Proc. Interspeech 2019 (2019) 1656–1660.

[90] M. Neumann, N. T. Vu, Improving speech emotion recognition with unsupervised representation learning on unlabeled speech, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 7390–7394.

[91] F. Bao, M. Neumann, N. T. Vu, Cyclegan-based emotion style transfer as data augmentation for speech emotion recognition, 2019 ISCA (2019) 35–37.