Automatic emotion recognition from speech signals: A Review

Shaikh Nilofer R. A.,Rani P. Gadhe, R. R. Deshmukh, V. B. Waghmare, P. P. Shrishrimal

Abstract— Automatic Speech emotion recognition has been a burning issue since last decade. Reserchers have been trying to develop a system more like human , emotion recognizing robots is an example of it. Speech has many parameters which have great weightage in recognizing emotion namely Prosodic and spectral features. Out of prosodic features namely pitch , energy and intensity are popularly used and out of spectral features formant Mel frequency cepstral coefficients are commonly used by the researchers worldwide. Further the classifiers are trained by using these features for classifying emotions accurately. This paper is an attempt to give a short review about the work on Emotion recognition from speech

Index Terms— Emotions, Speech emotion recognition, Elicited, Excitation source features, vocal tract features, MFCC, HMM.

1 INTRODUCTION

SPEECH peech is one of the basic and natural way of communication among human beings [1] Emotions makes speech more expressive and effective. Different ways like laughing , yelling, teasing, crying, etc, are used by humans to express their emotions [2]. Emotion detection can be an easy task for humans but a difficult one for machines. So there is a need of such emotion recognition systems that can make human computer interaction quite easy. Speech emotion recognition thus can be defined as the extraction of the emotional state of the speaker from his or her speech signal to make human machine interface more convenient. The widely used application of Automatic Speech emotion recognition is in the field of human machine interaction. Other applications of the Automatic speech emotion recognition system are Lie-Detection, Intelligent toys, phychiatric diagnosis and the most popular in Call center [3]

Till date many speech recognition systems have been proposed. Researchers have been using various techniques for identifying emotions. This includes accurate feature extraction and selection and further applying proper classifier These systems used various features viz. Prosodic and spectral where prosodic features included Pitch, Speech intensity glottal parameters and Spectral features included Mel-frequency cepstral coefficients(MFCC) and Linear Predictive cepstral coefficients LPCC. [4] The different classifiers used for emotion recognition are Hidden Markov Model(HMM), Gaussian Mixture Model (GMM), Support Vector Machine(SVM), Artificial Neural Network(ANN). A study using the features Pitch and energy and classifier HMM was performed in [5] where the accuracy rate achieved was 86%. Another study used SVM as a classifier with Berlin Database where the overall recognition rate was 82.5% [6].

The paper comprises of the following sections: Section two briefly describes the speech emotion recognition system followed by section three review of database. Section four gives idea about feature extraction and feature selection followed by a short review of classifiers used for emotion recognition finally the last section concludes the paper.

2 SPEECH EMOTION RECOGNITION SYSTEM

Speech Emotion recognition comprises of the steps as shown in Figure 1.

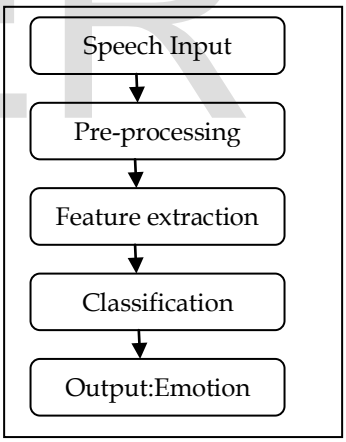

Fig1: Speech emotion recognition system

The speech samples are taken as input.The first thing to be done with the speech samples is the pre-processing where noise ffrom the sample is removed. Now from the noise free samples desired features are extraxted. These features are then further pass on to the cclassifier a. The classifier thus classifies the emotions accordingly and outputs the emotions.

3 DATABASE

Databases play a vital role for automatic emotion recognition as the rest statistical methods are learned using examples. The databases used till know in the research acted, Elicited and real life or natural emotions. As the naturalness of the database increases the complexity also increases. So at the

beginning of the research on automatic vocal emotion recognition, which started actually in the mid-90s, work began with acted speech [7] and shifts now towards more realistic data [8][9].

The most popular examples of acted database are Berlin Database of emotional speech [11] which comprised of 5 male and 5 female actresses and the Danish Emotional speech corpus (DES)[12].

Fig:2 Types of database used in emotion recognition and their difficulty level.[10]

Russian databse consists of ten pronounces sentences from61 speakers (12 male 49female) of age group 16-28years expressing six emotions viz., Happy, sad, angry, fear, neutral and disgust. The example of Induced database is SmartKom corpus [13] and the German Aibo emotion corpus [14] without knowing the people that their emotions are being recorded. The call center communication by Devillersand etal[8] is obtained from live recordings and is an example of real emotional database. Other examples includes Surrey Audio-Visual Expressed Emotion (SAVEE)[15]which comprised of 4 male actors expressing 7 different emotions. The Speech Under Simulate and Actual Stress (SUSA)[16] database of 32 speakers where speech was recorded in both simulated stress and actual stress

4 PREPROCESSING

Pre-Processing is to be used when no standard database is to be used.

Pre processing includes removing noise from the samples. Various softwares are available for pre processing like PRAAT and Audacity. Here we can see the spectrogram of the speech sample and remove the noise. Silence is also noise that too can be removed using Audacity software. The purpose of pre-processing is toboost the high frequencies of a signal and get flat frequency spectrum of signals and frequency characteristics. By using window function we get speech frames. Now a day's commonly used window functions are Hamming window and rectangular window.

5 FEATURE EXTRACTION AND SELECTION

Feature extraction and selection is an important step in Emotion Recognition System. As the features are to be choosen to reperesent information**.** After extracting the emotions one has to decide which features are to be selected accordingly.

Mainly features are classified as Elicited features Prosodic features and Spectral features many researchers used combination of features

5.1 EXCITATION SOURCE FEATURES

Speech features derived from excitation source signal are known as source features. Excitation source signal is obtained from speech, after suppressing vocal tract (VT) characteristics. In literature, very few attempts have been made to explore the excitation source information for developing any of the speech systems. Excitation source signal were used in [17] to discriminate emotions in continuous speech

5.2 PROSODIC FEATURES

Prosodic features include Intensity, Pitch, Energy. The mean Standard deviation, minimum, maximum, range and variance of Pitch, enery and other similar features are used for distinguishing emotions**. [**18**]**In another study the peaks and troughs of of fundamental frequency and intensity are studied which gave 55%accuracy for four emotions namely sad, fear, joy and anger.[19]

5.3 SPECTRAL FEATURES

The spectral features are also known as vocal tract, segmental or system features. Spectral features include formants MFCCs, LPCCs, and perceptual linear prediction coefficients (PLPCs). IN order to recognise anger,happy, boredom, sad and neutral emotions combination of PLP, RASTA, LPCC and MFCC and log frequency is used in [20][21]

A number of features can be extracted using feature extraction techniques. In order to achieve higher accuracy features should be selected wisely.

There are various feature selection algorithm present some of them are Forward selection and backward selection. In forward selection there is linear loss function to which a feature is added that provides the least error. Whereas in the backward selection all features are selected at first instance and then the feature that minimises loss function is removed.

6 CLASSIFIER

After feature extraction and feature selection the next step is to choose a suitable classifier. As classifier also contributes in accuracy of emotions recognised. There are number of classifiers available namely HMM, GMM, ANN, SVM etc.Combination of classifier can also be used making a hybrid model. Each one has some pros and cons over the other. Some are good at large database, some are good for some relevant features. Following are some of the commonly used classifiers

HMM

The Hidden Markow model has been widely used in the literature but its classification property is not upto the mark. Accuracy with above 70% Recognition rate was obtained by

using Hidden Markov Model used in [22] with the low-level features namely pitch and energy.

GMM

As the trainining and testing requirements for gmm are very less it is very efficient in modelling multi-modal distributions. GMM are suitable when large number of feature vectors are available. GMM work efficiently for spectral features. GMM are among the matured techniques for probability density function and clustering [23].

ANN

Artificial Neural Network is also popularly used for emotion recognition from speech A three layered(two hidden and one output) feed forward neural network is used in[24]. The hidden layers had 10-10 nodes each . the overall recognition rate obtained was about78.3% . the most recognizable emotion was anger with 90% accuracy whereas Fear was least recognized with an accuracy of 60%.

SVM

Another popular technique is Support Vector Machine(SVM). SVM creates an hyperplane in high or infinite space for classification. the distance to the nearest training data better the separation gained by hyperplane.[serhmmsvm] An overall accuracy of 94.2% is achieved using isolated SVM in[25].

BINARY SEARCH TREE

When Binary search tree is used for emotion recognition from speech a particular emotion is classified at each node of the tree. A study on two databases was done in [26] on German and Polish which gave a 72% for speaker independent recognition for Polish Database.

7 CONCLUSION

In this paper a brief idea of speech emotion recognition is illustrated. Various available databases, feature extraction techniques and classifiers are studied. This study may give a precise idea to the researchers working in the field of Emotion recognition from speech. The recognition rate depends upon the feature extracted and the database used to recognize emotions.

ACKNOWLEDGMENT

This work is supported by University Grants Commission as Major Research Project. The authors would like to thank the Department of Computer Science & IT, Dr. Babasaheb Ambedkar Marathwada University Authorities for providing the infrastructure to carry out the research.

REFERENCES

[1] Pukhraj Shrishrimal, R. R. Deshmukh, Vishal Waghmare, "Indian Language Speech Database: A Review," International Journal of Computer Application (IJCA) Vol 47, No.5 pp.17-21, July 2012

[2] Vishal B Waghmare, Ratnadeep R Deshmukh, Pukhraj P Shrishrimal "Development of Isolated Marathi Words Emotional Speech Database," International Journal of Computer Applications (0975 – 8887) Volume 94 – No 4, May 2014.

[3] Ayadi M. E., Kamel M. S. and Karray F., "Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases," Pattern Recognition, 44 (16), 572-587, 2011.

[4] Zhou y., Sun Y., Zhang J, Yan Y., "Speech Emotion Recognition using Both Spectral and Prosodic Features," IEEE, 23(5), 545-549, 2009.

[5] Schuller B., Rigoll G., Lang M., "Hidden Markov Model Based Speech Emotion Recognition," IEEE ICASSP, 1-3, 2003.

[6] Shen P., Changjun Z. and Chen X., "Automatic Speech Emotion Recognition Using Support Vector Machine," Proceedings of International Conference On Electronic And Mechanical Engineering And Information Technology, 621-625, 2011.

[7] Dellaert, F., Polzin, T., Waibel, A.: "Recognizing emotion in speech", In: Proceedings of ICSLP, Philadelphia, USA Automatic Recognition of Emotions from Speech 89 1996.

[8] Devillers, L., Vidrascu, L., Lamel, L.: "Challenges in real-life emotion annotation and machine learning based detection," Neural Networks, 18(4), 407–422 (2005).

[9] Litman, D.J., Forbes-Riley, K.: "Predicting student emotions in computer-human tutoring dialogues," In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL), Barcelona, Spain (2004).

[10] Thurid Vogt, Elisabeth Andr´e, and Johannes Wagner, "Automatic Recognition of Emotions from Speech: A Review of the Literature and Recommendations for Practical Realisation," C. Peter and R. Beale (Eds.): Affect and Emotion in HCI, LNCS 4868, pp. 75–91, 2008.c_Springer-Verlag Berlin Heidelberg 2008.

[11] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F., Weiss, B.: "A database of German emotional speech," In: Proceedings of Interspeech 2005, Lisbon, Portugal (2005).

[12] Engberg, I.S., Hansen, A.V.: "Documentation of the Danish Emotional Speech Database (DES)," Technical report. Aalborg University, Aalborg, Denmark (1996).

[13] Schiel, F., Steininger, S., T¨urk, U.: "The SmartKom multimodal corpus at BAS," In: Proceedings of the 3rd Language Resources & Evaluation Conference (LREC) 2002, Las Palmas, Gran Canaria, Spain, pp. 200–206 (2002).

[14] Batliner, A., Hacker, C., Steidl, S., N¨oth, E., D'Arcy, S., Russell, M., Wong, M.: "You stupid tin box" - children interacting with the AIBO robot: A cross-linguistic emotional speech corpus," In: Proceedings of the 4th International Conference of Language Resources and Evaluation LREC 2004, Lisbon, pp. 171–174 (2004).

[15] M. You, C. Chen, J. Bu, J. Liu, and J. Tao, "Getting started with susas: a speech under simulated and actual stress database," EuroSpeech, vol. 4, pp. 1743–1746, 1997.

[16] S. Haq, P. Jackson, and J. Edge, "Audio visual feature selection and reduction for emotion classification," AVSP, pp. 185–190, 2008.

[17] Hua, L. Z., Yu, H., & Hua, W. R. "A novel source analysismethod by matching spectral characters of LF model withSTRAIGHT spectrum,". Berlin: Springer.2005.

[18] Rao, K. S., & Yegnanarayana, B." Prosody modification using

instants of significant excitation," IEEE Transactions on Audio,Speech, andLanguage Processing, 14, 972–98, .2006.

[19] McGilloway, S., Cowie, R., Douglas-Cowie, E., Gielen, S.,Westerdijk, M., & Stroeve, S. (2000)."Approaching automatic recognition of emotion from voice," A rough benchmark. In ISCA workshop on speech and emotion, Belfast.

[20] Pao, T. L., Chen, Y. T., Yeh, J. H., & Liao, W. Y. "Combining acoustic features for improved emotion recognition in Mandarin speech," In J. Tao, T. Tan, & R. Picard (Eds.), ACII. LNCS (pp.279–285). Berlin: Springer.2005.

[21] Pao, T. L., Chen, Y. T., Yeh, J. H., Cheng, Y. M., & Chien, C. S. "Feature combination for better differentiating anger from neutral in Mandarin emotional speech," LNCS: Vol. 4738Berlin:Springer.2007.

[22] [Online] ww.cs.upc.edu/~nlp/papers/nogueiras01.pdf 02-01-2015.

[23] Dipti D. Joshi, Prof. M. B. Zalte "Speech Emotion Recognition: A Review," IOSR Journal of Electronics and Communication Engineering (IOSR-JECE)ISSN: 2278-2834, ISBN: 2278-8735. Volume 4, Issue 4 PP 34-37 (Jan. - Feb. 2013).

[24] [Online]http://desceco.org/O-COCOSDA2010/proceedings/paper_52.pdf.02-01-2015.

[25] Aastha Joshi,,"Speech Emotion Recognition Using Combined Features of HMM & SVM Algorithm," International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 8, pp. 387-393,August 2013.

[26] [Online]www.di.uniba.it/intint/DC-ACII07/Chicosz.pdf 02-01-2015.