Feature selection enhancement and feature space visualization for speech-based emotion recognition

Sofia Kanwal[1,2], Sohail Asghar[1] and Hazrat Ali[3]

[1] Department of Computer Science, Islamabad Campus, Comsats University, Islamabad, Pakistan
[2] Department of Computer Science, University of Poonch Rawalakot, Rawalakot, Azad Kashmir, Pakistan
[3] College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar

ABSTRACT

Robust speech emotion recognition relies on the quality of the speech features. We present speech features enhancement strategy that improves speech emotion recognition. We used the INTERSPEECH 2010 challenge feature-set. We identified subsets from the features set and applied principle component analysis to the subsets. Finally, the features are fused horizontally. The resulting feature set is analyzed using t-distributed neighbour embeddings (t-SNE) before the application of features for emotion recognition. The method is compared with the state-of-the-art methods used in the literature. The empirical evidence is drawn using two well-known datasets: Berlin Emotional Speech Dataset (EMO-DB) and Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) for two languages, German and English, respectively. Our method achieved an average recognition gain of 11.5% for six out of seven emotions for the EMO-DB dataset, and 13.8% for seven out of eight emotions for the RAVDESS dataset as compared to the baseline study.

INTRODUCTION

Speech is one of the most remarkable, natural, and immediate methods for individuals to impart their emotions (*Akçay & Oğuz, 2020*). The human voice reflects the emotional and psychological states of a person. The speech emotion recognition (SER) domain has been explored for more than two decades (*Schuller, 2018*). It is helpful in numerous applications that require human-machine interaction. SER is gaining attention in many fields like automatic speech processing, pattern recognition, artificial intelligence, and domains involving human-machine interaction (*Lee & Narayanan, 2005*). By recognizing speech emotions using artificial intelligence-based methods, the interaction between humans and machines can be made more natural and effective. SER is advantageous in specific areas like automatic SER systems could help reduce accident ratio by recognizing aggressive drivers and alerting other road users (*Ji, Lan & Looney, 2006*; *Yang et al., 2017*). In the health care field, SER systems could help physicians understand the actual emotional state (*Harati et al., 2018*). Automatic learning can be improved by recognizing learners'

Corresponding author
Sofia Kanwal,
sofiakanwal@upr.edu.pk

emotions, alerting tutors, and suggesting changes to reduce negative emotions of boredom, frustration, and anxiety (*Jithendran et al., 2020*). The computer game industry may gain more natural interaction among game players and computers by exploiting audio emotion detection systems (*Schuller, 2016*). Hence, SER has significant value in improving lives in the digital era.

One of the challenging tasks in the design of a speech-based emotion recognition framework is the identification and extraction of features that proficiently describe diverse emotions and perform reliably (*Zheng et al., 2019*). Humans have the ability to recognize emotions using verbal and non-verbal cues. However, in the case of machines, a comprehensive set of features is required to perform emotion recognition tasks intelligently. Speech has a large range of features, however, not all of them play a significant role in distinguishing emotions (*Özseven, 2019*) and less important features need to be discarded by dimensionality reduction methods. One more thing which makes defining emotion boundaries even more difficult, is the presence of wide range of languages, accents, sentences, and speaking styles of the different speakers.

Further, there is no standardized feature set available in the literature and existing work in SER is based on the experimental studies so far (*Akçay & Oğuz, 2020*) and also, many of the existing SER systems are not as much accurate to be fully relied upon (*Xu, Zhang & Khan, 2020*). Among two broad categories of speech features: linguistic and paralinguistic, the later is more advantageous. The reason is paralinguistic features are helpful to recognize speech emotions irrespective of the language being spoken, the text being spoken and the accent being used (*Hook et al., 2019*). Their only dependence is on characteristics of speech signals such as tone, loudness, frequency, pitch, *etc.* In past, many researchers used paralinguistic attributes for SER, however, a universal set based on such cues has not established yet (*El Ayadi, Kamel & Karray, 2011*; *Anagnostopoulos, Iliou & Giannoukos, 2015*). Therefore, it is required to make improvements regarding the effective use of paralinguistic attributes for the SER. Out of various paralinguistic features, the most commonly used are prosodic, spectral, voice quality and Teager energy operator (TEO) (*Akçay & Oğuz, 2020*). Prosodic features portray the emotional value of speech (*Tripathi & Rao, 2018*) and are related to pitch, energy and, formant frequencies (*Lee & Narayanan, 2005*). Spectral features are based on the distribution of spectral energy throughout the speech range of frequency (*Nwe, Foo & De Silva, 2003*). Good examples of spectral features are Mel Frequency Cepstrum Coefficients (MFCC) and linear predictor features. Qualitative features are related to perceiving emotions by the quality of voice, *e.g.*, voice level, voice pitch and temporal structure (*Gobl & Chasaide, 2003*). Teager Energy Operator (TEO) based features are related to vocal chord movement while speaking under stressful situations (*Teager & Teager, 1990*).

Some studies reported emotion recognition based on speech features combined with other modalities, for example, text (*Lee & Lee, 2007*), facial expressions (*Nemati et al., 2019*; *Bellegarda, 2013*) body signals, and outward appearances (*Yu & Tapus, 2019*; *Ozkul et al., 2012*) to fabricate multi-modal frameworks for emotion analysis. However in this work, we limit ourselves to emotion recognition explicitly from speech features only.

In this work, we use the features set of the INTERSPEECH 2010 challenge which is the combination of prosodic, spectral and energy-related features and have a large set of attributes, *i.e.*, 1,582. Such a huge feature-space requires a good dimensionality reduction technique. One such technique most effectively used in speech research is principle component analysis (PCA) (*Jolliffe & Cadima, 2016*). The efficient use of PCA requires the removal of outliers and scaling the data (*Sarkar, Saha & Agrawal, 2014*). To apply PCA appropriately, we normalized the chosen feature set and discovered three subsets based on the types of speech features. The PCA is applied separately to each subset. We analyzed the finalized feature set thoroughly, by drawing t-SNE graphs for the two datasets. For classification, support vector machine (SVM) for One-Against-All is used. The results highlighted that appropriate selection of features and successful application of the feature selection method impacted the classification performance positively. The method achieved a decision-level correct classification rate of 71.08% for eight emotions using the RAVDESS dataset (*Livingstone & Russo, 2018*) and 82.48% for seven emotions using the EMO-DB dataset (*Burkhardt et al., 2005*). Furthermore, the proposed method outperformed the results of *Yang et al. (2017)*, *Venkataramanan & Rajamohan (2019)*, *Tawari & Trivedi (2010)*.

Our contributions in this work are:

- An improved speech emotion classification method using robust features.
- An approach to apply features reduction technique on a subset of speech features.
- An evaluation mechanism of the selected speech features using t-SNE visualization before applying a machine learning classifier.

The remainder of the article is composed as follows. "Literature review" provides a review of two key modules of speech-based emotion classification framework, *i.e.*, speech features and classifiers. "Materials and methods" introduces the speech emotion datasets and assessment criteria utilized in this work. Evaluation results of the framework utilizing various databases and various situations are presented in "Experimental results and discussion". Finally, "Conclusion" concludes the article.

LITERATURE REVIEW

This section presents the previous work on speech-based emotion recognition with respect to selection of relevant speech features and machine learning methods used in SER systems.

Yang et al. (2017) opted for prosodic features mainly and used mutual information as a feature selection method to choose top-80 features for the recognition task. In *Venkataramanan & Rajamohan (2019)*, few features such as MFCC, pitch, energy, and Mel-spectrogram were used, and the top-2 and top-3 accuracy of 84.13% and 91.38% respectively were achieved for the RAVDESS dataset. Fourier coefficients of the speech signal were extracted and applied on EMO-DB dataset in *Wang et al. (2015)*, which came up with a recognition rate of 79.51%. *Lugger & Yang (2007)* used voice quality 85 and prosodic features in combination for emotion recognition. In *Tripathi & Rao (2017)*, a

speech-based emotion classification system was developed for the Bengali speech *corpus*. The classifiers used were ANN, KNN, SVMs, and naive Bayes. The maximum accuracy achieved was 89%. For robust emotion classification, an enhanced sparse representation classifier is proposed (*Zhao, Zhang & Lei, 2014*). The classifier performed better in both noisy and clean data in comparison to other existing classifiers. In *Al Machot et al. (2011)*, a probability-based classifier, Bayesian quadratic discriminate was used to classify emotions. Classification accuracy of 86.67% on EMO-DB dataset was reported by reducing the calculation cost and using a small number of features. Another study reported an ensemble classification method for speech emotion classification on Thai speech *corpus* (*Prasomphan & Doungwichain, 2018*). The ensembled algorithms were SVM, KNN, and ANN to improve the accuracy. To make a comprehensive feature set, capable of performing well in all situations and supporting multiple languages, there is a need to include all the important speech emotion attributes. One such set of features available in the literature is INTERSPEECH 2010 paralinguistic challenge feature-set (*Schuller et al., 2010*). This feature-set is an extension of the INTERSPEECH 2009 paralinguistic emotion challenge feature-set (*Schuller, Steidl & Batliner, 2009*).

MATERIALS AND METHODS

This section discusses the material and methods used in this research.

Data and preprocessing

The benchmark datasets used in this research are (EMO-DB) *Burkhardt et al. (2005)* and (RAVDESS) *Livingstone & Russo (2018)*.

EMO-DB

The EMO-DB is an acted dataset of ten experts (five male and five female) kept in the German language. It comprises seven emotion classes: anger, boredom, fear, happiness, disgust, neutral, and sadness. There are numerous expressions of a similar speaker. Ten sentences, which are phonetically unbiased, are picked for dataset development. Out of these 10 sentences, five sentences are short (around 1.5 s length) and five are long sentences (roughly 4 s term). Every emotion class has an almost equivalent number of expressions to avoid the issue of under-sampling emotion class. There is a sum of 535 expressions in this dataset. The validity of the dataset is ensured by rejecting the samples having a recognition rate less than 80% in the subjective listening test. The metadata of the EMO-DB dataset is given in Table 1.

RAVDESS

RAVDESS is multi-modular database of emotional speech and song. There are 24 professional actors each uttering 104 unique intonations with emotions: happiness, sad, angry, fear, surprise, disgust, calm, and neutral. The RAVDESS dataset is rich and diverse and does not experience gender bias, which means it has an equal number of male and female utterances, comprises a wide range of emotions, and at various levels of emotional intensity. Each actor uses two different statements with two different emotional intensities;

Table 1 Emotional speech databases (EMO-DB).

Database	EMO-DB Burkhardt et al. (2005)
No. of speakers	10 (five male, five female)
Language	German
Emotions	Anger, sadness, joy, fear boredom, disgust, neutral
No. of utterances	535
Style	Acted

Table 2 Emotional speech databases (RAVDESS).

Database	RAVDESS Database Livingstone & Russo (2018)
No. of speakers	24 (12 male, 12 female)
Language	English
Emotions	Anger, sadness, calm, surprise, happiness, fear disgust, neutral
No. of utterances	1,440
Style	Acted

normal and strong emotions. The total number of utterances is 1,440. The information about the RAVDESS dataset is in Table 2.

Preprocessing

The preprocessing step involved reading audio files, removing unvoiced segments, and framing the signal having 60 ms length.

Feature extraction

After preprocessing, we moved towards feature analysis which involved the extraction of useful features for speech emotion analysis. In this research, the OpenSMILE toolbox was used and the resulting feature-set was INTERSPEECH 2010 Challenge feature-set (*Schuller et al., 2010*). The reason for opting INTERSPEECH 2010 feature-set is because, it covers most of the features (namely, prosodic, spectral, and energy) effective for emotion recognition. The choice of initial features set is in line with the findings of *Özseven & Düğenci (2018)* which reported the effectiveness of these features for emotion recognition. INTERSPEECH 2010 feature-set consists of a total of 1,582 features. Upon a thorough analysis of the feature-set, we identified three subsets. The first subset consisted of 34 low-level descriptors (LLDs) with 34 corresponding delta co-efficients, having applied 21 functionals on each of it. This resulted in a subset of 1,428 features. The second subset was of 19 functionals applied to four pitch based LLDs with their corresponding four delta coefficients resulting in a total of 152 features. The third feature-set consisted only two features, which were pitch onset and duration.
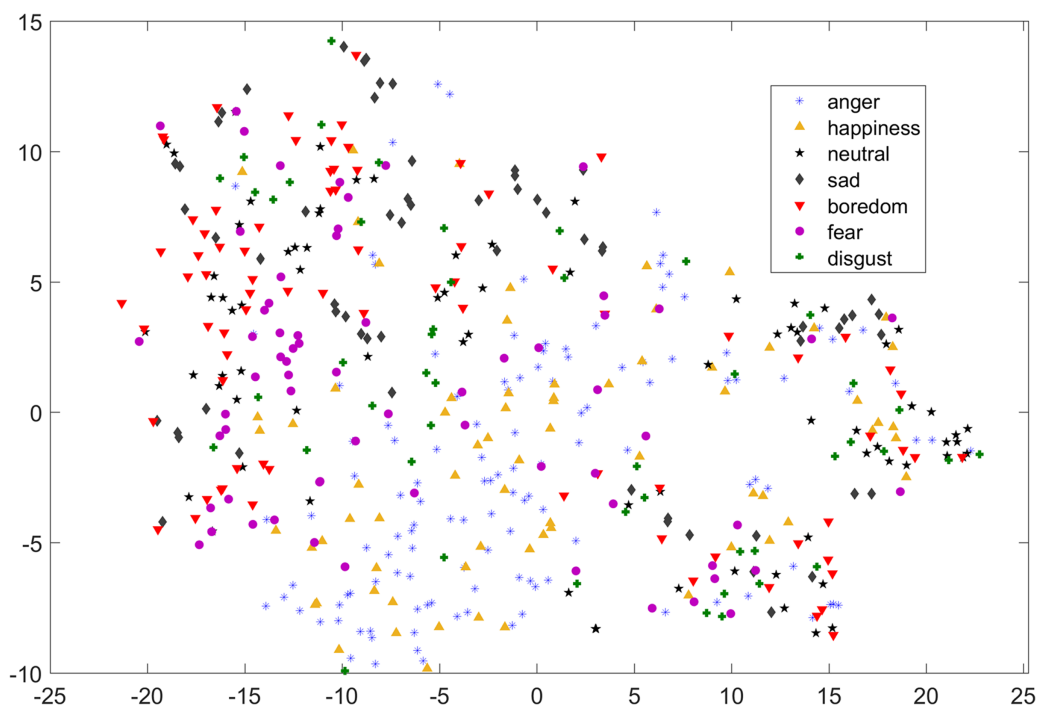
Figure 1 Two dimensional t-SNE graph of features used in *Yang et al. (2017)* for EMO-DB dataset.

Feature reduction technique

As the feature-set was huge, consisting of 1,582 features, we needed a good dimensionality reduction technique. For feature reduction in SER, some of the commonly used methods are forward feature selection (FFS), backward feature selection (BFS) (*Pao et al., 2005*), principle component analysis (PCA) and linear discriminant analysis (LDA) (*Haq, Jackson & Edge, 2008*; *Özseven, Düğenci & Durmuşoğlu, 2018*). Among these the most commonly used and applied in many studies is PCA (*Chen et al., 2012*; *Scherer et al., 2015*; *Patel et al., 2011*). PCA includes finding the eigenvalues and eigenvectors of the available covariance matrix and choosing the necessary number of eigenvectors comparing to the biggest eigenvalues to create a transformed matrix. We have fed the openSMILE INTERSPEECH 2010 feature-set of 1,582 features to the PCA and selected the top-100 features which we used for the classification task.

In Figs. 1 and 2, although we can see some clusters, however, most of the emotions are evenly distributed.

As compared to feature-set utilized by *Yang et al. (2017)*, when we used INTERSPEECH 2010 challenge feature-set and applied PCA for features reduction, a clearer representation of clusters could be seen. In Fig. 3 for the EMO-DB dataset, emotion categories of sad, neutral, and boredom were distributed in relatively compact clusters in the feature-space such that high classification accuracy was anticipated. It is also noticeable from Fig. 3, that the data points for emotions such as happiness and anger were overlapping with each other giving a chance of misclassifying, however, overall classification of happiness and anger was not going to degrade due to being clustered at one side of feature-space. Likewise, two-
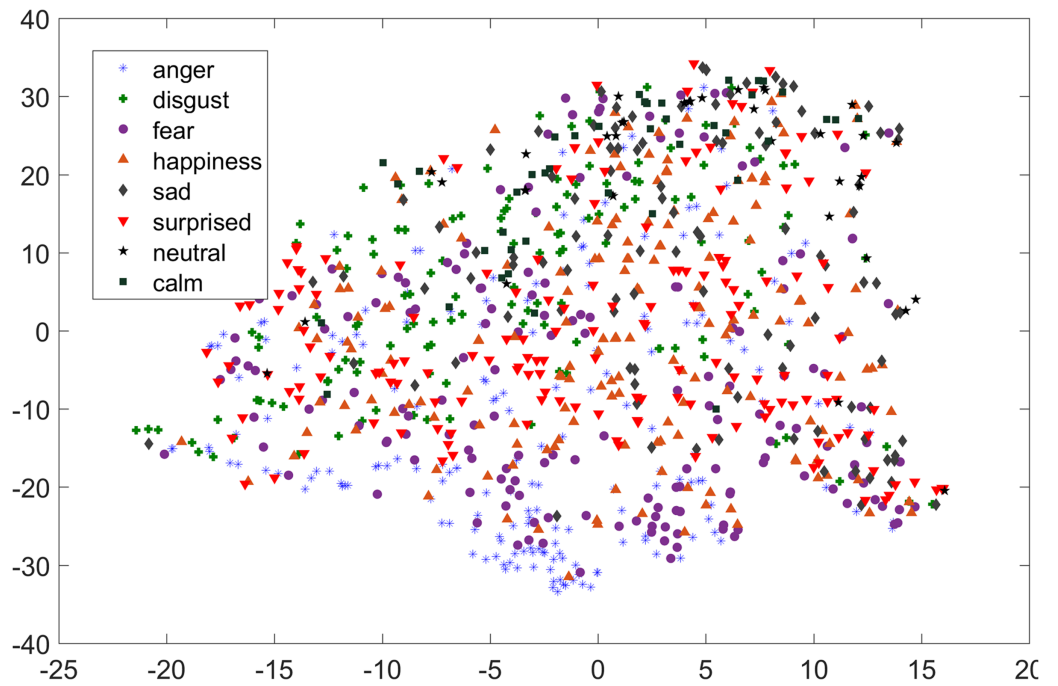
Figure 2 Two dimensional t-SNE graph of features used in *Yang et al. (2017)* for RAVDESS dataset.
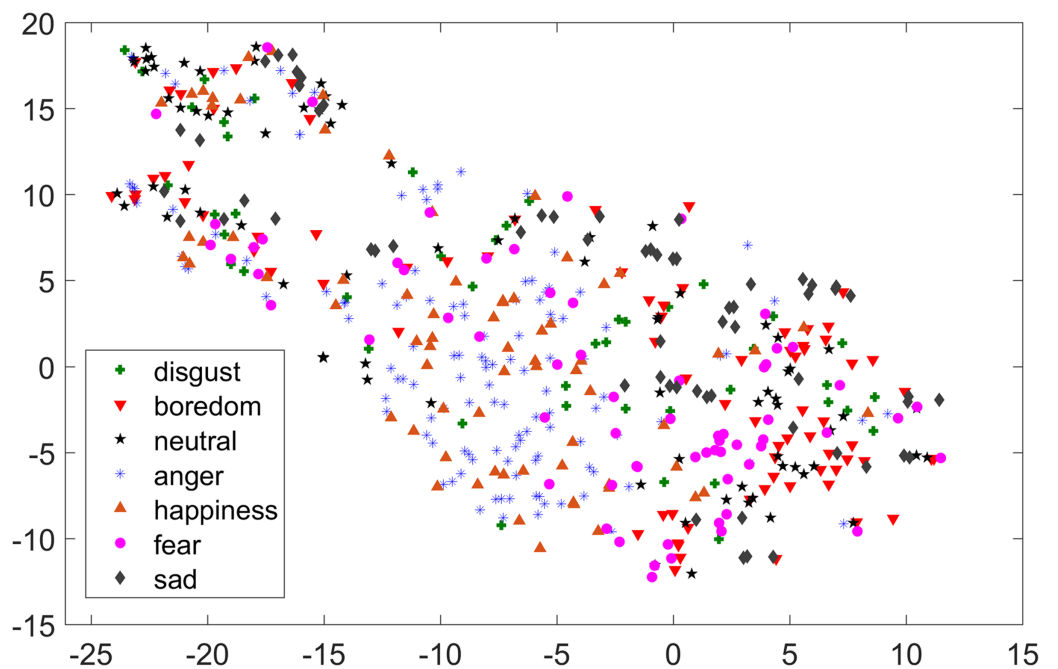
Figure 3 Two-dimensional t-SNE graphs of INTERSPEECH 2010 feature-set after applying PCA for feature reduction on EMO-DB dataset.
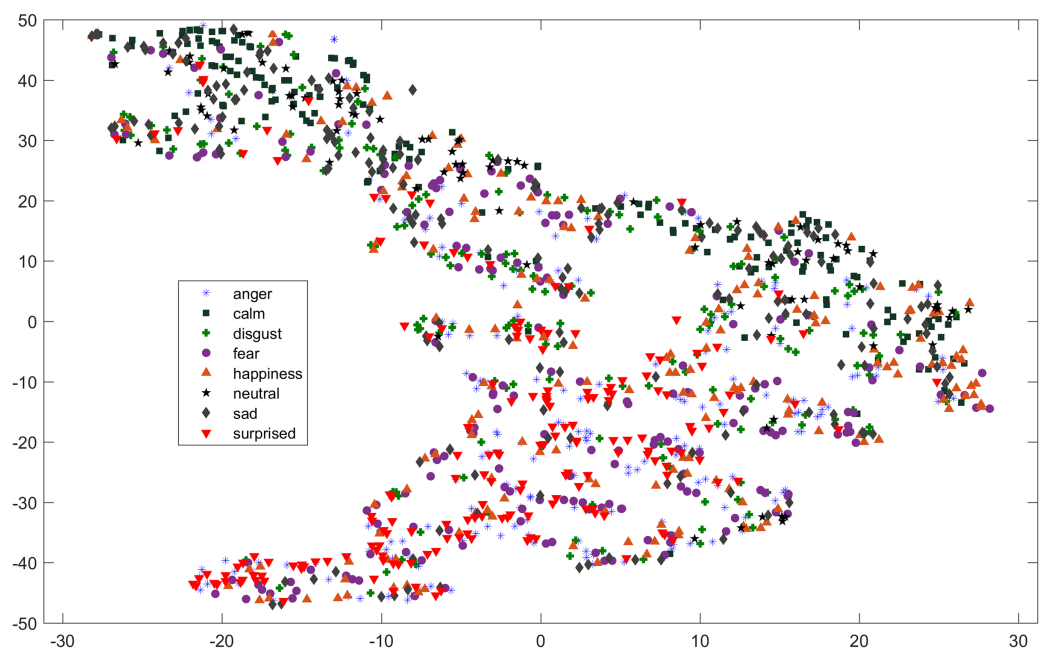
Figure 4 Two dimensional t-SNE graphs on RAVDESS dataset. Two dimensional t-SNE graphs of features after using INTERSPEECH 2010 feature-set and applying PCA for feature reduction on RAVDESS dataset.

dimensional representation of t-SNE of RAVDESS dataset is shown in Fig. 4. After using INTERSPEECH 2010 feature-set, there was a compact representation of clusters of anger, sadness, calm, and fear. Based on these observations, it was expected that overall classification accuracy for emotion classes would greatly improve.

Computational setup

To compare the performance of selected features, the process is divided into two phases: training and testing. The One-Against-All SVM classifier was trained and validated with its respective parameters to obtain the optimal model for 70% of total samples. In the testing phase, the remaining 30% of the samples were used to test the model. The entire framework covering preprocessing, feature selection, and model representation is visualized in Fig. 5.

EXPERIMENTAL RESULTS AND DISCUSSION

Performance evaluation scenarios

To evaluate the framework, along with metrics of accuracy and recall, two new metrics top-2 accuracy and top-3 accuracy are also introduced. Top-2 accuracy is the accuracy when the true class matches with any one of the two most probable classes predicted by the model and top-3 accuracy is the accuracy when the true class matches with any one of three most probable classes predicted by the model. The terminology used for evaluation is as follows:
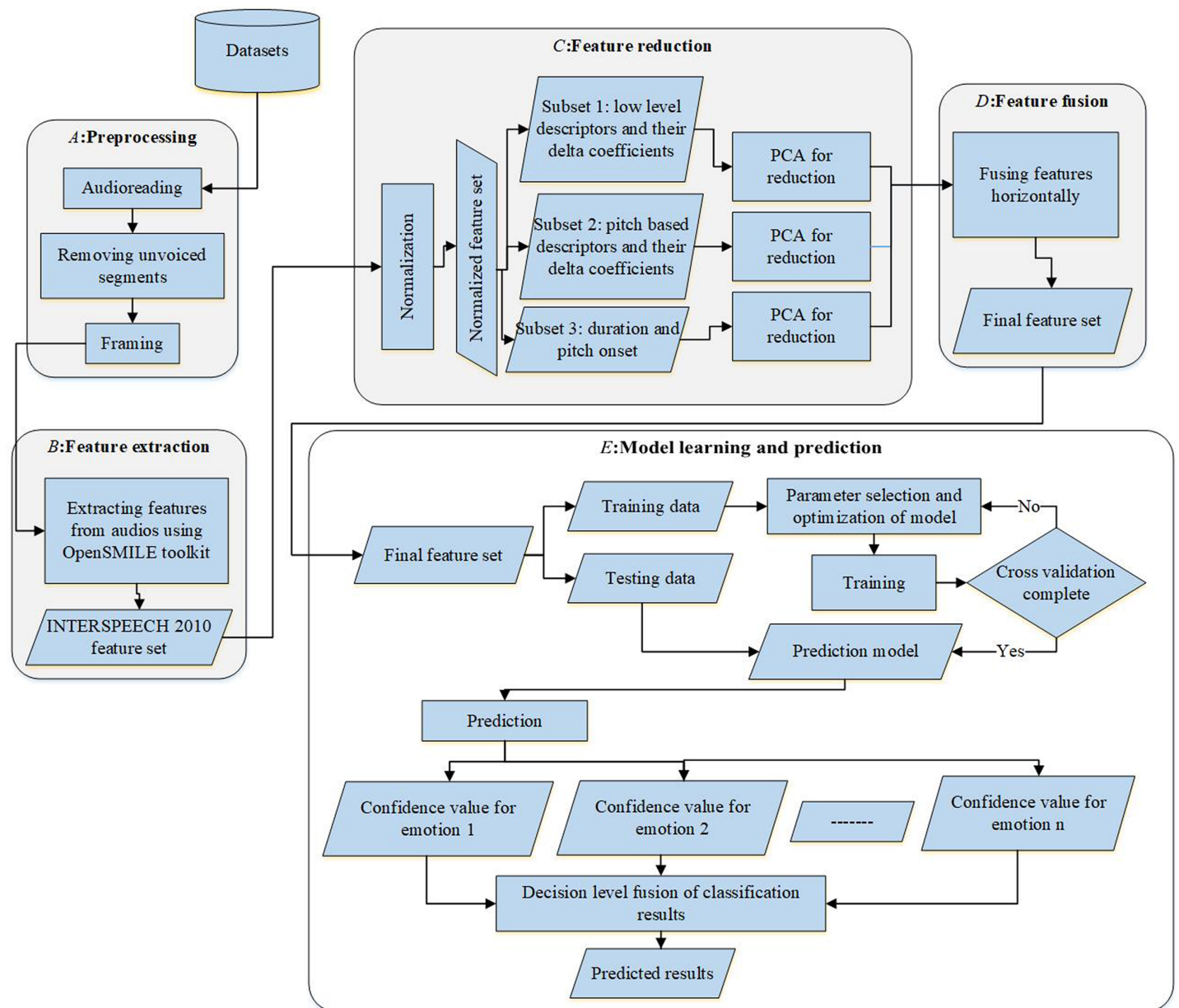
Figure 5 **The computational model.**

The ratio of unclassified samples over all samples in the test set is known as the rejection rate. Average classification performance for all emotions after fusion is taken as decision-level-classification-rate. It is defined in Eq. (1).

$$DL - Classification - rate = \frac{DL(tp)_{m1} + DL(tp)_{m2} + \ldots\ldots DL(tp)_{mn}}{N} \tag{1}$$

where $DL(tp)_{m1}$ denotes the number of true positive utterances of emotion class $m1$ to mn where n is the number of emotion classes and N is the total number of utterances.

Table 3 Comparison of the key characteristics of the proposed method with baseline methods of speech emotion recognition.

Method	Dataset	Classifier	Features	Feature fusion	Feature selection
Our extended method	RAVDESS EMO-DB	SVM	INTERSPEECH 2010 (prosodic, spectral energy)	Yes	Yes
Venkataramanan & Rajamohan (2019)	RAVDESS	2D CNN	Pure audio, log mel-spectrogram	No	No
Yang et al. (2017)	LDC	SVM	Prosodic features	No	Yes

To evaluate the emotion classification performance of each individual emotion, the metric we are using is decision level recall rate for emotion *mi*. Mathematically it is written as given in Eq. (2).

$$DL - emotion_i - recall - rate = \frac{DL(tp)_{mi}}{DL(tp)_{mi} + DL(tn)_{mi}} \tag{2}$$

where $DL(tp)_{mi}$ and $DL(tn)_{mi}$ denote the decision level true positive and true negative utterances for emotion *mi*, respectively.

Emotion classification performance

For analyzing the performance of our feature driven One-Against-All SVM based emotion classification method, the comparison with the state-of-the-art is performed. After that the performance improvement of the proposed method for general test and gender dependent test is presented. The baseline methods considered for comparison are *Yang et al. (2017)*, and *Venkataramanan & Rajamohan (2019)*. A summary of key features is shown in Table 3. In the later sections, we provide decision level emotion classification recall, top-2 accuracy and top-3 accuracy depending on which metric is provided by the reference method.

Performance comparison for general test

In general test performance comparison, the data samples from EMO-DB and RAVDESS datasets were used through seven-fold cross validation mechanism. In each round, one-seventh of the samples are kept for test. At the end, results are averaged over seven-folds.

The work of *Yang et al. (2017)* used LDC (Linguistic Data Consortium) dataset. The LDC Emotional Prosody Speech and Transcripts database is English database simulated by professional actors. The utterances consist of dates and numbers. Due to unavailability of LDC dataset freely, we employed the most frequently used freely available EMO-DB dataset (*Burkhardt et al., 2005*) and RAVDESS dataset *Livingstone & Russo (2018)* in this research. We reproduced the results of *Yang et al. (2017)* by using both datasets and reported the comparisons.

Comparison of average recall

The decision level average recall for each emotion of the two datasets for our method is compared with *Yang et al. (2017)* in Figs. 6 and 7.
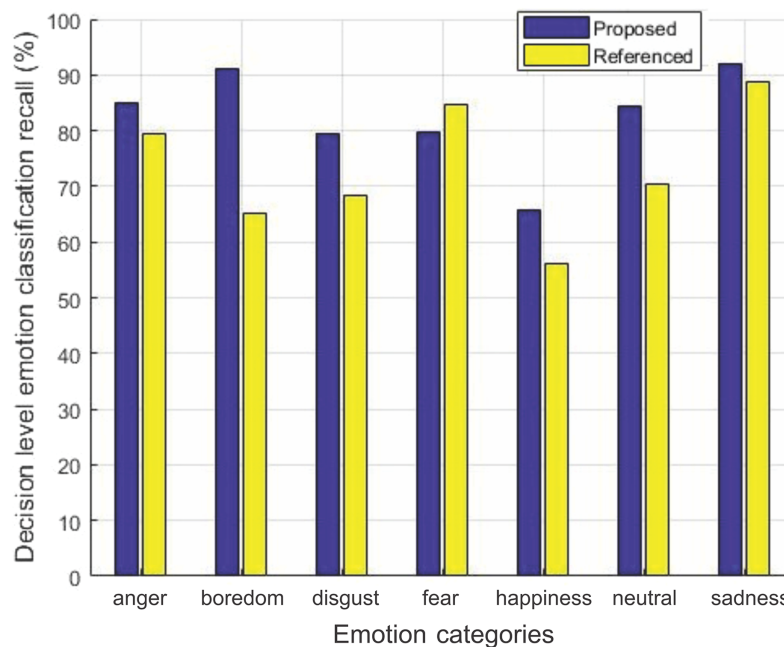
Figure 6 Decision-level emotion classification recall using EMO-DB dataset. Decision-level emotion classification recall for each individual emotion for our method without rejecting any samples and the method in *Yang et al. (2017)*, using EMO-DB dataset.

We can see from Fig. 6 that our method outperforms the baseline method in *Yang et al. (2017)* for six out of seven emotions using EMO-DB dataset. In the same way performance gain for seven out of eight emotions in case of RAVDESS dataset is achieved shown in Fig. 7. The average recall values for data samples of EMO-DB and RAVDESS datasets through seven rounds of cross validation and five upsampling are also given in Table 4.

Here in Table 4, the performance gain in terms of percentage is given. Our feature selection mechanism improved the recall rate for most of the emotions on both datasets as shown by highlighting the entries in bold.

Performance comparison using thresholding fusion

Here in this section, we are doing a comparison using a thresholding fusion mechanism. Thresholding fusion mechanism is an extension of the normal classification method (*Vapnik, 1998*). Here it is possible to output the class label only when it has a confidence score greater than certain threshold, otherwise it is rejected. In practice, such type of classification enhancement is needed when we are only concerned with samples having the high accuracy. For example, if a psychologist wanted to know when a patient, having some psychological disorder, at certain times feels emotionally high, thresholding fusion mechanism can be used.

General test comparison

Figures 8 and 9 show the decision-level correct classification rate when samples are rejected by gauging the confidence score.
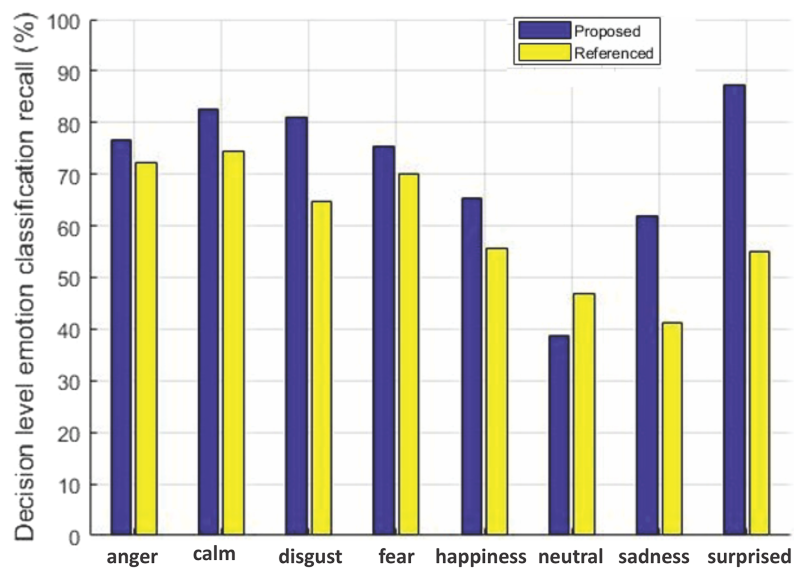
Figure 7 Decision-level emotion classification recall using RAVDESS dataset. Decision-level emotion classification recall for each individual emotion for our method without rejecting any samples and the method in *Yang et al. (2017)*, using RAVDESS dataset.

Table 4 Performance gain of our method with *Yang et al. (2017)* in terms of average recall for each emotion for EMO-DB and RAVDESS datasets. Our feature selection mechanism improved the recall rate for most of the emotions on both datasets as shown by highlighting the entries in bold.

Dataset	Emotion	*Yang et al. (2017)*	Our method
EMO-DB	Disgust	68.3673	**79.5918**
	Anger	79.4903	**84.9206**
	Happiness	56.2337	**65.7142**
	Fear	**84.6031**	79.6825
	Sadness	88.6904	**92.0634**
	Neutral	70.3463	**84.4155**
	Boredom	65.1515	**91.0173**
RAVDESS	Disgust	64.6593	**81.1208**
	Anger	72.3356	**76.5117**
	Happiness	55.7183	**65.2828**
	Fear	70.1058	**75.3212**
	Sadness	41.0866	**61.8233**
	Neutral	**46.7948**	38.6446
	Calm	74.4523	**82.5238**
	Surprised	54.9450	**87.3626**

It can be seen that by rejecting about 50% of the samples, we are able to gain above 90% accuracy rate. The performance gain as compared to the baseline study is given in Figs. 8 and 9 for EMO-DB and RAVDESS datasets respectively. This scheme of gaining a high accuracy rate at the cost of rejecting samples could be beneficial in situations where high
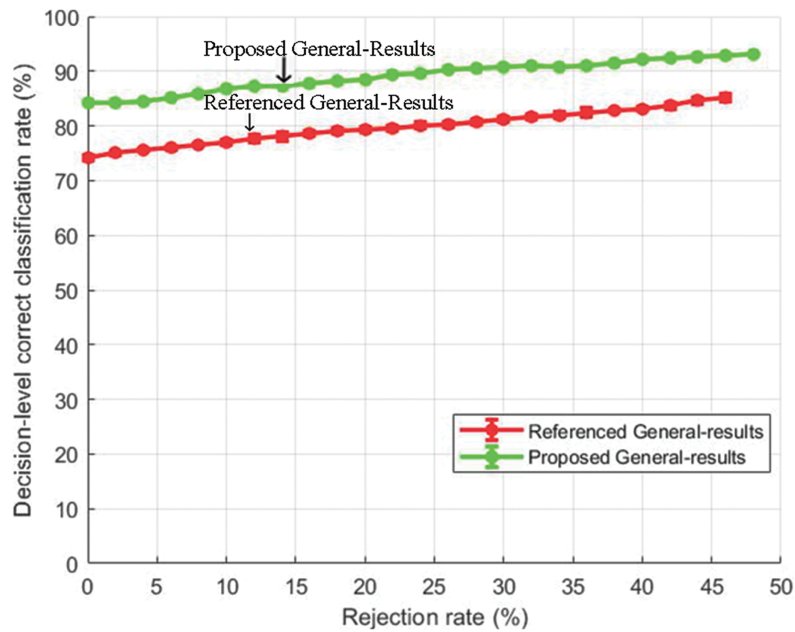
Figure 8 Decision-level correct classification rate *vs* rejection rate for general test of our method and the method of *Yang et al. (2017)* using EMO-DB dataset.
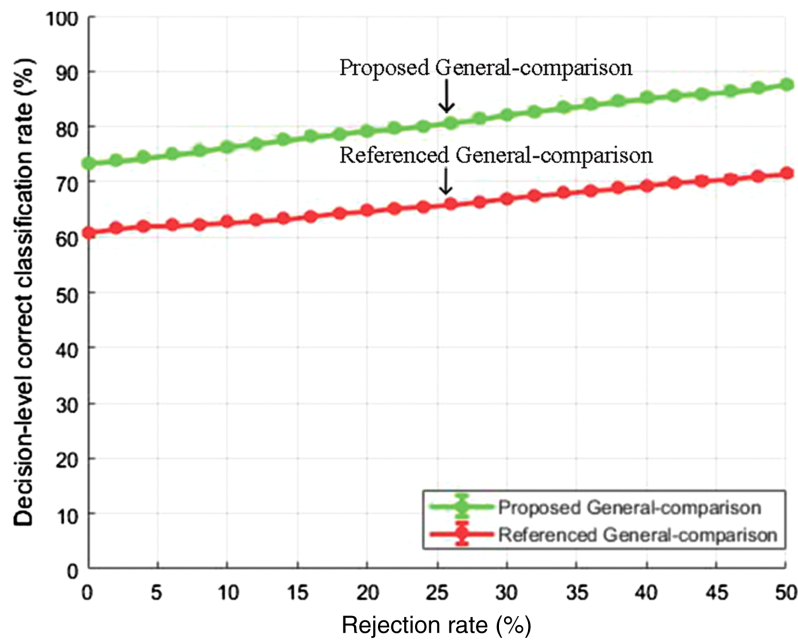
Figure 9 Decision-level correct classification rate *vs* rejection rate for general test of our method and the method in *Yang et al. (2017)* using RAVDESS dataset.
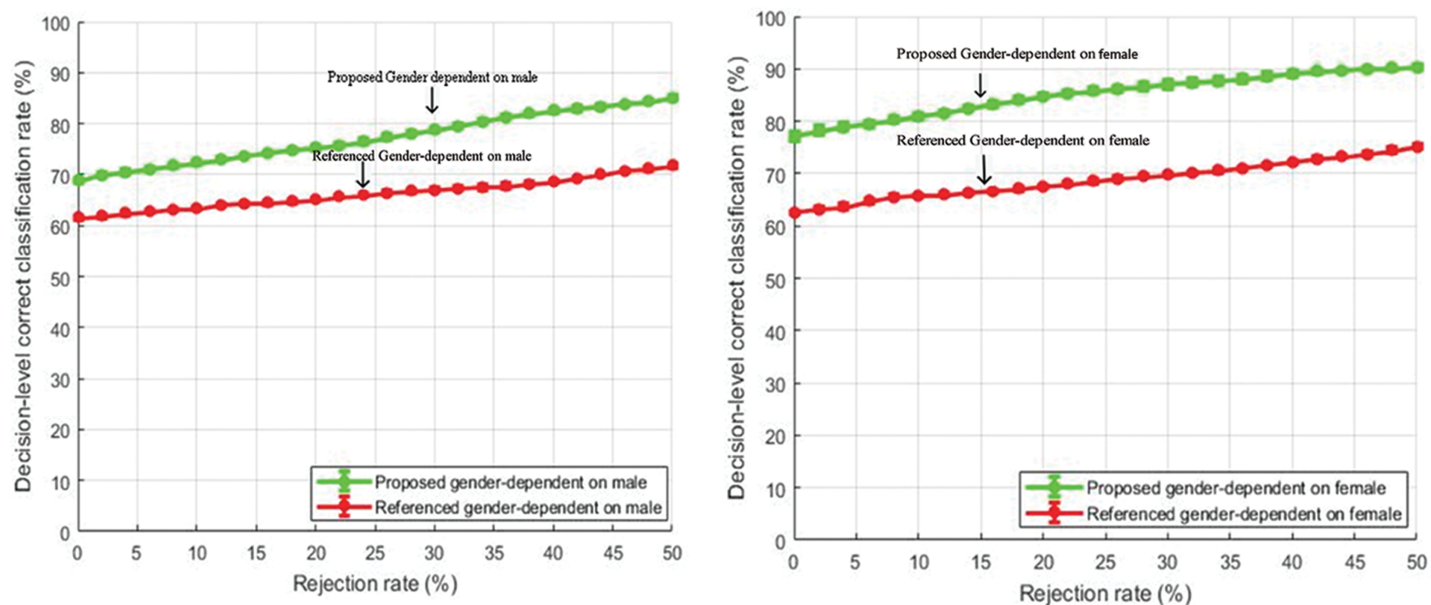
Figure 10 Comparison of decision-level correct classification rate against rejection rate for gender dependent test on EMO-DB.

accuracy for precise samples is required as compared to including all samples with low accuracy.

Gender-dependent comparison

Decision level correct classification rate for gender-dependent tests for male and female speakers are shown in Figs. 10 and 11 respectively.

The features selected for gender-specific tests were based on male and female utterances and were not the same as the general tests. The performance gain of about 10% is achieved by rejecting 50% of the total samples. Our feature enhancements worked well and outperformed the baseline study of *Yang et al. (2017)*. Further, it is also observed from Figs. 10 and 11 that the emotion classification rate for female speakers was high as compared to male speakers. One of the reasons could be that female speakers are more accurate in decoding emotions as compared to males (*Lausen & Schacht, 2018*). The same observation is also seen in the study of *Yang et al. (2017)*.

State-of-the-art comparison

In Table 5, top-2 and top-3 accuracy of our method for RAVDESS is compared with the one reported in *Venkataramanan & Rajamohan (2019)*. Even though, *Venkataramanan & Rajamohan (2019)* used a neural network model, the top-2 accuracy for our method improved from 84.13% to 87.86% having the performance gain of 3.73%, The performance gain of 1.51% was achieved for top-3 which is still higher then the baseline method. The top-2 and top-3 accuracy measures for the EMO-DB dataset are also given in Table 5. The gender-dependent accuracy score achieved on EMO-DB dataset is also compared with the results reported in *Tawari & Trivedi (2010)*. The accuracy results averaged for both
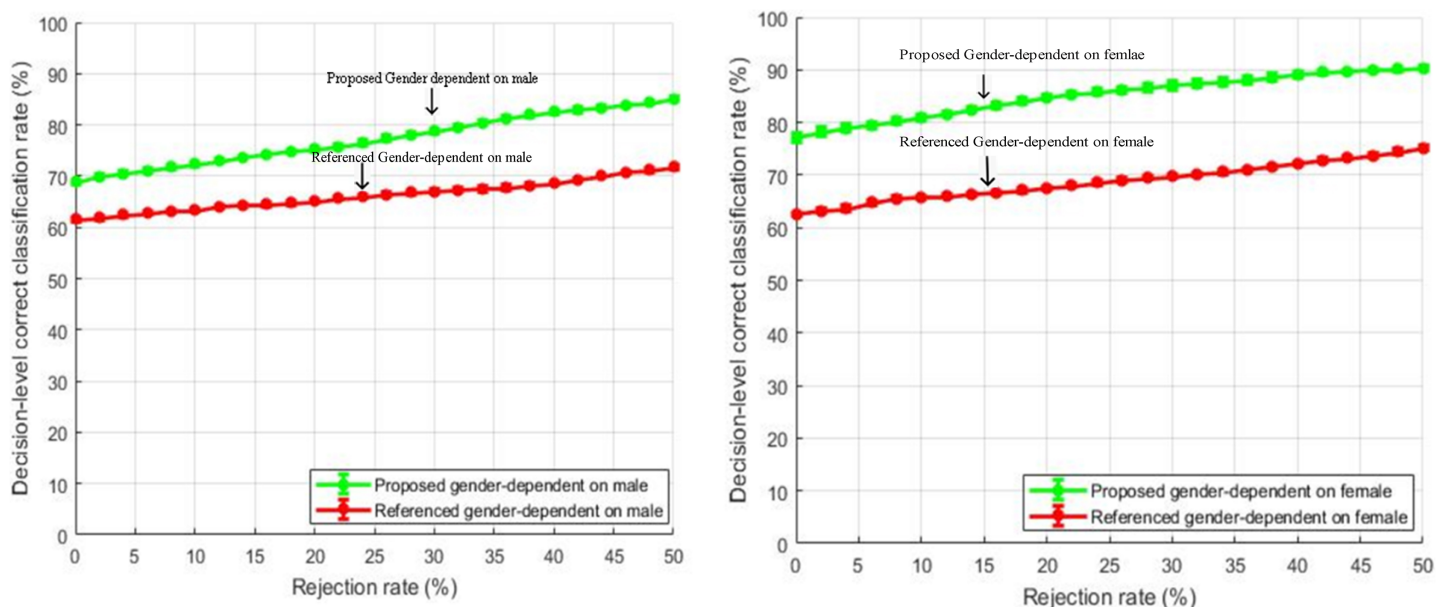
Figure 11 Comparison of decision-level correct classification rate against rejection rate for gender dependent test of our method with the baseline study using RAVDESS dataset with speaker normalization, feature selection, and over-sampled training sets.

Table 5 Performance comparison of our method with *Venkataramanan & Rajamohan (2019)* with respect to top-2 and top-3 accuracy for RAVDESS and EMO-DB datasets. The bold values indicate the effectiveness of using appropriately tuned features along with classifier.

Dataset	Metric	*Venkataramanan & Rajamohan (2019)*	*Tawari & Trivedi (2010)*	Proposed extensions
RAVDESS	Top-2 accuracy	84.13		**87.86**
	Top-3 accuracy	91.38		**92.89**
EMO-DB	Top-2 accuracy			**92.35**
	Top-3 accuracy			**94.89**
	Unweighted accuracy (gender-wise)		84	**85.47**

male and female speakers in our work are 85.47%, whereas 84% in the work by *Tawari & Trivedi (2010)*. The bold values in Table 5 indicate the effectiveness of using appropriately tuned features along with classifier.

CONCLUSION

In this article, we presented a strategy to improve the performance of speech emotion recognition frameworks. We used a feature selection strategy based on INTERSPEECH 2010 challenge feature-set with PCA. Here we reanalyzed and extended the existing model of multiclass SVM with thresholding fusion for speech-based emotion classification. Our in-depth analysis of feature-space through t-SNE showed that selected features represent most of the emotion-related information. The datasets used for evaluation were EMO-DB and RAVDESS having emotions of anger, calm, disgust, fear, happiness, sad, surprise, and

neutral. We used PCA, where subsets of features were fed separately and finally fused to reduce the features from 1,582 to 100. We used the One-Against-All SVM model for emotion classification. The performance of feature-driven SER method was compared with the baseline methods. The experimental results demonstrate the effectiveness of the selected features as the method improved the performance of speech emotion recognition. In future, it is anticipated that further investigation of various features will bring additional improvement to the recognition accuracy.

Funding
The authors received no funding for this work.

Competing Interests
Hazrat Ali is an Academic Editor for PeerJ.

Author Contributions
- Sofia Kanwal conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, and approved the final draft.
- Sohail Asghar conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, authored or reviewed drafts of the article, and approved the final draft.
- Hazrat Ali conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, authored or reviewed drafts of the article, and approved the final draft.

Data Availability
Raw data is available at the following:
Database Name: EMO-DB
Link: https://www.kaggle.com/piyushagni5/berlin-database-of-emotional-speech-emodb.
Database name: RAVDESS
Link: https://www.kaggle.com/uwrfkaggler/ravdess-emotional-speech-audio.

Supplemental Information
Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.1091#supplemental-information.

REFERENCES

Akçay MB, Oğuz K. 2020. Speech emotion recognition: emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication* **116(12)**:56–76 DOI 10.1016/j.specom.2019.12.001.

Al Machot F, Mosa AH, Dabbour K, Fasih A, Schwarzlmuller C, Ali M, Kyamakya K. 2011. A novel real-time emotion detection system from audio streams based on Bayesian quadratic discriminate classifier for Adas. In: *Nonlinear Dynamics and Synchronization (INDS) & 16th International Symposium on Theoretical Electrical Engineering (ISTET), 2011 Joint 3rd International Workshop on.* Piscataway: IEEE, 1–5.

Anagnostopoulos C-N, Iliou T, Giannoukos I. 2015. Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Artificial Intelligence Review* **43(2)**:155–177 DOI 10.1007/s10462-012-9368-5.

Bellegarda JR. 2013. Data-driven analysis of emotion in text using latent affective folding and embedding. *Computational Intelligence* **29(3)**:506–526 DOI 10.1111/j.1467-8640.2012.00457.x.

Burkhardt F, Paeschke A, Rolfes M, Sendlmeier WF, Weiss B. 2005. A database of German emotional speech. In: *INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology.* Lisbon, Portugal.

Chen L, Mao X, Xue Y, Cheng LL. 2012. Speech emotion recognition: features and classification models. *Digital Signal Processing* **22(6)**:1154–1160 DOI 10.1016/j.dsp.2012.05.007.

El Ayadi M, Kamel MS, Karray F. 2011. Survey on speech emotion recognition: features, classification schemes, and databases. *Pattern Recognition* **44(3)**:572–587 DOI 10.1016/j.patcog.2010.09.020.

Gobl C, Chasaide AN. 2003. The role of voice quality in communicating emotion, mood and attitude. *Speech Communication* **40(1–2)**:189–212 DOI 10.1016/S0167-6393(02)00082-1.

Haq S, Jackson PJ, Edge J. 2008. Audio-visual feature selection and reduction for emotion classification. In: *Proceedings of the International Conference on Auditory-Visual Speech Processing (AVSP'08).* Tangalooma, Australia.

Harati S, Crowell A, Mayberg H, Nemati S. 2018. Depression severity classification from speech emotion. In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).* Picataway: IEEE, 5763–5766.

Hook J, Noroozi F, Toygar O, Anbarjafari G. 2019. Automatic speech based emotion recognition using paralinguistics features. *Bulletin of the Polish Academy of Sciences. Technical Sciences* **67(3)**:2019 DOI 10.24425/bpasts.2019.129647.

Ji Q, Lan P, Looney C. 2006. A probabilistic framework for modeling and real-time monitoring human fatigue. *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans* **36(5)**:862–875 DOI 10.1109/TSMCA.2005.855922.

Jithendran A, Karthik PP, Santhosh S, Naren J. 2020. Emotion recognition on E-learning community to improve the learning outcomes using machine learning concepts: a pilot study. In: Somani AK, Shekhawat RS, Mundra A, Srivastava S, Verma VK, eds. *Smart Systems and IoT: Innovations in Computing.* Singapore: Springer, 521–530.

Jolliffe IT, Cadima J. 2016. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **374(2065)**:20150202 DOI 10.1098/rsta.2015.0202.

Lausen A, Schacht A. 2018. Gender differences in the recognition of vocal emotions. *Frontiers in Psychology* **9**:882 DOI 10.3389/fpsyg.2018.00882.

Lee C, Lee GG. 2007. Emotion recognition for affective user interfaces using natural language dialogs. In: *RO-MAN 2007-The 16th IEEE International Symposium on Robot and Human Interactive Communication.* Picataway: IEEE, 798–801.

Lee CM, Narayanan SS. 2005. Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing* **13(2)**:293–303 DOI 10.1109/TSA.2004.838534.

Livingstone SR, Russo FA. 2018. The Ryerson audio-visual database of emotional speech and song (RAVDESS): a dynamic, multimodal set of facial and vocal expressions in north American English. *PLOS ONE* **13(5)**:e0196391 DOI 10.1371/journal.pone.0196391.

Lugger M, Yang B. 2007. The relevance of voice quality features in speaker independent emotion recognition. In: *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*. Vol. 4. Picataway: IEEE, IV-17–IV-20.

Nemati S, Rohani R, Basiri ME, Abdar M, Yen NY, Makarenkov V. 2019. A hybrid latent space data fusion method for multimodal emotion recognition. *IEEE Access* **7**:172948–172964 DOI 10.1109/ACCESS.2019.2955637.

Nwe TL, Foo SW, De Silva LC. 2003. Speech emotion recognition using hidden Markov models. *Speech Communication* **41(4)**:603–623 DOI 10.1016/S0167-6393(03)00099-2.

Ozkul S, Bozkurt E, Asta S, Yemez Y, Erzin E. 2012. Multimodal analysis of upper-body gestures, facial expressions and speech. In: *Procceedings of the 4th International Workshop on Corpora for Research on Emotion Sentiment and Social Signals*. Istanbul, Turkey.

Özseven T, Düğenci M. 2018. Speech acoustic (SPAC): a novel tool for speech feature extraction and classification. *Applied Acoustics* **136(1)**:1–8 DOI 10.1016/j.apacoust.2018.02.009.

Özseven T, Düğenci M, Durmuşoğlu A. 2018. A content analaysis of the research approaches in speech emotion recognition. *International Journal of Engineering Sciences & Research Technology* **7(1)**:1–26 DOI 10.5281/zenodo.1133547.

Özseven T. 2019. A novel feature selection method for speech emotion recognition. *Applied Acoustics* **146(3)**:320–326 DOI 10.1016/j.apacoust.2018.11.028.

Pao T-L, Chen Y-T, Yeh J-H, Chang Y-H. 2005. Emotion recognition and evaluation of mandarin speech using weighted D-KNN classification. In: *Proceedings of the 17th Conference on Computational Linguistics and Speech Processing*. Tainan, 203–212.

Patel S, Scherer KR, Björkner E, Sundberg J. 2011. Mapping emotions into acoustic space: the role of voice production. *Biological Psychology* **87(1)**:93–98 DOI 10.1016/j.biopsycho.2011.02.010.

Prasomphan S, Doungwichain S. 2018. Detecting human emotions in a large size of database by using ensemble classification model. *Mobile Networks and Applications* **23(4)**:1–6 DOI 10.1007/s11036-018-1074-3.

Sarkar J, Saha S, Agrawal S. 2014. An efficient use of principal component analysis in workload characterization-a study. *AASRI Procedia* **8(1)**:68–74 DOI 10.1016/j.aasri.2014.08.012.

Scherer KR, Sundberg J, Tamarit L, Salomão GL. 2015. Comparing the acoustic expression of emotion in the speaking and the singing voice. *Computer Speech & Language* **29(1)**:218–235 DOI 10.1016/j.csl.2013.10.002.

Schuller B. 2016. Emotion modelling via speech content and prosody: in computer games and elsewhere. In: Karpouzis K, Yannakakis G, eds. *Emotion in Games*. Cham: Springer, 85–102.

Schuller BW. 2018. Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the ACM* **61(5)**:90–99 DOI 10.1145/3129340.

Schuller B, Steidl S, Batliner A. 2009. The INTERSPEECH 2009 emotion challenge. In: *Proceedings INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association*.

Schuller B, Steidl S, Batliner A, Burkhardt F, Devillers L, Müller C, Narayanan SS. 2010. The INTERSPEECH 2010 paralinguistic challenge. In: *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association*. Makuhari.

Tawari A, Trivedi MM. 2010. Speech emotion analysis: exploring the role of context. *IEEE Transactions on Multimedia* **12(6)**:502–509 DOI 10.1109/TMM.2010.2058095.

Teager H, Teager S. 1990. Evidence for nonlinear sound production mechanisms in the vocal tract. In: Hardcastle WJ, Marchal A, eds. *Speech Production and Speech Modelling.* Dordrecht: Springer, 241–261.

Tripathi K, Rao KS. 2017. Improvement of phone recognition accuracy using speech mode classification. *International Journal of Speech Technology* **21(3)**:1–12 DOI 10.1007/s10772-017-9483-4.

Tripathi K, Rao KS. 2018. Improvement of phone recognition accuracy using speech mode classification. *International Journal of Speech Technology* **21(3)**:489–500 DOI 10.1007/s10772-017-9483-4.

Vapnik V. 1998. *Statistical learning theory.* Hoboken, New Jersey: John Wiley & Sons.

Venkataramanan K, Rajamohan HR. 2019. Emotion recognition from speech. *ArXiv preprint* DOI 10.48550/arXiv.1912.10458.

Wang K, An N, Li BN, Zhang Y, Li L. 2015. Speech emotion recognition using fourier parameters. *IEEE Transactions on Affective Computing* **6(1)**:69–75 DOI 10.1109/TAFFC.2015.2392101.

Xu M, Zhang F, Khan SU. 2020. Improve accuracy of speech emotion recognition with attention head fusion. In: *2020 10th Annual Computing and Communication Workshop and Conference (CCWC).* Picataway: IEEE, 1058–1064.

Yang N, Yuan J, Zhou Y, Demirkol I, Duan Z, Heinzelman W, Sturge-Apple M. 2017. Enhanced multiclass SVM with thresholding fusion for speech-based emotion classification. *International Journal of Speech Technology* **20(1)**:27–41 DOI 10.1007/s10772-016-9364-2.

Yu C, Tapus A. 2019. Interactive robot learning for multimodal emotion recognition. In: *International Conference on Social Robotics.* Cham: Springer, 633–642.

Zhao X, Zhang S, Lei B. 2014. Robust emotion recognition in noisy speech via sparse representation. *Neural Computing and Applications* **24(7–8)**:1539–1553 DOI 10.1007/s00521-013-1377-z.

Zheng C, Wang C, Sun W, Jia N. 2019. Research on speech emotional feature extraction based on multidimensional feature fusion. In: *International Conference on Advanced Data Mining and Applications.* Cham: Springer, 535–547.