A Real-time Emotion Recognition from Speech using Gradient Boosting

Aseef Iqbal
Department of Computer Science and Engineering
Chittagong Independent University
Chittagong, Bangladesh
aseef@ciu.edu.bd

Kakon Barua
Department of Computer Science and Engineering
Chittagong Independent University
Chittagong, Bangladesh
kakonbarua1@gmail.com

***Abstract*—Emotion recognition from speech is one of the research fields for emotional human-computer interaction. In this contribution, a real-time emotion recognition system is presented which recognizes emotions from live recorded speech by analyzing tonal properties. 34 audio features are extracted including MFCCs, energy, spectral entropy etc. Basically this system classifies emotions using models trained by Gradient Boosting. Other two classifiers such as Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) are also applied to observe the accuracies of them on test audio files. Two databases are employed for training the system like RAVDESS and SAVEE database. This system examines four basic emotions - anger, happiness, sadness and neutral.**

Keywords—RAVDESS, SAVEE, SVM, KNN, Gradient Boosting.

I. INTRODUCTION

Emotions can be recognized in various ways such as by analyzing tonal properties, facial expressions, body gestures etc. To enrich interaction, one needs to know and understand the correct emotion of other person and how to react on it. Emotional expressions play an important role not only in human interaction but also in computing world. Automatic emotion recognition has now become a vast research field involving more and more scientists majored in different areas like psychology, computer vision, physiology, artificial intelligence etc. Emotion recognition from speech is one of the methods of emotion recognition. For recognizing emotions, appropriate feature extraction is first and foremost work. Previous works related to emotion recognition from speech extracted different audio features to classify emotions. MFCCs, energy, pitch, spectral entropy are the most common audio features for emotion recognition. Feature selection and normalization are also equally important. Feature normalization helps to boost the accuracy of an emotion recognition system in a significant manner.

In this work, recognizing emotions from live speech data was the primary goal. The aim of the contribution was to develop a system capable of automatically recognize emotions from real-time speech of a person speaking in English. The objective was achieved by training a model using machine learning techniques utilizing available verbal emotion database. Finally the performance of the system is benchmarked in terms of its accuracy in detecting correct emotion from live test data.

In the next section, summary of the some significant relevant literature is presented. Section III gives an insight on the databases selected for the research. Section IV describes the methodology followed by a summary of results achieved in section V. Section VI gives an analysis of the results. Finally, section VII concludes the paper with brief discussion on the contribution of the work as well as future scopes of improving upon the results achieved.

II. RELATED WORK

Emotion recognition from speech has been a popular field of study. Several works have explored ways to improve this field. For feature extraction, Ghai et al. [1] chose to take frame sample of the sound signals at 16000 Hz and the selection duration of each frame was 0.25 seconds. In [3] the sample rate was 22050 Hz which were encoded through 16-bit PCM in two-channel. Different works selected different frame size like 10-20 ms [1], 0.2 s [5], 10 ms [13] etc. The most common features for audio speech data are MFCC, entropy and spectral entropy, ZCR (zero crossing rate), pitch, energy, formants etc. Most of the papers gave emphasis on pitch and energy [1,10-12,6,13,2]. Most of the papers calculated derivation and statistical features such as mean, standard deviation, etc. to increase accuracy. Previous works proposed different methods to classify emotions from speech such as Support Vector Machine (SVM), Gradient Boosting, K-Nearest Neighbor (KNN), Random Forest and Neural Network. They used different emotional speech databases to build their systems. Summary of some previous works are presented in Table I.

III. DATABASES

A. RAVDESS Database

Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is an audio, facial and multimodal database [7]. It contains 7356 files including calm, neutral, happiness, sad, anger, fear, disgust and surprise emotions. There are 24 subjects where 12 female and 12 male. It has speech and song files under each subjects except actor_18. All the recordings are in American English. In this work, high intensity audio files of four emotions like anger, happiness, neutral and sadness are used. There are total 192 audio files which are examined by considering female, male and combining both. Those files are divided as training and testing files.

978-1-5386-9111-3/19/$31.00 ©2019 IEEE

TABLE I. SUMMARY OF SOME PREVIOUS WORKS

Ref. No.	Classification	Database	No. of emo-tions	Accurac-y
Ghai et al. [1]	SVM, Gradient Boosting, Random Decision Forest	EMO-DB	7	55.89% 65.23% 81.05%
Schuller et al. [2]	GMM, HMM	Speech Sample	7	86.8% 77.8%
Rong et al. [3]	MDS, ISOMap, ERFT	Chinese (Mandarin)	5	61.18% 60.40% 69.21%
Chen et al. [4]	Fisher+SVM, PCA+SVM, Fisher+ANN, PCA+ANN	BHUDES	6	50.17% 43.15% 40.43% 39.17%
Anagnostopoulos et al. [5]	ANN, SVM	EMO-DB	7	49.19% 76.75%
Kwon et al. [6]	HMM, GSVM	SUSAS, AIBO	5	70.1% 42.3%

B. SAVEE Database

Surrey Audio-Visual Expressed Emotion (SAVEE) [8] is also a multimodal database. This database contains recordings of 4 male actors in seven emotions like anger, happiness, sadness, neutral, fear, disgust and surprise. There are total 480 utterances which were recorded in British English. For this project, 180 audio files of 3 male actors are chosen. Those files are categorized into four emotional expression class - anger, happiness, sadness and neutral. They are divided for both training and testing.

IV. METHODOLOGY

This system works as real-time emotion recognition environment. The complete system architecture is presented in Fig 1. The system is divided into two sections training module and output module. In training module, audio files of databases are collected for training the classifiers. RAVDESS database is divided into three sets such as RAVDESS male, RAVDESS female and RAVDESS combined (male and female both). The audio files of each datasets are divided into training files and testing files. Training files are used to train the system and the testing files are applied to see the accuracies of the classifiers on each dataset. In output module the system is tested by the live recorded speech data. The processes of training and output module are illustrated in Fig 2. Basically, this work has three stages such as feature extraction, feature normalization and classification.

C. Feature Extraction

In this work, 34 audio features are extracted from each audio file using pyAudioAnalysis[9]. Zero crossing rate (ZCR), energy, entropy of energy, spectral centroid, spectral spread, spectral entropy, spectral flux, spectral rolloff, 13 MFCCs, 12 chroma vectors and chroma deviation are the extracted features. That library excerpts both short-term and mid-term features. Short-term features are calculated by 0.05s frame size and 0.025s step size. For mid-term the duration is 1.0s for both cases. For this system, only mid-term features are considered. In mid-term features, it calculates mean and standard deviation over each short-term features.

D. Feature Normalization

This part of implementation is really important as accuracy result depends on it mostly. Without normalization extracted features may not be in a specific range. And these unstructured data may hamper severely in recognition rate. Normalization makes features equally weighted. In this work, normalization is calculated by subtracting mean from each feature and divided by standard deviation. After collecting features, those are analyzed and normalized by that method for bringing them into a shape. Those newly oriented data are ready to train the system after normalization.

E. Classifiers

1) Gradient Boosting: Gradient boosting produces a prediction model in the form of an ensemble of weak prediction models. It involves three elements like loss function, weak learner and additive model. Gradient boosting framework is a general framework in which any differentiable loss function can be utilized. In gradient boosting the weak learner are the decision trees. To minimize the cost, trees are assembled by greedy manner by splitting in best purity scores points. It is an over fitting algorithm. Weak learners can be constrained in some specific ways like nodes, splits, leaf nodes or a maximum number of layers. Additive model means trees are added for once at a time. The aim of the Gradient boosting procedure is to minimize the loss while adding trees. In this procedure, loss is calculated in each step and after calculating the loss, a tree which reduces the loss must be added and weights are updated. For gradient boosting classifier, scikit-learn ensemble is applied in this emotion recognition system. One parameter is given named n_estimators which is the number of stages to perform training section. Gradient Boosting performs training by over-fitting, so large number of n_estimators usually gives better performance. To find the best estimator, some values are tried and tested on training data. After finding best value, the classifier is fitted with the training data applying the evaluated n_estimator.

2) SVM: SVM is a supervised machine learning algorithm. It is used for both regression and classification problems. SVM is a classifier which is defined by a separating hyper-plane. The main theme of this algorithm is to choose hyper-planes which segregate multiple classes better. Most classification tasks are not that much simple which are included with scattered data. An accurate hyper-plane will be hard to find which divides the classes more specifically. To solve these types of problems SVM has different kernel methods such as polynomial and radial. Polynomial kernel allows curved lines in the input space. It is applied to make the data best fitted. Radial kernel can generate complex regions within input space like closed polygon in 2D space. Without kernel there are other tuning parameters like regularization and gamma. Regularization parameter C has very significant effect on SVM training. There are several techniques to catch perfect C and gamma

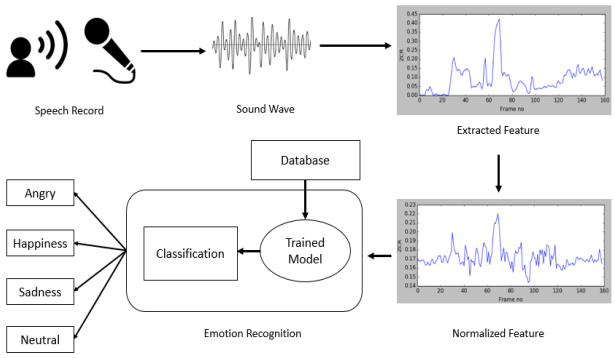

Fig. 1. Overall system architecture of real-time emotion recognition from speech

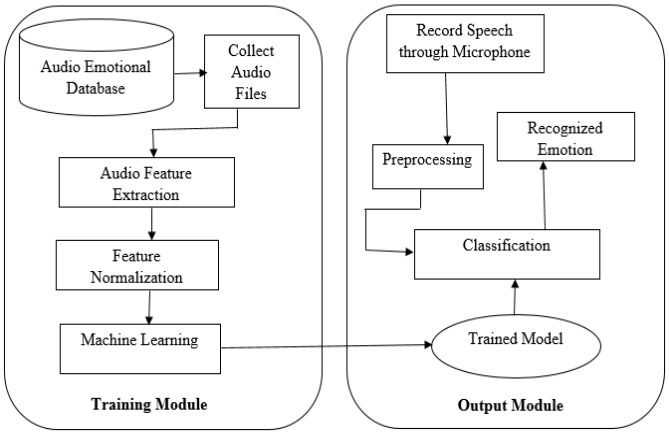

Fig. 2. Block diagram of real-time emotion recognition system

value for the training data. For this work, Grid_search method of scikit-learn is applied for tuning these parameters. It applies multiple values for C and gamma on the training features. After executing that function, grid_search returns the best C and gamma value which give best accuracy results over other given values on training data. Finally the best C and gamma value are passed to the main SVM program for producing trained model. For completing the training procedure python scikit-learn SVM library with linear kernel is employed.

3) KNN: Another machine learning algorithm which is numerously touched for building classifier models is K-nearest neighbors. It is also implemented for regression predictive problems. It generally stores the entire dataset (features), no training is required unlike SVM. For a new instance, predictions are made by examining throughout the entire dataset to find the K-most similar neighbors (instances) and summing up the output variable for those k-neighbors. To determine the most similar K instances in the training data to the new input, a distance measure needs to be calculated. For real value, Euclidian distance (1) is the most prevalent distance measure which is applied in this system. For finding the best number of neighbors, an evaluation program is run which tested some given K values on the training data by training and testing. After finding best value, all the test necessary information are dumped into a file which are used during testing.

$$\text{Euclidian distance (q, p)} = \sqrt{\sum_{j=1}^{n}(q_j - p_j)^2} \quad (1)$$

V. RESULT

The three classifiers Gradient Boosting, SVM and KNN are applied for classifying the emotions. At first, testing files of databases have been tested by the all classifiers. In this case both databases have been participated. The summary of the results are given in the following tables and graphs. Table II, III and IV explain the confusion matrix of all classifiers on RAVDESS female dataset. In addition, Table V, VI and VII show the confusion matrix of the RAVDESS male dataset. For RAVDESS combined dataset the confusion matrix are presented in Table VIII, IX and X. And the last database called SAVEE test results are shown in Table XI, XII and XIII.

Then a real-time experiment has been performed where five subjects have been chosen for recording emotional expressions. The recorded speech has been classified by the classifiers. For live test experiment only RAVDESS gender based trained models are applied. For each emotional expression the recording time is around 7s. The male expressions are classified by the trained models which are trained by RAVDESS male dataset. Similarly, RAVDESS female dataset is used for female expressions. Table XIV, XV and XVI represent the accuracy results of all classifiers on each emotions of real-time female data. Similarly, Table XVII, XVIII and XIX show confusion matrix for real-time male data.

Fig 3 illustrates the average accuracy results of all classifiers on all datasets. And the average recognition rates of all classifiers on real-time data are represented in Fig 4.

TABLE II. CONFUSION MATRIX OF RAVDESS FEMALE SVM

	Anger	**Happiness**	**Sadness**	**Neutral**
Anger	1.0	0.0	0.0	0.0
Happiness	0.17	0.83	0.0	0.0
Sadness	0.0	0.33	0.5	0.17
Neutral	0.0	0.0	0.17	0.83

TABLE III. CONFUSION MATRIX OF RAVDESS FEMALE KNN

	Anger	**Happiness**	**Sadness**	**Neutral**
Anger	0.87	0.17	0.0	0.0
Happiness	0.5	0.33	0.17	0.0
Sadness	0.0	0.5	0.33	0.17
Neutral	0.0	0.0	0.0	1.0

TABLE IV. CONFUSION MATRIX OF RAVDESS FEMALE GRADIENT BOOSTING

	Anger	**Happiness**	**Sadness**	**Neutral**
Anger	0.33	0.5	0.17	0.0
Happiness	0.17	0.66	0.17	0.0
Sadness	0.0	0.33	0.67	0.0
Neutral	0.0	0.0	0.5	0.5

TABLE V. CONFUSION MATRIX OF RAVDESS MALE SVM

	Anger	**Happiness**	**Sadness**	**Neutral**
Anger	1.0	0.0	0.0	0.0
Happiness	0.0	0.66	0.17	0.17
Sadness	0.0	0.0	0.5	0.5
Neutral	0.0	0.0	0.0	1.0

TABLE VI. CONFUSION MATRIX OF RAVDESS MALE KNN

	Anger	Happiness	Sadness	Neutral
Anger	1.0	0.0	0.0	0.0
Happiness	0.17	0.66	0.0	0.17
Sadness	0.17	0.0	0.33	0.5
Neutral	0.0	0.0	0.0	1.0

TABLE VII. CONFUSION MATRIX OF RAVDESS MALE GRADIENT BOOSTING

	Anger	Happiness	Sadness	Neutral
Anger	0.87	0.0	0.17	0.0
Happiness	0.0	0.87	0.17	0.0
Sadness	0.0	0.0	0.67	0.33
Neutral	0.0	0.17	0.17	0.66

TABLE VIII. CONFUSION MATRIX OF RAVDESS COMBINED SVM

	Anger	Happiness	Sadness	Neutral
Anger	1.0	0.0	0.0	0.0
Happiness	0.07	0.66	0.2	0.07
Sadness	0.0	0.0	0.67	0.33
Neutral	0.0	0.07	0.0	0.93

TABLE IX. CONFUSION MATRIX OF RAVDESS COMBINED KNN

	Anger	Happiness	Sadness	Neutral
Anger	0.93	0.07	0.0	0.0
Happiness	0.33	0.47	0.07	0.13
Sadness	0.07	0.27	0.33	0.33
Neutral	0.0	0.07	0.0	0.93

TABLE X. CONFUSION MATRIX OF RAVDESS COMBINED GRADIENT BOOSTING

	Anger	Happiness	Sadness	Neutral
Anger	0.53	0.27	0.2	0.0
Happiness	0.13	0.6	0.2	0.07
Sadness	0.0	0.13	0.6	0.27
Neutral	0.0	0.13	0.07	0.8

TABLE XI. CONFUSION MATRIX OF SAVEE SVM

	Anger	Happiness	Sadness	Neutral
Anger	0.56	0.11	0.0	0.33
Happiness	0.11	0.78	0.11	0.0
Sadness	0.0	0.0	1.0	0.0
Neutral	0.0	0.0	0.11	0.89

TABLE XII. CONFUSION MATRIX OF SAVEE KNN

	Anger	Happiness	Sadness	Neutral
Anger	0.56	0.11	0.11	0.22
Happiness	0.22	0.67	0.0	0.11
Sadness	0.0	0.0	1.0	0.0
Neutral	0.0	0.0	0.11	0.89

TABLE XIII. CONFUSION MATRIX OF SAVEE GRADIENT BOOSTING

	Anger	Happiness	Sadness	Neutral
Anger	0.56	0.33	0.0	0.11
Happiness	0.22	0.78	0.0	0.0
Sadness	0.0	0.0	1.0	0.0
Neutral	0.11	0.0	0.11	0.78

TABLE XIV. CONFUSION MATRIX OF SVM FOR FEMALE REAL-TIME TEST DATA

	Anger	Happiness	Sadness	Neutral
Anger	0.0	0.67	0.0	0.33
Happiness	0.0	1	0.0	0.0
Sadness	0.0	0.5	0.33	0.17
Neutral	0.0	0.0	0.0	1

TABLE XV. CONFUSION MATRIX OF KNN FOR FEMALE REAL-TIME TEST DATA

	Anger	Happiness	Sadness	Neutral
Anger	0.0	0.17	0.83	0.0
Happiness	0.0	0.5	0.5	0.0
Sadness	0.0	0.0	1	0.0
Neutral	0.0	0.17	0.17	0.66

TABLE XVI. CONFUSION MATRIX OF GRADIENT BOOSTING FOR FEMALE REAL-TIME TEST DATA

	Anger	Happiness	Sadness	Neutral
Anger	0.67	0.33	0.0	0.0
Happiness	0.17	0.5	0.33	0.0
Sadness	0.0	0.5	0.33	0.17
Neutral	0.0	0.0	0.0	1

TABLE XVII. CONFUSION MATRIX OF SVM FOR MALE REAL-TIME TEST DATA

	Anger	Happiness	Sadness	Neutral
Anger	0.0	0.6	0.0	0.4
Happiness	0.0	1	0.0	0.0
Sadness	0.0	0.0	1	0.0
Neutral	0.0	0.0	0.4	0.6

TABLE XVIII. CONFUSION MATRIX OF KNN FOR MALE REAL-TIME TEST DATA

	Anger	Happiness	Sadness	Neutral
Anger	0.8	0.2	0.0	0.0
Happiness	0.0	0.4	0.6	0.0
Sadness	0.2	0.4	0.4	0.0
Neutral	0.0	0.0	0.4	0.6

TABLE XIX. CONFUSION MATRIX OF GRADIENT BOOSTING FOR MALE REAL-TIME TEST DATA

	Anger	Happiness	Sadness	Neutral
Anger	0.4	0.6	0.0	0.0
Happiness	0.2	0.8	0.0	0.0
Sadness	0.0	0.0	1	0.0
Neutral	0.0	0.6	0.0	0.4

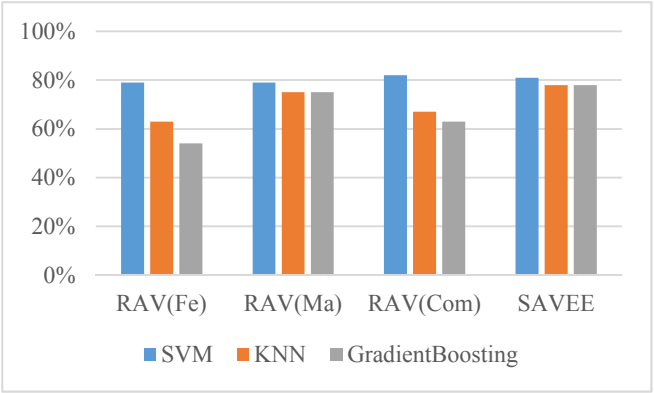

Fig. 3. Average accuracies of the classifiers on databases

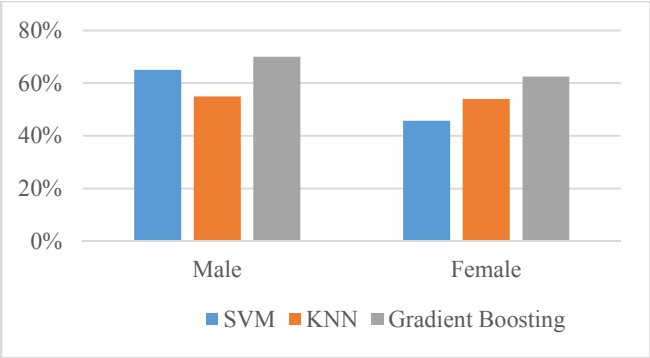

Fig. 4. Average accuracies of the classifiers of real-time test

RAV (RAVDESS), Fe (Female), Ma (Male), Com (Combined) [Fig 3]

VI. Analysis

In the previous section, all types of accuracy results in both databases and live recorded data are shown. The proposed classifiers performed differently in different datasets. There are four types of datasets like RAVDESS (male), RAVDESS (female), RAVDESS (combined) and SAVEE. In RAVDESS (male) SVM and KNN have 100% accuracy in both anger and neutral [Table V and VII]. But in happiness and sadness Gradient Boosting performs better than SVM and KNN. In RAVDESS (female) SVM achieves 100% accuracy in anger as same as male part [Table II]. SVM has overall good performance except in sadness. Performance of KNN is also good in anger and neutral like 87% and 100% respectively [Table III]. In anger and neutral, Gradient Boosting performs poorly. KNN performance is very poor in happiness and sadness comparing with other classifiers. In male and female combined dataset, performances of SVM and KNN are really good in anger and neutral rather than Gradient Boosting. KNN's performance is really poor in happiness and sadness. Average performances of classifiers in male dataset are better than female dataset except SVM. In combined database, SVM get high accuracy than gender based datasets. In SAVEE database, sadness emotion has 100% accuracy in all classifiers. But in anger, all classifier's performances are too poor - around 56%. Performance of SVM is dominating over other classifiers in all datasets [Fig-3]. And SAVEE gets better accuracy in all classifiers than all other datasets because SAVEE is examined with only three subjects. In real-time test, male testing performance is better than female in all classifiers. Gradient Boosting classification accuracies are more than other classifiers in both male and female. This classifier shows overall good performance in all emotions rather than SVM and KNN. Now, this emotion recognition system predicts emotions from live recorded data through Gradient Boosting trained model.

VII. Conclusion

In summary, it can be said that different classifiers performed differently in different databases. That was also found in previous works. High intensity audio files are selected from RAVDESS database. But SAVEE audio files have less intense than RAVDESS selected files. SAVEE training and testing files are under 3 subjects but RAVDESS had 24 subjects. That's why SAVEE has higher accuracy than RAVDESS. SVM is the dominating classifier in databases. It performs better than KNN and Gradient Boosting. In live data, accuracies of SVM are poor. In that case, performance of gradient boosting is better than SVM and KNN. There are some limitations in real-time test. Microphone adds noise during recording of real-time test which affects to gain better accuracy. Another is trained models. The applied trained models are trained with RAVDESS database. Audio files of RAVDESS were made in American English language. The tone of talking is not as same as the Asian tone specially Bangladesh. So it may cause problems to match the features. To improve the accuracy, feature selection should be considered more specifically. Features have great impact on increasing or decreasing accuracy. Today deep learning is the most pronounced machine learning algorithm. For implementing better accurate system of emotion recognition from speech, next step will be to apply deep learning method. This system can be improved by noise reduction and using large and appropriate database for producing trained model. Further study will be considering more emotions like fear, disgust, surprise, boredom etc. This emotion recognition system would be extended to multimodal by applying video data with audio to recognize emotions.

References

[1] M. Ghai, S. Lal, S. Duggal, and S. Manik, "Emotion recognition on speech signals using machine learning," *2017 Int. Conf. Big Data Anal. Comput. Intell.*, pp. 34–39, 2017.

[2] B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov model-based speech emotion recognition," *2003 IEEE Int. Conf. Acoust. Speech, Signal Process. 2003. Proceedings. (ICASSP '03).*, vol. 2, p. II--1--4, 2003.

[3] J. Rong, G. Li, and Y. P. P. Chen, "Acoustic feature selection for automatic emotion recognition from speech," *Inf. Process. Manag.*, vol. 45, no. 3, pp. 315–328, 2009.

[4] L. Chen, X. Mao, Y. Xue, and L. L. Cheng, "Speech emotion recognition: Features and classification models," *Digit. Signal Process. A Rev. J.*, vol. 22, no. 6, pp. 1154–1160, 2012.

[5] C. N. Anagnostopoulos and T. Iliou, "Towards emotion recognition from speech: Definition, problems and the materials of research," *Stud. Comput. Intell.*, vol. 279, pp. 127–143, 2010.

[6] O. Kwon, K. Chan, J. Hao, and T. Lee, "Emotion Recognition by Speech Signals," in *In Proceedings of International Conference EUROSPEECH*, 2003, pp. 125–128.

[7] S. R. Livingstone and F. A. Russo, The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic , multimodal set of facial and vocal expressions in North American English. 2018.

[8] 'Surrey Audio-Visual Expressed Emotion (SAVEE) database' [online]. Available: http://kahlan.eps.surrey.ac.uk/savee/. [Accessed: 5-March-2018]

[9] T. Giannakopoulos, "pyAudioAnalysis : An Open-Source Python Library for Audio Signal Analysis," pp. 1–17, 2015.

[10] N. Sebe, I. Cohen, T. Gevers, and T. S. Huang, "Emotion recognition based on joint visual and audio cues," in *Proceedings - International Conference on Pattern Recognition*, 2006, vol. 1, pp. 1136–1139.

[11] P. Anandan, "A computational framework and an algorithm for the measurement of visual motion," *Int. J. Comput. Vis.*, vol. 2, no. 3, pp. 283–310, 1989.

[12] B. E. Mart and J. C. Jacobo, "An improved characterization methodology to efficiently deal with the speech emotion recognition problem," no. Ropec, 2017.

[13] V. Chernykh, G. Sterling, and P. Prihodko, "Emotion Recognition From Speech With Recurrent Neural Networks," 2017.