International Journal of Multimedia and Ubiquitous Engineering
Vol. 14, No. 2 (2019), pp.15-22
http://dx.doi.org/10.21742/ijmue.2019.14.2.03

Speech Emotion Recognition: A Survey

Swarna. kuchibhotla

Associate Professor, Department of CSE, Koneru Lakshmaiah Education Foundation, Guntur
drkswarna@kluniversity.in

Abstract

Speech Recognition is a tremendous application from the history that is identification and conversion of spoken words into text. The performance and quality of work had increased a lot. This performance lead to the research work on Emotion Recognition based on the language spoken that is obtaining the kind of emotion from the spoken speech which is an application based on human-robot interactions. Emotions can be recognized in a better way using Speech processing, Artificial Intelligence techniques and linguistic semantics. Systems are given training in such a way to detect the emotions from the spoken utterances. This paper contains about the survey from the history to the present works that took place in the speech emotion recognition and also the experiment results. The survey contains about the works that took place from the by different scientists and their usage of different features, classifiers etc. The paper also holds three categories, one is different databases that are involved, second is what features are involved for representation of speech and third about the classification schemes. The survey also includes the conclusions of performances and limitations of current speech emotion recognition.

Keywords: *Emotion, Features, Classification Techniques, Speech Recognition*

1. Introduction

The way of communication between humans in natural way is through speaking. The scientists have used this as a source and thought of speech recognition that is interaction between human and a system. There was tremendous work happened on this speech recognition from many years back. Now the system is capable of converting the spoken speech into text. But speech alone cannot completely create the natural way of communication as how it happens between humans. This slowly lead scientists to think of recognizing the type of emotion from the spoken words. Currently speech emotion recognition have become the most popular works in the research field. Still the scientists are working on this to increase the accuracy and performance level of the system by using different training ways extracting the emotions from the orators. The speech emotion recognition is used for many human-computer applications such as computer tutorial applications, web movies [1], here response of the systems depends on the emotions detected. However, recognizing emotions from speech is very challenging because features sometimes are unclear to detect the correct emotions.

The diagram of Speech emotion Recognition system is shown in Fig1 in which five major important steps are involved to implement the system. Initially, the speech signal is given as an input to the system. For this database of certain emotions is to be created. Second step is features

Article history:
Received (March 6, 2019), Review Result (June 15, 2019), Accepted (October 11, 2019)

Print ISSN: 1975-0080, eISSN: 2652-1954 IJMUE
Copyright ⓒ 2019 GV Press

extraction where different features such as pitch, MFCC, formants, log energy, LPC [1][5][7] etc., are obtained. Large number of features can be extracted from a given set of data samples. As the dimensionality of the feature vector is very large only fewer features are extracted. Next step is choose of particular classifier, there are many classifiers such as Support Vector Machine (SVM), Gaussian Mixture Model (GMM), and Hidden Markov Model (HMM) etc. Last step is output that is obtained which includes accuracy level of the system.[13]

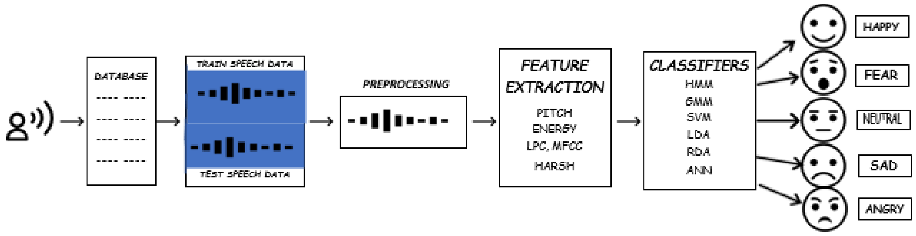

Figure 1. Process involved in Speech Emotion Recognition System

One best applications is it's usage in car-board system [1]the driver's mood or mental state can be detected by the system. So that will ensure the driver's safety and stop accidents. It is also used in call center applications [1][10][11], mobile communications and also in automatic translation systems where speaker's emotional state can be known. Recognition of speech from emotions has become very challenging now for the scientists to increase the performance levels. Some of the challenging issues they come across are, what speech features are required to distinguish the emotions, some of the people undergo same kind of emotions like sadness for many days, weeks and months in that case it is difficult to predict that other emotions may or may not exists for longer time. Sometimes the speaker's way of speaking, his style of speaking may affect during features extractions. Emotions are categorized into 2 dimensions they are activation and valence [1].

Activation is defined as the amount of energy required to exhibit emotions. Activation levels are of again two types they are low activation which are otherwise called as low arousal emotions and high activation can be called as high arousal emotions. Some of the high arousal emotions like happiness and anger have different affect and they can be shown by using valence dimensions. Linguists plays a major role to invent emotional states that mostly encountered in our lives. Linguistic information through words detect emotions in qualitative patterns [12]. But the drawback of linguistic information is cross culture diversities. A data set which contains 300 emotional states given by Schubiger and O'connor[6]. It is very difficult to deal with such huge data. So there is a Palette theory [1][6] which suggests that the emotions set can be decomposed into primary emotions such as color contains come basic colors but have only common set of colors. Here the common set of emotions are Anger, Happy, Sad, Fear, Surprise, and Neutral [1][9]. These emotions are called archetypal emotions.

2. Literature Survey

The survey contains about the study of emotions recognition through speech processing from the past times. The Information more precisely can be viewed from the Table 1. It all started in the year 1930 where the important voice feature vectors such as time-energy distribution vector, MFCC, fundamental frequency, LFCC etc., as their feature sets and classifiers. In 1972, Williams and Stevens compared the acted speech data to the spectrograms of real emotional speech. In 1993, Murray and Arnott[2] done an analysis on qualitative correlation between emotion and speech features such as pitch, intensity and timing of utterances. During the period of 1998 to 1999, Petrushin distinguished between agitation (Anger, Happy, Fear) and calm (Sad, Neutral) a type of emotions using RELIEF-F algorithm [3], K-NN, ANN classifiers obtained 43 features such as min, max, range, sd etc., but selected only top 14 features and the accuracy level was about 77% for normal state and sadness state. During the period of 2001, Nwe made analysis taking six emotions exhibited by two speakers with 12 MFCC features as input to discrete Hidden Markov Model (HMM) and the accuracy was 70%. In 2002, Yuet.al done analysis on four distinct emotions using support vector machines (SVM) as binary classifiers and got an accuracy of 73%. During the same period, Lee distinguished between negative and positive emotions in call center environment. He used K-NN, Linear Discrimination, and SVM classifiers and the accuracy is 75%. Owkwon made analysis on emotional expressions using pitch, log-energy, mel bond energies, formant, MFCC's (base features) and Gaussian mixture model, SVM in 2003. The accuracy was about 96.3% [3]. In the same year Gobl and Chasaide said that voice quality is reasonable for certain emotions. Batliner, made research o 4 class problem with emotions in spontaneous speech.

In 2004, Busso made a statement that statistics that relate MFCCs possess emotional information. In 2009, there are two types of emotion recognitions were analyzed one is Emotion recognition neural networks (ERNN). In this 128 input nodes, 20 hidden neurons, 3 summing output nodes, 97920 training set and 24480 testing sets are used. MATLAB, ANN, Hidden Markov Models, Radial basis function networks are it's classifiers and the performance was 100%. The other model is Gram-Charlier emotion recognition neural network (GERNN) that used 20 hidden neurons, three output nodes and the accuracy was about 33% only. In the year 2010, Manolis Wallao done analysis taking 133 sound or speech features of acted speech with seven emotions as feature set and the accuracy was about 51% only. By considering Arousal emotions, the accuracy for high arousal emotions is 100% and low arousal emotions is 87% [12] and it was successful. Later SVM classifier was used and the accuracy was 78%. In 2012, Schuller took Persian language and made analysis for emotion recognition taking 2400 wave clips with different emotions. MFCC, pitch, rate, energy, ANN as features and classifiers. The accuracy was about 78%. In 2013, Swarna etal[18][20] used MFCC, pitch and energy features and RDA classifier. The accuracy was 81%. In 2015, Koteswara rao Anne etal used same features and classifiers and obtained accuracy of 85%. There are many more scientists worked on SER research field to obtain emotions from speech using different classifiers and features set.

3. Speech Emotion Databases

Emotional speech databases play a vital role in the recognition of emotions. The databases that are chosen to evaluate and assess the performance of the emotional speech recognizer. Outcome performance shows whether the data base used is high-quality or low-quality. Emotions are classified as infant-directed [1] which include soothing, prohibition etc., and adult –directed which include joy, anger, happy etc., There are some criteria where the databases will simulate the real – world environment. Some of the studies suggests that databases need to be selected from the real life situations. Many of the recordings from radio channels, News channels, [1] etc., we can get the natural emotions through them. Some of the cases involve where the actors are allowed to express their emotions and then these emotion are collected. Examples for such databases are Danish Emotional Speech. In most cases the spoken utterances may not be natural. Even computer games can give natural emotional speech. Most of the scientific and experiment analysis use balanced utterances but they may sometimes reduce the validity of data. So unbalanced and valid utterances are used. Neutral emotions are part of our life in such cases databases on these emotions must be created.

Some of the people shows same emotions for different statements which shows the effect when comes to human-system interactions. Bilingual people we can see such type of cases. Their way of accent also differs. Most of the databases are private oriented and are not public free. Some of the databases can be viewed from Table 2. KISMET and BabyEars are infant-directed emotions and most of the databases are adult-oriented databases. In the human-system interaction mostly infant-oriented databases are used. There are many problems that occur in databases. As mentioned earlier some of the cases where most databases don't simulate in a natural and clear way. So that performance gradually decreases. Infant-directed emotions don't provide phonetic transcriptions in that case it is difficult to extract linguistic content. Many speech emotional databases are created by the recording made by professional and nonprofessional actors.

4. Speech Emotion Features

Extraction of particular features to characterize various emotions takes a huge part in speech emotion recognition. The performance of the system is based on the s elected features only. There can be many features in a speech signal can be seen from Table 3. During the features extraction some of the issues are considered. The region of analysis to be used properly. But some researchers divide the speech signal into frames (small intervals)[1]. Based on the frames the researchers select local or global vectors of features. Best features need to be selected that was also an issue. During the features extraction sometimes the emotions need to be combined with other type of features like facial expressions, linguistics etc., and these issues can be solved by taking up certain measures.

The features can be divided into 2 types local and global [1]. As mentioned earlier when the frames are obtained in each frame the signals are constant. Pitch, energy, etc. are local features. Global features can be all speech features that are obtained from speech signals. Based on accuracy and average time it is said that global features are more applicable than local. Global features are very less in number and also the algorithms or functioning can be much easier. Global features are executed for high- arousal emotions [11] like anger, joy, fear etc., but they cannot work for similar arousals. Temporary information is also lost by the use of global features. Classifiers such as HMM, SVM cannot work for global features but only for local features. Features Extraction can also be done using phoneme based approach which is combination of segment based and global features. Use of global features increased the accuracy up to 5%. The best speech features for a particular task cannot be explored. Because there can be many features obtained for a speech signal. Speech features are classified into 4 types.

5. Speech Emotion Classifiers

The important module in speech emotion recognition system is classification schemes. Extraction of features is the front end part of speech emotion recognition where classifiers lies in the back end part of the system. This the stage where the features extracted are given as input to these classifiers. The main use of classifiers is to classify the emotional states from the speech samples using databases and features. There are many classifiers can be seen from Table 4 that are used in this system. Some of them are gaussian mixture model (GMM), support vector machine (SVM), hidden markov model (HMM), k nearest neighbor (KNN), artificial neural networks (ANN), linear discriminant analysis (LDA), quadratic discriminant analysis (QDC) ,regularized discriminant analysis (RDA), Bayes classifier, maximum a posteriori (MAP) [4], Multiple classifier system (MCS) and many more. Different scientists used different classifiers to recognition different emotions and increase the accuracy. The speech set is divided training set and test set. Classifiers are given training using training speech set and are tested to obtain optimal solution using test set.

5.1 Hidden Markov Model (HMM)

Most used classifier in speech emotion recognition (SER) is hidden markov model. This is doubly stochastic process [6] in which system is designed using markov process with hidden states. HMM is represented in the form of dynamic Bayesian network. In markov model the states are not visible to the observer but the outputs that are dependent on the states can be visible at the end to the observer. The states that are hidden captures the temporal structure of the data. HMM parameters use Machine Learning principle that is maximizing the likelihood function. HMM are used in applications such as reinforcement learning, temporal pattern recognitions such as speech, handwriting, gestures etc., In SER, some part of database is used for training the HMM and other part is used for testing. Left to right strategy [8] is used in HMM system. Here the number of states are the initial probabilities of HMM. Based on ML criteria each emotion is recognized. The output generated from HMM has probability distribution and the maximum probability is chosen for every emotion. In automatic speech recognition (ASR), HMM is the widely used model.

5.2 Gaussian Mixture Model (GMM)

GMM is an unsupervised classifier where the training set is not labelled and the usage of training, testing sets are limited. In GMM we calculate probability density function. As of HMM, GMM don't contain multiple hidden states but contains only one state. In GMM, the observation vector using different classes can be designed by linearly combining the multi normal densities. It doesn't talk about temporal structure [4] in order to model temporal structure, GMM is combined to vector auto regressive and resulted in Gaussian mixture vector auto regressive model (GMVAR). GMM is used in various field such as image pattern recognition, MIR, speech recognition etc. GMM is known as state-of-art[1][14] classifier for verification, classification, speech identification etc. Here are the outcome of GMM determines the optimized Gaussian components for each emotion. GMM is most efficient over global features.

5.3 Support Vector Machine (SVM)

SVM is supervised learning model that analyze data for regression and classification. They are the best examples for general discriminant classifier. SVM perform non-linear classification with kernel function[2][4]mapping normal features set into high dimensional feature space that leads to optimal classification. In SER, emotional states can be obtained with high level accuracy compared to other classifiers. For speaker independent classification the accuracy by using SVM was 75% whereas for speaker dependent classification the accuracy is 80%[4][15]. The support vectors are measurement vectors which defines the boundaries of the margin. SVM classifiers are originally designed for two class problems but they can be used for more classes also. SVM systems are given risk minimization oriented training to have high generalizing capability.

5.4 Artificial Neural Networks (ANN)

Artificial neural networks (ANN) is an example of supervised learning that finds non-linear boundaries separating the emotional states. It is used in many pattern recognitions. There are many types of neural networks such as feed forward neural networks, recurrent neural networks (RNN), Multilayer perceptron (MLP), radial basis function (RBF) etc. Multilayer perceptron is easy for implementation in speech emotion recognition and have well defined learning algorithms. ANN models are based on normalized temporal features vector and static feature vector to recognize emotions from speech. ANN's performance depends on parameters such as the number of hidden layers, neuron activation function and the number of neurons in each layer. But the classification accuracy was low compared to other classifiers[16][17].

5.5 K-Nearest Neighbor (KNN)

K-Nearest Neighbor is a non-parametric method that classifies speech dataset based on closest training samples in the feature space [2]. In the classification phase, an unlabeled vector is classified by assigning the label which is most frequent among the k training samples nearest to the k point where k is a user-defined constant, Unclassified speech samples are sent into the system to extract speech coefficients and uses model file to classify the speech emotion[19].

4. Conclusion

To brief about the paper, it contains literature survey from the past to the current study work in speech emotion recognition. In the development of Speech emotion recognition system feature extraction, speech database, classifiers play a major role. We can also see the average accuracies by using different classifiers in the given task. Most used classifiers are Gaussian mixture model (GMM), hidden markov model (HMM), as the average accuracy is high on using these for the given task. Many new features are also developed and speech data bases are created using the emotions uttered by different professional actors in different languages like Spanish, English, etc. SVM classifiers increased the accuracy nearly to 80% [1]in this domain. Most of the scientists now-a-days are preferring multiple classifier systems by combining different classifiers. The performance of speech emotion recognition using classifiers can be obtained with feature fusion technique. In future many algorithms will be developed to achieve 100% accuracy rate in recognizing emotions.

References

[1] El Ayadi, Moataz, Mohamed S. Kamel, and Fakhri Karray. "Survey on speech emotion recognition: Features, classification schemes, and databases." Pattern Recognition 44, no. 3 (2011): pp.572-587.DOI: 10.1016/j.patcog.2010.09.020

[2] Kuchibhotla, Swarna, H. D. Vankayalapati, R. S. Vaddi, and Koteswara Rao Anne. "A comparative analysis of classifiers in emotion recognition through acoustic features." International Journal of Speech Technology 17, no. 4 **(2014)**: pp.401-408.DOI: 10.1007/s10772-014-9239-3

[3] Khanna, Preeti, and M. Sasikumar. "Recognizing emotions from human speech." In Thinkquest~ 2010, pp. 219-223. Springer, New Delhi, **(2011)**.DOI: 10.1007/978-81-8489-989-4_40

[4] Gadhe, Rani P., R. A. Shaikh Nilofer, V. B. Waghmare, P. P. Shrishrimal, and R. R. Deshmukh. "Emotion recognition from speech: a survey." Int. J. Sci. Eng. Res 6, no. 4 **(2015)**: pp.632-635.

[5] Joshi, Aastha, and Rajneet Kaur. "A Study of speech emotion recognition methods." Int. J. Comput. Sci. Mob. Comput.(IJCSMC) 2, no. 4 **(2013)**: pp.28-31.

[6] Ingale, Ashish B., and D. S. Chaudhari. "Speech emotion recognition." International Journal of Soft Computing and Engineering (IJSCE) 2, no. 1 **(2012)**: pp.235-238.

[7] Kuchibhotla, Swarna, Hima Deepthi Vankayalapati, and Koteswara Rao Anne. "An optimal two stage feature selection for speech emotion recognition using acoustic features." International journal of speech technology 19, no. 4 **(2016)**: 657-667.DOI: 10.1007/s10772-016-9358-0

[8] Davletcharova, Assel, Sherin Sugathan, Bibia Abraham, and Alex Pappachen James. "Detection and analysis of emotion from speech signals." Procedia Computer Science 58 **(2015)**: 91-96.

[9] Nanavare, V. V., and S. K. Jagtap. "Recognition of human emotions from speech processing." Procedia Computer Science 49 **(2015)**: 24-32.DOI: 10.1016/j.procs.2015.04.223

[10] Utane, Akshay S., and S. L. Nalbalwar. "Emotion recognition through Speech." In 2nd National Conference on Innovative Paradigms in Engineering & Technology, International Journal of Applied Information Systems, pp. 5-8. **(2013)**.

[11] Vogt, Thurid, Elisabeth André, and Johannes Wagner. "Automatic recognition of emotions from speech: a review of the literature and recommendations for practical realisation." In Affect and emotion in human-computer interaction, pp. 75-91. Springer, Berlin, Heidelberg, **(2008)**.DOI: 10.1007/978-3-540-85099-1_7

[12] Anagnostopoulos, Christos-Nikolaos, and Theodoros Iliou. "Towards emotion recognition from speech: definition, problems and the materials of research." In Semantics in Adaptive and Personalized Services, pp. 127-143. Springer, Berlin, Heidelberg, **(2010)**.DOI: 10.1007/978-3-642-11684-1_8

[13] Anne, Koteswara Rao, Swarna Kuchibhotla, and Hima Deepthi Vankayalapati. "Emotion recognition using spectral features." In Acoustic Modeling for Emotion Recognition, pp. 17-26. Springer, Cham, **(2015)**.DOI: 10.1007/978-3-319-15530-2_3

[14] Atal, Bishnu S., and Suzanne L. Hanauer. "Speech analysis and synthesis by linear prediction of the speech wave." The journal of the acoustical society of America 50, no. 2B **(1971)**: 637-655.DOI: 10.1121/1.1912679

[15] Barbu, Tudor. "Discrete speech recognition using a Hausdorff-based metric." In Proceedings of the 1st International Conference of E-Business and Telecommunication Networks, ICETE 2004, vol. 3, pp. 363-368. Setubal, **(2004)**.

[16] Batliner, Anton, Richard Huber, Heinrich Niemann, Elmar Nöth, Jörg Spilker, and Kerstin Fischer. "The recognition of emotion." In Verbmobil: Foundations of speech-to-speech translation, pp. 122-130. Springer, Berlin, Heidelberg, **(2000)**.

[17] Burges, Christopher JC. "A tutorial on support vector machines for pattern recognition." Data mining and knowledge discovery 2, no. 2 **(1998)**: 121-167.DOI: 10.1023/a:1009715923555

[18] Kuchibhotla, Swarna, H. D. Vankayalapati, R. S. Vaddi, and Koteswara Rao Anne. "A comparative analysis of classifiers in emotion recognition through acoustic features." International Journal of Speech Technology 17, no. 4 **(2014)**: 401-408.

[19] Kuchibhotla, Swarna, B. S. Yalamanchili, H. D. Vankayalapati, and Koteswara Rao Anne. "Speech emotion recognition using regularized discriminant analysis." In Proceedings of the International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2013, pp. 363-369. Springer, Cham, **(2014)**.

[20] Kuchibhotlaa, Swarna, Hima Deepthi Vankayalapati, BhanuSree Yalamanchili, and Koteswara Rao Anne. "Analysis and evaluation of discriminant analysis techniques for multiclass classification of human vocal emotions." In Advances in Intelligent Informatics, pp. 325-333. Springer, Cham, **(2015)**.