Review on Speech Emotion Recognition

S.Praveena
ECE Dept, M.G.I.T,
Hyderabad.

Abstract: Speaker's emotion is one of the latest challenges in speech technologies. There are three important issues in speech emotion recognition. The first one is selection of database. The second one is choice of features to recognize the emotion from the speech signals and the third one is selection of classifier. In this paper, a review of existing work on emotional speech recognition is discussed.

Keywords: ***Speech Emotion Recognition, feature extraction, classifier***

I. INTRODUCTION

The main objective of speech emotion recognition is to get different emotions from a given speech signal. Automatic emotion recognition from speech is challenging when training data and test data are drawn from different domains due to different recording conditions, languages, speakers and many other factors[1]. Despite the great progress made in artificial intelligence, we are still far from being able to naturally interact with machines, partly because machines do not understand our emotion states. Recently, speech emotion recognition, which aims to recognize emotion states from speech signals, has been drawing increasing attention. Speech emotion recognition is a very challenging task of which extracting effective emotional features is an open question [2, 9].

II. SPEECH EMOTION RECOGNITION

The speech emotion recognition system contains five main stages emotional speech input, feature extraction, feature selection, classification, and recognized emotional output.

The need to find out a set of the significant emotions to be classified by an automatic emotion recognizer is a main concern in speech emotion recognition system [2]. Primary emotions are anger, disgust, fear, joy, sadness and surprise.

The prosodic features are known as the primary indicator of the speakers emotional states. Research on emotion of speech indicates that pitch, energy, duration, formant, Mel frequency cepstrum coefficient (MFCC), and linear prediction cepstrum coefficient (LPCC) are the important features [3, 4]. With the different emotional state, corresponding changes occurs in the speak rate, pitch, energy, and spectrum. Typically anger has a higher mean value and variance of pitch and mean value of energy. In the happy state there is an improvement in mean value, variation range and variance of pitch and mean value of energy. On the other hand the mean value, variation range and variance of pitch is decreases in sadness, also the energy is weak, speak rate is slow and decrease in spectrum of high frequency components. The feature of fear has a high mean value and variation range of pitch, improvement of spectrum in high frequency components. Therefore statistics of pitch, energy and some spectrum feature can be extracted to recognize emotions from speech [3, 4].

The choice of classifier plays an important role in classifying emotion from given speech signal. There are two types of classifiers linear and non linear classifier. Linear classifiers

include Naïve Bayes classifier, linear support vector machine, Least square methods etc. Non linear classifiers include ANN,GMM, HMM& K-NN etc.

III. Review of Literature

Kun Han, Dong Yu, Ivan Tashev[5] have utilized deep neural networks (DNNs) to extract high level features from raw data and show that they are effective for speech emotion recognition. In this work,an emotion state probability distribution for each speech segment using DNNs is produced and then constructed utterance-level features from segment-level probability distributions. These utterance level features are then fed into an extreme learning machine (ELM), a special simple and efficient single-hidden-layer neural network, to identify utterance-level emotions. The experimental results demonstrate that the proposed approach effectively learns emotional information from low-level features and leads to 20% relative accuracy improvement compared to the stateof-the-art approaches.

Kunxia Wang, Ning An et al.[6] have used harmony features for speech emotion recognition. They found that that the first- and second-order differences of harmony features also play an important role in speech emotion recognition They proposed a new Fourier parameter model by using the perceptual content of voice quality, the first- and second-order differences for speaker-independent speech emotion recognition. Their results have shown that the proposed Fourier parameter (FP) features are effective in identifying various emotion states in speech signals. They improve the recognition rates over the methods using Mel Frequency Cepstral Coefficient (MFCC) features by 16.2 points, 6.8 points and 16.6 points on the German database (EMODB), the Chinese language database (CASIA) and the Chinese elderly emotion database (EESDB). In particular, if combining FP with MFCC, the recognition rates can be further improved by 17.5 points, 10 points and 10.5 points on the aforementioned databases, respectively.

Zhengwei Huang et al.[7] have proposed a novel feature transfer approach with PCANet (a deep network), which extracts both the domain-shared and the domain-specific latent features to facilitate performance improvement. The proposal attempts to learn multiple intermediate feature representations
along an interpolating path between the source and target domains using PCANet by considering the distribution shift between source domain and target domain, and then aligns
other feature representations on the path with target subspace to control them to change in the right direction towards the target. The work is done on INTERSPEECH 2009 Emotion Challenge's FAU Aibo Emotion Corpus as the target database and two public databases (ABC and Emo-DB) as source set. Experimental results demonstrated that the proposed feature transfer learning method outperforms the conventional machine learning methods and other transfer learning methods on the performance.

Seyedmahdad Mirsamadi, Emad Barsoum, Cha Zhang[8] have used a deep recurrent neural network, we can learn both the short-time frame-level acoustic features that are emotionally relevant, as well nas an appropriate temporal aggregation of those features into a compact utterance-level representation. Moreover, they proposed a novel strategy for feature pooling over time which uses local attention in order to focus on specific regions of a speech signal that are more emotionally salient. The proposed solution is evaluated on the IEMOCAP corpus, and is shown to provide more accurate predictions compared to existing emotion recognition algorithms.

Feature extraction is a challenging task in speech emotion recognition. Due to the lack of discriminative acoustic features, classical approaches based on traditional acoustic features

could not provide satisfactory performances. They proposed a novel type of feature related to prominence, which, together with traditional acoustic features, are used to classify seven typical different emotional states. To this end, the author group produces a Chinese Dual-mode Emotional Speech Database (CDESD), which contains additional prominence and paralinguistic annotation information. Then, a consistency assessment algorithm is presented to validate the reliability of the annotation information of this database. The results show that the annotation consistency on prominence reaches more than 60% on average. The proposed prominence features are validated on CDESD through speaker dependent and speaker-independent experiments with four commonly used classifiers. The results show that the average recognition rate achieved using the combined features is improved by 6% in speaker dependent experiments and by 6.2% in speaker-independent experiments compared with that achieved using only acoustic features.

IV. Conclusion

The important issues in speech emotion recognition system are the signal processing unit in which appropriate features are extracted from available speech signal and another is a classifier which recognizes emotions from the speech signal. Classification accuracy depends on features and classifier. Hence choice of features and classifier plays an important role in speech emotion recognition system.

V. References

[1] Seyedmahdad Mirsamadi, Emad Barsoum, | Cha Zhang, automatic speech emotion ecognition using recurrent neural networks with local attention

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

[3] A. Nogueiras, A. Moreno, A. Bonafonte, Jose B. Marino, "Speech Emotion Recognition Using Hidden Markov Model", Eurospeech, 2001.

[4] P.Shen, Z. Changjun, X. Chen, "Automatic Speech Emotion Recognition Using Support Vector Machine", International Conference On Electronic And Mechanical Engineering And Information Technology, 2011.

[5] *Kun Han, Dong Yu, Ivan Tashev* speech emotion recognition using deep neural network and extreme learning machine, 2014 isca

[6] Kunxia Wang, Ning An, Bing Nan Li, Yanyong Zhang, Lian Li, Speech Emotion Recognition Using Fourier Parameters, IEEE transactions on affective COMPUTING,2015

[7] Zhengwei Huang1 ·Wentao Xue1 · Qirong Mao1·Yongzhao Zhan1 Unsupervised domain adaptation for speech emotion recognition using PCANet, Multimed Tools Appl,Springer

[8] Seyedmahdad Mirsamadi1, Emad Barsoum2, cha zhang2 automatic speech emotion recognition using recurrent neural networks with local attention 2017 ieee international conference on acoustics, speech and signal processing (icassp)

[9] ShaolingJing, XiaMao, Lijiang Chen Prominence features: Effective emotional features for speech emotion recognition, Digital Signal Processing ,2017

[10] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognisin grealistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 9, pp. 1062–1087, 2011