Jurnal
Teknologi

Full Paper

A REVIEW ON SPEECH EMOTION FEATURES

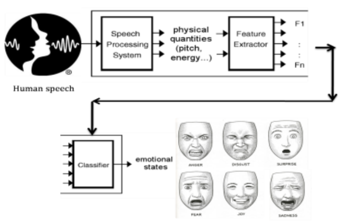
Noor Aina Zaidan*, Md Sah Hj. Salam*

Department of Computer Science, Faculty of Computing, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia

Article history
Received
3 December 2013
Received in revised form
2 July 2014
Accepted
25 November 2014

*Corresponding author
sah@utm.my

Graphical abstract



Abstract

Research works on combining emotions in intelligent machines are expanding and improving. Human's speeches basically have various emotional states. The finding of reliable speech features is an ongoing research. Specific features in the speech signal that contribute to emotional information are uncertain, extremely challenging problem and continue being explored. The recognition rate of emotion in speech signal is inconsistent depending on the features used in the experiment and also the database itself. Prosodic, spectral and wavelet features are mostly being used to determine which of these features or its hybrid carry more information about emotions. This paper intends to summarize previous work and make reviews about single and hybrid features based on prosodic, spectral and wavelet feature.

Keywords: Emotion, features, prosodic, wavelet, spectral, hybrid

1.0 INTRODUCTION

In human communication, information conveyed through speech usually comes along with the emotion of the speaker. Effective interaction is not just focused on what the contents of the information delivered but also how the information is presented to accomplish the natural communication [1]. Human's speech basically has various emotional states such as happiness, sadness, fear, anger, boredom, disgust, surprise and neutral. Emotions are very important in producing effective communication. Improper emotions delivered may lead to wrong perception and thus give ineffective interaction. With the advancement of today's technology, computer application can be programmed to understand human emotion in the better way of HCI.

There are a lot of applications used dealing with human emotion such as automatic speech services, lie detector, intelligence toys, psychiatric diagnosis, call center applications and car board system for driver safety. For example, in a call center application, the system can detect customer's dissatisfaction according to the conversation that occur during that time while in medication, the emotion of the patient need to be examined to detect mental illness [2].

The specific features in the speech signal that contribute to emotional information are uncertain, extremely challenging problem and continue to be sought [3, 4, 5].The extraction process of these features is very important to ensure the result of emotion recognition system is more precise and efficient [4, 6].

Studies on emotion recognition have been carried out to identify the presence of several emotions in speech. However, many studies were done using a single language database. Therefore, although in demanding studied field, language dependency in some emotion recognition system makes it less accessible for most of the languages. Ali et.al in [6] found that database resource in multiple languages is lacking in emotion recognition area. Most systems are limited to only few standard languages. Emotional expression is said to depend on the speaker, culture and the environment. Culture and environment influence the speaking style as the language used may be different.

Emotion expresses the psychological state of humanity. The emotional speech recognition system is related to the recognition of human emotion conveyed in speech. Speech production is substantially affected by the different emotional states of the speaker [21]. The presence of different emotions can be determined by several parameters in speech signal features.

Ayadi *et al*. [9] did a survey of speech emotion classification dealing with three key aspects of speech emotion recognition system design which are the selection of appropriate features for represents speech, classification scheme design and preparation of emotional speech database. They claimed that data that are collected from real life situations is more realistic to use. Also stated that most researchers admit that global features (all features extracted from an utterance except prosodic features) are spectacular in term of classification accuracy and time consume.

This review paper is organized into 7 sections. Section II contains the explanation about feature extraction and the features category. Section III describes about the integrated features. Section IV discusses about the classification method used in previous studies. The using of various emotional speech databases is described in section V. Summary of prior works is presented in section VI and finally the conclusion in section VII.

2.0 FEATURE EXTRACTION

Speech features, containing the emotional information that mostly used for speech emotion recognition are spectral features and prosodic features. Ingale in [4] state that speech energy, formant frequency, fundamental frequency and Mel frequency cepstrum were used by several researchers because of its effectiveness in differentiates certain states of emotion. To excite the different emotions, prosodic features such as speech intensity, glottal parameters, fundamental frequency, pitch and loudness were used [8].

Based on the results of previous studies, spectral and prosodic are two features category that carry most emotion information. From a survey done by Ayadi [9], prosody continuous features such energy and pitch, contain most emotional information of an utterance. Moreover, the combination of spectral and prosodic features is also said to be able to improve the performance of emotion recognition systems since both of them contain emotion information [8]. Apart from that, wavelet feature is also believed to be more efficient to use compared to spectral features in recognizing emotion. Since wavelet features could give better results, the idea of combining it into spectral and prosodic features is expected to improve the recognition rate.

Based on previous studies, there are many speech features that can be used for recognition. Studies in the field of speech recognition, Automatic Speech Recognition (ASR) has been widely carried out, but the optimal speech features are still being sought. To improve the performance of emotion recognition systems, some features from speech signal have been used: prosodic (energy, pitch frequency-related, formant frequency-related), and spectral features (MFCC) [11]. Besides, a wavelet features, Subband based coefficient was used by some researcher to carry out the study.

2.1 Prosodic Features

A prosodic feature is a vocal aspect of speech such as pitch, energy, speaking rate, fundamental frequency and it has been used in existing emotion recognition systems [3]. Most of the researchers believe that prosodic features carry the emotion information. Intonation is one of prosodic features that can distinguish human emotion because of its variation. Zhou *et al*. [8] said that prosodic features can describe the emotion effectively.

Other research about multilingual anger recognition carried out by Polzehl *et al.* [20] to analyze different combinations of prosodic and acoustic features to recognize the occurrence of anger emotional state in noisy environment. The databases used in the study are real-life American English and German Interactive Voice Response systems including the background sound, cross- and off- talk, and speak in conventional style. The features used are pitch, formants frequency, intensity, loudness and MFCC while the classification method is SVM with kernel function. They were successfully developing a single bi-lingual anger recognition system and it performs just like two separate mono-lingual systems on the test data with 75.6% f1 on multi-lingual train set and 74.5% f1 on multi-lingual test set.

2.2 Spectral Features

Previous researcher found that information about emotion and expressivity was present in spectral features. The spectral features that mostly used for emotion recognition are MFCC [5]. Zhou *et al.* [8] also state that spectral features like LPC and MFCC have been widely used in speech emotion recognition. From research done by Bozkurt [12], they proved that MFCC features did a better job in recognizing emotion compare to prosody features because it carries lots of spectral information. They found that dynamic features could enhance the emotion recognition rate.

Besides, Bitouk *et al.* [5] did a research to investigate features performance in speaker-independent task and introduced a more fine and robust set of spectral features. They used two publicly available databases: English emotional speech database from the Linguistic Data Consortium (LCD) and a Berlin database of German emotional speech. They found that classification accuracies are consistently higher for spectral features compared to prosodic or utterance-level spectral features. The result of the experiment shows that LDC gave 23.1% and Berlin gave 68.1% of classification rate using prosodic features while using spectral consonant, LDC gave 40.2% and Berlin gave 72.50% of classification rate.

2.3 Wavelet Features

Moreover, besides spectral and prosodic features, wavelet feature is considerably to give better recognition rate. Wavelet feature is believed to be more efficient to use compared to spectral features in terms of recognizing emotion. Wavelet coefficient is used to capture important time and frequency features.

Kishore *et al.* in [10] said that environment in real time speech recognition is diverse from recognition system that using a speech database. So, emotional database in noisy environments is necessary for the study. They do the comparison using spectral MFCC and wavelet features, Subband Based Cepstral (SBC) using Surrey Audio-Visual Expressed Emotion (SAVEE) database to recognize six emotions: anger, disgust, fear, happy, neutral and sad. According to the result, SBC parameters are less sensitive towards the noisy data and produce a good performance and better recognition accuracy than MFCC. The classifier used for this study is GMM. To estimate the GMM parameters like mean, covariance matrix of each component and probability, they used Expectation Maximization Algorithm (EA). The result shows SBS method gave 70% of recognition while MFCC gave 51% of recognition.

A comparison about sensitivity towards noisy data had been done by Kishore [10] using spectral MFCC and wavelet features, Subband Based Cepstral (SBC). According to the result, SBC parameters are less sensitive towards the noisy data and produce a good performance and better recognition accuracy than MFCC. Speech features selection really gives a huge contribution to the emotion recognition system performance. There a several study about integrating speech features to get a more accurate result. Integrating features is a task of combining two or more features together to produce hybrid features. Han *et al.* [19] and Bitouk *et al.* [5] claimed that the use of integrating features instead of single features is much more robust and leads to further improvement.

3.0 HYBRID FEATURES

Currently, research in emotion recognition field become progressive to attain the best performance of emotion recognition system. Seehapoch *et al.* in [2] had done a research to investigate the integrated features to get the highest accuracy performance of emotion recognition system using SVM. They used both prosodic and spectral features to classify the speech emotion from three language databases which are German (Berlin Database of Emotional Speech), Japan (Japanese Emotional Speech Database) and Thai (Audiovisual Thai Emotion). The result of the study shows that the combination of features fundamental frequency (F0), energy and MFCC gave higher average accuracy with German – 89.80%, Japan – 93.57% and Thai – 98.0%. They concluded that Japan and Thai gave higher recognition because they are not variant tone language like German. Besides, German database gave difficulties in recognizing emotion because the database was recorded by sentences.

There is also a research about a speech emotion recognition system based on integrating feature and improved HMM done by Han *et al.* [19]. The researchers claimed most existing emotion recognition systems use only one type of feature causes the system may not be robust. So they came out with the idea to integrate prosodic features: fundamental frequency, energy, durational aspect, with quality features: formants, bandwidth spectral energy distribution, to enhance the system performance. The quality features were obtained using phonetic analysis software (PRAAT). They also optimized the HMM classifier by presenting a new algorithm for training HMM based on Genetic Algorithm (GA). Seven emotional states: anger, disgust, fear, joy, neutral, sadness and surprise were successfully classified with accuracy higher than 77%. They have contributed the ideas of finding the most optimal HMM and proved that integrating features is much more robust instead of single features.

Selection of classification technique is also a key aspect in emotion recognition system design. Several classification methods are widely used to identify the different types of sounds like the sound of human speech (Speech Recognition), human voice (Voice Recognition), as well as the noise generated by certain objects (Sound Recognition). Although using different methods, but the purpose is the same, to process, identify and classify sounds.

4.0 CLASSIFICATION METHOD

In the speech recognition field, there are accessible classifiers can be used such as Hidden Markov Model (HMM), Artificial Neural Network (ANN), k-nearest neighbors (KNN), Gaussian Mixtures Model (GMM), and Support Vector Machine (SVM). Among the classification methods used by previous researchers for emotion recognition task are HMM, SVM and GMM [17].

The HMM is modeling a stochastic process defined by a set of states and transition probabilities between those states, where each state describes a stationary stochastic process and the transition from one state to another state describes how the process changes its characteristics in time [18]. A statistical parametric speech emotion recognition system based on hidden Markov models (HMM) has rapidly expanded and gain concern among researcher over the last few years.

SVM provides an easy machine learning algorithm that is well-organized and has been widely used in classification tasks [6, 4]. By using kernel functions, SVM can transform the original input set to a high dimensional feature space. From a survey done by

Ayadi *et al.* [9], SVM is shown to overtake other classifier technique because of the existence of excellent data-dependent generalization bounds and the global optimal of training algorithm. In situations where the training data should be limited, SVM can be used because of its good performance of classification compared to other classifier [6].

Artificial Neural Networks (ANN) is one of the approaches in the field of recognition and has been used in the fields of research including speech recognition field. ANN method was developed based on neural network in the human brain [17]. As an example of the human brain activities in delivering information and process that information to control all functions of the body through a connected neuron. This design process has been implemented in ANN topology.

The third key aspect is the selection of a good database. It is essential to ensure the emotion recognition system provides the maximum performance. There are lots of databases from various languages developed by speech processing community in collaboration with professional actors available for use by researchers [1]. The database is available in many languages such as English, German, Spanish, Chinese and Japanese, Korean and also Thai language.

5.0 EMOTIONAL SPEECH DATABASE

Previous research had found that it is necessary to record accurate emotional speech database because the precision of the system, is extremely depends on the emotional speech database used in the system [4]. Database resource in multiple languages is lacking in emotion recognition area. The important thing to take note when assessing the performance of a speech emotion recognition system is the originality of the database used. If the low-quality database is used, less satisfactory results will be obtained [9].

Berlin Emotional Speech Database (EMO-DB) is well-known and easy access because it is freely available. The database use German language and covers seven emotions: anger, sadness, disgust, fear, boredom, joy/happiness and neutral [13].

Other than that, English emotional speech database from the Linguistic Data Consortium (LCD) also available to be used for research and development purposes [14]. The LDC database contains records of native English and covers up to 15 emotional states: hot anger, cold anger, neutral, happy, sad, panic, interest, disgust, fear, despair, shame, elation, pride, contempt and boredom [5].

In addition, the database in English language (Surrey Audio-Visual Expressed Emotion (SAVEE)) is free to use for research and education purpose. It has been recorded as a pre-requisite for the development of an automatic emotion recognition system. The database consists of recordings from 4 male actors in 7 different emotions, 480 British English utterances in total [15].

Moreover, the database in Spanish language also can be used for research purpose. The recorded database allows wide-ranging studies on speech conversion, far-field speech recognition and speech, emotional speech synthesis, video-based emotion identification and also prosodic modeling. The database covers up to six emotions categories: happiness, sadness, fear, anger, disgust and surprise [16].

Because of the limitation of emotional speech resource in multiple languages, Ali *et al.* in [7] have developed their own speech emotion database using Pakistan language. They used a Berlin database of emotional speech (EMO-DB) along with their own Pakistan database known as Emotion-Pak to perform a cross linguistic analysis. The result of the experiment shows that human emotions are independent of language, gender and cultural backgrounds. They claimed that females show a deeper emotion than males whereas men doing well in linguistic utterances.

6.0 PRIOR WORK

Many researchers use prosodic, spectral and wavelet feature in their study to recognize emotion in speech. Some of them are using the integration of these features to obtain higher recognition rate. The review of previous study is summarized in Table 1 below.

Table 1 Prior work on speech emotion recognition

Author	Features	Classifier	Database	Result
2013. Manav Bhaykar et.al.	• MFCC	• GMM • HMM	• IITKGP-SESC • IITKGP-SESHC • Indian and Telegu language.	• Text dependent:89.20% • Text independent: 84.57%
2013. Thapanee Seehapoch *et al.*	• Fundamental Frequency (F0) • Energy • ZCR • LPC • MFCC	• SVM	• Berlin Database of Emotional Speech • Japanese Emotional Speech Database • Audiovisual Thai Emotion Database.	• German-89.80% • Japan – 93.57% • Thai – 98.0%
2013. Krishna Kishore et.al.	• MFCC • SBC	• GMM • Expectation Maximization (EA) algorithm.	• SAVEE database	• SBS – 70% of recognition • MFCC algorithm – 51% of recognition.
2012. Mehmet Cenk Sezgin *et al.*	• Perceptual quality metrics that are given in the perceptual evaluation of audio quality -ITU BS.1387 recommendation. • Perceptual features	• SVM and GMM • Soft Majority Voting	• German emotional speech database (EMO-DB) • Vera am Mittag German audio visual emotional speech database –(VAM)	• 16% for EMO-DB and 7-11% in VAM for "all" and "valence" tasks.
2012. Ashish B. Ingale *et al.*	• Fundamental frequency • Energy • MFCC	• HMM • SVM	• Use new database - Recording the voice samples from various speakers (male and female). • Using one common statement.	• HMM - Anger – 83.33% Happy – 57.14% Sad – 62.50% Surprise – 71.4% Neutral – 75% • SVM- Anger – 71.42% Happy – 57.14% Sad – 71.43% Surprise – 63.3% Neutral – 75%
2012. Mohammad Khadem Safdarkhani *et al.*	• Formant frequency-related • Pitch frequency-related • Energy • MFCC	• 2 types of neural network -MLP - RBF • Gaussian mixture model (GMM)	• 20 speakers – uttered 285 sentences. • Used rich features set to the size 55. • 5 distinct feature subsets.	• Emotion recognition rate above 82% by employing a small-size feature set using AUC_S based feature ranking method and GMM recognizer.
2012, Han Zhiyan *et al.*	• Prosody – Logarithma F0, Energy, Durational aspects • Quality –16 features, describing the first three formants, bandwidths, harmonic to noise ratio, spectral energy distribution, voice to unvoiced energy ratio, glottal flow.	• A new algorithm for training HMM based on a Genetic Algorithm (GA) is presented.	• Chinese language	• Improved HMM is more effective than HMM, all emotions recognized with an accuracy higher than 77%.
2012. Ashish B. Ingale *et al.*	• Energy • Pitch • LPCC • MFCC	• GMM • ANN • K-NN • HMM • SVM	• Database- Not Stated	• GMM overall – 78.77% SD- 89.12%, SI -75% • ANN SD- 51.19%, SI -52.87% • K-NN – 64% in four emotional states. • HMM- SD- 76.12%, SI -64.77% • SVM SD- 80%, SI -75% *speaker dependent (SD) *speaker independent (SI)

Author	Features	Classifier	Database	Result
2011. Moataz El Ayadi *et al*.	• Prosodic speech features such as pitch and energy. • Global features	• 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 this task because different emotional corpora with different experimental setups were applied.	• Problems with existing emotional speech databases: 1. Most speech emotional databases do not well enough simulate emotions in a natural and clear way. 2. Some databases such as KISMET, the quality of the recorded utterances is not so good and the sampling frequency is low (8 kHz). 3. Phonetic transcriptions are not provided with some databases like BabyEars. Difficult to extract linguistic content from the utterances.	• 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.
2010. Tim Polzehl et.al.	• Pitch • Formants • Intensity • Loudness • MFCC • Spectral.	• SVM with linear kernel function.	• Real-life American English and German Interactive Voice Response systems.	• 75.6% f1 on multilingual train set. • 74.5% f1 on multilingual test set.
2010. Dmitri Bitouk *et al*.	• MFCC	• SVM with radial basis kernels constructed using the LIBSVM library • Used Balance Accuracy (BAC)	**1.** English emotional speech database from Linguistic Data Consortium (LDC) **2.** Berlin database of German emotional speech	• Classification accuracies are consistently higher for spectral features compared to prosodic or utterance-level spectral features.
2009. Yutai Wang et.al.	• Time • Amplitude • Fundamental frequency (F0) • Formant.	• Principle Component Analysis (PCA) is used to recognize the state of emotion in multilingual speech.	• Chinese, English, Russian, Korean, Japanese.	• Chinese emotion recognition rate is higher than multi-language (Above 80%).
2009. Elif Bozkurt1, et.al	• Prosodic Features • Spectral Features – MFCC and (line spectral frequency) – LSF • Dynamic Features • HMM-based Features	• GMM	• Employ the FAU Aibo Emotion Corpus, which is distributed through the INTERSPEECH 2009 Emotion Challenge.	• MFCC features perform better than prosody features since they capture rich spectral information. Similarly, the LSF features do well in emotion recognition.

7.0 CONCLUSION

Improper emotions delivered may lead to wrong perception and thus give ineffective interaction. That is why the existence of emotion in one speech is crucially important to deliver the right message.

From the review of previous research done, it is proven that the recognition rate depends on the features, data and classification method used. Apart from that, integrated features will give better recognition rate compare to a single feature. There are still more hybrid features that were not studied. Future research on hybrid speech features will be expected to enhance the knowledge about the variety of features in human speech and contribute to a better speech emotion recognition rate. Furthermore, it will be able to analyze the efficiency of some methods chosen in recognizing the emotion occurrence in particular speech signal along with reliable emotional speech database.

Effective communication is not just focused on what the contents of the information delivered but also how the information is presented to accomplish the natural communication. In addition of emotion in the application will make the speech more perfectly natural with the human speech. Advances in this area dramatically improved emotion recognition system using multiple languages and can be used by people worldwide.

References

[1] Lanjewar, R. B. & Chaudhari, D.S., 2013. Speech Emotion Recognition : A Review. *International Journal of Innovative Technology and Exploring Engineering*. 2(4): 68-71.

[2] Seehapoch, T. & Wongthanavasu, S. 2013. Speech Emotion Recognition Using Support Vector Machines. *5th International Conference on Knowledge and Smart Technology (KST)*.

[3] Sezgin, M., Gunsel, B. & Kurt, G., 2012. Perceptual Audio Features for Emotion Detection. *EURASIP Journal on Audio, Speech, and Music Processing*. 2012(1): 16.

[4] Ingale, A. & Chaudhari, D. 2012. Speech Emotion Recognition. *International Journal of Soft Computing and Engineering (IJSCE)*. 2(1): 235-238.

[5] Bitouk, D., Verma, R. & Nenkova, A., 2010. Class-level Spectral Features for Emotion Recognition. *Speech Communication*. 52(7-8): 613-625.

[6] Shen, P., Changjun, Z. & Chen, X., 2011. Automatic Speech Emotion Recognition using Support Vector Machine. *International Conference on Electronic and Mechanical Engineering and Information Technology*. 621-625.

[7] Ali, S. A. *et al*. 2013. Development and Analysis of Speech Emotion Corpus Using Prosodic Features for Cross Linguistics. *International Journal of Scientific & Engineering Research*. 4(1): 1-8.

[8] Zhou, Y. *et al*. 2009. Speech Emotion Recognition Using Both Spectral and Prosodic Features. *2009 International Conference on Information Engineering and Computer Science*. 1-4.

[9] Ayadi, M. El, Kamel, M. & Karray, F. 2011. Survey on Speech Emotion Recognition: Features, classification schemes, and databases. *Pattern Recognition*. 44(3): 572-587.

[10] Kishore, K. V. K. & Satish, P. K. 2013. Emotion Recognition in Speech using MFCC and Wavelet Features. *Advance Computing Conference (IACC), 2013 IEEE 3rd International*. 842-847.

[11] Safdarkhani, M. K. *et al*. 2012. Emotion Recognition of Speech Using ANN and GMM. *Australian Journal of Basic and Applied Sciences*. 6(9): 45-57.

[12] Bozkurt, E. & Erzin, E. 2009. Improving Automatic Emotion Recognition from Speech Signals. In *Interspeech 2009: 10th Annual Conference of the International Speech Communication Association*.

[13] Schuller, B. & Burkhardt, F. 2010. Learning with Synthesized Speech for Automatic Emotion Recognition. *Speech and Signal Processing*. 5150-5153.

[14] Ahtaridis, E., Cieri, C. & DiPersio, D. 2009. LDC Language Resource Papers: Building a Bibliographic Database. *8th International Conference on Language Resources and Evaluation, Istanbul*. (5).

[15] Haq, S. & Jackson, P. 2009. Speaker-dependent Audio-visual Emotion Recognition. In *International Conference on Auditory-Visual Speech Processing (AVSP)*.

[16] Barra-Chicote, R. *et al.* 2008. Spanish Expressive Voices: Corpus for Emotion Research in Spanish. *6th conference of Language Resources & Evaluation (Workshop on Corpora for Research on Emotion and Affect)*. 2.

[17] Caponetti, L., Buscicchio, C. & Castellano, G., 2011. Biologically Inspired Emotion Recognition from Speech. *EURASIP Journal on Advances in Signal Processing*. 1): 10.

[18] Plannerer, B. 2005. *An Introduction to Speech Recognition.* Munich, Germany.

[19] Han, Z., Lun, S. & Wang, J. 2012. Speech Emotion Recognition System Based on Integrating Feature and Improved HMM. *Proceedings of the 2nd International Conference on Computer Application and System Modeling*. 571-574.

[20] Polzehl, T., Schmitt, A. & Metze, F. 2010. Approaching Multi-Lingual Emotion Recognition from Speech-On Language Dependency of Acoustic/Prosodic Features for Anger Recognition. *Speech Prosody 2010-Fifth.* 1-4.

[21] Iliev, A. 2009. *Emotion Recognition Using Glottal and Prosodic Features*. University of Miami.