International Journal Online of Sports Technology& Human Engineering.
ISSN: 2349-0772| Volume III Issue II April 2016| www.ijosthe.com

A Review on Speech Emotion Recognition

Nomaan Khan
M.Tech Scholar
TIT, Bhopal
nomaankhan034@gmail.com

Abstract

Emotion recognition from Audio signal Recognition is a recent research topic in the Human Computer Interaction. The demand has risen for increasing communication interface between humans and digital media. Many researchers are working in order to improve their accuracy. But still there is lack of complete system which can recognize emotions from speech. In order to make the human and digital machine interaction more natural, the computer should able to recognize emotional states in the same way as human. The efficiency of emotion recognition system depends on type of features extracted and classifier used for detection of emotions. There are some fundamental emotions such as: Happy, Angry, Sad, Depressed, Bored, Anxiety, Fear and Nervous. These signals were preprocessed and analyzed using various techniques. In feature extraction various parameters used to form a feature vector are: fundamental frequency, pitch contour, formants, duration (pause length ratio) etc. These features are further classified into different emotions. This research work is the study of speech emotion classification addressing three important aspects of the design of a speech emotion recognition system. The first one is the choice of suitable features for speech representation. The second issue is the design of an appropriate classification scheme and the third issue is the proper preparation of an emotional speech database for evaluating system performance.

Keywords – Speech corpus, Speech features, Emotion recognition, Classifiers.

1. Introduction

The speech signal is the fastest and the most natural method of communication between humans. This fact has motivated researchers to think of speech as a fast and efficient method of interaction between human and machine. However, this requires that the machines should have the sufficient intelligence to recognize human voices. Since the late fifties, there has been tremendous research on speech recognition, which refers to the process of converting the human speech into a sequence of words [1].

However, despite the great progress made in speech recognition, we are still far from having a natural interaction between man and machine because the machine does not understand the emotional state of the speaker. This has introduced a relatively recent research field, namely speech emotion recognition, which is defined as extracting the emotional state of a speaker from his or her speech. It is believed that speech emotion recognition can be used to extract useful semantics from speech, and hence, improves the performance of speech recognition systems.

Emotion recognition from speech is a challenging problem in audio signal processing. Lot of information like: age, gender, emotion, person, and action can be estimated from a speech signal, emotion recognition is one of them. Emotion depends on voice generated from different parts of human vocal system. These systems can be helpful in detecting customers' emotion, medical entertainment, crime detection, robotics voice and may other cases. Speech communication contains paralinguistic information of the speaker. Although enormous efforts are invested in recognizing the emotions from speech but still much research is needed.

Existing human-machine interaction systems can identify "what is said" and "who said it" using speaker identification and speech recognition techniques. These machines can evaluate "how it is said" to respond more correctly and make the interaction more natural, if provided with emotion recognition techniques. The emotion recognition using speech signals have wide applications.

Applications of speech emotion detection are:

- Human-computer intelligent interaction (HCII) for make machines more user friendly, Project can be implemented as a Lie Detector, Designing intelligent Robotics, Develop learning environments and consumer relations, Entertainment, etc.
- Emotion recognition is useful for applications such as Entertainment, e- Learning, and

diagnostic tool for therapists, call centre applications etc.

Usually in emotion classification, researchers consider the acoustic features alone. Though features like pitch, energy and speaking rate change with emotional state, strong emotions such as anger and surprise have high pitch and energy. In that case, it is very difficult to distinguish the emotions such as anger and surprise using acoustic features alone. But, if we classify speech solely on its textual component, we will not obtain a clear picture of the emotional content [2].

2. Primary Emotions in Speech

Anger: Anger requires high energy to be expressed. Definition meaning of the anger is simple extreme displeasure. In case of anger, aggression increases in which control parameter weakens. Anger is stated to have the highest energy and pitch level when compared with the emotions disgust, fear, joy and sadness. The widest observed pitch range and highest observed rate of pitch change are other findings about the emotion label anger when compared with other emotions.

Fear: In emotional dimension, fear has similar features to anger. High pitch level and raised intensity level are correlated with fear. It is stated that fear has a wide pitch range. Highest speech rate is observed in fear speeches. The pitch contour trend separates fear from joy. Although the pitch contour of fear resembles the sadness having an almost downwards slope, emotion of joy have a rising slope.

Sadness: In emotional dimension, sadness requires very low energy. In addition, valence degree is negative. Sadness exhibits a pattern that is normal or lower than normal average pitch, a narrow pitch range and slow tempo. Speech rate of a sad person is lower than the neutral one.

Joy/Happiness: Joy/happiness exhibit a pattern with a high activation energy, and positive valence. Strength of the happiness emotion may vary. In the emotional state happiness or joy, pitch mean, range and variance increases.

Disgust: Disgust is stated the lowest observed speech rate and increased pause length.

Boredom: Boredom is a negative emotion with negative valence and low activation level same as sad. A lowered mean pitch and a narrow pitch range with a slow speech rate are defined as the properties of a bored expression.

3. Emotion Recognition from Speech

Speech is a complex signal containing information about message, speaker, language, emotion and so on. Most existing speech systems process studio recorded, neutral speech effectively, however, their performance is poor in the case of emotional speech. This is due to the difficulty in modeling and characterization of emotions present in speech. Presence of emotions makes speech more natural. In a conversation, non-verbal communication carries an important information like intention of the speaker. In addition to the message conveyed through text, the manner in which the words are spoken, conveys essential non-linguistic information. The same textual message would be conveyed with different semantics (meaning) by incorporating appropriate emotions. Spoken text may have several interpretations, depending on how it is said. For example, the word ‘OKAY’ in English, is used to express admiration, disbelief, consent, disinterest or an assertion. Therefore understanding the text alone is not sufficient to interpret the semantics of a spoken utterance. However, it is important that, speech systems should be able to process the non-linguistic information such as emotions, along with the message. Humans understand the intended message by perceiving the underlying emotions in addition to phonetic information by using multi-modal cues.

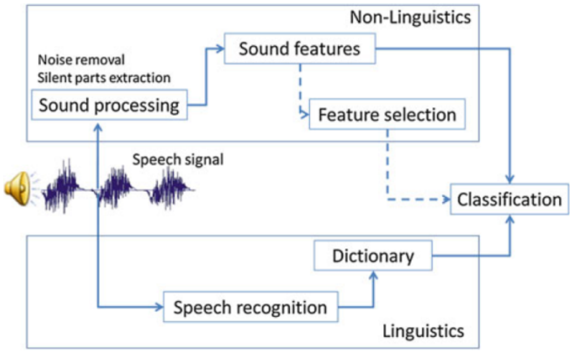

Figure 1: Speech Emotion Recognition System

The purpose of this paper is to present a comprehensive survey of emotion recognition systems from speech in order to provide pattern recognition and speech processing researchers with basic information, theoretical background, materials and methods and current trends of this field. In this survey three important issues of speech emotion recognition are presented:

- Available emotion databases and their usability in speech emotion recognition.
- Various feature selection methods on previously extracted sound features and their specific

contribution in speech emotion recognition performance.

- Numerous classifiers that have been used in speech emotion recognition portraying their classification rate as reported in the literature.

4. Feature Vector in Speech Recognition

Vector features are categorized as short-time (segmental) or long-time (suprasegmental) according to their temporal structure. Segmental features are calculated once for every small time frame (usually 25–50 msec using windowing techniques), allowing the analysis of their temporal evolution. In contrast, suprasegmental features are calculated over the entire utterance duration (as seen in Fig. 2). A quantitative feature-type-wise comparison between short time and suprasegmental analysis is carried out for the recognition of interest in human conversational speech in [4].

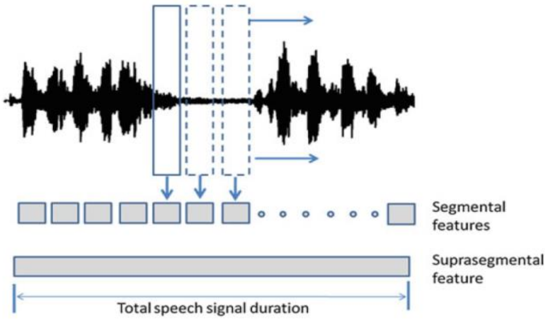

Figure 2: Segmental and suprasegmental features in a speech signal

Speech features and their description are discussed in below table:

Table 1: Speech Features and their Description

Features	Description
Mel-Frequency Cepstral Coefficients (MFCCs)	Derive from spectrum, which is the inverse spectral transform of the logarithm of the spectrum
Linear Prediction Cepstral Coefficients (LPCCs)	Derive from Spectrum
Noise-to-harmonic ratio, jitter, shimmer, amplitude quotient, spectral tilt, spectral balance	Measurements of Signal (voice) quality
Energy, short energy	Are measurements of intensity
Pitch	Are measurements of frequency
Duration, Time Stamps	Are measurements of time

In an early research [5], author suggested that apart from prosodic features in the speech signal, syntactic and behavioral hints like repetitions in a dialogue and part of speech features, should be taken into consideration. The research revealed that nouns and adjectives are more useful in emotion categorization arguing the point that generally, content words are more salient and more prone to be emotionally marked than function words (i.e. verbs).

In [6], a combination of three sources (acoustic, lexical and discourse) was used for emotion recognition. To capture emotion information at language level, an information-theoretic notion of emotional salience in two hyper-classes (negative and non-negative emotion) was introduced. The salience of a word in emotion recognition can be defined as mutual information between a specific word and emotion category.

Similarly,in [7], author examined the utility of speech and lexical features for predicting student emotions in computer-human tutoring dialogues. Emotion annotation was performed for negative, neutral, positive and mixed emotions. Prosodic features are then extracted from the speech signal and lexical items (words) from recognized speech.

Moreover, in a very recent paper [8], author fuses the results of semantic label classification with acoustic-prosodic information to boost emotion recognition in affective speech.

There is also a special sub-category in non-linguistic information that relates to human vocalizations (often referred to as non-linguistic vocalizations). Laughs, cries, sighs, yawns and other similar vocal outbursts seem at first to be good examples of expressions of discrete (although not necessarily basic) emotions. A funny joke elicits amusement, which produces a laugh; a loss elicits sadness, which produces crying; an uninspired lecture elicits boredom, which produces a yawn.

5. Feature selection in Speech Recognition

Prior to classification, feature selection, also known as variable selection or feature reduction, is often used in speech emotion recognition in order to speed up the learning process and minimize the problem known as “the curse of dimensionality”. Popular feature selection methods that have been implemented include Principal Component Analysis (PCA) [9] and Canonical Correlation Analysis (CCA)

[10]. Moreover, Correlation-based Sub Set Evaluators have also been used for feature selection, where several search methods evaluate a subset of features for the optimal subset.
Such search methods include BestFirst, correlation-based analysis, Genetic Algorithms, Support Vector Machine-Sequential forward floating search Mutual Information (MI) between the class Y and an attribute X.

6. Classification for Speech Recognition

Usually, classification evaluations are carried out using a single database or dataset. In this case, several testing frameworks appear based on the dependency or not on the speaker (speaker dependent/independent) as well as the context (context dependent or independent).

However, every speech database is created on the basis of fixed recording conditions and noise levels and specific room acoustics, while the data is recorded only in one language.

For emotional state modeling, a variety of pattern recognition methods are utilized to construct a classifier, such as Gaussian Mixture Models (GMMs), Hidden Markov Models (HMMs), Support Vector Machines (SVMs), Artificial Neural Networks (ANNs), Decision Trees or k-Nearest Neighbor distance classifiers (kNNs).

Several classifiers may not perform well on all emotional states. For example, a GMM-based classifier may fail to correctly recognize the neutral emotion, while theMLP-based classifier is clearly superior in neutral emotion recognition. Therefore, hybrid classifiers and ensembles were proposed in order to achieve higher recognition performance than individual classifiers.

Some of the classifiers are discussed in table 2.

Table 2: Overview of Classification Performance

Author	Technique	Result
Wu et al. [11]	SVM	88.6% in Berlin EMO database
Wu and Liang [12]	GMMs	72.61%, 4 emotions in 2 unknown datasets
Yun and Yoo [13]	HMMs	89% in Berlin EMO database
Rong J [14]	RF	80.6%, 3 emotions in 2 unknown dataset
Pao et al. [15]	k-NN	72.2%, 5 emotions (unknown dataset)
Wu and Liang [12]	SVM, GMM, MLP	83.55%(speaker independent with linguistic information), 4 emotions in 2 unknown datasets
Prajakta et al. [16]	RBF kernel function based SVM	84%, Polish database
Akash Shaw et al. [17]	ANN	86.87% accuracy
Zhiyan Han Jian Wang [18]	PSVM	86.75% accuracy on 4 emotions
Nam Kyun Kim et al.[19]	MTL-CNN	F1-score improvement of 3.64% for a task on a Berlin database
Nattapong Kurpukdee et al.[20]	ConvLSTM-RNN	65.13% on IEMOCAP database

7. Emotional Datasets/Databases

Surveying the literature, it becomes evident that emotion recognition in speech is mostly assessed using digital sources that are datasets rather than databases. Datasets are smallscale collections of material created to focus on a specific research and most importantly they are not widely available. Generally, it is extremely difficult to produce a database representing the natural speech of a man or a woman in completely natural conversation. Many examples of humans talking exist, but very few of them illustrate speech in a natural environment. In the latter case, some databases use corpora (i.e. large collections) of spontaneous speech, usually consisting of clips from live television, radio programs or call centers, with natural speech recorded in real-world situations. On the other hand, such databases are not distributed easily, since their assessment and processing could raise serious ethical or copyright issues.

Table 3: Databases of Emotional Speech

Databases	**Links**
Berlin database of emotional speech	http://pascal.kgw.tu-berlin.de/emodb/index-1280.html
KISMIT database	http://www.ai.mit.edu/projects/sociable/expressive-speech.html
Bavarian archive for speech	http://www.bas.uni-muenchen.de/Bas/
HUMAINE database	http://universal.elra.info/product_info.php?cPath=25\&products_id=2063
SmartKom	http://www.smartkom.org
Socrates emotional speech database	http://www.wcl.ece.upatras.gr/ai/resources/demo-emotion-recognition-from-speech
DSPLAB emotional speech	http://wwwbox.uni-mb.si/eSpeech/speech.html
BELFAST naturalistic emotion database	http://www.idiap.ch/mmm/cor\discretionary-pora/emotion-corpus
Kids audio speech corpus	http://techexplorer.cusys.edu/show_NCSum.cfm?NCS=258629
LDC Emotional prosody speech-transcripts	http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002S28
eNTERFACE	http://www.enterface.net/
CASIA Mandarin emotional corpus	http://www.chineseldc.org

8. Conclusion

In this study, the overview of different speech emotion recognition methods are discussed for extracting audio features from speech sample, various classifier algorithms are explained briefly. Speech Emotion Recognition has a promising future and its accuracy depends upon the combination of features extracted, type of classification algorithm used and the correct of emotional speech database. This study aims to provide a simple guide in the field of speech emotion recognition systems.

Research papers that investigate emotion recognition from audio channels were surveyed and classified mostly based on: (i) the features extracted and selected for training the classifiers (linguistic or non-linguistic) and (ii) their classification methodology. It should be emphasized that there is a lack of uniformity in the way methods are evaluated and, therefore, it is inappropriate to make direct comparisons and to explicitly declare which methods demonstrate the highest performance. Indeed, one of the main conclusions of this survey is to highlight that the evaluation of the proposed methods is often not performed in common test sets consequently a common reference point for algorithmic assessment cannot be achieved.

REFERENCES

[1] Nicholson J., Takahashi K., Nakatsu R., "Emotion Recognition in Speech using Neural Networks", IEEE Trans. Neural Information Proc., Vol. 2, pp. 495-501, 2009.

[2] Shashidhar G. Koolagudi, Ramu reddy,Jainath Yadav , K.Sreenivasa Rao. "IITKGP-SEHSC:Hindi speech corpus for emotion analysis." IEEE 2011.

[3] Wankhade, Sujata B., and YashpalsingChavhan PritishTijare. "Speech Emotion Recognition System Using SVM AND LIBSVM." International Journal of Computer Science And Applications, vol. 4, no. 2, 2011.

[4] Schuller B, Rigoll , “Recognising interest in conversational speech–comparing bag of frames and supra-segmental features”In Proceedings of INTERSPEECH, pp 1999–2002, 1999.

[5] Batliner A, Fischer K, Huber R, Spilker J, Nolth E, “How to find trouble in communication”, Speech Commun 40:117–143, 2003.

[6] Lee CM, Narayanan SS, “Toward detecting emotions in spoken dialogs”, IEEE Trans Speech Audio Process 13:293–303, 2005.

[7] Litman DJ, Forbes-Riley K, “Predicting student emotions in computer-human tutoring dialogues”, In: Proceedings of 42nd annual meeting on association for computational linguistics, 2004.

[8] Wu CH, LiangWB, “Emotion recognition of affective speech based on multiple classifiers using acoustic- prosodic information and semantic labels”, IEEE Trans Affect Comput 2:10–21, 2011.

[9] Wang S, Ling X, Zhang F, Tong J, “Speech emotion recognition based on principal component analysis and back propagation neural network”, In: Proceedings of international conference on measuring technology and mechatronics automation, pp 437–440, 2010.

[10] Cheng XM, Cheng PY, Zhao L, “ A study on emotional feature analysis and recognition in speech signal.n”, In: Proceedings of international conference on measuring technology and mechatronics automation, pp 418–420, 2009.

[11] Wu S, Falk TH, ChanWY, “Automatic recognition of speech emotion using long-term spectro-temporal features”, In: Proceedings of 16th international conference on digital signal processing, 2009.

[12] Wu CH, LiangWB, “Emotion recognition of affective speech based on multiple classifiers using acoustic prosodic information and semantic labels”, IEEE Trans Affect Comput 2:10–21, 2011.

[13] Yun S,Yoo CD, “Speech emotion recognition via amax-margin framework incorporating a loss function based on the Watson and Tellegen’s emotion model. In: Proceedings IEEE international conference on acoustics, speech and signal processing, pp 4169–4172, 2009.

[14] Rong J, Chen YPP, Chowdhury M, Li G, “Acoustic features extraction for emotion recognition”, In: Proceedings 6th IEEE/ACIS international conference on computer and information science, pp 419–424, 2007.

[15] Pao TL, Liao WY, Chen YT, Yeh JH, Cheng YM, Chien CS, “Comparison of several classifiers for emotion recognition from noisy mandarin speech. In: Proceedings of 3rd international conference on international information hiding and multimedia signal processing, pp 23–26, 2007.

[16] Prajakta P. Dahake, Kailash Shaw, P. Malathi, “Speaker Dependent Speech Emotion Recognition using MFCC and Support Vector Machine”, International Conference on Automatic Control and Dynamic Optimization Techniques, IEEE, pp. 1080-1084, 2016.

[17] Akash Shaw, Rohan Kumar Vardhan, Siddharth Saxena, “Emotion Recognition and Classification in Speech using Artificial Neural Networks”, International Journal of Computer Applications vol 145 – No.8, pp. 5-9, 2016.

[18] Zhiyan Han, Jian Wang, “Speech emotion recognition based on Gaussian kernel nonlinear proximal support vector machine”, IEEE, 2017.

[19] Nam Kyun Kim, Jiwon Lee, Hun Kyu Ha, Geon Woo Lee, Jung Hyuk Lee, Hong Kook Kim, “Speech emotion recognition based on multi-task learning using a convolutional neural network”, IEEE, 2017.

[20] Nattapong Kurpukdee, Tomoki Koriyama , Takao Kobayashi, Sawit Kasuriya, Chai Wutiwiwatchai, Poonlap Lamsrichan, “Speech emotion recognition using convolutional long short-term memory neural network and support vector machines”, IEEE, 2017.