A Review on Emotion Recognition using Speech

Saikat Basu*, Jaybrata Chakraborty‡, Arnab Bag† and Md. Aftabuddin§

*Member IEEE, School of Medical Science and Technology, Indian Institute of Technology Kharagpur and
Department of Computer Science and Engineering, Maulana Abul Kalam Azad University of Technology, West Bengal
‡Department of Information Technology, Maulana Abul Kalam Azad University of Technology, West Bengal
†Department of Electronics and Electrical Communication Engineering, Indian Institute of Technology Kharagpur
§Maulana Abul Kalam Azad University of Technology, West Bengal

Abstract

Emotion recognition or affect detection from speech is an old and challenging problem in the field of artificial intelligence. Many significant research works have been done on emotion recognition. In this paper, the recent works on affect detection using speech and different issues related to affect detection has been presented. The primary challenges of emotion recognition are choosing the emotion recognition corpora (speech database), identification of different features related to speech and an appropriate choice of a classification model. Different types of methods to collect emotional speech data and issues related to them are covered by this presentation along with the previous works review. Literature survey on different features used for recognizing emotion from human speech has been discussed. The significance of various classification models has been presented along with some recent research works review. A detailed description of a prime feature extraction technique named Mel Frequency Cepstral Coefficient (MFCC) and brief description of the working principle of some classification models are also discussed here . In this paper terms like affect detection and emotion recognition are used interchangeably.

KEYWORDS: Affect Detection, Corpora, Features, MFCC (Mel Frequency Cepstral Coefficient), LPCC (Linear Prediction Cepstral Coefficients), LPC (Linear Prediction Coefficients), Classifier, Neural Network, GMM (Gaussian Mixture Model), HMM (Hidden Markov Model), KNN (K-Nearest Neighbors), MLP (Multi Layer Perceptron), RNN (Recurrent Neural Network), Back Propagation.

I. Introduction

Human machine interaction are widely used nowadays in many applications. One of the medium of interaction is speech. The main challenges in human machine interaction is detection of emotion from speech. When two persons interact to each other they can easily recognize the underlying emotion in the speech spoken by the other person. The objective of emotion recognition system is to mimic the human perception mechanisms. There are several application in speech emotion recognition. Emotion can play an important role in decision making. Emotion can be detected from different physiological signal also [2], [3]. If emotion can be recognized properly from speech then a system can act accordingly. An efficient emotion recognition system can be useful in the field of medical science [1], [7], robotics engineering, call center application etc. Human can easily recognize emotion of speaker. This can be achieved by many years of practice and observation. Human first analyzes different characteristics of particular speech and then using previous experience or observation he recognizes the emotion of the speaker. There is a need to build a human like system that can detect emotions effectively and efficiently. Identification of emotion can be done by extracting the features or different characteristics from the speech and a training is needed for a large number of speech database to make the system accurate. The steps towards building of an emotion recognition system are, an emotional speech corpora is selected or implemented then emotion specific features are extracted from those speeches and finally a classification model is used to recognize the emotions.

II. Review of Databases (Speech corpora)

A suitable choice of speech database (corpora) plays a very important role in the field of affect detection. A context rich emotional speech database are preferred for a good emotion recognition system. Mainly three types of corpora are used for developing a speech system [11], [5] they are,

1) Elicited emotional speech database: This type of corpora are collected from speaker by creating artificial emotional situation. Advantage of this type of database is that it is very close to the natural database but there are some problems also. All emotions may not be available and if the speaker is aware of that they are being recorded, then the emotion expressed by him may be artificial.

2) Actor based speech database: This type of speech data set collected from professional and trained artists. Collecting of these type of data are very easy and a wide variety of emotion are available in the corpora.The main problem of this type of database are it is episodic in nature and it is very much artificial in nature.

3) Natural speech database: This type of database created from real world data. These type of data are completely natural and very useful for real world emotion recognition.

The problem is that, all emotions may not be present and it consists of background noise.

Some important speech corpora being used for emotion recognition system are briefly discussed in TABLE I.

TABLE I
LITERATURE SURVEY OF SPEECH CORPORA USED FOR EMOTION RECOGNITION

SL	Ref.	Emotions	Type of Corpora	Description
1	Daniel J.France et al.(2000)[7]	Depression and Neutral	Natural	Emotions are recognised from speech to identify depressed and suicide prone patients.
2	Q.Jin et al. (2015)[10]	Anger, Happy, Sad, Neutral	Elicited	Generate features from both acoustic and lexical level to identify emotion like anger, sad, happy, neutral.
3	S.Yildirim et al. (2004) [24]	Angry, Happy, Sad, Neutral	Actor based	Look into acoustic features of speech related with emotions like sadness, anger, happiness, and neutral.
4	Chul Min Lee et al.(2005)[13]	Negative and Nonnegative	Natural	Explore domain specific emotion recognition from the data collected from call center.
5	Jiahong Yuan et al.(2002)[25]	Anger, Fear, Joy, Sadness, Neutral	Elicited	Articulation, prosody, and phonation properties are explored to identify emotions.
6	Reda Elbarougy et al.(2013) [6]	valence, activation and dominance.	Actor based	Investigate emotions from more than one language.

Implementation of emotional speech database depends on objective of the research. For efficient affect detection system, it is important that the corpora must consists of real and natural emotional speech spoken by a large number of male and female persons. Though there are different corpora exists, there is no standard, globally approved speech database available for emotion recognition. In Indian context, there are different speech corpora for speaker recognition but there is a lack of corpora for emotion recognition [11].

A major challenge for speech emotion recognition is the reduction of noise which incorporated with the speech data sample. Specifically when the second and third approaches are considered for collecting speech data, noise reduction will be necessary. Different low pass filters can be applied for this purpose. Though this processes can vary depending on actual input.

III. NOISE REDUCTION

If speech samples are collected under real life condition then speech signal corrupted with several noise. To recover from this problem, a noise reduction phase is performed before analyzing emotional speech. There are different approaches for noise reduction from speech signal.

A. Wiener filter

Consider $x(t)$ is a clean speech signal affected by a Gaussian noise $n(t)$. If $y(t)$ be the noisy speech at time t then,

$$y(t) = x(t) + n(t) \tag{1}$$

The estimated clean speech can be obtained from the observed signal through a filtering process. The filter can be defined by the coefficients h_i. So, the estimate of clean signal is given by

$$\hat{x}(t) = \sum_{i=0}^{L-1} h_i y(t-i) \tag{2}$$

Where L is the number of observation. In other words this expression can be expressed as $\hat{x}(t) = h^T y(t)$ where vector $h = [h_1\ h_2\ h_3\ .....h_{L-1}]^T$ represents the finite impulse filter and vector $y(t) = [y(t-1)y(t-2).....y(t-L+1)]^T$ is L observed signals.

Now the error signal can be generated by

$$e(t) = x(t) - \hat{x}(t) \tag{3}$$

The mean squared error can be defined as

$$E[e^2(t)] = E[(x(t) - \hat{x}(t))^2] \tag{4}$$

Wiener filter provide the minimum mean squared error in terms of coefficient vector. So Wiener filter is an optimum filter. It is used to reduce noise from speech signal. Minimization of mean squared error can be achieved by applying derivative on the above equation and when it is equal to zero, error will be minimized [14].

B. Spectral subtraction

This approach subtracts noise from corrupted signal in the power spectral density domains [4].

If $x(t)$, $y(t)$ and $n(t)$ represents the clean speech signal, corrupted signal and noise signal respectively in time domain then,

$$y(t) = x(t) + n(t) \tag{5}$$

Now if $X(f)$, $Y(f)$ and $N(f)$ are the Fourier transformation of the clean speech signal, corrupted signal and noise signal respectively then estimated noise free signal can be generated by subtracting noise spectra from corrupted signal spectra.

$$|X(f)|^n = |Y(f)|^n - |N(f)|^n \tag{6}$$

Where n is the power exponent.

IV. REVIEW OF FEATURES

Features of a speech can be used to identify the difference between several emotional statements. Characteristics of a human vocal tract and hearing system is represented by different features of speech signal [17]. To build an emotion recognition system it is very much important to extract

different prosodic and acoustic features from speech signal. Some important features of speech signal are pitch, amplitude, formants and spectral features.

Pitch of the signal is generated from tremble of the vocal cord. Pitch can be measured by the change of frequency. The time duration between two consecutive vocal chord vibration is called period of pitch and the number of vibration in one unit time is called the fundamental frequency or pitch frequency [6]. Pitch of the signal can be calculated well by auto correlation with center clip function [21], [18]. Auto correlation function for periodic signal is as follows,

$$A(i) = \lim_{M\to\infty} \frac{1}{2M+1} \sum_{n=-M}^{M} x(n)x(n+i) \tag{7}$$

where M is the number of sample and $x(n)$ is the value of the signal at n^{th} instance.

Amplitude of the signal in different intervals represent the loudness (energy) of a sound perceived by the human ear [20]. Formants are the distinguishable frequency peaks in the speech signal. A spectral root group delay function approach can be applied for formant extraction [16].

Speech signals are produced as a result of excitation in the vocal tract by the source signal. Features that are measured from the vocal tract system are called system features or spectral features. The most popular spectral features used by various emotion recognition systems are Linear prediction coefficients (LPCs), Mel frequency cepstral coefficients (MFCCs) and Linear prediction cepstral coefficients (LPCC) [17].

Linear prediction coefficients method approximate n^{th} speech sample $s(n)$ with a weighted combination of previous m speech samples. Linear prediction coefficient can be calculated with the formula,

$$E(n) = s(n) - \sum_{k=1}^{m} a_k s(n-k) \tag{8}$$

Where $E(n)$ is the error signal and a_k are the filter coefficients.

Mel frequency cepstral coefficients are computed on the basis of human hearing ability. In Mel frequency cepstral coefficients (MFCC) method, two types of filter are used. Some filter are spaced linearly at low frequency below 1kHz and other are spaced logarithmically at high frequency above 1kHz [17], [8]. This process consists of a few steps as discussed below,

Pre-emphasis: Pre-emphasis is required to increase signal energy. In this process, speech signal is passed through a filter which increase the energy of signal. This increment of energy level gives more information.

Framing: In this process, speech sample is segmented into 20-40 ms frames. The length of human voice may vary, so for fixing the size of speech this processes is necessary. Although the speech signal is non-stationary in nature (i.e. frequency can be changed over the time period), but for a

short duration of time, signal behave like a stationary signal.

Windowing: After framing process, the windowing process is performed. Windowing function reduce the signal discontinuities at the start and end of each frame. In this process, frame is shifted with a 10ms span. That means each frame contains some overlapping portion of previous frame.

Fast Fourier Transform (FFT): FFT is used to generate the frequency spectrum of each frame. Each sample of each frame converted from time domain to frequency domain by the FFT. FFT is used to find all frequencies present in the particular frame.

Mel scale filter bank: This is a set of 20-30 triangular filters applied to each frame. The mel scale filter bank identify how much energy exists in a particular frame. The mathematical equation to convert the normal frequency f to the Mel scale m is as follows,

$$m = 2595 \log\left(1 + \frac{f}{700}\right) \tag{9}$$

Log energy computation: After getting the filter bank energy of each frame, log function is applied to them. It is also inspired by human hearing perception. A human do not listen loud volume on a linear scale. If the volume of the sound is high human ear can not recognize large variations in energy. Log energy computation give those features for which human can listen clearly.

DCT: In the final step discrete cosine transformation (DCT) is calculated of the log filter bank energies .

Linear prediction cepstral coefficients (LPCC) can be measured by using all those steps which have been used to calculate LPC. LPCC is more reliable than LPC and it have been widely used for feature extraction [22]. LPCC can be calculated from LPC with the formula,

$$LPCC_i = LPC_i + \sum_{k=1}^{i-1} \frac{k-i}{i} LPCC_{(i-k)} LPC_k \tag{10}$$

Some important acoustic features being used for emotion recognition system are briefly discussed in TABLE II. Recognizing emotions from speech signals requires few steps that makes the emotion recognition system as a whole. Step wise activities can be listed as follows.

At first various acoustic and prosodic features have been extracted from the speech signal, and a class label is associated with the extracted features. A collection of such data (features with emotion class label) will be then segmented in two parts (usually a 60%-40% division) appropriate for training a Neural Network with train data set and to evaluate performance with test dataset. Then a suitable neural network architecture can be selected to form the classifier.

V. Review of Classifiers

A classification system is an approach to set each speech to a proper emotion class according to the extracted features from speech. There are different classifiers available for emotion recognition. There is no thumb rule for choosing a proper classifier. Most of the cases the choice of classifier

made based on past references. Features extracted from each speech sample (feature vector) supplied as an input to classifiers with a linear combination of real weight vector *W*. This weight vector then adjusted with a proper training method. An activation function is then used to generate the output from the model which mapped each input to a predefined emotion class. This

TABLE II LITERATURE SURVEY ON FEATURES OF SPEECH

SL	Ref.	Features	Description
1	Daniel J. France et al. (2000)[7]	Pitch, Amplitude Modulation, Formant	Analyze and compare different speeches of male and female patients and diagnose the depression level.
2	H.K. Palo et al.(2015)[17]	MFCC, LPC, LPCC	Analyze two emotion classes like low arousal and high arousal.
3	Qirong Mao et al.(2014)[15]	Pitch, energy	Use convolutional neural network to identify emotions like anger, joy, sadness, and neutral
4	Lee Chul Min (2005)[13]	Pitch, Energy, Duration, Formant.	Works on classification of negative and non negative emotion with data collected from call center.
5	Chung-Hsien Wu et al. (2011)[23]	Pitch, intensity, formants, shimmer, MFCC.	Use multiple classifier to recognize emotion like neutral, happy, angry and sad.
6	Yuanlu Kuang and Lijuan Li(2013)[12]	Pitch, energy, formant, LPCC, MFCC	Propose a Dempster–Shafer evidence theory based decision fusion technique among two classifiers to classify emotion like angry, sad, surprise and disgust.

activation function may be linear or non linear. According to the nature of activation function classifiers can be grouped into two different categories, which are linear classifier and non linear classifier. Linear classifier will classify accurately if the feature vectors are linearly separable. In real life scenario most of the feature vectors are not linearly separable so a nonlinear classifier is a better choice [11]. There are various nonlinear classifiers available for emotion recognition, namely SVM (Support Vector Machine), GMM (Gaussian Mixture Model), MLP (Multi Layer Perceptron), RNN (Recurrent Neural Network), KNN (K-Nearest Neighbors), HMM (Hidden Markov Model).

Support vector machine is a supervised learning method whose objective is to find minimal number of separating hyperplane which have maximum margin from data.

KNN classification is the most fundamental and simple classification method. KNN is used when we do not have any prior knowledge of data distribution. This method is based on Euclidean distance between the features of test sample and train sample.

The Gaussian Mixture Model is used to generate the probability density function of feature vector, $\sim x$, which is *D*dimensional continuous valued data vector, by the linear combination of multivariate Gaussian distribution [22].

$$p(\sim x|\alpha) = \sum_{i=1}^{M} w_i b_i(\sim x) \tag{11}$$

with

$$b_i = \frac{1}{\sqrt{(2\pi)^D |\Sigma_i|}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu_i})^T \Sigma_i^{-1} (\vec{x}-\vec{\mu_i})} \tag{12}$$

where α is the model describe by

$$\alpha = \{ w_i, \mu\sim_i, \Sigma_i \} \tag{13}$$

i is the mixture index ($1 \le i \le M$), w_i is the mixture weight such that $\sum_{i=1}^{M} w_i = 1$ and $b_i(\sim x)$ is a multivariate Gaussian distribution with means $\mu\sim_i$ and diagonal covariance matrix Σ_i.

Hidden Markov model is a double random process with an underlying process which is hidden, but they can be observed by using another set of random processes which are responsible for producing the sequence of observed labels. HMM is widely used in case of speech processing.

Multi Layer Perceptron (MLP) model is used to compute an appropriate output from a sets of input data. MLP is a neural network model. A MLP model made up with three layers. These layers are input layer, hidden layer and output layer where hidden layer may be more than one. The architecture of MLP model is like a connected graph where nodes of each layer is fully connected with a weighted edge to the nodes of next layer. Each layer consists of multiple nodes. Each node consist of two functions one is input function and another is output function. MLP uses back propagation for training the network which is a supervised learning technique.

Different features of the speech are given as an input in the input layer and each node of next layer take a input as a weighted sum of each node of previous layer. In MLP each node uses a nonlinear function which is called activation function to generate output. This output function can be designed in several ways.

Mainly sigmoid functions are used as activation functions, some of them are described by:

$f(x) = \tanh(x)$ and $f(x) = (1+e^{-x})^{-1}$, in which the former function produce output ranges from -1 to 1, and the latter function produce the output ranges from 0 to 1.

Back Propagation: The objective of back propagation is to optimize the weight matrix. These weight matrix is responsible for mapping any arbitrary inputs to proper output class. To do this error is calculated for each output node using the squared error function. The total error E can be measured by adding up all these errors. A learning rule is then used to optimized the weight matrix. For n number of training samples,

$$E = \sum_{i=1}^{n} (T_i - O_i) \tag{14}$$

Where T_i and O_i are the target and output for i^{th} sample. And each weight W_i is adjusted with the learning rule,

$$W_i = W_i + \mu * \frac{\partial E}{\partial W_i} \quad (15)$$

Where μ is called learning rate.

Though standard feed forward MLP is powerful tool for classification problems, extremely sparse matrix may not yield favorable result but however experiment with a properly tuned MLP network should be interesting. It is noteworthy that for input vectors of different length training of a Recurrent Neural Network (RNN) can be a better option. A recurrent neural network can be defined as a feed forward network with a feedback connection to hidden layer. RNN works on time step basis. Input feature vector segmented into a number of sequence vectors, each associated with different time step t. If the inputs and outputs vectors of neural network are $x(t)$ and $y(t)$, and the three connection weight matrices are W_i, W_h and W_o, and the activation functions for hidden and output node are f_h and f_o, the behavior of the recurrent neural network can be described by the pair of nonlinear equations:

$$y(t) = f_o(W_o h_i(t)) \quad (16)$$

$$h_i(t) = f_h(W_i x(t) + W_h h_i(t-1)) \quad (17)$$

where, $h_i(t-1)$ is the net-input to the i^{th} node at time $(t-1)$ [18]. A list of classifiers used for emotion recognition in various research is provided in TABLE III.

TABLE III
LITERATURE SURVEY ON CLASSIFIERS USED IN EMOTION RECOGNITION USING SPEECH

SL	Ref.	Classifiers	Description
1	Yuanlu Kuang and Lijuan Li et al. (2013)[12]	HMM, ANN	Propose a Dempster–Shafer evidence theory based decision fusion technique among two classifiers to classify emotion like angry, sadness, surprise and disgust. It produced 83.86% recognition rate.
2	Bjorn Schuller et al. (2004)[20]	GMM	System produce 86% recognition rate, where human judge the same corpus at 79.8% recognition rate.
3	Wen-Yi Huang, Tsang-Long Pao (2012)[9]	KNN, HMM, GMM, SVM	Propose inclusion of keyword as a feature rather than only speech signal. Final result is obtained by a fusion technique.
4	Amiya Kumar et al. (2015)[19]	SVM	Multilevel SVM is used to identify seven emotions and it is observed that the recognition rate is 82.26%
5	Chung-Hsien Wu et al.(2011)[23]	HMM, SVM, MLP	Propose meta decision tree(MDT) for the fusion of the outcome of multiple classifiers to recognize emotions. They also employed a personality trait of a specific speaker obtained from the Eysenck personality questionnaire and integrated into classifier to obtain emotion. Obtain 85.79% accurate result
6	Chang-Wun Park et al. (2002)[18]	RNN	Proposed that pitch as an important feature. And give an idea of emotion recognition using RNN.

VI. CONCLUSION

The study reveals the fact that identification of emotion of a person is a task yet to have complete and general solution. Till now, most of the work has been done on the fixed size speech segment for classification of emotion, that means on the off line speech. The problem arises when speech samples are of different size, for this type of data, the input features matrix may be mostly sparse. Though standard feed forward MLP is powerful tool for classification problems, extremely sparse matrix may not yield favorable result however experiment with a properly tuned MLP network should be interesting. It is noteworthy that for input vectors of different length training, recurrent neural network (RNN) can be a better option. Moreover as human emotion is not only related to voice but also other physical gestures like facial expression or body parts movement. For this reason voice related to emotion may often be ambiguous also due to nature of a person. Thus emotion recognition using machine intelligence still have a various difficulties to overcome and a long way to run.

REFERENCES

[1] S. Basu, A. Bag, M. Mahadevappa, J. Mukherjee, and R. Guha. Affect detection in normal groups with the help of biological markers. In *India Conference (INDICON), 2015 Annual IEEE*, pages 1–6. IEEE, 2015.

[2] S. Basu, A. Bag, Md. Aftabuddin, A. Mahadevappa, J. Mukherjee, and J. Guha. Effects of emotion on physiological signals. In *2016 IEEE Annual India Conference (INDICON)*, pages 1–6, Dec 2016.

[3] S. Basu, N. Jana, A. Bag, M. Mahadevappa, J. Mukherjee, S. Kumar, and R. Guha. Emotion recognition based on physiological signals using valence-arousal model. In *Image Information Processing (ICIIP), 2015 Third International Conference on*, pages 50–55. IEEE, 2015.

[4] F. Chenchah and L. Zied. Speech emotion recognition in noisy environment. *2nd International Conference on Advanced Technologies for Signal and Image Processing*, pages 788–792, 2016.

[5] M. El Ayadi, M. S. Kamel, and F. Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587, 2011.

[6] R. Elbarougy and M. Akagi. Cross-lingual speech emotion recognition system based on a three-layer model for human perception. *2013 AsiaPacific Signal and Information Processing Association Annual Summit and Conference*, pages 1–10, 2013.

[7] D. J. France and R. G. Shiavi. Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE Transactions on Biomedical Engineering*, 47(7):829–837, 2000.

[8] R. Hibare. Feature Extraction Techniques in Speech Processing : A Survey. *International Journal of Computer Applications*, 107(5):1–8, 2014.

[9] W. Huang and P. Tsang-Long. A Study on the Combination of Emotion Keywords to Improve the Negative Emotion Recognition Accuracy. *Information Science and Service Science and Data Mining (ISSDM), 2012 6th International Conference on New Trends in Date of Conference: 2325 Oct. 2012*, pages 496–500, 2012.

[10] Q. Jin, C. Li, and S. Chen. Speech emotion recognition with acoustic and lexical features. *PhD Proposal*, 1:4749–4753, 2015.

[11] S. G. Koolagudi and K. S. Rao. Emotion recognition from speech: A review. *International Journal of Speech Technology*, 15(2):99–117, 2012.

[12] Y. Kuang and L. Li. Speech emotion recognition of decision fusion based on DS evidence theory. *Proceedings of the IEEE International Conference on Software Engineering and Service Sciences, ICSESS*, pages 795–798, 2013.

[13] C. M. Lee and S. S. Narayanan. Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*, 13(2):293– 303, 2005.

[14] I. A. Lima, M. S. Alencar, W. T. A. Lopes, and F. Madeiro. Evaluation of optimal and sub-optimal speech noise reduction wiener filters. In *IWT 2015 - 2015 International Workshop on Telecommunications*, pages 1–5. IEEE, jun 2015.

[15] Q. Mao, M. Dong, Z. Huang, and Y. Zhan. Learning Salient Features for Speech Emotion ¡newline/¿Recognition Using Convolutional ¡newline/¿Neural Networks. *IEEE Transactions on Multimedia*, 16(8):2203– 2213, dec 2014.

[16] H. A. Murthy and B. Yegnanarayana. Formant extraction from group delay function. *Speech Communication*, 10(3):209–221, 1991.

[17] H. K. Palo, M. N. Mohanty, and M. Chandra. Computational Vision and Robotics. *Advances in Intelligent Systems and Computing*, 332:63–70, 2015.

[18] C. Park, D. Lee, and K. Sim. Emotion recognition of speech based on RNN. (November):4–5, 2002.

[19] A. K. Samantaray and K. Mahapatra. A novel approach of speech emotion recognition with prosody, quality and derived features using SVM classifier for a class of North-Eastern Languages. *Recent Trends in Information Systems (ReTIS), 2015 IEEE 2nd International Conference on*, pages 372–377, 2015.

[20] B. Schuller, G. Rigoll, and M. Lang. Hidden Markov model-based speech emotion recognition. *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, 2:1–4, 2003.

[21] D. Ververidis and C. Kotropoulos. Emotional speech recognition: Resources, features, and methods. *Speech Communication*, 48(9):1162–1181, 2006.

[22] E. Wong and S. Sridharan. Comparison of linear prediction cepstrum coefficients and mel-frequency cepstrum coefficients for language identification. *Proceedings of 2001 International Symposium on Intelligent Multimedia, Video and Speech Processing. ISIMP 2001 (IEEE Cat. No.01EX489)*, pages 95–98, 2001.

[23] C. H. Wu and W. B. Liang. Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. *IEEE Transactions on Affective Computing*, 2(1):10–21, 2011.

[24] S. Yildirim, M. Bulut, and C. Lee. An acoustic study of emotions expressed in speech. *Proceedings of InterSpeech*, pages 2193–2196, 2004.

[25] J. Yuan, L. Shen, and F. Chen. The acoustic realization of anger, fear, joy and sadness in Chinese. *Proceedings of ICSLP*, pages 2025–2028, 2002.