Advances in Computer Science and Information Technology (ACSIT)
Print ISSN: 2393-9907; Online ISSN: 2393-9915; Volume 1, Number 3; November, 2014 pp. 135-139

http://www.krishisanskriti.org/acsit.html

A Survey on Speech Emotion Recognition

Mr. N. Ratna Kanth[1] and Dr. S. Saraswathi[2]

[1]Ph. D. Scholar, Pondicherry University Pondicherry
[2]Professor & Head, Department of Information Technology Pondicherry Engineering College
E-mail: [1]nratnakanth@yahoo.co.in, [2]swathimuk@yahoo.com

Abstract: *Emotion is one of the most basic factors with respect to the communication among humans. It would be ideal to have human emotions automatically recognized by machines for improving human machine interaction. This is the motive behind the constantly increasing attention that this particular scientific field has been receiving lately.This paper presents an up-to-date survey of Emotion Recognition from Speech addressing the important aspects of the design of a Speech Emotion RecognitionSystem. The first issue is the choice of suitable features for speech representation, the second issue is the proper preparation of an emotional speech database and the third issue is the design of an appropriate classification method. Conclusions about the performance and limitations of current speech emotion recognition systems are discussed in the last section. This section also suggests possible ways of improving speech emotion recognition systems.*

1. INTRODUCTION

Speech is the fastest and the most natural method ofcommunication between humans. This fact has motivatedresearchers to think of speech as a fast and efficient method ofinteraction between human and machine. However, this requiresthat the machine should have sufficient intelligence to recognize human voices. Since the late fifty's, there has been tremendousresearch on speechrecognition, which refers to the process ofconverting the human speech into a sequence of words. However,despite the great progress made in speech recognition, we are stillfar from having a natural interaction between man and machinebecause machinedoes notunderstandthe emotionalstate ofthespeaker. This has introduced a relatively recent research field,namely Speech EmotionRecognition (SER), which is defined as extractingthe emotional state of a speaker from his or her speech. It isbelieved that speech emotion recognition can be used to extractuseful semantics from speech, and hence, improves the performance of speech recognition systems[1].

The task of speech emotion recognition is very challenging for the following reasons. First, it is not clear which speech features are most powerful in distinguishing between emotions. The acoustic variability introduced by the existence of different sentences, speakers, speaking styles, and speaking rates adds another obstacle because these properties directly affect most of the common extracted speech features such as pitch, and energy contours[2]. Moreover, there may be more than one perceived emotion in the same utterance; each emotion corresponds to a different portion of the spoken utterance. In addition, it is very difficult to determine the boundaries between these portions. Another challenging issue is that how a certain emotion is expressed generally depends on the speaker, his or her culture and environment. Most work has focused on monolingual emotion classification, making an assumption there is no cultural difference among speakers. However, the task of multi-lingual classification has been investigated. Another problem is that one may undergo a certain emotionalstate such as sadness for days, weeks, or even months. In such a case, other emotions will be transient and will not last for more than a few minutes. As a consequence, it is not clear which emotionthe automatic emotion recognizer will detect: the long-term emotion or the transient one. Emotion does not have a commonly agreed theoretical definition. However, people know emotions when they feel them. For this reason, researchers were able to study and define different aspects of emotions. It is widely thought that emotion can be characterized in two dimensions: activation and valence[3].

An important issue in speech emotion recognition is the need to determine a set of the importantemotions to be classified by an automatic emotion recognizer. Linguists have defined inventoriesof the emotional states, most encountered in our lives. A typical set is given by Schubiger [4]and O'Connor and Arnold [5], which contains 300 emotional states. However, classifying such a largenumber of emotions is very difficult. Many researchers agree with the 'palette theory', which states that any emotion can be decomposed into primary emotions similar to the way that any color is a combination of some basic colors. Primary emotions are Anger, Disgust, Fear, Joy, Sadness, and Surprise. These emotions are the most obvious and distinct emotions in our life. They are called the archetypal emotions.

In this paper, a review of speech emotion recognition systems is presented. We survey three important aspects in speech emotion recognition: (1) important design criteria of emotional speech databases, (2) impact of speech features on the classification performance and (3) classification systems

employed in speech emotion recognition. The paper is divided into five sections. In Section 2, important issues in the design of an emotional speech database are discussed. Section 3 reviews in detail speech feature extraction methods. Classification techniques applied in speech emotion recognition are discussed in Section 4. Finally, important conclusions are drawn in Section 5.

2. EMOTIONAL SPEECH DATABASES

An important issue to be considered in the evaluation of an emotional speech recognizer is the degree of naturalness of the database used to assess its performance. Incorrect conclusions may be established if a low-quality database is used. Moreover, the design of the database is critically important to the classification task being considered. For example, the emotions being classified may be infant-directed; e.g. soothing and prohibition [6], or adult-directed; e.g. joy and anger[7]. In other databases, the classification task is to detect stress in speech. The classification task is also defined by the number and type of emotions included in the database.

2.1. Design Criteria

There should be some criteria that can beused to judge how well a certain emotional database simulates a real-world environment. According to some studies[7, 8], the following are the most relevant factors to be considered: Real-world emotions or acted ones? It is more realistic to use speech data that are collected from real life situations. Such recordings contain utterances with very natural conveyed emotions. Unfortunately, there may be some legal and moral issues that prohibit the use of them for research purposes. Alternatively, emotional sentences can be elicited in sound laboratories as in the majority of the existing databases. It has always been criticized that acted emotions are not the same as real ones.

Who utters the emotions? In most emotional speech databases, professional actors are invited to express pre-determined sentences with the required emotions. However, in some of them semi-professional actors are employed instead in order to avoid exaggeration in expressing emotions and to be closer to real world situations.

*How to simulate the utterances?*The recorded utterances in mostemotional speech databases are not produced in a conversationalcontext. Therefore, utterances may lack some naturalness since itis believed that most emotions are outcomes of our response todifferent situations. Generally, there are two approaches for elicitingemotional utterances. In the first approach, experienced speakers act asif they were in a specific emotional state, e.g. being glad, angry, or sad. Inmany developed corpora such experienced actors were notavailable and semi-professional or amateur actors were invited to utterthe emotional utterances. In a recent study, it was proposed to usecomputer games to induce natural emotional speech. Voice sampleswere elicited following game events whether the player won or lost thegame and were accompanied by either pleasant or unpleasant sounds.

Utterances are uniformly distributed over emotions? Some corpusdevelopers prefer that the numberof utterances for each emotion isalmost the same in order to properly evaluate the classificationaccuracy such as in the Berlin corpus. On the other hand, manyother researchers prefer that the distribution of the emotions in thedatabase reflects their frequency in the world. Forexample, the neutral emotion is the most frequent emotion inour daily life. Hence, the number of utterances with neutralemotion should be the largest in the emotional speech corpus.

Same statement with different emotions? In order to study theexplicit effect of emotions on the acoustic features of the speechutterances, it is common in many databases to record the same sentence with different emotions. One advantage of such adatabase is to ensure that the human judgment on the perceivedemotion is solely based on the emotional content of the sentenceand not on its lexical content.

3. FEATURES USED FOR SPEECH EMOTION

Recognition

An important issue in the design of a speech emotion recognition system is the extraction of suitable features that efficiently characterize different emotions. Since pattern recognition techniques are rarely independent of the problem domain, it is believed that a proper selection of features significantly affects the classification performance.

Four issues must be considered in feature extraction. The first issue is the region of analysis used for feature extraction. While some researchers follow the ordinary framework of dividing the speech signal into small intervals, called frames, from each which a local feature vector is extracted, other researchers prefer to extract global statics from the whole speech utterance. Another important question is what are the best feature types for this task are, e.g. pitch, energy, zero crossing, etc.? A third question is what is the effect of ordinary speech processing such as post-filtering and silence removal on the overall performance of the classifier? Finally, whether it suffices to use acoustic features for modeling emotions or is it necessary to combine them with other types of features such as linguistic, discourse information, or facial features.

4. SPEECH EMOTION RECOGNITION

In this section a survey of most recent works in Speech Emotion Recognition and their results are discussed. Hidden Markov Models (HMMs), Gaussian Mixture Models (GMMs), Artificial Neural Networks (ANNs), Support Vector Machines (SVMs) and Decision Trees are the most preferred methods used for Speech Emotion Recognition. Speech Emotion

Recognition based on Hidden Markov Model has been presented in [9]. Seven emotion classes namely anger, disgust, fear, surprise, joy, neutral and sad are considered. In the speech corpus all the phrases were collected in German and English Languages from five speakers totaling 5250 samples. Two methods have been used: in the first method a global framework of an utterance is classified by Gaussian mixture models using derived features of the raw pitch and energy contour of the speech signal, and the second method introduced increased temporal complexity applying continuous hidden Markov models considering several states using low-level instantaneous features instead of global statistics. Using the Gaussian mixtures models an overall recognition rate of 86.8% is achieved and with continuous hidden Markov models 77.8% overall recognition rate has been achieved. The recognition rate of human judges over the same test set is 81.3% on average.

A language-independent emotion recognition system based on Neural Network approach is presented in [10]. Six emotional classes were considered and seven subjects speaking four different languages are selected. The speech utterances were recorded in English, Chinese, Urdu and Indonesian. 580 speech utterances each delivered with one of six particular emotions were used for training and testing. From these samples 435 utterances were selected for training the network and the rest were used for testing. A total of 17 prosodic features are extracted by analyzing the speech spectrogram and out of these 12 features are selected using the combined Sequential Forward Selection/General Regression Neural Network and Consistency-Based Feature Selection as input to the Modular Neural Network (MNN). The Modular Neural Network using the 12 selected features achieved overall classification accuracy of 83.31% on the test set.

In [11] a hybrid scheme that combines the Probabilistic Neural Network and the Gaussian Mixture Model (GMM) is used. To handle mismatches more effectively, the Universal Background Model (UBM) is incorporated into the GMM. In the hybrid scheme the strengths of the PNN and the UBM-GMM are combined through a novel conditional-probability based fusion algorithm. The speech emotion database used is extracted from the Linguistic Data Consortium Emotional Prosody Speech Corpus. Eight emotions are tested in the experiment and the system is trained in speaker-independent mode. The hybrid scheme achieved a higher average accuracy of 80.75% when compared with the average accuracies of 72.50% and 58.88% achieved by PNN and UBM-GMM respectively.

To deal with emotionally ambiguous utterances, a classification scheme based on Emotional Profiles (EPs) is presented in [12]. EPs can interpret the emotion content of ambiguous utterances. This is an approach to interpret the emotional content of naturalistic human expression by providing multiple probabilistic class labels rather than a single hard label. The data set used in this study is the USC IEMOCAPwhich is an audio-visual database. Audio features extracted include prosodic and spectral envelope features. Support Vector Machines are used with Radial Basis Function as kernel with sigma of eight. This approach is able to attain an accuracy of 68.2%.

Detecting emotion from nonverbal features of speech and applying the method to assess the public speaking skills using Support Vector Machines is proposed in [13]. The database used is the Mind Reading Corpusconsisting of 2,927 acted everyday sentences, covering 442 different concepts of emotions each with 5 to 7 sentences. For the classifier a subset of nine categories representing a large variety of emotions is chosen. This subset consisted of 548 samples spoken by 10 different actors. OpenSMILE is used for feature extraction. Since a large feature set will be extracted from the speech there will be some irrelevant and redundant data that will not improve SVM prediction performance. Using the predefined openSMILE set emo_large with 6,552 features, the ten most relevant features are selected using Correlation-based Feature Selection algorithm. In the SVMs the kernel used is Radial Basis Function and Grid Search Algorithm is used to identify good (C, ϒ). Pair-wise classifiers are constructed for nine classes and achieved an average cross-validation accuracy of 89% for the pair-wise machines and 86% for the fused machine. When applied for assessing public speaking skills this approach achieved an average cross validation accuracy of 81% and a leave-one-speaker-out classification accuracy of 61%.

Emotion recognition based on multiple classifiers using Acoustic-Prosodic (AP) information and Semantic Labels (SLs) is proposed in [14]. Three base-level classifiers consisting of GMMs, SVMs, and MLPs are used for emotion detection based on AP features. A Meta Decision Tree (MDT) is then employed for the fusion of the three classifiers. For emotion recognition using semantic labels the maximum entropy model (MaxEnt) is used. Finally a weighted product fusion model is used to integrate the results from AP-based and SL-based approaches to output the recognized emotional state. For evaluation, 2033 utterances for four emotional states- Neutral, happy, angry, and sad- are collected. Emotion recognition performance based on MDT achieved 80%, which is better than each individual classifier, while average recognition accuracy of 80.92% is obtained for SL-based recognition. Combining acoustic-prosodic information and semantic labels achieved 83.55%, which is superior to either AP-based or SL-based approaches.

Speech based emotion detection using only acoustic data without using any linguistic or semantic information in general suffers from the fact that acoustic data is speaker dependent, and can result in inefficient estimation of the statistics modeled byclassifierssuch as HMMs and GMMs. This can be overcome bythe use of speaker-specific feature

warping [15] as ameans of normalizing acoustic features. Feature warping is a technique that maps each feature to a predetermined distribution.HMM based classifier and a short four-dimensional feature vector composed of pitch, energy, zero crossing rate and energy slope is used in this approach. For the experiments Linguistic Data Consortium's Emotional Prosody Speech Corpus is used. The mean emotion classification accuracy for neutral verses anger without warping is 91.6% and with warping is 95.2%. Performance comparison of various classifiers discussed in this section is given in Table 1.

Table 1: Performance comparison of various classifiers

S. No	Classifier	No. of Emo-tions	Speech Database	Average Recog-nition Accuracy
1	Hidden Markov Model (HMM)	7	Custom Database	77.8%
2	Modular Neural Network (MNN)	6	Custom Database	83.31
3	Hybrid Model PNN & GMM	8	Linguistic Data Consortium's Emotional Prosody Speech Corpus	80.75%
4	Support Vector Machine (SVM)	-	USC IEMOCAP	68.2%
5	Support Vector Machine (SVM)	9	Mind Reading Corpus	86%
6	Hybrid Model GMM,SVM,MLP Fused using Meta Decision Tree (MDT)	4	Custom Database	83.55%
7	Hidden Markov Model (HMM)	2	Linguistic Data Consortium's Emotional Prosody Speech Corpus	91.6%

5. APPLICATIONS

Speech emotion recognition is particularly useful for applications which require natural man–machine interaction such as web movies and computer tutorial applications where the response of those systems to the user depends on the detected emotion[16]. It is also useful for in-car board system where information of the mental state of the driver may be provided to the system to initiate his/her safety[16]. It can be also employed as a diagnostic tool for therapists[17]. It may be also useful in automatic translation systems in which the emotional state of the speaker plays an important role in communication between parties. In aircraft cockpits, it has been found that speech recognition systems trained to stressed-speech achieve better performance than those trained by normal speech[18]. Speech emotion recognition has also been used in call center applications and mobile communication. The main objective of employing speech emotion recognition is to adapt the system response upon detecting frustration or annoyance in the speaker's voice[19].

6. CONCLUSION

In this paper, a survey of current research work in speech emotion recognition system has been given. Three important issues have been studied: the features used to characterize different emotions, the classification techniques used in previous research, and the important design criteria of emotional speech databases. There are several conclusions that can be drawn from this study.

The first one is that while high classification accuracies have been obtained for classification between high-arousal and low-arousal emotions, N-way classification is still challenging. Moreover, the performance of current stress detectors still needs significant improvement. The average classification accuracy of speaker-independent speech emotion recognition systems is less than 80% in most of the proposed techniques. In some cases,it is as low as 50%. For speaker-dependent classification, the recognition accuracy exceeded 90% only in few studies. Many classifiers have been tried for speech emotion recognition such as the HMM, the GMM, the ANN, and the SVM. However, it is hard to decide which classifier performs best for thistask because differentemotional corpora with different experimental setups were applied.

Most of the current body of research focuses on studying many speech features and their relations to the emotional content of the speech utterance. New features have also been developed such as the TEO-based features. There are also attempts to employ different feature selection techniques in order to find the best features for this task. However, the conclusions obtained from different studies are not consistent. The main reason may be attributed to the fact that only one emotional speech database is investigated in each study. Most of the existing databases are not perfect for evaluating the performance of a speech emotion recognizer. In many databases, it is difficult even for human subjects to determine the emotion of some recorded utterances; e.g. the human recognition accuracy was 67% for DED[20], 80% for Berlin [21], and 65% in [22]. There are some other problems for some databases such as the low quality of the recorded utterances, the small number of available utterances, and the unavailability of phonetic transcriptions. Therefore, it is likely that some of the conclusions established in some studies cannot be generalized to other databases. To address this problem, more cooperation across research institutes in developing bench mark emotional speech databases is necessary.

In order to improve the performance of current speech emotion recognition systems, the following possible extensions are proposed. The first extension relies on the fact

that speaker-dependent classification is generally easier than speaker-independent classification. At the same time, there exist speaker identification techniques with high recognition performance such as the GMM-based text-independent speaker identification system proposed by Reynolds. Thus, a speaker-independent emotion recognition system may be implemented as a combination of a speaker identification system followed by a speaker-dependent emotion recognition system. It is also noted that the majority of the existing classification techniques do not model the temporal structure of the training data. The only exception may be the HMM in which time dependency may be modeled using its states. However, all the Baum–Welch re-estimation formulae are based on the assumption that all the feature vectors are statistically independent. This assumption is invalid in practice. It is sought that direct modeling of the dependency between feature vectors, e.g. through the use of autoregressive models, may provide an improvement in the classification performance.

REFERENCES

[1] J. Nicholson, K. Takahashi, R. Nakatsu, "Emotion recognition in speech using neural networks", *Neural Compuing & Applications*, 2000, Vol. 9, No. 4, pp. 290-296.

[2] R. Banse, K. Scherer, "Acoustic profiles in vocal emotion expression", *Journal of Personality and Social Psychology*, 1996, Vol.70, No. 3, pp.614-636.

[3] R. Fernandez, "*A computational model for the automatic recognition of affect in speech*", Ph.D. Thesis, Massachusetts Institute of Technology, February 2004.

[4] M. Schubiger, "*English intonation: its form and function*", Tübingen, M. Niemeyer Verlag, Germany, 1958.

[5] J. O'Connor, G. Arnold, "*Intonation of Colloquial English*", second ed., Longman, London, UK, 1973.

[6] C. Breazeal, L. Aryananda, "*Recognition of affective communicative intent in robot-directed speech*", Autonomous Robots, 2002, Vol. 12, No. 1, pp. 83-104.

[7] W. Campbell, "Databases of emotional speech",in *Proceedings of the International Speech Communication and Association (ISCA) ITRW on Speech and Emotion*, 2000, pp. 34-38.

[8] C. Lee, S. Narayanan, "Toward detecting emotions in spoken dialogs", *IEEE Transactions on Speech and Audio Processing*,2005, Vol. 13, No. 2,pp. 293-303.

[9] Björn Schuller, Gerhard Rigoll, and Manfred Lang, "Hidden Markov Model-Based Speech Emotion Recognition", *International Conference on Acoustics, Speech and Signal Processing(ICASSP),* 2003, IEEE.

[10] Muhammed Waqas Bhatti, Yongjin Wang and Ling Guan, "A Neural Network Approach for Human Emotion Recognition in Speech", *International Symposium on Circuits and Systems (ISCAS),* 2004, IEEE.

[11] Wee Ser, Ling Cen and Zhu Liang Yu, "A Hybrid PNN-GMM Classification Scheme for Speech Emotion Recognition", in *Proceedings of International Conference on Pattern Recognition*,IEEE,2008, pp. 1-4

[12] Emily Mower, Maja J Matarić and Shrikanth Narayanan, "A Framework for Automatic Human Emotion Classification Using Emotion Profiles",*IEEE Transactions on Audio, Speech, and Language Processing*, July 2011, vol. 19, No. 5, pp. 1057-1070.

[13] Tomas Pfister and Peter Robinson,"Real-Time Recognition of Affective States from Nonverbal Features of Speech and Its Application for Public Speaking Skill Analysis",*IEEE Transactions Affective Computing*, April-June 2011, vol. 2, No. 2, pp. 66-78.

[14] Chung-Hsien Wu, Wei-Bin Liang, "Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels", *IEEE Transactions on Affective Computing*, Jan-Mar 2011, Vol. 2, No. 1, pp. 10-21.

[15] Vidhyasaharan Sethu, Eliathamby Ambikairajah and Julien Epps, "Speaker Normalisation for Speech-Based Emotion Detection", in *Proceedings of the 2007 15th International Conferences on Digital Signal Processing*, pp. 611-614.

[16] B. Schuller, G. Rigoll, M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture", in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing, 2004*, Vol. 1, pp. 577-580.

[17] D.J. France, R.G. Shiavi, S. Silverman, M. Silverman, M. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk", *IEEE Transactions on Biomedical Engineering*, 2000, Vol. 47, No. 7, pp. 829-837.

[18] J. Hansen, D. Cairns, Icarus,"source generator based real-time recognition of speech in noisy stressful and Lombard effect environments", *Speech Commun*., 1995, Vol. 16, No. 4, pp. 391-422.

[19] J. Ma, H. Jin, L. Yang, J. Tsai, "Ubiquitous Intelligence and Computing", Third International Conference, UIC 2006, Vol. 4159.

[20] I.Engberg,A.Hansen,"DocumentationoftheDanishemotionalspeech database",*International AAU Report*, Center for Person Kommunikation, Denmark, 1996.

[21] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, B. Weiss, "A database of German emotional speech", in *Proceedings of the Interspeech 2005*, 2005, pp. 1517-1520.

[22] T. Nwe, S. Foo, L. De Silva, "Speech emotion recognition using hidden Markov models", *Speech Commun*., 2003, Vol. 41, pp. 603-623