International Journal of Engineering & Technology, 7 (4) (2018) 5517-5520

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i4.28292
Research paper

A review of classification methods and databases used for speech emotion recognition

Shrikala Madhav Deshmukh [1] *, Sita Devulapalli [2]

[1] Research Scholar, Amity School of Engineering & Technology, Amity University Mumbai, Maharashtra, India
[2] Professor, CSE, Amity School of Engineering & Technology, Amity University Mumbai, Maharashtra, India
**Corresponding author E-mail: shrikaladeshmukh89@gmail.com*

Abstract

In today's world speech is the ideal way to interact with people. Speech emotion recognition (SER) has an increasingly significant role in the interactions among human beings and computers. For improving human machine interaction, it is very ideal to recognize emotions automatically because attention is aimed at study of the emotions. This paper is a review of classification methods and databases used for speech emotion recognition. Here two important fields in speech emotion recognition are addressed. First is the choice of appropriate classification method and second is the creation of emotional speech database or choosing appropriate database. The main purpose behind this review paper is to analyze the efficiency of several techniques widely used among the field of speech emotion recognition.

Keywords: *Artificial Neural Networks (ANN); Convolutional Neural Networks (CNNs); Classification Methods; Database; Gaussian Mixture Model (GMM); Hidden Markov Model (HMM); Neural Network Classifier; Recurrent Neural Network (RNN); Speech Emotion Recognition (SER); Support Vector Machine (SVM).*

1. Introduction

Speech is very faster and easier way of communication between human beings. This fact has motivated many researchers to work in this field. Speech is a fast medium of interaction between humans and machines. Speech Emotion Recognition (SER) is introduced recently. SER is extraction of speaker's emotions from his or her speech.
The speech emotion recognition is a highly challenging task for some reasons such as variety of sentences, different speakers, various speaking styles and speaking rates, etc. Many researchers focused on monolingual emotion classification rather than multilingual emotion classification, because there should not be any cultural difference amongst speakers.
There are mainly two types of speech emotion recognition: Speaker Dependant and Speaker Independent. In speaker dependant speech emotion recognition system, energy and pitch are used as features and in speaker independent speech emotion recognition system focus is on "What was said" regardless of "Who said it".

1.1. Classification approaches

For correct identification of emotion, it is very necessary to choose appropriate classification approach. As per the literature survey, various classification approaches are used such as Gaussian Mixture Model (GMM) [2] [7] [18], Hidden Markov Model (HMM) [27], Support Vector Machine (SVM) [8] [9] [11] [12] [16] [22] [24], Neural Network classifier [19] [21], Artificial Neural Networks (ANN) [25], Convolutional Neural Networks (CNNs) [1], Recurrent Neural Network (RNN) [4], etc.

1.2. Speech database

The result of study basically depends on database used to assess its performance. As per our literature survey, some authors have used available database and some create database by their own. The various databases used are RECOLA [1] [15], Berlin emotional speech database [2] [4] [8] [11] [12] [21] [24], IEMOCAP database [14] [23] [26], eNTERFACE database [13], German database (EMODB) [19], etc. The most widely used databases are Berlin emotional speech database [2] [4] [8] [11] [12] [21] [24] and IEMOCAP database [14] [23] [26].

2. Literature review

The literature review is based on studies carried out and presented at various levels on emotion recognition. This literature survey is divided according to techniques used for emotion recognition.
In this paper [2] for automatic emotion recognition from speech signals authors proposed use of harmony features [2]. It is based on music theory. The system will measure two pitch intervals. In this paper, emotion recognition experiments use these harmony parameters with state of the art features to improve performance. GMM [2] [7] [18] is used for classification. The average recognition rate is improved by 2% [2].
In this paper [7] authors recognizes gender-dependant age and emotions. This system works in two-stages. At first system will identify the gender and then focus on recognition of emotions. This system performs noise elimination from voice signal. Authors used Mel-Frequency Cepstral Coefficients (MFCCs) as a feature extraction method. For the classification of large data, SVM [8] [9] [11] [12] [16] [22] [24], and GMM [2] [7] [18] are used [7].
[18] In this paper, authors analyzed performance of speaker dependent, speaker independent and cross language emotion recognition from speech. For classification Gaussian Mixture Model

(GMM) [2] [7] [18], and Hidden Markov Model (HMM) [27] are used. IITKGP-SESC and IITKGP-SEHSC databases are used. For identifying emotions Mel Frequency Cepstral Coefficients (MFCCs) features are used. Authors found that performance of emotion recognition from speaker dependant system is good than others [18].

[27] In this paper, authors developed a system to recognize the emotions based on gender of the speaker. Hidden Markov Model used for gender identification. Jahmm library is adopted for implementing HMM [27].

[8] In this paper, authors proposed a method which uses fractal dimension features to recognize the emotion from speech signals. For classification and recognition authors used Support Vector Machine algorithm. Authors used Berlin Emotional Speech Database. Recognition rate is approximately 77% [8].

[9] In this paper, authors used 2 subsystems: a) Gender Recognition. Authors used various gender recognition algorithms such as Thresholding method using Pitch, gender recognition by binary SVM [8] [9] [11] [12] [16] [22] [24] classifier and gender recognition by ANN [25] and b) Emotion Recognition. In this various emotions such as anger, boredom, disgust, fear, happiness, sadness and neutral state are used. The aim of the system is to provide "a priori" information about the gender. This system is implemented by 3 methods, a Pitch Frequency Estimation method and thresholding to perform gender recognition, gender recognition using SVM [8] [9] [11] [12] [16] [22] [24], and ANN [25], the latter by two Support Vector Machine (SVM) classifiers. Overall emotion recognition accuracy is increased from 74.28 % to 88.57 % by priori knowledge of the speaker gender [9].

[11] In this paper, a system is proposed which allows recognition of person's emotional state from audio signals registrations. This system is used to improve human-computer interaction. This system recognizes six emotions such as anger, boredom, fear, disgust, happiness and sadness and the neutral state. There are 2 subsystems: Emotion recognition and gender recognition. Priori knowledge of speaker gender will increase the performance. Pitch Frequency Estimation method and two Support Vector Machine (SVM) [8] [9] [11] [12] [16] [22] [24] classifiers are used [11].

[12] In this paper, authors describes the emotion recognition using beagle board OMAP 3530 in linux platform. Pitch, Energy, Mel-Frequency Cepstral Coefficients (MFCC) are extracted from speech. In this paper, Support Vector Machine (SVM) [8] [9] [11] [12] [16] [22] [24] is used as a classifier.

[16] In this paper, Mel Frequency Cepstrum Coefficient(MFCC) is used for extracting samples and used to train Support Vector Machine (SVM) [8] [9] [11] [12] [16] [22] [24]. English and Telugu languages are used. It is implemented in MATLAB 12b environment. This system shows results such as, 85.77% for speaker dependent, 55.51% for speaker independent and 49.83% for cross language emotion recognition [16].

[22] In this paper, authors have used Chinese speech. From speech some features are extracted like Mel Frequency Cepstrum Coefficient (MFCC), pitch, formant, short-term zero-crossing rate and short-term energy. Authors proposed a new method which combines DBN (Deep Belief Network) and SVM (Support vector machine) [8] [9] [11] [12] [16] [22] [24]. New classification approach achieves 95.8% accuracy, which is higher than using either DBN or SVM separately [22].

[24] In this paper, Mel Frequency Cepstral coefficients (MFCC) and energy of the speech signals are used as a feature inputs. Berlin database of emotional speech is used. The features extracted from speech are converted into a feature vector, then it is used to train different classification algorithms namely, Support Vector Machine (SVM) [8] [9] [11] [12] [16] [22] [24], Random Decision Forest and Gradient Boosting. Random forest was found to have the highest accuracy [24].

[17] In this paper, authors have used emotions from speech and facial gestures for speaker dependant emotion recognition. Authors have used English audio-visual emotional database. In this method, LDA and PCA [17] are used. For classification Gaussian classifiers are used. The system will give higher performance for LDA features than PCA features. [17].

[19] In this paper, authors used harmony features for speech emotion recognition. Use of Fourier parameter (FP) features for speaker independent system is proposed. By adding FP and MFCC, the recognition rates are increased by 17.5 on EMODB, 10 on CASIA and 10.5 on EESDB respectively [19]. SVM [8] [9] [11] [12] [16] [22] [24] classifier and Bayesian classifier are used [19].

[21] In this paper, for intelligent emotion recognition, features from speech wave are extracted, Open smile is the feature extract tool used to get features from speech recordings. For recognition of emotions, a supervised neural network is used. Berlin emotional database is used [21].

[25] In this paper, emotion recognition from female speech is done. Authors have created database of 340 samples and used four emotions such as neutral, happy, sad and anger. Malayalam language is used for this work. For feature extraction Daubechies8 wavelet was used. Artificial neural network [25] was used for pattern recognition. An overall recognition accuracy of 72.055% obtained from this experiment [25].

[15] In this paper, authors have decided to use Semi-supervised learning (SSL) algorithms for unlabelled data. Self-training of conventional SSL has an error of self-accumulation. To overcome this issue authors proposed an enhanced learning strategy. In this method re-evaluation of previously automatically labeled samples is done, in this training set is updated by correcting mislabeled samples. Authors have used multiple modalities and models of SSL system to increase the performance [15].

[23] In this paper, authors proposed to utilize Deep Neural Networks (DNNs) [23] to estimate emotion states for each speech segment in an utterance, construct an utterance level feature from segment-level estimations, and then employ an ELM to recognize the emotions for the utterance. The system will improve accuracy by 20% [23].

[26] In this paper, authors used para-lingual information from speech, this system is based on a Deep Neural Network [23], and it is applied to spectrograms. This method gives better results compared to other methods. Convolution-only deep network with lower complexity achieves a prediction accuracy of 66% on IEMOCAP, while a combination of convolution-LSTM higher-complexity model achieves 68% [26].

[1] In this paper, authors presented a model for continuous emotion recognition from speech. This paper is based on Deep Neural Networks (DNN). The model trained end-to-end and it uses Convolutional Neural Networks (CNN), this extracts features from the raw signals, and stacked this data on top of a 2-layer Long Short-Term Memory (LSTM). This model performs very well in terms of concordance correlation coefficient and the state-of-the-art methods for the RECOLA database [1].

[4] In this paper authors, have used speech emotion recognition system; important speech features are used in this paper. It contains emotion information such as pitch, energy, formant frequency, Mel Frequency ceptrum coefficients (MFCC), linear prediction cepstrum coefficients (LPCC) and Modulation Spectral Features are used. Recurrent neural network (RNN) classifier is used to classify emotions found in the Berlin and Spanish databases. Its performances are compared to Multivariate linear regression (MLR) and Support vector machine (SVM) [8] [9] [11] [12] [16] [22] [24] classifiers. Berlin emotional speech database and Spanish emotional database is used. In this paper, authors combined MFCC and MS features for RNN model in Spanish emotional database, the authors achieved the result of 90.05%. For Berlin emotional database using MLR classifier, authors achieved results of 82.41% [4].

[5] In this paper, authors done analysis of emotions using Formant frequencies (F1, F2, F3), instantaneous fundamental frequency using Zero Frequency Filtering, signal energy and dominant frequencies. Authors used German and Telugu Emotion Databases [5].

[3] In this paper, authors used neutral, anger, joy and sadness emotions. In this, classifications are performed for different classifiers.

Authors says that, for increasing accuracy, data should be collected from one person rather than group of people [3].

[10] In this paper, Hybrid feature extraction method is used for 100% accuracy for all emotions. The speaker dependent Emotion Recognition system gave 100% accuracy for all emotions such as happy, sad, surprise, anger and neutral. For speaker independent ER system, Mel Frequency Cepstrum Coefficient (MFCC) gives 100% accuracy for surprise emotion, Perceptual Linear Predictive (PLP) gives 100% accuracy for Anger emotion and LPCC features give 100% accuracy for fear emotion [10].

[13] In this paper, for improving accuracy authors combined evolutionary algorithm (EA) with Empirical Mode Decomposition (EMD). In this method, emotional speeches are decomposed into several Intrinsic Mode Functions (IMFs) by use of EMD processes. Emotion from speech is extracted using IMFs. MFCC are computed and used [13].

[14] In this paper, authors used iterative feature normalization (IFN) framework. This is speaker dependant approach. In this method, normalization is applied to all samples. Performance of emotion detection based on IFN framework gives better accuracies. The IFN improves the accuracy in detecting emotional speech obtained from real life and unconstrained recordings [14].

[20] In this paper, authors have used acoustic and spectral features and studied automatic recognition of human emotional states. Authors reviewed acoustic features and proposed features from time-frequency representation. This system shows Recognition accuracy of 94.6% for SES database of emotional speech of Spanish language [20].

Table 1: Summary

Paper No.	Classifiers	Database
1	Convolutional Neural Network (CNNs) [1]	RECOLA
2	Bayesian learning framework, Gaussian mixture model (GMM) [2] [7] [18]	Berlin Emotion Database
3	Mel-frequency cepstral coefficients (MFCC) and AUC	Manually created in Russian language
4	Recurrent neural network (RNN) [4], Multivariate linear regression (MLR) and Support vector machine (SVM) [8] [9] [11] [12] [16] [22] [24]	Spanish emotional database, Berlin emotional database
5	Signal energy, Zero-crossing rate (ZCR)	IITKGP-SESC Telugu speech database, German emotion database
6	MLB (Maximum-Likelihood Bayes), NN (Nearest Neighbor) and HMM (Hidden Markov Model) [27]	Manually created
7	GMM [2] [7] [18], SVM [8] [9] [11] [12] [16] [22] [24]	Database created from various audio clips
8	Support Vector Machine(SVM) [8] [9] [11] [12] [16] [22] [24]	Berlin Emotional Speech Database
9	Support Vector Machine (SVM) [8] [9] [11] [12] [16] [22] [24], Artificial Neural Networks (ANN)	Employed Reference Database (DB)
10	Hidden Markov Model (HMM) [27]	Manually created
11	Support Vector Machine(SVM) [8] [9] [11] [12] [16] [22] [24]	Reading-Leeds Database, Belfast Database, CREST-ESP, Berlin Emotional Speech (BES) Database
12	Support Vector Machine(SVM) [8] [9] [11] [12] [16] [22] [24]	Berlin Emotional Speech Database
13	Evolutionary algorithm (EA) with Empirical Mode Decomposition	eNTERFACE 2005 emotion database
14	Iterative Feature Normalization (IFN)	IEMOCAP database
15	Semi-supervised learning (SSL) approaches	RECOLA
16	Support Vector Machine (SVM) [8] [9] [11] [12] [16] [22] [24]	Manually created
17	PCA, LDA [17]	British English audio-visual emotional database
18	Gaussian Mixture Model (GMM) [2] [7] [18], Hidden Markov Model (HMM) [27]	IITKGP-SESC and IITKGP-SEHSC emotional speech database
19	Support Vector Machine(SVM) [8] [9] [11] [12] [16] [22] [24], Bayesian classifier	EMODB, EESDB database and CASIA database
20	Gabor Transform and Discrete Wavelet (DWT)	SES database of emotional speech in Spanish language
21	Neural network approach	Berlin Database of Emotional Speech
22	Deep Belief Network (DBN) and SVM (support vector machine) [8] [9] [11] [12] [16] [22] [24]	Chinese Academy of Sciences emotional speech database
23	Deep Neural Networks (DNNs)[1][23][26]	IEMOCAP database
24	Support Vector Machine [8] [9] [11] [12] [16] [22] [24], Random Decision Forest and Gradient Boosting.	Berlin Database of Emotional Speech
25	Artificial Neural Network (ANN) [25]	Elicited emotional database
26	Deep Neural Network [1] [23] [26]	IEMOCAP database
27	Hidden Markov Model [27]	Not mentioned

3. Conclusion

This paper presents survey of Emotion Recognition from Speech addressing use of various algorithms of Speech Emotion Recognition System. In this paper, we reviewed and discussed various speech emotion recognition systems. We had also seen its performance in terms of various classifiers and datasets. Authors had used variety of classifiers like Gaussian Mixture Model (GMM) [2] [7] [18], Hidden Markov Model (HMM) [27], Support Vector Machine (SVM) [8] [9] [11] [12] [16] [22] [24], Neural Network classifier, Artificial Neural Networks (ANN) [25], Convolutional Neural Networks (CNNs) [1], Recurrent Neural Network (RNN) [4], etc. Well-designed classifiers have obtained high accuracies for various emotions. Result varies for different databases. It is found that some emotions are recognized very correctly but there is problem for other emotions. So, there are so many contributions needed in this area.

References

[1] Panagiotis Tzirakis, Jiehao Zhang, Bjorn W. Schuller, End-To-End speech emotion recognition using deep neural networks, IEEE (ICASSP 2018), pages 5089-5093.

[2] B. Yang, M. Lugger, Emotion recognition from speech signals using new harmony features, Elsevier Signal Processing 90 (2010), pages1415–1423.

[3] Assel Davletcharovaa, Sherin Sugathanb, Bibia Abrahamc, Alex Pappachen Jamesa, Detection and analysis of emotion from speech signals, Elsevier Procedia Computer Science 58 (2015), pages 91-96. https://doi.org/10.1016/j.procs.2015.08.032.

[4] Leila Kerkeni, Youssef Serrestou, Mohamed Mbarki, KosaiRaoof and Mohamed Ali Mahjoub, Speech emotion recognition: methods and cases study, International Conference on Agents and Artificial Intelligence (ICAART 2018), ISBN: 978-989-758-275-2, Volume 2, pages 175-182.

[5] Esther Ramdinmawii, Abhijit Mohanta and Vinay Kumar Mittal, Emotion recognition from speech signal, Proc. of the 2017 IEEE

Region 10 Conference (TENCON), 2017, pages 1562-1567. https://doi.org/10.1109/TENCON.2017.8228105.
[6] Bong-Seok Kang, Chul-Hee Han, Sang-Tae Lee, Dae-HeeYoun and Chungyong Lee, Speaker dependent emotion recognition using speech signals, International Conference on Spokan language processing (ICSLP 2000).
[7] Shivaji J. Chaudhari, Ramesh M. Kagalkar, Automatic speaker age estimation and gender dependent emotion recognition, International Journal of Computer Applications, Volume 117 – No. 17, May 2015, pages 5-10.
[8] Jun-Seok Park and Soo-Hong Kim, Emotion recognition from speech signals using fractal features, International Journal of Software Engineering and Its Applications, Vol.8, No.5, 2014, pages 15-22.
[9] Nisha Chandran, Mahesh B. S., emotion recognition of speech signals using priori information of speaker's gender, International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol. 5, Issue 6, June 2016, pages 4509-4520.
[10] Puja Ramesh Chaudhari and John Sahaya Rani Alex, Selection of features for emotion recognition from speech, Indian Journal of Science and Technology, Vol.9 (39), October 2016.
[11] Igor Bisio, Alessandro Delfino, Fabio Lavagetto, Mario Marchese and Andrea Sciarrone, Gender-driven emotion recognition through speech signals for ambient intelligence applications, IEEE Transactions on Emerging Topics in Computing. 1, No. 2, December 2013, pages 244-257. https://doi.org/10.1109/TETC.2013.2274797.
[12] Sravani Nellore, A. Ramesh Kumar, V. Naveen Kumar, Emotion recognition from speech using embedded board OMAP 3530, International Journal for Advance Research in Engineering and Technology, Vol. 1, Issue IX, Oct.2013 ISSN 2320-6802, pages 38-44.
[13] Shing-Tai Pan, Chih-Hung Wu, Chen-Sen Ouyang, Ying-Wei Lee, Emotion recognition from speech signals by using evolutionary algorithm and empirical mode decomposition, Proceedings of EVA London 2018, UK, 2018, pages 140-147. https://doi.org/10.14236/ewic/EVA2018.29.
[14] Carlos Busso, Angeliki Metallinou, Iterative feature normalization scheme for automatic emotion detection from speech, IEEE Transactions on Affective Computing, June 2012.
[15] Zixing Zhang, Jing Han, Jun Deng, Xinzhou Xu, Fabien Ringeval, Björn Schuller, Leveraging unlabeled data for emotion recognition with enhanced collaborative semi-supervised learning, IEEE, 2018, pages 22196-22209.
[16] E. Sarath Kumar Naik, K. Suvarna, Comparative analysis of speaker dependent, speaker independent and cross language emotion recognition from speech using SVM, IRACST - International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN: 2249-9555, Vol.6, No. 4, July-August 2016.
[17] SanaulHaq and Philip J.B. Jackson, Speaker-dependent audio-visual emotion recognition, International Conference on Audio-Visual Speech Processing, 2009, pages 53-58.
[18] Manav Bhaykar, Jainath Yadav and K. Sreenivasa Rao, Speaker dependent, speaker independent and cross language emotion recognition from speech using GMM and HMM, IEEE, 2013.
[19] Nisha Beegum S, Wavelet and Fourier features based emotion recognition of speech signals, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 5, Issue 1, January 2016.
[20] M. Morales-Perez, J. Echeverry-Correa, A. Orozco-Gutierrez and G. Castellanos-
[21] Dominguez, Feature extraction of speech signals in emotion identification, International IEEE EMBS Conference, 2008, pages 2590-2593.
[22] A Tickle, S Raghu and M Elshaw, Emotional recognition from the speech signal for a virtual education agent, Journal of Physics, 2013, pages 1-6.
[23] Lianzhang Zhu, Leiming Chen, Dehai Zhao, Jiehan Zhou and Weishan Zhang, Emotion recognition from chinese speech for smart affective services using a combination of SVM and DBN, Sensors MDPI, 2017.
[24] Kun Han, Dong Yu, Ivan Tashev, Speech emotion recognition using deep neural network and extreme learning machine, ISCA, 2014, pages 223-227.
[25] Mohan Ghai, Shamit Lal, Shivam Duggaand Shrey Manik, Emotion recognition on speech signals using machine learning, IEEE,2017, pages 34-39.
[26] Firoz Shah A., Raji Sukumar A., Babu Anto P., Speaker and text dependent automatic emotion recognition from female speech by using artificial neural networks, IEEE, 2009, pages 1411-1413. https://doi.org/10.1109/NABIC.2009.5393712.
[27] Aharon Satt, Shai Rozenberg, Ron Hoory, Efficient emotion recognition from speech using deep learning on spectrograms, ISCA,2017, pages 1089-1093.
[28] K. Sathiyamurthy, T. Pavidhra, B. Monisha and K. VishnuPriya, Hidden Markov model approach towards emotion detection from speech signal, CSCP, 2015, pages 13-19.