Information Sciences 563 (2021) 309–325

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

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], 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

Article history:
Received 17 February 2020
Received in revised form 6 February 2021
Accepted 12 February 2021
Available online 19 February 2021

Keywords:
Speech
Emotion recognition
Formants extraction
Phonemes
Clustering
Cross-corpus

ABSTRACT

Speech Emotion Recognition (SER) has numerous applications including human-robot interaction, online gaming, and health care assistance. While deep learning-based approaches achieve considerable precision, they often come with high computational and time costs. Indeed, feature learning strategies must search for important features in a large amount of speech data. In order to reduce these time and computational costs, we propose pre-processing step in which speech segments with similar formant characteristics are clustered together and labeled as the same phoneme. The phoneme occurrence rates in emotional utterances are then used as the input features for classifiers. Using six databases (EmoDB, RAVDESS, IEMOCAP, ShEMO, DEMoS and MSP-Improv) for evaluation, the level of accuracy is comparable to that of current state-of-the-art methods and the required training time was significantly reduced from hours to minutes.

© 2021 Elsevier Inc. All rights reserved.

1. Introduction

Human speech carries many different signals that convey important affective information. Speech can be colored by emotions that have obscure differences and adjectives that define emotions can vary by meaning and intensity for different individuals. Consequently, there is always uncertainty in emotional annotations as well as their respective speech signals. Although computers are getting better at understanding human communication, some of these affective signals are unintelligible to computers, and indeed, humans cannot always validate them with certainty. This makes Speech Emotion Recognition (SER) an area of research gaining increased attention in the Human–Computer Interaction (HCI) field.

An SER system predicts the type and/or intensity of emotion being conveyed in speech signals. This is usually achieved by first extracting some useful features (e.g., pitch) 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 a few hundreds or thousands of annotated examples of speech signals. Therefore, the availability of speech

☆ 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.

* Corresponding author at: School of Automation, China University of Geosciences, Wuhan 430074, China.
E-mail addresses: liuzhentao@cug.edu.cn (Z.-T. Liu), abdulrehman@cug.edu.cn (A. Rehman), wumin@cug.edu.cn (M. Wu), weihuacao@cug.edu.cn (W.-H. Cao), haoman@cug.edu.cn (M. Hao).

https://doi.org/10.1016/j.ins.2021.02.016
0020-0255/© 2021 Elsevier Inc. All rights reserved.

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, there could be up to 24 unique emotional labels. The higher the number of classes, the higher the chances of overlapping definitions of emotional labels, which increases the level of complexity for classifiers.

Major challenges reported for SER systems are robustness and autonomy (i.e., an SER system needs to adapt to different speakers). Due to constraints in feature definitions, one model that works well for one group of speakers performs poorly for another group. The training of a spectrogram based on deep learning models can take a few hours or up to 14 days of training data. Then if a new training data arrives (in the case of active learning or speaker dependent learning), the model needs to be retrained again which can again take another few hours. Part of the reason for the large required input size is the importance of modulation along the feature temporal axis. Indeed, features such as Mel Frequency Cepstral Coefficients (MFCCs) are difficult to condense into fewer variables because each contextual frame has individual importance, therefore all frames of MFCCs or Mel-filter energies 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 labeled phoneme type for each contextual frame of 10–50 ms by K-means clustering. We aim to narrow the distinguishable features to only a few components such that a classifier achieves 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 pre-processing of speech signals 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 for different emotional classes are selected as input features for the classifiers. This whole process takes 1–25 min of training and matches the accuracy of other state-of-the-art methods.

Phonemes are the smallest phonetic units which are often labeled as the parts of words or syllables during lexical recognition of speech. Formants are one of many ways to characterize a phoneme. Formants are the several frequency regions of relatively great intensity in a sound spectrum caused due to acoustic resonance of the human vocal tract. Cues in speech signal 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 nuances 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.

Our approach to 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 similar sounding words for emotional symbols [1]. 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 [2]. The first phoneme of a word is shown to have a higher importance for predicting a word's emotional valance and lexical meaning regardless of the phonetic features (e.g., intonation, tone, stress, and rhythm) [3]. Similarly, particular phonological units are shown to convey basic affective tone of poems to readers [4]. 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; first, 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) and, second, clustering of phonemes using 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 fewer but compact variables thus speed up the training process and increasing the generalizability. Our method actively refines speech data by focusing on 12 specific formant characteristics, while 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. Different systems vary in the types of speech signal features and the method of intermediate mappings. Deep Neural Network (DNN) use spectrograms or periodograms as the pre-processing input [5], whereas machine learning classifiers prefer 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). Formant based features, such as Linear Predictive Coding (LPC) coefficients, are quite helpful to consolidate speech information into a few variables, and therefore widely used to detect emotion and phoneme cues [6]. MFCCs are a popular features for detecting many different speech cues and recognizing emotion in speech. MFCCs are derived from Mel spectrograms [7], and some studies have suggested that using log-Mel spectrograms as CNN input creates a more efficient architecture [8]. While phoneme based methods are widely used for Automatic Speech Recognition (ASR) [9], there are relatively few examples of phoneme based methods for emotion recognition. SERs have used a variety of inputs including phoneme-class specific HMM models [10], combinations of acoustic features and occurrence rates of prominent syllables [11], and combinations of phoneme labels with spectrograms [12]. These studies had better performance for phonetically aware methods. Indeed, a phonetically aware acoustic feature set was also shown to make significant improvement in emotional arousal and valence recognition [13].

Most SER research explores different AI tools such as convolution neural network (CNN) [14], recurrent neural network (RNN) [15,16], and single or multiple machine classifiers [17]. A few brain-inspired SER models have also been proposed and work similarly to the human limbic system [18]. By far, CNN is the preferred choice. While research has mostly focused on the front-end of the classification model, back-end speech signal preprocessing methods are borrowed from decades old speech recognition methods [19].

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 [20]. Similarly, an extreme learning machine was used to classify emotions using a fusion of personalized and non-personalized speech features [21,22].

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. Latency and accuracy of SER can be improved by cleaning the input signal from noise [8] or focusing the attention on salient regions of the spectrogram by learning the prominent discriminant regions [23]. Attention-based feature learning has shown improvements in recognition accuracy of LSTM architectures [24].

3. Phoneme Type Convergence Method

In traditional SER methods, spectrograms and/or audio features are used as the inputs for a deep learning classifier, generally by segmenting the audio signal into 20–150 ms frames. The RNN based neural networks use LSTM or a similar technique to learn the temporal order of frames, but there are a few other methods to learn the temporal cues (e.g., pyramid matching by [25]). While there are thousands of audio features, usually only a few of them are used, most noticeably are MFCCs and LPC. There are typically 20 to 100 audio features 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 [26]. Spectrogram based methods use image sizes between 2–65 KB 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 this traditional approach, a phoneme type convergence method would reduce the Mel-frequency cepstrum to only 12 variables per frame. Feature information per frame is further reduced to <10 variables by generating phoneme labels for the frames based on the similarity of formant characteristics. This size reduction not only helps process large amounts of data, but also helps in focusing the attention on the useful information and discarding the rest. An overview of the phoneme type convergence method is shown in Fig. 1.

Clustering of phonemes is proposed in order 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 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. Since phonemes come in different shapes and sizes, the phoneme cluster types are mainly divided into two categories: 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.

Formant characteristics are scaled according to importance of phoneme perception. For example, a small difference in fundamental frequency changes the phoneme label, while a relatively 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 related work exploring *just noticeable differences* in hearing perception [27]. These scaling factors ensure that the similarity level of phoneme cluster members is 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 speech with the minimum number of variables. Speech usually has two or more formants which carry most of the energy, while 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

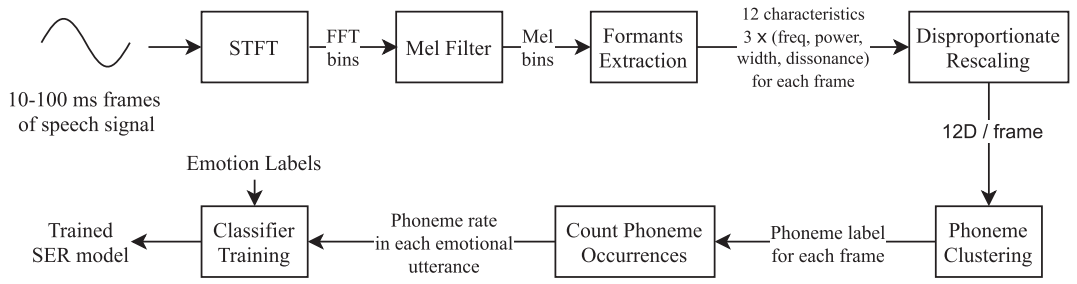

Fig. 1. Overview of the proposed phoneme type convergence method.

music, the emotional quality of speech is conveyed partially through the differences in timbre. 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, measuring the most defining components of the sound provides enough information to quantitatively judge timbre. Therefore, according to the proposed method, four variables are taken as the 12 defining characteristics of the signal window: (1) frequency, (2) power (amplitude), (3) width, and (4) dissonance of the top three formants with the highest amplitude.

The process of extracting the characteristic features starts with dividing the speech signal into a few millisecond frames such that each frame has minimum variation within its time-frame. 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

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

where s_t is the input signal of frame t, x_t is the windowed frame, $0 \leqslant n \leqslant W-1$, 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, and 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 its similarity with human perception. The Mel-filter banks have a triangular shape and gets wider as the frequency increases. The Mel scale frequency can be converted to Hertz scale by

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

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 frequencies have the highest magnitude as compared to the rest of the 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 the top three Mel-filter banks with the highest magnitude as the top three formants. Formants 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

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

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. Fig. 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 (the highest peak between two valleys) are ordered by rank as

$$\phi_h \in \left\{\phi_0, \phi_1, \phi_2, \ldots c, \phi_{N_h - 1}\right\} \tag{5}$$

where ϕ_h gives the index l of mel-filter bank that is ranked h based on its 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

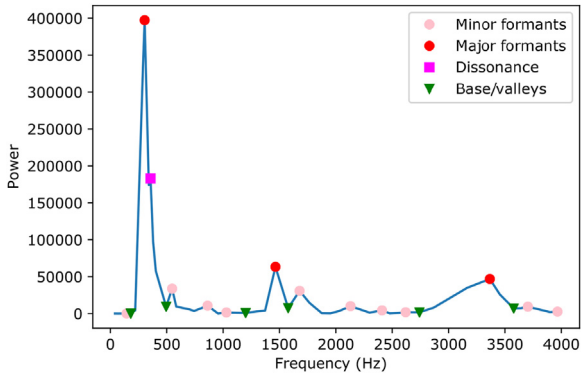

Fig. 2. A sample of a 25 ms window frame showing the power of 256 bins of Mel-filter at their central frequencies (converted from Mel-scale to Hz). Formants (the highest local maxima) are separated by valleys (local minima) in between them. Other peaks within the peak-valley threshold around the highest 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.

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 $k \leqslant 1$ is the decay constant that decreases the scale of power for each proceeding formant ranked from highest to lowest in power. Similarly, the frequency of formants are rescaled again using

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

where f_h is the rescaled central frequency, and $f_c(\phi_h)$ is Hertz-scale central frequency of formant h of the current frame. Rescaling increases the scaling ratio of each predeceasing formant over its succeeding (lower in rank) formant such that the higher ranking formant has a larger component in the Euclidean distance when 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 the consonance within it. Formant width can be guessed by the shape of the instrument that produced it. For example, small and narrow instruments (i.e., flutes or clarinets) have narrow formant widths whereas bigger instruments (i.e., horns or bassoons) have wider formant widths [28]. Similar to formant frequency and power, formant width w_h is calculated and rescaled as

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

where f_{h,v_0} and f_{h,v_1} are the amplitude local minima points (valleys) on the lower side and the higher side of the current frame's formant h on the Hertz frequency scale, respectively. There is a threshold condition for the local minima to be considered as valleys (i.e., their amplitude should be lower than a 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 [29]. A sound or voice is perceived as pleasant when there is a higher ratio of harmonic components to non–harmonic components. Dissonance does not occur regularly around major formants, but it is recorded as a feature since it can carry important emotional signals. If the signal is noisy, this method will not guarantee a valid measurement of unpleasantness.

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 temporal 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 $\phi_{h,j} \in \left\{\phi_{h,0}, \phi_{h,1}, \phi_{h,2}, \ldots c, \phi_{h,M_h-1}\right\}$ 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 is an important indicator of temporal cues. These features include the frequency difference and the power difference between the first and second formants. The number of adjacent frames can be varied from 6 to 12 depending on the context frame size, with best results acquired at a 25 ms window. Therefore, a minimum of 60 ms (i.e., 6 frames at 10 ms 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 of the frames and the second half of the frames, one can measure the transitions that have happened in that window. 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

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

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

where $0 \leqslant t \leqslant T-g, g$ 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

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

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

Eqs. (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. While only three characteristic features, the frequencies of first two formants (f_{t0},f_{t1}) and power of first formant (p_{t0}) are recommended 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 the differential feature values are more similar. However, this difference is crucial for detecting the transitional gradient in phonemes.

3.2. Phoneme clustering

Different permutations of the 12 formant characteristic features can create innumerable varieties of phonemes. Therefore, we limit the scope of phoneme heterogeneity by clustering according to the similarity of formant features. However, 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 have a drastic drop in power after the second or third formant [30]. 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 clusters. Fig. 3 shows the 16 clusters created from formant characteristic features of more than 67 thousand frames 25 ms long window. 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. (See Fig. 4).

Another group of clustering models is created using the differential features described in Section 3.1.3. The number of clusters (N_m) for both groups can be varied from 16 to 128 depending on the size and variability of the dataset or a set of different cluster sizes can be used together. The size parameters of clustering are a function of phonetic similarity. We recommend using a set of clustering models covering sizes from 16 to 128 such that large cluster models group sparingly similar frames together whereas small cluster models differentiate between phonemes at a finer 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. Other algorithms such as Agglomerative and BIRCH clustering were pretested for method selection, but did not show any improvement in the accuracy of the whole model and instead actually increased the duration of training. BIRCH clustering did not distribute the clusters mem-

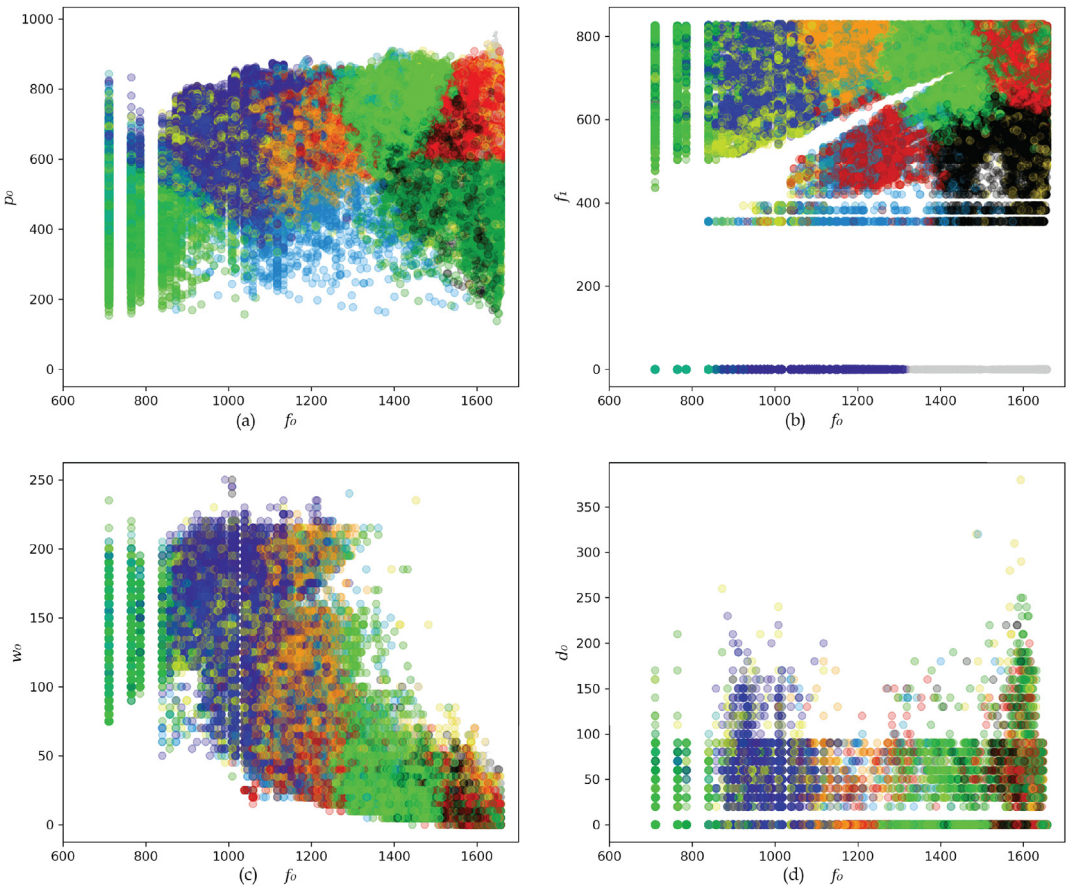

Fig. 3. Scatter plots of clustering data points across 5 out of total 12 dimensions extracted from EmoDB. Different colors of points represent the cluster labels of total 16 clusters created by K-Means clustering (parameters same as No. 1 in Table 1). All features are scaled differently according to 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.

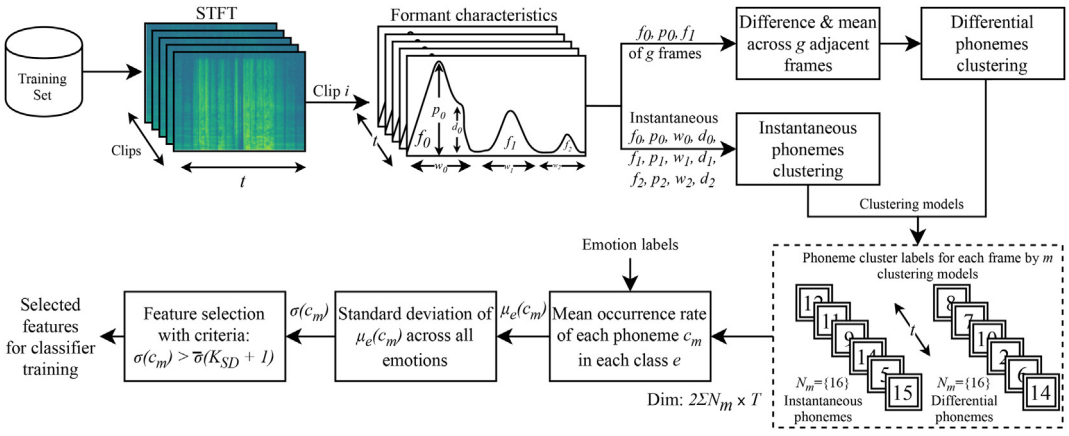

Fig. 4. An overview of the feature extraction and selection process. First, formant based characteristic features are extracted from raw WAV files, then at least 2 types of phonemes are created using K-means clustering. Then cluster labels, regarded here as phoneme labels, are counted to calculate the 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.

bers uniformly, but rather all members were clustered together in one or two dense groups. Other hierarchical models (Agglomerative Clustering) were not able to converge on larger datasets due to memory restraints. Another advantage of K-means clustering is the batch-by-batch processing of the dataset, this allows it to be used in active applications for

real-time updates of the classification model for big databases. Table 1 gives Silhouette coefficient (S_{Sil}), Calinski-Harabasz index (S_{CH}), and Davies-Bouldin index (S_{DB}) as metrics of the clustering quality of K-means and Agglomerative clustering algorithms with different parameters. K-Means clustering performed better than Hierarchical clustering when comparing computation time and Calinski-Harabasz index (S_{CH} is higher when clusters are dense and well separated). 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 judgment 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 all utterances in the dataset are calculated. Only a few of these phonemes are selected as input features for a classifier based on the standard deviation of the occurrence rate across all emotion classes in the 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 $0 \leqslant c_m < N_m, N_m$ 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 the 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

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

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. Eq. (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 the 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 very large and there is a significant imbalance, then all phonemes are more likely to occur in classes with the largest sample size [32]. If one class has comparatively more samples, then the phoneme clustering is likely to be biased towards that class, meaning a high number of phonemes will occur 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. Eq. (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.

Table 1
Computation time, Silhouette coefficient (S_{Sil}), Calinski-Harabasz index (S_{CH}) and Davies-Bouldin index (S_{DB}) as metrics of clustering quality of K-Means and Hierarchical clustering algorithms [31] 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'

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 experiments are carried out in Python 3.7.4 (64-bit) environment on a computer with 64-bit Windows 10 OS 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 [33], Ryerson audiovisual database of emotional speech and song (RAVDESS) [34], IEMOCAP (Interactive emotional dyadic motion capture database) [35], Sharif Emotional Speech Database (ShEMO) [36], DEMoS (Database of Elicited Mood in Speech) [37] and MSP-Improv [38]. 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 provided in Table 3.

Emotions in most of these databases were acted, but the data collectors have tried their best to verify the naturalness ([33]) or genuineness ([34]) 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, 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% [33].

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, Cohen's kappa was 0.64 among 12 annotators [36].

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 had at least 67% agreement among raters. In IEMOCAP, while 1,405 (out of 2,943) utterances met this criterion, we used all 2,943 utterances to keep the experimentation samples similar to the other comparative works [39,40].

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) the effect of change in feature selection parameter K_{SD} on the 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 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 in two or more clustering models and their respective generated phoneme labels for each frame of each clip. Training and testing of the dataset is split according to the validation scheme before clustering and only the training set is used for creating clustering models.

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

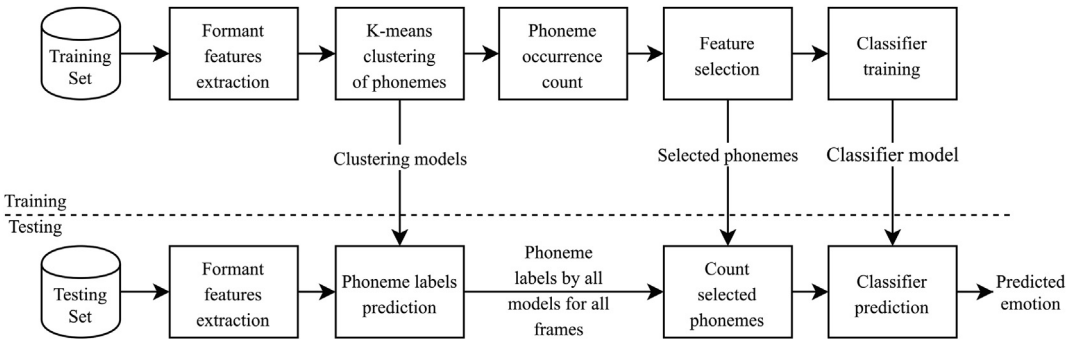

Fig. 5. Overview of the experiment pipeline.

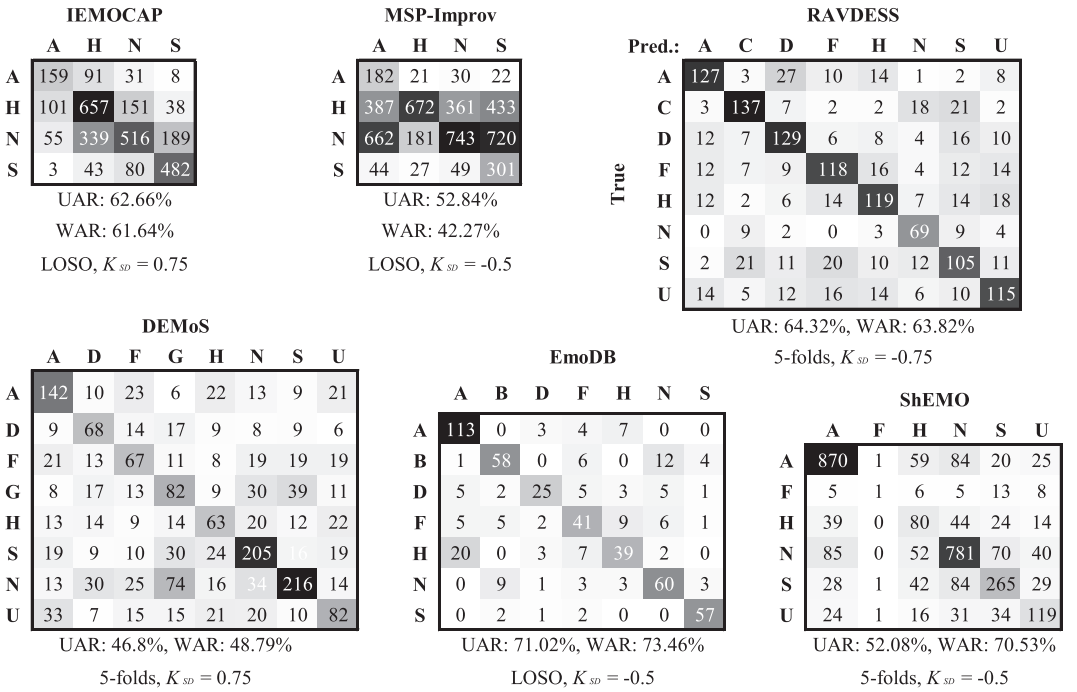

Fig. 6. Confusion matrices of emotion recognition using phoneme clustering models $N_m = \{16, 32, 64, 128\}$ and SVM classifier. Emotion labels are given in Table 3.

Table 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)	5 M:5F	12 M:12F	5 M:5F	56 M:31F	38 M:21F	6 M:6F
Environment	Script	Script	Improv	Radio drama	Elicited	Improv
Emotions	7	8	4	6	8	4
Total duration	24.8 m	1 h 16 m	3 h 28 m	3 h 25 m	1 h 25 m	5 h 7 m
Trimmed duration	23.5 m	46 m	3 h 26 m	3 h 19 m	1 h 23 m	5 h 3 m
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	[1]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		

[1] Excitement considered as happiness.

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 s	16 s	52 s	49 s	15 s	1.4 min
Occurrence count & feature selection	3 s	7 s	50 s	20 s	6 s	59 s
SVM training	1 s	2 s	4 s	2 s	2 s	5 s

Step 3. Phoneme occurrence rates are calculated in each emotion class of the 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 the cluster size model comparisons (Fig. 7), for both 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 three metrics: unweighted Average Recall (UAR), number of classifier input features, and training duration. Confusion matrices, UAR and WAR (Weighted Average Recall or accuracy), of within corpus validation are shown in Fig. 6. The training duration of each step is shown in Table 4. In the following subsections, we review the results from each step of the proposed method.

5.2.1. Formant features extraction

The top three formants are preferred because the proposed method had a slightly less accuracy for one or two formants based features, therefore only three formants 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, which was 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 a 16-bit unsigned integer to store formant characteristics, but the maximum value (after scaling according to Section 3.1) was within range of 1,000 for all formant characteristic features, which means there is more room for compression. Moreover, if a marginal increase in UAR is not an important factor for method selection, then the model with half the phoneme cluster size is comparable in performance to the model with the highest UAR (i.e., $N_m = \{16, 32, 64, 128\}$).

5.2.2. Phoneme clustering

Phoneme cluster sizes (N_m) ranging from 16 to 128 were used for experiments. 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 as shown in Fig. 7. The best results were achieved by using a set of cluster sizes for both instantaneous and differential phonemes with

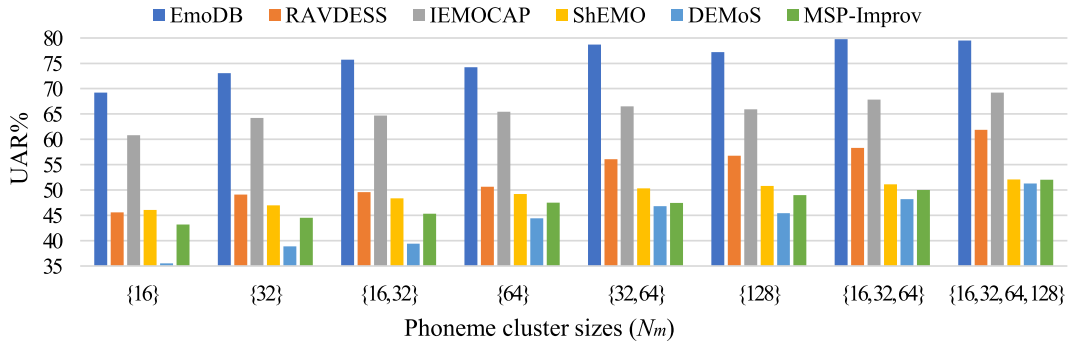

Fig. 7. SVM prediction UAR using different phoneme cluster size sets (N_m). Parameter of feature selection is constant at $K_{SD} = -0.5$ for all databases.

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

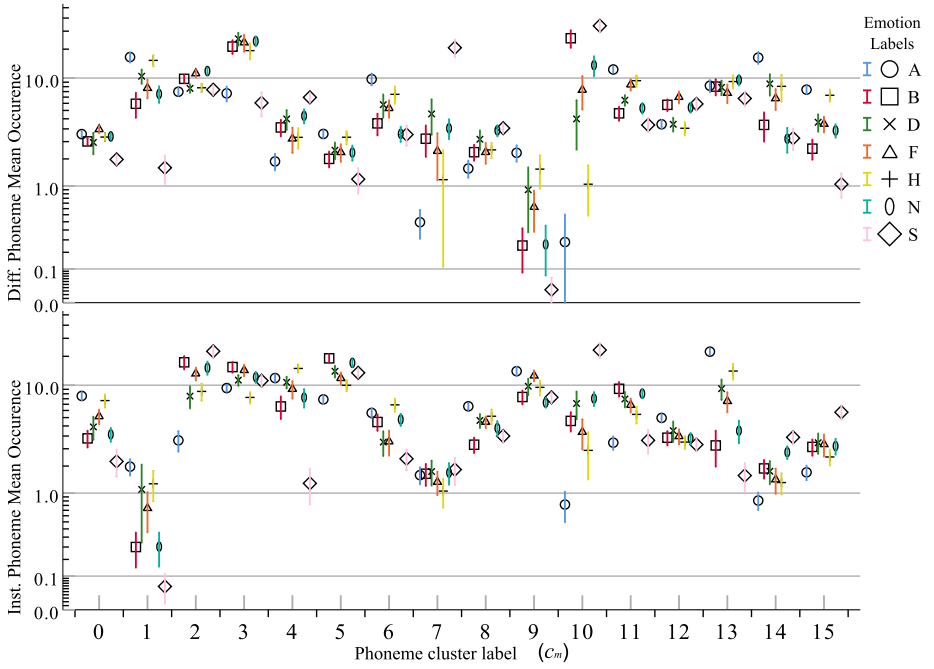

Fig. 8. Mean occurrence rate of phonemes per 100 frames in EmoDB. Error bars confidence interval is 95%. $N_m = \{16\}$, i.e., 16 clusters for both instantaneous and differential types. Legends represent emotions as A = anger, B = boredom, D = disgust, F = fear, H = happy, N = neutral, and S = sad. Phonemes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

$N_m = \{16, 32, 64, 128\}$. The difference between the lowest ($N_P = 32$) and highest ($N_P = 480$) number of clusters was less than 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 increased 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.

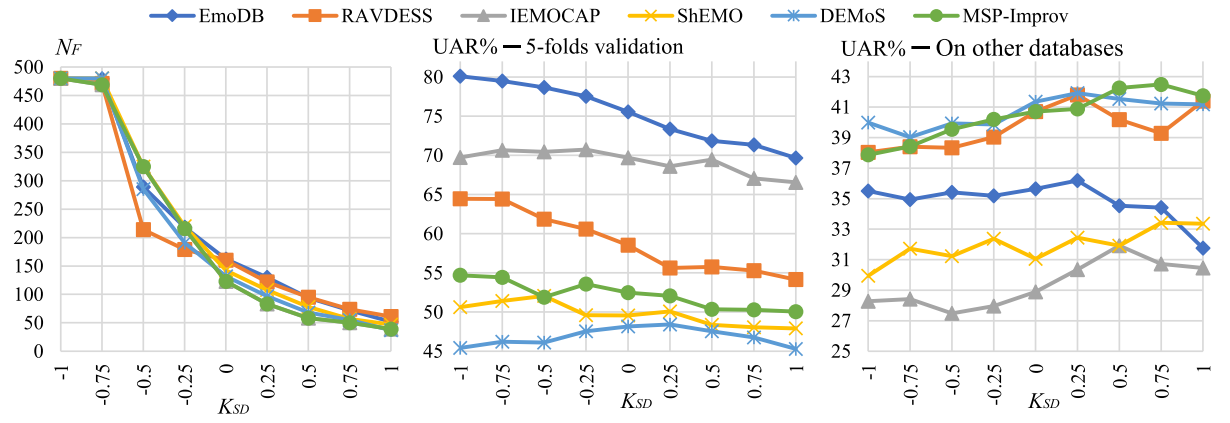

Fig. 9. These plots show the effects of parameter K_{SD} on the average number of selected features (N_F out of total 480), validation UAR% and the cross-corpus UAR%. For cross-corpus tests, models are trained on one dataset and tested for robustness on other 5 datasets using only the 4 common emotions.

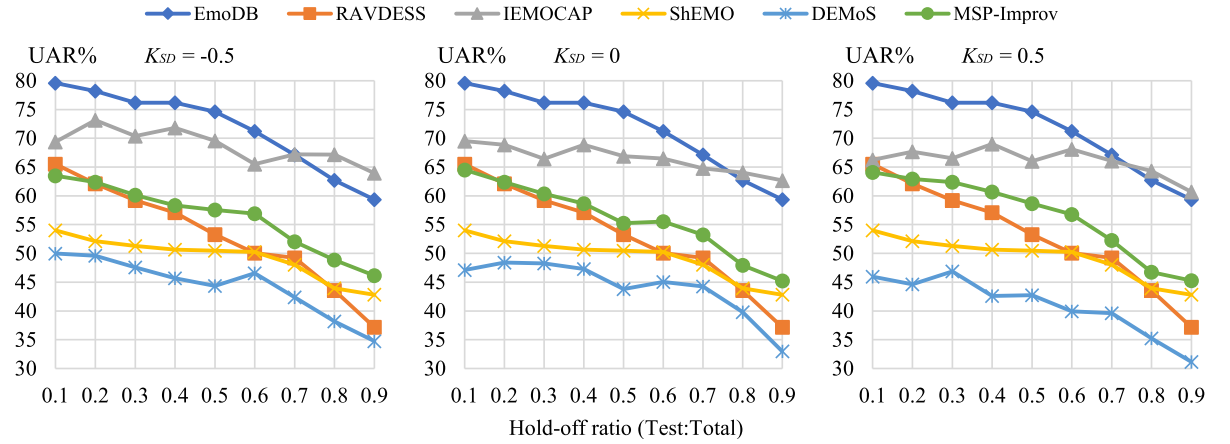

Fig. 10. The effect of decreasing the training set size (and increasing the testing set size) on the validation UAR of models with more features ($K_{SD} = -0.5$) verses models with fewer features ($K_{SD} = 0.5$) shows that the number of features has bigger effect on the UAR for databases with more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

5.2.3. 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. For the 16 clusters extracted from training set (80%) of EmoDB, Fig. 8 shows the phoneme occurrence rate for different emotions. 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 the case of Fig. 8). These occurrence rates of selected phonemes are then used as the input features for SVM training. The number of selected features is dependent on parameter K_{SD}, which has different optimum values for different datasets. The plots in Fig. 9 show the change in UAR and the number of selected features (N_F) as the number K_{SD} changes from -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 small change given the large 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 [31] (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. When the classifier comparison experiment was performed, 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-

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. Increase in the number of features N_F shows a decrease in 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

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. Fig. 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.

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 Fig. 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 the number of speakers, emotions, recordings, the variation of sentences, naturalness, and inter-rater agreement. According to all these measures, the RAVDESS database is far more complex than all other databases, indicated by the steeper slope for UAR in Fig. 10. But the number of samples per label in RAVDESS are less than IEMOCAP which performs better at generalization due to the increased variance within the same labels.

5.3.1. Generalizability evaluation

The relation between generalization capability and the 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) shows relatively little 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 decrease 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. For IEMOCAP database, the relatively small change in UAR is an indicator that the model doesn't learn drastically more from a large portion 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 and overlapping sample spaces between the training and testing sets are given. As we increased the number of features (N_F), the UAR for test with mutually exclusive sets decreases, but the UAR for 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 the number of features per utterance is a big factor for regularization. Fewer features imply a higher sensitivity to those selected, 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 the training set, the selected phonemes become less domain dependent and, consequently, perform better on out of domain corpora, meaning the UAR was higher when we

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	[41]	All 7	LLDs, SVM	LOSO	84.6	85.6	N/A	>6 K/u
EmoDB	[25]	All 7	CNN	LOSO	**86.3**	87.3	4096/t	N/A
EmoDB	[40]	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	[42]	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	[43]	A,H,N,S	CNN, LSTM	LOSO	60.89	64.78	1600/t	N/A
IEMOCAP	[39]	A,H,N,S	CNN, LSTM	LOSO	**68.0**	65.0	16 K/t	N/A
IEMOCAP	[44]	A,H,N,S	CNN, LSTM	LOSO	60.23	N/A	16 K/t	N/A
IEMOCAP	[40]	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	[45]	A,H,N,S	LLDs, BLSTM	5-folds	63.9	62.8	32/t	N/A
IEMOCAP	[8]	A,H,N,S	CNN, LSTM	5-folds	59.4	68.8	60 K/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	[46]	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	[36]	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**

trained the model on a variety of datasets (and languages) and tested on EmoDB. However, when we tested the models on IEMOCAP, the UAR was lower than when we trained the model with a variety of datasets (and languages) as compared to training only on MSP-Improv, meaning the variety in languages can have the opposite effect.

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. The UAR of our method was not drastically different from other state-of-the-art methods, but there is a big difference in the number of features used as inputs for 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 inputs using deep learning. Indeed, a deep spectrum features learning strategy performed slightly better than low-level descriptors (LLDs) and SVM based models [39]. These methods appear to have given the best UAR of 68% for IEMOCAP dataset. Quantum-behaved particle swarm optimization (QPSO) has also been proposed for dimension reduction, and increased the accuracy of Gaussian elliptical basis function (GEBF) neural network classifier from 74.55% to 79.94% for EmoDB [40].

Related works on SER rarely report the associated training time and computational costs. For some perspective, one evaluation reported a training duration of 2 to 14 days for a deep neural network using spectrogram inputs extracted from the IEMOCAP dataset [43]. Recently, Daneshfar et al. reported that despite the higher accuracy of their method, high computational costs and low convergence speed were the major drawbacks for their method [40]. In contrast, our method finishes training within minutes. The prediction latency is also not limited by any computational constraints other than the length of speech itself. The average utterance length for all datasets was less than 5 s (see Table 2), which means a latency of 5 s should be expected from the proposed method.

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 [47,48]. An analysis of LSTM, CNN, and CNN-LSTM was given where deep learning architectures were trained using the 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) [48]. Gideon et al. proposed an adversarial discriminative domain generalization method which was trained on the MSP-IMPROV dataset and tested on IEMOCAP (UAR $47 \pm 1\%$) for recognizing three valance classes [47], and their model made 1.42 times more correct predictions compared to random chance. Although not comparable, our model makes 1.84 times more correct predictions compared to random chance for four categorical classes in IEMOCAP. Recently, some studies have proposed adversarial networks that were trained on IEMOCAP and tested on the MSP-IMPROV dataset, autoencoding and attentive CNN [49], and CycleGAN (Cycle-consistent generative adversarial networks) [50], both of which gave UAR figures of $45 \pm 1\%$.

As shown in Fig. 10, when the hold-off ratio is 90% (i.e., testing set is nine times the training set) adding more features (by decreasing K_{SD}) does not 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 the 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 decrease the time taken with the feature finding process by directly looking at formant 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 reduce the computational cost of the overall SER model. We used Mel-filter banks to extract each frame's formant characteristics using the phoneme labels that were generated by K-means clustering. Then the occurrence count of the 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 a UAR of 62% on the IEMOCAP database and 71% on EmoDB with fewer features and a training time of just a few minutes, but still yielding a computational-friendly SER system with 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 improve the current method to effectively minimize supervision in noisy and unpredictable environments. In the current model, we tried to ensure 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 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.

CRediT authorship contribution statement

Zhen-Tao Liu: Validation, Writing - review & editing, Supervision, Resources, Data curation, Project administration. **Abdul Rehman:** Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing - original draft, Visualization. **Min Wu:** Supervision, Funding acquisition. **Wei-Hua Cao:** Validation, Resources. **Man Hao:** Validation, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] E. Finegan, Language: Its structure and use, Cengage Learning, 2014.
[2] V. Slavova, Towards emotion recognition in texts–a sound-symbolic experiment, Int. J. Cognitive Res. Sci., Eng. Educ. 7 (2) (2019) 41–51.
[3] J.S. Adelman, Z. Estes, M. Cossu, Emotional sound symbolism: Languages rapidly signal valence via phonemes, Cognition 175 (2018) 122–130.
[4] 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..
[5] 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..
[6] N. Dave, Feature extraction methods lpc, plp and mfcc in speech recognition, Int. J. Adv. Res. Eng. Technol. 1 (6) (2013) 1–4.
[7] 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..
[8] A. Satt, S. Rozenberg, R. Hoory, Efficient emotion recognition from speech using deep learning on spectrograms, INTERSPEECH (2017) 1089–1093.
[9] 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, Inf. Sci. 257 (2014) 115–126.
[10] 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.
[11] S. Jing, X. Mao, L. Chen, Prominence features: Effective emotional features for speech emotion recognition, Digital Signal Processing 72 (2018) 216–231.
[12] P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, J. Vepa, Speech emotion recognition using spectrogram & phoneme embedding, Interspeech (2018) 3688–3692.
[13] Z. Huang, J. Epps, An investigation of partition-based and phonetically-aware acoustic features for continuous emotion prediction from speech, IEEE Trans. Affective Computing doi:10.1109/TAFFC.2018.2821135..
[14] 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, Neurocomputing 391 (2020) 42–51.
[15] O.V. Verkholyak, H. Kaya, A.A. Karpov, Modeling short-term and long-term dependencies of the speech signal for paralinguistic emotion classification, Proc. SPIIRAS 18 (1) (2019) 30–56.

[16] A. Rehman, Z.-T. Liu, D.-Y. Li, B.-H. Wu, Cross-corpus speech emotion recognition based on hybrid neural networks, in, 39th Chinese Control Conference (CCC), IEEE 2020 (2020) 7464–7468.
[17] D. Kamińska, Emotional speech recognition based on the committee of classifiers, Entropy 21 (10) (2019) 920.
[18] 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.
[19] B.J. Shannon, K.K. Paliwal, A comparative study of filter bank spacing for speech recognition, in: Microelectronic engineering research conference, Vol. 41, 2003..
[20] 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, Inf. Sci. 509 (2020) 150–163.
[21] 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.
[22] Z.-T. Liu, A. Rehman, M. Wu, W. Cao, M. Hao, Speech personality recognition based on annotation classification using log-likelihood distance and extraction of essential audio features, IEEE Transactions on Multimedia (2020), https://doi.org/10.1109/TMM.2020.3025108.
[23] Q. Mao, M. Dong, Z. Huang, Y. Zhan, Learning salient features for speech emotion recognition using convolutional neural networks, IEEE Trans. Multimedia 16 (8) (2014) 2203–2213.
[24] Y. Xie, R. Liang, Z. Liang, L. Zhao, Attention-based dense lstm for speech emotion recognition, IEICE Trans. Inform. Syst. 102 (7) (2019) 1426–1429.
[25] S. Zhang, S. Zhang, T. Huang, W. Gao, Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching, IEEE Trans. Multimedia 20 (6) (2017) 1576–1590.
[26] D. O'Shaughnessy, Linear predictive coding, IEEE Potentials 7 (1) (1988) 29–32.
[27] A. Jongman, Z. Qin, J. Zhang, J.A. Sereno, Just noticeable differences for pitch direction, height, and slope for mandarin and english listeners, J. Acoust. Soc. Am. 142 (2) (2017) EL163–EL169..
[28] 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.
[29] K.M. Liew, Meaningful noise: auditory roughness and dissonance predict emotion recognition and cross-modal perception, Ph.D. thesis (2018)..
[30] B. Prica, S. Ilić, Recognition of vowels in continuous speech by using formants, Facta universitatis-series: Electronics and Energetics 23 (3) (2010) 379–393.
[31] 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, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[32] Z.-T. Liu, B.-H. Wu, D.-Y. Li, P. Xiao, J.-W. Mao, Speech emotion recognition based on selective interpolation synthetic minority over-sampling technique in small sample environment, Sensors 20 (8) (2020) 2297.
[33] F. Burkhardt, A. Paeschke, M. Rolfes, W.F. Sendlmeier, B. Weiss, A database of german emotional speech, in, in: Ninth European Conference on Speech Communication and Technology, 2005.
[34] 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.
[35] 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 Resour. Eval. 42 (4) (2008) 335.
[36] O.M. Nezami, P.J. Lou, M. Karami, Shemo: a large-scale validated database for persian speech emotion detection, Language Resour. Eval. 53 (1) (2019) 1–16.
[37] E. Parada-Cabaleiro, G. Costantini, A. Batliner, M. Schmitt, B.W. Schuller, Demos: an italian emotional speech corpus, Language Resour. Eval. (2019) 1–43.
[38] 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 Trans. Affective Comput. 8 (1) (2016) 67–80.
[39] 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..
[40] 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, Appl. Acoust. 166 (2020) 107360.
[41] 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.
[42] 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.
[43] H.M. Fayek, M. Lech, L. Cavedon, Evaluating deep learning architectures for speech emotion recognition, Neural Networks 92 (2017) 60–68.
[44] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Epps, Direct modelling of speech emotion from raw speech, Proc. Interspeech 2019 (2019) 3920–3924.
[45] J. Lee, I. Tashev, High-level feature representation using recurrent neural network for speech emotion recognition, in, in: Sixteenth annual conference of the international speech communication association, 2015.
[46] Y. Zeng, H. Mao, D. Peng, Z. Yi, Spectrogram based multi-task audio classification, Multimedia Tools Appl. 78 (3) (2019) 3705–3722.
[47] J. Gideon, M. McInnis, E. Mower Provost, Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (addog), IEEE Trans. Affective Comput. (2019), https://doi.org/10.1109/TAFFC.2019.2916092.
[48] 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.
[49] M. Neumann, N.T. Vu, Improving speech emotion recognition with unsupervised representation learning on unlabeled speech, in, in: ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 7390–7394.
[50] F. Bao, M. Neumann, N.T. Vu, Cyclegan-based emotion style transfer as data augmentation for speech emotion recognition, 2019 ISCA (2019) 35–37..