Knowledge-Based Systems 184 (2019) 104886

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Bagged support vector machines for emotion recognition from speech ☆

Anjali Bhavan [a], Pankaj Chauhan [b], Hitkul [c], Rajiv Ratn Shah [c,*]

[a] Delhi Technological University, New Delhi, India
[b] St. Francis Institute of Technology, University of Mumbai, Mumbai, India
[c] Indraprastha Institute of Information Technology, New Delhi, India

ARTICLE INFO

Article history:
Received 24 December 2018
Received in revised form 14 June 2019
Accepted 27 July 2019
Available online 2 August 2019

Keywords:
Speech emotion recognition
Machine learning
Ensemble learning

ABSTRACT

Speech emotion recognition, a highly promising and exciting problem in the field of Human Computer Interaction, has been studied and analyzed over several decades. It concerns the task of recognizing a speaker's emotions from their speech recordings. Recognizing emotions from speech can go a long way in determining a person's physical and psychological state of well-being. In this work we performed emotion classification on three corpora — the Berlin EmoDB, the Indian Institute of Technology Kharagpur Simulated Emotion Hindi Speech Corpus (IITKGP-SEHSC), and the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). A combination of spectral features was extracted from them which was further processed and reduced to the required feature set. Ensemble learning has been proven to give superior performance compared to single estimators. We propose a bagged ensemble comprising of support vector machines with a Gaussian kernel as a viable algorithm for the problem at hand. We report the results obtained on the three datasets mentioned above.

1. Introduction

Speech is one of the primary means of communication among human beings. One can convey their emotions, state of mind *etc.* through speech, and speech related applications have sprung up in numerous areas such as personal digital assistants, text-to-speech models, sensors and others. Thus, the natural next step is to teach a computer to interact just like humans, in that it could learn to understand the emotions underlying spoken language and respond appropriately. This is why it becomes important to train a machine to recognize the emotions of people from their speech.

The task of recognizing emotions in speech (both speaker-dependent and independent) has been a subject of considerable interest for quite some time. This is a problem that is highly challenging and multi-dimensional, because various emotions can be conveyed differently in different forms of speech. Also, the task of determining what all features to extract from speech to analyze its inherent emotions is a different problem in itself.

The existing approaches to this problem mostly make use of support vector machines (SVMs), hidden Markov models (HMMs) or neural networks. While SVMs provide reasonably good estimates with lesser effort, neural networks and hidden Markov models are difficult to build and train, and require high computational power and time. There thus needs to be a method to enhance the performance of support vector machines on the problem. This is where ensemble learning comes into the picture.

Ensemble learning [1] comprises of training multiple estimators, and aggregating their outcomes using particular rules. Some of the prominent ways of building ensembles include bagging (bootstrap aggregating) and boosting. Both these methods usually comprise of ensembling similar learners. Bagging, however, is a parallel mechanism, while boosting is an iterative procedure.

Our approach comprises of examining the performance of these ensemble methods on the problem of emotion recognition from speech. Particularly, we wish to assess the performance of ensembles of support vector machines. We compare the bagged and boosted ensembles prepared from the same, and observe that the bagging estimator demonstrates a better performance as compared to boosting.

Section 2 summarizes the previous research done in speech emotion recognition and ensemble learning methods. Section 3 gives a general overview of the system, including the model description and the process of feature extraction. Section 4 gives a thorough description of the datasets used, the experimental setup and procedure. Section 5 subsequently reports observations and compares results with some state-of-the-art systems, and Section 6 then proceeds to derive conclusions from the observations.

☆ No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.knosys.2019.104886.

* Corresponding author.
E-mail address: rajivratn@iiitd.ac.in (R.R. Shah).

https://doi.org/10.1016/j.knosys.2019.104886
0950-7051/

2. Prior research

This section is divided into two parts: one covering research on emotion recognition from speech, and the other covering advances in ensemble learning methods.

2.1. Emotion recognition from speech

To correctly recognize the emotion from speech data, it is very important to extract the features which accurately represent the emotional aspect of speech signals. One of the biggest challenges in this field is to extract efficient features for the best classification of emotions. Some notable works in this area include analysis and synthesis of emotional speech [2,3].

Mel Frequency Cepstral Coefficients (MFCCs) have been studied and applied frequently for tasks like speech recognition and speaker identification [4]. Existing studies have found that MFCCs are a far preferable way of analyzing emotions compared to other commonly used speech features (*e.g.*, loudness, formants, linear predictive coefficients *etc.*) [5]. Bou-Ghazale and Hansen [6] demonstrated that the features based on cepstral analysis, outperform the linear features of Linear Predictive Coding (LPC) in detecting emotion in speech. Liu [7] showed that a feature set comprising of Gammatone Frequency Cepstral Coefficients (GFCCs) gave an average increase of 3.6% in accuracy over MFCCs for emotion detection. In addition, voice quality features such as jitter and shimmer, glottal parameter, *etc.* are also related to emotion in speech [8,9]. Li et al. [10] extracted jitter and shimmer as voice quality parameters mixed with MFCC features to identify emotions on SUSAS database.

Recently, the combination of different kinds of features has been widely used for emotion recognition in speech. Pan et al. [11] showed that the combination of MFCCs, Mel-energy spectrum dynamic coefficients (MEDCs) and energy with a SVM classifier on a self-constructed Chinese emotional database and the EmoDB. The difference between MEDCs and MFCCs is that MEDCs are calculated as the logarithmic average of energies after the filter bank, while MFCCs are calculated as the logarithmic after the filter bank. Chen et al. [12] used a three-level speech emotion recognition model to solve the speaker independent emotion recognition problem and extracted the energy, zero crossing rate (ZCR), pitch, the first to third formants, spectrum centroid, spectrum cut-off frequency, correlation density, fractal dimension, and five Mel-frequency bands energy. The three levels classify the six emotions pairwise, with each level providing finer classification than the last.

Schuller et al. [13] proposed the usage of a multiple-stage classifier with a support vector machine over 7 emotional classes, with the aim of employing both acoustic and linguistic features for emotion classification. A deep belief network was used for spotting emotional key-phrases. Various classifiers (Gaussian Mixture Models (GMMs), SVMs, Neural Networks, Nearest Neighbors) were used for training on the acoustic features, and then combined with the belief network using a neural network and their performances evaluated.

Liu et al. [14] proposed a feature selection method based on correlation analysis and Fisher criterion and used extreme learning machine (ELM) decision trees as the classification method on the Chinese speech database from the Institute of Automation of Chinese Academy of Sciences (CASIA). The idea behind using the Fisher correlation coefficient was to remove redundant features, which was a possibility considering that features are extracted from the same audio sources for emotion recognition.

Fahad et al. [15] used a DNN-HMM speaker adaptive model on IEMOCAP and IITKGP-SEHSC databases. Features based on glottal closure instants (also called epochs) were used in combination with MFCCs. The epoch features that were used were instantaneous pitch, strength of excitation (SoE) and instantaneous phase. The idea behind the extraction and usage of these features was that speech features tend to change very rapidly due to the vibrations of the vocal chords of the speaker; such changes are not captured in general prosodic features, which assume the speech signal to be static. Hence, excitation source features in the form of GCIs were extracted from the signals and studied (see Table 1).

2.2. Ensemble learning

Ensemble learning constitutes the process of combining the learning procedures of multiple models in other to give a final, (usually) stronger learner. Such methods have been used in a wide range of application areas — including credit scoring [16], medical diagnosis [17] and accent prediction [18]. Many kinds of ensemble techniques have been proposed [19], of which the primary ones are bagging [20] and boosting [21].

Hypotheses generated using ensembles made of diverse base estimators have been demonstrated to be far superior to single hypotheses [22]. Quinlan [23] conducted trials over a diverse dataset collection and demonstrated bagging and boosting ensembles as performing noticeably well.

Bagging comprises of training several estimators on subsets of the dataset chosen randomly. If drawn with replacement, the samples are known as bootstrap samples. This approach has been used in several areas of study, for instance traffic forecasting [24] and credit card fraud detection [25].

Ensemble methods have been applied to audio data as well. Schuller et al. [26] presented an analysis of ensemble machine learning on speaker-independent speech emotion recognition, and reported improved accuracy on data scraped from movie content. Morrison et al. [27] ensembled various classifiers using an unweighted vote rule, and used it on emotion recognition in call-center speech.

Particularly, bagged ensembles of support vector machines have been analyzed in a few works [28]. Hu et al. [29] used such an ensemble for the problem of fault detection in rotating machinery. However, such works are few and far in between, and we hope to further analyze this model and apply it for emotion recognition from speech.

3. System overview

This section will cover the system overview — that is, the nature and quantity of features extracted and the structure and design of the model used.

3.1. Feature extraction

The primary aspect of analyzing emotions inherent in speech data is the set of features extracted from the same. The right set of features extracted could go a long way in developing sound speech emotion analysis. Many works have been proposed in this direction [2,11], with Mel-Frequency Cepstral Coefficients (MFCCs) emerging as some of the most popular [4].

In order to analyze the speech data using machine learning techniques, we extracted spectral features from the datasets and prepared a feature vector from each. The following features were extracted:

1. Mel-Frequency Cepstral Coefficients (MFCCs) [30]: These coefficients are a better way of representing sound as heard by the human ear. Since the cochlea in human ears perceives frequency of sounds by vibrating according to the present frequencies (information on which then travels to

Table 1
Existing research in SER.

Authors	Features	Model	Accuracy
Pan et al.	MFCC, MEDCs and energy	SVM	91.3% on Chinese corpus and 95.1% on EmoDB
Chen et al.	Prosodic + spectral features	3-level system based on SVMs	86.5%, 68.5% and 50.2% on each level resp.
Schuller et al.	Acoustic + linguistic features	Linear classifiers with a belief network	81.19%
Liu et al.	Prosodic + spectral features	ELM decision trees	89.6%
Fahad et al.	MFCCs and epoch features	DNN-HMM speaker adaptive model	64.2%

the brain by nerve firings), it makes more sense to quantify perceived frequency according to the actual measured frequencies — which is where the Mel scale is used. The formula for converting from frequency to the Mel scale is:

$$M(f) = 1125ln(1 + f/700) \tag{1}$$

2. Delta and Delta–Delta MFCCs: These coefficients are also known as differential and acceleration coefficients, respectively, and characterize the trajectories of the MFCCs over time.
3. Spectral Centroids: These coefficients are the spectral sub-band centroids of each frame, and are usually 26 in number.

Each audio file (signal) was first divided into frames of length 25 ms each, with frame step 10 ms — these values being the usual standard ones used in speech emotion recognition works [31]. Then for each frame:

- The Discrete Fourier Transform is calculated. A 512 point FFT (number can be varied according to data) is calculated and the first 257 points are kept.
- The periodogram-based power spectral estimate for each frame is calculated by squaring the result of the absolute value of the complex Fourier transform calculated previously.
- The Mel-space filterbank is calculated by applying 26 filters to the periodogram-based power spectral estimate calculated previously. This gives us 26 numbers describing the energy of each frame.
- The logarithm of each of the 26 numbers is calculated to give us 26 log filterbank energies, and a Discrete Cosine Transform is performed on these to give us 26 cepstral coefficients, of which we keep the first 13 as the final Mel-Frequency Cepstral Coefficients.

The delta and delta-delta coefficients are next calculated using the following equation:

$$d_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2} \tag{2}$$

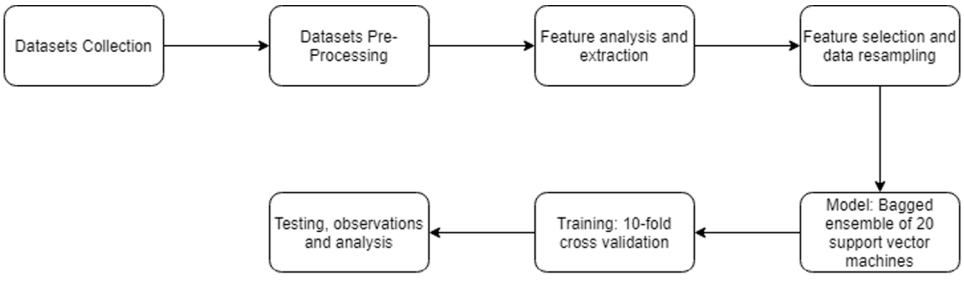
where d_t is a delta coefficient from frame t computed in terms of c_{t+n} to c_{t-n}. N is the number of samples, and c_{t+n} to c_{t-n} constitute the static coefficients. The delta-delta coefficients are calculated as the delta of the delta coefficients using the same formula.

The spectral sub-band centroids are calculated next, 26 for each frame. Since the length of the audio files varies, the above coefficients alone cannot give us a uniform feature vector — because the number of frames vary due to the varying audio file lengths. In order to get a proper feature vector from the above features, we calculated seven values for each audio file based on the values of each frame constituting the file: the mean, variance, maximum value, minimum value, skewness, kurtosis and inter-quartile range. These values were calculated for each audio file over all the frames and for each coefficient, which gave us a feature vector of size $(13 + 13 + 13 + 26) * 7 = 455$.

3.2. Model description

We use a bagging ensemble method as our model for the data. Bagging, short for bootstrap aggregating, consists of training samples (drawn at random, hence called bootstrap samples) fed into the various base estimators of the ensemble, then combining and deciding on the final predictions by using a majority voting rule.

Our base estimator was a support vector machine with a Gaussian kernel, penalty term 100 and kernel coefficient 0.1. We combined 20 of these in a bagging ensemble, and prepared it



so samples were drawn from the training set as subsets of the feature set as well as the training examples. This took care of the correlation factor that could arise when similar estimators are trained on samples drawn with replacement.

In our bagging ensemble, 20 sample sets were drawn from the training set with replacement, and then trained on each base estimator in a parallel manner. The results obtained are then aggregated using averaging to give the final predictions.
We compared this with an AdaBoost [32] ensemble of support vector machines. AdaBoost, developed by Freund and Schapire, works by training estimators in series, unlike in bagging where the estimators are trained separately in parallel. AdaBoost gives a final, strong prediction by iteratively improving upon the errors made in each step by the estimator. It is a forward stagewise additive model with an exponential loss function; weights are adjusted accordingly after each iteration so that misclassifications are penalized and the classification ability improves.

4. Experimental setup

4.1. Datasets description

We present results on three emotional speech corpora, the details of which are described below.

4.1.1. Berlin EmoDB

The Berlin EmoDB [33] is an emotional corpus in the German language consisting of ten actors (5 male, 5 female) speaking ten German utterances in various emotions. The corpus consists of 7 emotions: Happy, Sad, Angry, Boredom, Fear, Neutral, Disgust and a total of 535 audio files in the .wav format with a sampling rate of 16,000 Hz.

4.1.2. RAVDESS

The Ryerson Audio-Visual Database of Emotional Speech and Song [34] is an emotional song and speech corpus in the English language consisting of 24 actors (12 male, 12 female). Each expression is produced at two levels of emotional intensity (normal, strong). This experiment uses the speech part of the corpus, which consisted of 8 emotions: Happy, Sad, Angry, Calm, Fear, Neutral, Disgust, Surprise and a total of 1440 audio files in the .wav format with a sampling rate of 48,000 Hz.

4.1.3. IITKGP-SEHSC

The Indian Institute of Technology Kharagpur Simulated Emotion Hindi Speech Corpus (IITKGP-SEHSC) [35] is an emotional song and speech corpus in the Hindi language consisting of 10 speakers (5 male, 5 female) speaking 15 utterances in 10 sessions. The corpus consisted of 8 emotions: Happy, Sad, Angry, Sarcastic, Fear, Neutral, Disgust, Surprise and a total of 1200 audio files per speaker in the .wav format. The sampling rate was 16,000 Hz.

4.2. Data pre-processing

The feature vector having been obtained, the data was first scaled to the range (0, 1), and then split into training and test data with a 90:10 proportion. Next, Boruta [36], a wrapper-based all-relevant feature selection method was applied on the data in order to reduce the size of the feature vector. Table 2 gives the size of the feature vector for each dataset after feature selection using Boruta.

Since the EmoDB and RAVDESS datasets were highly imbalanced, data resampling techniques were applied on them so as to have better training and results. We used the imblearn package [37] for the same.

Data resampling for unbalanced classes can be done in two ways: over-sampling (increasing the number of samples in the smallest class to bring it to par with the other classes) and under-sampling (decreasing the number of samples in the largest class to bring it to par with the other classes). Combinations of both, which would oversample the smaller class and undersample the larger class are also used. We use the combination of SMOTE over-sampling and Tomek Links under-sampling in our work.

Table 2
No. of features for each database after feature selection.

Dataset	Features after feature selection
EmoDB	147
RAVDESS	183
IITKGP-SEHSC	403

Table 3
Experimental Observations.

Dataset	Training accuracy	Holdout set accuracy
EmoDB	96.25%	92.45%
RAVDESS	79.85%	75.69%
IITKGP-SEHSC	85.72%	84.11%

4.3. Model training

The entire procedure was carried out using the scikit-learn package for machine learning algorithms and resources [44]. Training and evaluation on the datasets was performed using 10-fold cross validation, with accuracy chosen as the cross-validation metric. The model was trained in the one versus rest fashion. A similar procedure was followed for the AdaBoost ensemble as well.

For bagging, 20 subsets of the data are sampled uniformly and with replacement from the dataset. In our experiments, these are drawn as subsets of both samples and features. One SVM is trained on each subset, and the results from the 20 models are then aggregated using averaging. In the case of AdaBoost, the base estimators are trained sequentially on repeatedly modified versions of the dataset. The final estimation is given by the weighted aggregate of the individual predictions. In each iteration, the sample weights are adjusted and the learner is reapplied. The weights are adjusted such that the misclassified samples gain more weight compared to the correctly classified ones, so that the focus on them is increased with each further iteration (see Fig. 1).

5. Results and discussion

We first extracted 455 features from the datasets for emotion recognition and then reduced their dimensionality using Boruta. As per 4.3 the recognition rate remained approximately same, while the number of features required is reduced. In case of the EmoDB dataset the reduction is almost 68%, whereas the recognition rate improved by 7% compared to results obtained by using all the 455 features.

The proposed method was also evaluated using only MFCC features, but their recognition rate was as low as roughly 66% on RAVDESS database. It suggests that a feature set comprising of MFCCs alone is not descriptive enough for speech emotion recognition, which has also been shown in [35].

Table 3 shows the training and test accuracies with MFCCs and spectral centroid features on EmoDB, RAVDESS and IITKGP-SEHSC databases. Tables 4–6 demonstrate the veracity of our results in a more quantified manner. Table 7 reports the performances of the model on each emotion in the datasets.

Since the works with which we compare our results have mostly not specified the split ratio which was used for the training and test sets, we have proceeded to present our results

Table 4
Comparison of proposed method based on recognition performance on EmoDB corpus.

Reference	Classifier used	Features used	Obtained accuracy	Split ratio
Wang et al.[38]	SVM (Gaussian kernel)	Fourier parameters, MFCCs and derivatives	88.9%	90:10
Kotti et al. [39]	Linear SVM	Pitch and cepstrum-based features	87.7%	Not mentioned
Wang et al.[40]	SVM (Gaussian kernel)	Wavelet packet coeff.	79.5%	Not mentioned
Guo et al. [41]	CNN	Amplitude spectrogram and phase information	91.28%	Not mentioned
Proposed model	Bagged ensemble of SVMs	MFCCs, spectral centroids and MFCC derivatives	**92.45%**	90:10
AdaBoost model	AdaBoost ensemble of SVMs	MFCCs, spectral centroids and MFCC derivatives	87.32%	90:10

Table 5
Comparison of proposed method based on recognition performance on RAVDESS corpus.

Reference	Classifier used	Features used	Obtained accuracy	Split ratio
Zeng et al. [42]	DNNs	Spectrograms	64.52%	Not mentioned
Shegokar and Sircar [43]	SVMs	Continuous wavelet transform, prosodic coefficients	60.1%	Not mentioned
Proposed model	Bagged ensemble of SVMs	MFCCs, spectral centroids and MFCC derivatives	**75.69%**	90:10
AdaBoost model	AdaBoost ensemble of SVMs	MFCCs, spectral centroids and MFCC derivatives	72.10%	90:10

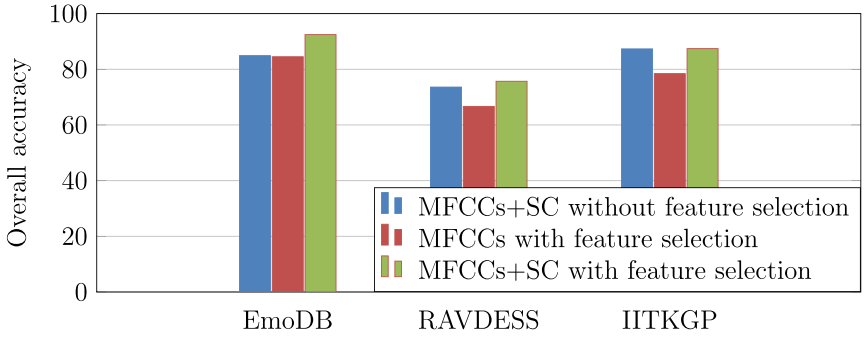


Fig. 1. Comparisons of recognition rates (where SC: spectral centroids).

with a 90:10 split ratio only. The only exception to this is the works with which the performance on IITKGP-SEHSC dataset is compared, which have mentioned a split ratio of 70:30. For that case, we split our dataset (IITKGP-SEHSC in this case) identically, and compare results. We have also produced results with 90:10 split ratio for the IITKGP-SEHSC dataset to see how much is the difference in performance between 70:30 and 90:10. We found not much difference in performance for these two split rations (see Table 6). For all other datasets and comparisons, the split ratio is 90:10 as noted above. Thus, we believe that despite the exact split ratio for different related works are not known the difference in performance between 90:10 and the split ratio these experiments used is not significant and we may compare our results for 90:10 with results reported in these papers. Note that for the IITKGP-SEHSC dataset, our proposed methods outperform on both 70:30 and 90:10 split rations, thus we believe the same trends for other two datasets as well.

Wang et al. [38] explored Fourier parameter-based features for their emotion recognition model. The idea behind capturing and using Fourier parameter features was steeped in music theory that stated that harmony structures of intervals were responsible for the overall perception of the musical piece by the listeners. They concluded that a combination of Fourier parameter features with MFCCs produced the best results compared to either of the feature sets when used individually. This concurs with our observation that individual MFCC features do not demonstrate as strong a performance as when combined with other feature sets.

Kotti et al. [39] presented a binary cascade classification schema which focused on classifying between pairs of emotion categories instead of all emotions at once, so easily-confused emotions can be easily separable. They extracted numerous features based on prosody, formants, energy, pitch, jitter and TEO-autocorrelation, and also took the difference that would arise in speech signal patterns owing to gender into consideration.

Table 6
Comparison of proposed method based on recognition performance on IITKGP-SEHSC corpus.

Reference	Classifier used	Features used	Obtained accuracy	Split ratio
Bhaykar et al.[45]	GMMs	MFCCs	73.68%	Not mentioned
Proposed model	Bagged ensemble of SVMs	MFCCs, spectral centroids and MFCC derivatives	**84.11%**	70:30
AdaBoost model	AdaBoost ensemble of SVMs	MFCCs, spectral centroids and MFCC derivatives	77.19%	90:10

Our experiments do not take gender into consideration and uses a smaller number and limited kind of features, but still demonstrate better performance than the results reported by the authors in the former. Extra features and gender-based analysis could possibly be an additional direction of research and even an enhancement, and would likely be explored in further experiments.

In another set of experiments, Wang et al. [40] studied wavelet packet transforms as features for emotion recognition. Wavelet packet transforms provide for better frequency resolution at low frequencies, and even somewhat resemble auditory perception in humans. They reported an accuracy of 79.5%, which is a 6% improvement over their results without wavelet packet analysis. This result was obtained with the combination of wavelet packet analysis and sequential forward feature selection, a wrapper-based feature selection method such as the one used in our experiments. A case could be made that wrapper-based feature selection algorithms tend to help with the recognition rates in SER.

In the case of the RAVDESS database, substantial work was difficult to be found, owing to its recent introduction into the field. Owing to this, only a few works could be found and analyzed. Zeng et al. [42] experimented with various deep neural networks (including gated residual networks), and reported a best accuracy of 64.52%. Shegokar and Sircar [43] used a SVM with a feature set obtained from prosodic coefficients and continuous wavelet transforms, and reported a best accuracy of 60.1%. In contrast, our best performance is 75.69%.

In order to understand the effect of the classification method, our emotion recognition system is also evaluated using a simple support vector machine classifier which has achieved overall accuracies of 86.69% and 72.91% for EmoDB and RAVDESS databases respectively. With the use of a bagged ensemble comprising of support vector machines for classification, this accuracy is further enhanced by roughly 5%.

Finally, from the observations of the AdaBoost ensemble's performance compared with the others (in Tables 4–6), it can be inferred that our bagged model gives a better performance than the AdaBoost ensemble on all the three datasets. It could be possible that AdaBoost is not a suitable learner for this class of problems, although further experiments can help prove or disprove this claim more rigorously.

6. Conclusion and future work

In this work, we proposed a bagged ensemble comprising of support vector machines with a Gaussian kernel for SER. We firstly extracted MFCCs along with spectral centroids to represent emotional speech followed by a wrapper-based feature selection method to retrieve the best feature set. Experiments on the EmoDB, RAVDESS and IITKGP-SEHSC databases show the superiority of our proposed approach compared with the state-of-the-art in terms of overall accuracy.

Table 7
Performance on the three datasets according to emotion.

Emotion	EmoDB	RAVDESS	IITKGP-SEHSC
Happiness	100%	78.81%	87.5%
Neutral	100%	74.32%	81.23%
Anger	100%	82.69%	92.76%
Fear	66.67%	67.91%	84.4%
Sadness	77.77%	71.88%	83.32%
Disgust	100%	75.02%	82.00%
Boredom	87.5%		
Calm		79.45%	
Surprise		80.28%	83.98%
Sarcastic			85.51%

Many questions and potential avenues for further research have emerged during and after our experiments. Some of them are listed below.

- Currently, our work focuses on acoustic features only; could linguistic features in combination with acoustic ones improve performance and help gain new insights in SER? How much can semantic features extracted from speech help determine the emotions inherent in it?
- How much and in what way does gender influence the patterns observed in speech signals, and how best can it be studied for recognizing emotions?
- In recent times, deep learning methods have been introduced for emotion recognition, which learn the features from speech in the network itself instead of using hand-crafted features. Would such features fare well in comparison with carefully studied and hand-crafted features?

Acknowledgments

This article has received funding from Infosys Center for AI, IIIT-Delhi and ECRA Grant (ECR/2018/002449) by SERB, Government of India.

References

[1] Cha Zhang, Yunqian Ma (Eds.), Ensemble Machine Learning, in: Methods and Applications, Springer Science & Business Media, 2012.
[2] Jeong-Sik Park, Ji-Hwan Kim, Yung-Hwan Oh, Feature vector classification based speech emotion recognition for service robots, IEEE Trans. Consum. Electron. 55 (3) (2009).
[3] Eun Ho Kim, Kyung Hak Hyun, Soo Hyun Kim, Yoon Keun Kwak, Improved emotion recognition with a novel speaker-independent feature, IEEE/ASME Trans. Mechatronics 14 (3) (2009) 317–325.
[4] Md Rashidul Hasan, Mustafa Jamil, M.G.R.M.S. Rahman, Speaker identification using mel frequency cepstral coefficients, Variations 1 (4) (2004).
[5] Namrata Dave, Feature extraction methods LPC, PLP and MFCC in speech recognition, Int. J. Adv. Res. Eng. Technol. 1 (6) (2013) 1–4.
[6] Sahar E. Bou-Ghazale, John HL. Hansen, A comparative study of traditional and newly proposed features for recognition of speech under stress, IEEE Trans. Speech Audio Process. 8 (4) (2000) 429–442.

[7] Gabrielle K. Liu, Evaluating Gammatone frequency cepstral coefficients with neural networks for emotion recognition from speech, 2018, arXiv preprint arXiv:1806.09010.

[8] Christer Gobl, AilbheNı Chasaide, The role of voice quality in communicating emotion, mood and attitude, Speech Commun. 40 (1–2) (2003) 189–212.

[9] Bin Yang, Marko Lugger, Emotion recognition from speech signals using new harmony features, Signal Process. 90 (5) (2010) 1415–1423.

[10] Xi Li, Jidong Tao, Michael T. Johnson, Joseph Soltis, Anne Savage, Kirsten M. Leong, John D. Newman, Newman stress and emotion classification using jitter and shimmer features, in: Acoustics, Speech and Signal Processing, ICASSP 2007, IEEE International Conference On, Vol. 4, IEEE, 2007, pp. IV–1081.

[11] Yixiong Pan, Peipei Shen, Liping Shen, Speech emotion recognition using support vector machine, Int. J. Smart Home 6 (2) (2012) 101–108.

[12] Lijiang Chen, Xia Mao, Yuli Xue, Lee Lung Cheng, Speech emotion recognition: features and classification models, Digit. Signal Process. 22 (6) (2012) 1154–1160.

[13] Björn Schuller, Gerhard Rigoll, Manfred Lang, Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture, in: Acoustics, Speech, and Signal Processing, Proceedings, ICASSP'04, IEEE International Conference on, Vol. 1, IEEE, 2004, pp. I–577.

[14] Zhen-Tao Liu, Min Wu, Wei-Hua Cao, Jun-Wei Mao, Jian-Ping Xu, Guan-Zheng Tan, Speech emotion recognition based on feature selection and extreme learning machine decision tree, Neurocomputing 273 (2018) 271–280.

[15] Md Fahad, Jainath Yadav, Gyadhar Pradhan, Akshay Deepak, DNN-HMM based speaker adaptive emotion recognition using proposed epoch and MFCC features, 2018, arXiv preprint arXiv:1806.00984.

[16] Gang Wang, Jinxing Hao, Jian Ma, Hongbing Jiang, A comparative assessment of ensemble learning for credit scoring, Expert Syst. Appl. 38 (1) (2011) 223–230.

[17] Akin Ozcift, Arif Gulten, Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms, Comput. Methods Programs Biomed. 104 (3) (2011) 443–451.

[18] Xuejing Sun, Pitch accent prediction using ensemble machine learning, in: Seventh International Conference on Spoken Language Processing, 2002.

[19] Thomas G. Dietterich, Ensemble methods in machine learning. in: International workshop on multiple classifier systems, Springer, Berlin, Heidelberg, 2000, pp. 1-15.

[20] L. Breiman, Bagging predictors, Machine Learn. 24 (2) (1996) 123–140.

[21] R.E. Schapire, The strength of weak learnability, Machine Learn. 5 (2) (1990) 197–227.

[22] Hendrik Jacob van Veen, Le Nguyen The Dat, Armando Segnini, Kaggle ensembling guide, 2015, https://mlwave.com/kaggle-ensembling-guide/ (accessed 6 February 2018).

[23] J.R. Quinlan, Bagging, boosting, and C4 5, in: AAAI/IAAI, Vol. 1 1996, pp. 725-730.

[24] Fabio Moretti, Stefano Pizzuti, Stefano Panzieri, Mauro Annunziato, Urban traffic flow forecasting through statistical and neural network bagging ensemble hybrid modeling, Neurocomputing 167 (2015) 3–7.

[25] Masoumeh Zareapoor, Pourya Shamsolmoali, Application of credit card fraud detection: Based on bagging ensemble classifier, Procedia Comput. Sci. 48 (2015) 679–685.

[26] Björn Schuller, Stephan Reiter, Ronald Muller, Marc Al-Hames, Manfred Lang, Gerhard Rigoll, Speaker independent speech emotion recognition by ensemble classification, in: Multimedia and Expo, ICME 2005, IEEE International Conference on, IEEE, 2005, pp. 864–867.

[27] Donn Morrison, Ruili Wang, Liyanage C. De Silva, Ensemble methods for spoken emotion recognition in call-centres, Speech Commun. 49 (2) (2007) 98–112.

[28] Hyun-Chul Kim, Shaoning Pang, Hong-Mo Je, Daijin Kim, Sung-Yang Bang, Support vector machine ensemble with bagging, in: Pattern Recognition with Support Vector Machines, Springer, Berlin, Heidelberg, 2002, pp. 397–408.

[29] Qiao Hu, Zhengjia He, Zhousuo Zhang, Yanyang Zi, Fault diagnosis of rotating machinery based on improved wavelet package transform and SVMs ensemble, Mech. Syst. Signal Process. 21 (2) (2007) 688–705.

[30] Steven B. Davis, Paul Mermelstein, Mermelstein Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, in: Readings in speech recognition, 1990, pp. 65-74.

[31] Qifeng Zhu, Abeer Alwan, On the use of variable frame rate analysis in speech recognition, in: Acoustics, Speech, and Signal Processing, ICASSP'00, Proceedings 2000, IEEE International Conference on, Vol. 3, IEEE, 2000, pp. 1783–1786.

[32] Yoav Freund, Robert E. Schapire, Experiments with a new boosting algorithm, in: Icml, Vol. 96, 1996, pp. 148-156.

[33] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F. Sendlmeier, Benjamin Weiss, A database of German emotional speech, in: Ninth European Conference on Speech Communication and Technology, 2005.

[34] Steven R. Livingstone, Frank 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] Shashidhar G. Koolagudi, Ramu Reddy Jainath Yadav, K. Sreenivasa Rao, IITKGP-SEHSC: Hindi speech corpus for emotion analysis, in: Devices and Communications, ICDeCom 2011, International Conference on, IEEE, 2011, pp. 1–5.

[36] Miron B. Kursa, Witold R. Rudnicki, Feature selection with the boruta package, J. Stat. Softw. 36 (11) (2010) 1–13.

[37] Guillaume Lemaître, Fernando Nogueira, Christos K. Aridas, Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning, J. Mach. Learn. Res. 18 (1) (2017) 559–563.

[38] Kunxia Wang, Ning An, Bing Nan Li, Yanyong Zhang, Lian Li, Speech emotion recognition using fourier parameters, IEEE Trans. Affect. Comput. 6 (1) (2015) 69–75.

[39] Margarita Kotti, Fabio Paternó, Speaker-independent emotion recognition exploiting a psychologically-inspired binary cascade classification schema, Int. J. Speech Technol. 15 (2) (2012) 131–150.

[40] Kunxia Wang, Ning An, Lian Li, Speech emotion recognition based on wavelet packet coefficient model, in: Chinese Spoken Language Processing, ISCSLP 2014, 9th International Symposium on, IEEE, 2014, pp. 478–482.

[41] Lili Guo, Longbiao Wang, Jianwu Dang, Linjuan Zhang, Haotian Guan, Xiangang Li, Speech Emotion Recognition by Combining Amplitude and Phase Information Using Convolutional Neural Network, in: Proc. Interspeech 2018, 2018, pp. 1611-1615.

[42] Yuni Zeng, Hua Mao, Dezhong Peng, Zhang Yi, Spectrogram based multi-task audio classification, in: Multimedia Tools and Applications, 2017, pp. 1–18.

[43] Pankaj Shegokar, Pradip Sircar, Continuous wavelet transform based speech emotion recognition, in: Signal Processing and Communication Systems, ICSPCS 2016, 10th International Conference on, IEEE, 2016, pp. 1–8.

[44] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al., Scikit-learn: machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[45] Manav Bhaykar, Jainath Yadav, K. Sreenivasa Rao Speaker dependent, Speaker dependent speaker independent and cross language emotion recognition from speech using GMM and HMM, in: Communications, NCC 2013, National Conference On, IEEE, 2013, pp. 1–5.