Noname manuscript No.
(will be inserted by the editor)

Bagging With Strong Learners for Emotion Recognition from Speech

Anjali Bhavan · Pankaj Chauhan · Rajiv Ratn Shah

Received: date / Accepted: date

Abstract Speech emotion recognition, a highly promising and exciting problem in the field of Human Computer Interaction, has been studied and analyzed over several years, and 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, and also provides us with valuable information regarding language etc. In this work we analyzed 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), and extracted spectral features from them which were further processed and reduced to the required feature set by using Boruta, a wrapper-based feature selection method. We propose a bagged ensemble comprising of Support Vector Machines with a Gaussian kernel as a viable algorithm for the problem at hand, and report the results obtained on the three datasets above.

Keywords Speech Emotion Recognition · Machine Learning · Ensemble Learning

F. Anjali Bhavan
Delhi Technological University, New Delhi
Tel.: +91-9654647823
E-mail: anjalibhavan98@gmail.com

F. Pankaj Chauhan
St. Francis Institute of Technology, University of Mumbai, Mumbai
Tel.: +91-9769492087
E-mail: cpankajr@gmail.com

S. Dr.Rajiv Ratn Shah
Indraprastha Institute of Information Technology, New Delhi

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 so on. 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 - which is why it becomes important to train a machine to recognize the emotions of a person 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, and the task of determining what all features to extract from speech to analyze its inherent emotions correctly is a challenge in itself.
There is also the point to note that comparison with previous works in this field is usually not stringent - because the features extracted, the nature of the corpus used, the proportion of data used for training and the model settings etc. are different in every case. We will present here the research done on various datasets, extracting various sets of features and employing different models and data proportions over the years.

2 Prior Research

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 and they should be speaker independent. 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 [1],[2]. Mel Frequency Cepstral Coefficients (MFCCs) are proved to be very useful in the field of speech recognition. And existing studies have found that MFCCs are also performing well in the field of SER as compare to other commonly used speech features (e.g. loudness, formants, linear predictive coefficients). Bou-Ghazale and Hansen [6] found that the features based on cepstral analysis, such as LPCC and MFCC, outperform the linear features of LPC in detecting speech emotions.
Experimental results founded by Gabrielle K. Liu [12] shows that Gammatone Frequency Cepstral Coefficients (GFCCs) gives an average increase in accuracy over MFCCs of 3.6% for emotion detection. In addition, voice quality features such as jitter and shimmer, glottal parameter, etc. are also related to emotion in speech [4],[5]. X. Li [7] extracted jitter, shimmer as voice quality parameters mixed with MFCC features to identify emotions on SUSAS database, and compared to MFCC, the accuracy increased by 4%.
Recently, the combination of different kinds of features has been widely used for

ASER. Pan [3] used the combination of MFCC, Mel-energy spectrum dynamic coefficients and Energy and obtained overall accuracy of 91.3% on Chinese emotional database and 95.1% on Berlin emotional database (EMODB). Chen [8] used a three-level speech emotion recognition model to solve the speaker independent emotion recognition problem and has 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. Average accuracies for each level were 86.5%, 68.5% and 50.2%, respectively [8].
B. Schuller [11] used a multiple stage classifier with support vector machine (SVM) for speech emotion recognition problem and reported an accuracy of 81.19% with 7 emotional classes. Liu and Wu [10] proposed feature selection method based on correlation analysis and Fisher criterion and obtained 89.6% recognition rate on average using extreme learning machine (ELM) decision tree as classification method on Chinese speech database from institute of automation of Chinese academy of sciences (CASIA). J. Yadav and Md. Fahad [9] used DNN-HMM speaker adaptive model using combined features (MFCC and epoch features) on IMPCAP and IITKGP-SEHSC database. And they observed 5.34% increase in recognition rate compared to model developed using MFCC features only[9].

3 Materials and experimental method

3.1 Datasets description

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

3.1.1 Berlin EmoDB

The Berlin EmoDB [17] 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 16000 Hz. Table 1 gives the number of samples of each emotion in the corpus.

3.1.2 RAVDESS

The Ryerson Audio-Visual Database of Emotional Speech and Song [18] 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 paper 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

Table 1 Class distribution of Berlin EmoDB database

Emotion	Number of samples
Happiness (Freude)	71
Neutral (Neutral)	79
Anger (rger)	127
Fear (Angst)	69
Sadness (Trauer)	62
Disgust (Ekel)	46
Boredom (Langeweile)	81

Table 2 Class distribution of RAVDESS database

Emotion	Number of samples
Happiness	192
Neutral	192
Anger	192
Fear	192
Sadness	192
Disgust	192
Calm	192
Surprise	96

Table 3 Class distribution of IITKGP-SEHSC database

Emotion	Number of samples
Happiness	1500
Neutral	1500
Anger	1500
Fear	1500
Sadness	1500
Disgust	1500
Sarcastic	1500
Surprise	1500

audio files in the .wav format with a sampling rate of 48000 Hz. Table 2 gives the number of samples of each emotion in the corpus.

3.1.3 IITKGP-SEHSC

The Indian Institute of Technology Kharagpur Simulated Emotion Hindi Speech Corpus (IITKGP-SEHSC) [19] 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 15 utterances * 8 emotions * 10 sessions = 1200 audio files per speaker in the .wav format, with a sampling rate of 16000 Hz. Table 3 gives the number of samples of each emotion in the corpus.

3.2 Feature Extraction

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) [20]: 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 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:
$$M(f) = 1125ln(1 + f/700)$$
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 25ms each, with frame step 10ms. 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:

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

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

Table 4 No. of features for each database after dimensionality reductione

Dataset	Features after dimensionality reduction
EmoDB	147
RAVDESS	183
IITKGP - SEHSC	403

The spectral sub-band centroids are calculated next, 26 for each frame. Since the length of the audio files vary, 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 dimension $(13 + 13 + 13 + 26) * 7 = 455$.

3.3 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, a wrapper-based all-relevant feature selection method was applied on the data in order to reduce the dimensionality of the feature vector. Table 4 gives the size of the feature vector for each dataset after feature selection using Boruta.

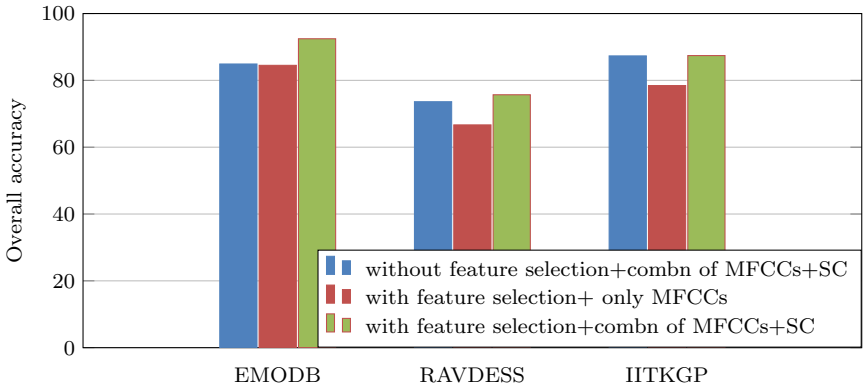
Now since the EmoDB and RAVDESS datasets were high imbalanced, data resampling techniques were applied on them so as to have better training and results. We used the imblearn package [21] for the same.

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.

3.4 Model building and training

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.

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



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 set bootstrap features as True - so samples are drawn from the training set with replacement. This entire procedure was carried out using the scikit-learn Python package. [22]

We trained and evaluated on the datasets using 10-fold cross validation, with accuracy chosen as the cross-validation metric (as the datasets are now balanced). A small holdout set (10*percent*) was kept aside untouched, and the cross-validation procedure was performed on the remaining data. The model was trained in the one versus rest fashion.

3.5 Results and Conclusions

We first extracted 455 features from the datasets for emotion recognition and then reduced their dimensionality using Boruta, a wrapper-based feature selection method to get all relevant features. As per fig. 1 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% reduction in features whereas the recognition rate improved by 7%.

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 [13] [9].

Table 5 shows the training and test accuracies with MFCCs + Spectral centroids features on EmoDB, RAVDESS and IITKGP-SEHSC databases.

Kotti and Paterno [14] extracted a total of 2327 features for the same problem and reported an average accuracy of 83.3 %, with individual emotions and recognition rates being happiness (89.7%), neutrality (90.5%), anxiety (87.7%), anger (90.1%), sadness (88.6%) and boredom (89.3%). In contrast, in

Table 5 Experimental Observations

Dataset	Training Accuracy	Holdout Set Accuracy
EmoDB	96.00 %	92.45%
RAVDESS	79.85%	75.69%
IITKGP - SEHSC	87.73%	87.42%

this paper, by using only 147 features we report recognition rates for happiness (78%), anger (92%), sadness (100%), neutrality (88%) and boredom (100%). In other words, our proposed method is able to get almost same amount accuracy on individual emotion states and improved average recognition rate about 9.15% with almost 90% less number of features.

[13] proposed a new FP (Fourier parameters) model to extract emotion related features from speech signals, and validated it on EmoDB and other regional databases. Overall accuracy of their emotion recognition system by combining FP and MFCC features on EMODB was around 89%, whereas for our proposed method, the average accuracy was 92.42 % on EMODB database.

In [9], a new feature set called epoch features was proposed for emotion recognition and evaluated along with MFCCs by using DNN-HMM classifier on IITKGP Hindi emotion database. The observed accuracy of the proposed model using MFCC and epoch features together is around 72%. We also evaluated our model on IITKGP Hindi emotion database and by using a combination of MFCC and spectral centroid features, we have achieved an overall accuracy of 87.42%. In other words, the proposed combination of features and method is able to improve the recognition rate about 15 % than the best result of [9].

In [15], Mel-Frequency Cepstral Coefficients (MFCCs) features have been extracted from IITKGP-SEHSC for defining the emotions and three type of Gaussian Mixture Models (GMMs) used: 8-centered, 16-centered and 32-centered. Overall accuracies of 73.68, 70.43, and 65.96 % were obtained respectively. In contrast, in this paper, a bagged ensemble comprising of Support Vector Machines is used to evaluate our system and has achieved overall accuracy of 78.443% and 87.42% by using only MFCCs and a combination of MFCCS & spectral centroids respectively. This is a better performance in both cases, with a 14.74% improvement in the second case over [15].

In table 5, we can see that the rates of emotion recognition rate varies between German and Hindi. It is almost understandable because different countries have different cultures, and the way by which they express their emotion is also different [16].

In order to understand the effect of 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 Suppor Vector Machines for classification, this accuracy is further enhanced by roughly 5%.

In this paper, 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 baselines in terms of overall accuracy.

References

1. J. S. Park, J. H. Kim and Y. H. Oh, Feature vector classification based speech emotion recognition for service robots, IEEE Trans. On Consumer Electronics, 55(3), 1590-1596, 2009.
2. E. H. Kim, K. H. Hyun, S. H. Kim and Y. K. Kwak , Improved emotion recognition with a novel speaker-independent feature, IEEE/ASME Trans. on Mechatronics, 14(3), 317-325, 2009.
3. Pan, Yixiong, Peipei Shen, and Liping Shen. "Speech emotion recognition using support vector machine." International Journal of Smart Home 6.2 (2012): 101-108.
4. C. Gobl and A.N Chasaide, The role of voice quality in communicating emotion, mood and attitude, Speech Communication, pp.189212, 2003.
5. B.Yang and M. Lugger, Emotion recognition from speech signals using new harmony features, Signal Processing, pp. 14151423, 2010.
6. S.E. Bou-Ghazale, J. Hansen, A comparative study of traditional and newly proposed features for recognition of speech under stress, IEEE Transactions on Speech and Audio Processing, vol. 8, no. 4, pp. 429442, 2000
7. Li, X., Tao, J., Johnson, M.T., Soltis, J., Savage, A., Leong, K.M., Newman, J.D.: Stress and emotion classification using jitter and shimmer features. In: IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP 2007, vol. 4, IEEE IV-1081 (2007)
8. Chen, L., Mao, X., Xue, Y., Cheng, L.L.: Speech emotion recognition: features and classification models. Digit. Signal Process. 22(6), 11541160 (2012)
9. Fahad, Md, et al. "DNN-HMM based Speaker Adaptive Emotion Recognition using Proposed Epoch and MFCC Features." arXiv preprint arXiv:1806.00984 (2018).
10. Liu, Zhen-Tao, et al. "Speech emotion recognition based on feature selection and extreme learning machine decision tree." Neurocomputing 273 (2018): 271-280.
11. Schuller, B., Rigoll, G., Lang, M., 2004. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP04, pp. I-577I-580.
12. Liu, Gabrielle K. "Evaluating Gammatone Frequency Cepstral Coefficients with Neural Networks for Emotion Recognition from Speech." arXiv preprint arXiv:1806.09010 (2018).
13. Wang, Kunxia, et al. "Speech emotion recognition using Fourier parameters." IEEE Transactions on Affective Computing 6.1 (2015): 69-75.
14. M. Kotti and F. Patern, Speaker-independent emotion recognition exploiting a psychologically-inspired binary cascade classification schema, International Journal of Speech Technology, pp.131-150, 2012.
15. Nayak, Biswajit, and Manoj Kumar Pradhan. "Text-Dependent Versus Text-Independent Speech Emotion Recognition." Proceedings of the Second International Conference on Computer and Communication Technologies. Springer, New Delhi, 2016.
16. K. Norhaslinda, W. Abdul, and Q. Chai, Cultural dependency analysis for understanding speech emotion, Expert Systems with Applications, pp. 51155133, 2012.

17. Burkhardt, Felix, Astrid Paeschke, Miriam Rolfes, Walter F. Sendlmeier, and Benjamin Weiss. "A database of German emotional speech." In Ninth European Conference on Speech Communication and Technology. 2005.
18. Livingstone SR, Russo FA "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. (2018).
19. Koolagudi, Shashidhar G., Ramu Reddy, Jainath Yadav, and K. Sreenivasa Rao. "IITKGP-SEHSC: Hindi speech corpus for emotion analysis." In Devices and Communications (ICDeCom), 2011 International Conference on, pp. 1-5. IEEE, 2011.
20. Davis, S. Mermelstein, P. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. In IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 28 No. 4, pp. 357-366 (1980)
21. Lematre, Guillaume, Fernando Nogueira, and Christos K. Aridas. "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning." The Journal of Machine Learning Research 18.1 : 559-563 (2017)
22. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), 2825-2830. (2011).