Acoustic Features Extraction for Emotion Recognition

Jia Rong, Yi-Ping Phoebe Chen, Morshed Chowdhury, Gang Li*
School of Engineering and Information Technology
Deakin University, Melbourne, Australia
{jrong, phoebe, muc, gang.li}@deakin.edu.au

***Abstract*— In the last decade, the efforts of spoken language processing have achieved significant advances, however, the work with emotional recognition has not progressed so far, and can only achieve 50% to 60% in accuracy [1]. This is because a majority of researchers in this field have focused on the synthesis of emotional speech rather than focusing on automating human emotion recognition. Many research groups have focused on how to improve the performance of the classifier they used for emotion recognition, and few work has been done on data pre-processing, such as the extraction and selection of a set of specifying acoustic features instead of using all the possible ones they had in hand. To work with well-selected acoustic features does not mean to delay the whole job, but this will save much time and resources by removing the irrelative information and reducing the high-dimension data calculation. In this paper, we developed an automatic feature selector based on a `RF2TREE` algorithm and the traditional `C4.5` algorithm. `RF2TREE` applied here helped us to solve the problems that did not have enough data examples. The ensemble learning technique was applied to enlarge the original data set by building a bagged random forest to generate many virtual examples, and then the new data set was used to train a single decision tree, which selects the most efficient features to represent the speech signals for the emotion recognition. Finally, the output of the selector was a set of specifying acoustic features, produced by `RF2TREE` and a single decision tree.**

Keyword: Feature Extraction, Machine Learning, Ensemble Learning, Twice Learning, Random Forest, Decision Tree

I. Introduction

Since antiquity, people have realized the importance of vocal cues in the expression of emotion, and the powerful effects of vocal emotion expression on interpersonal interaction and social influence. However, in the speech signal processing field, emotional speech recognition is usually considered as a difficult task, because of the lack of comprehensive theoretic knowledge from psychologists or psycholinguistics and the limitation of widely available emotion data sets.

Table I gives a summary of the previous study of acoustic and prosodic features which can be used to encode emotion states by the psychologists and human behavior biologists since 1930's. The correlation between the speech signals and the archetypal states of negative and positive emotions is also shown in Table I.

TABLE I
Emotion States and Acoustic Features

Features	Negative Emotion	Positive Emotion
Pitch	mean ↑ [2] [3] [4] [5], median ↑ [7], range ↑ [7] [4] [8] variability ↑ [7] [4] [9], below normal mean F0 [11] [5], below normal range [3]	mean ↑ [2] [3] [6], range ↑ [9] [10], variability ↑ [9] [11],
Intensity	↑ [2] [4] [6] [12] , ↓ [2], normal [15],	↑ [6] [13] [14]
Duration	rate ↑ [2] [3] [12] [5], rate ↓ [16] [17], slightly slow [3] [18], long pitch falls [2]	rate ↑ [2] [5], slow tempo [6], fast rate [14],
Spectral	high-frequency energy ↓ [15]	high-frequency energy ↑ [6]

Although there are many questions which need to be answered in psychology, it is still widely used as the most important theoretic foundation by a large number of computing-background research groups who have strenuously engaged with the emotional speech recognition during the past decade. The basic acoustic features were the primary choices in the early days. Most of the feature vectors were composed with the simple extracted *pitch*-related, *intensity*-related, and duration-related attributes, such as the maximum, minimum, median, range, and variability values [19] [20] [21] [22]. Other groups preferred to use the pre-processed attributes from different mathematic transforms, such as LPCC [23], LFPC [23] and MFCC [23] [24]. However, it is evident that there are some contradictory results in the existing work, for example, the increase change of the pitch mean value (see *mean* ↑ in Table I) may cause a negative emotion in some cases [4] [5], while it can also cause a positive emotion in the others [6].

Though the items in the candidate acoustic feature vector increase, there is still no conclusive evidence to tell which set of features can provide the best prediction accuracy, because of the different understanding or the various data sets used. One difficulty is that neither psychologists nor psycholinguistics can provide us information on how those features affect emotion states. Another problem is that the large number of potential features and a limited number of emotion data samples makes it very difficult to use empirical method to identify useful features.

With the advances of data mining and machine learning, it is possible to overcome this problem by combining hybrid

* Corresponding author.

6th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2007)
0-7695-2841-4/07 $25.00 © 2007 IEEE

ensemble learning and classification tree together, and extracting the most effective features to represent speech signals for emotion recognition.

In this work, we collected a data set that involved both basic acoustic features and the transformed ones. A well-designed feature selector based on a combined method of a simple decision tree and a random forest ensemble learning proposed in [25] was applied in this paper instead of the well-known feature selection methods: Promising First Selection (PFS), Forward Selection (FS) and Principal Component Analysis (PCA), which all need a huge training data set that tried to cover all possible situations in real-life.

The rest of the paper is organized as follows: the related previous efforts in emotional speech recognition from other computing research group are reviewed in section 2; the details of the data corpus used in the study are depicted in section 3; in section 4, the extraction of the candidate acoustic features is introduced; section 5 discusses the new method used for feature selection, which is combined the simple decision tree and the random forest ensemble learning from `RF2TREE` algorithm; the practical experiments and the results are reported in section 6; finally, in Section 7, we conclude the paper with a perspective analysis of possible future work.

II. Related Works

As mentioned in the section above, the first important task of emotion recognition, is to determine what emotions will be classified in the work. About 50 years ago, researchers began their exploration to represent complex human emotions using a set of basic emotion states [26], also called *primary emotions*.

However, the arguments among scholars about the nature of these primary emotions, or the appropriate amount that should be considered in research work, have continued since the very beginning of the studies. According to the various knowledge and understandings of different research groups, people often assumed a particular set of primary emotions by themselves for their own convenience. Table II shows the primary emotions that were used in previous emotion recognition work.

Due to the uncertainties in the definition of emotion states, we considered two basic ones, *negative* and *positive*. The *neutral* state was also included for comparison.

The same thing occurred in the feature extraction and selection phase (see Table III). Some of the groups only used the raw acoustic features extracted directly from the pure speech signals, such as pitch, intensity and duration parameters without any pre-processing applied. Other people tried to do more. They transformed the original speech wave from the time-domain scale to frequency-domain scale, in order to highlight the potential change trend [23] [24].

TABLE II
Emotional States used in Emotional Speech Recognition

Study Group	Emotion States
McGilloway S. [27]	happiness, anger, fear, sadness
Nicholson J. [28]	joy, anger, fear, sadness, teasing, disgust, surprise, neutral
Nwe T. L. [29]	anger, fear, sadness, dislike, disgust
Scherer K. R. [30]	joy, anger, fear, sadness, disgust, surprise, neutral
Kwon O. [31]	happiness, anger, sadness, bored, neutral
Cai L. [32]	happiness, anger, sadness, surprise
Park C. [33]	anger, laugh, surprise, neutral
Bhatti M. W. [34]	happiness, anger, fear, sadness, surprise, disgust
Hyun K. H. [35]	joy, anger, sadness, neutrality
Lee C. M. [24]	negative, positive, neutral

TABLE III
Feature Vectors used in Emotional Speech Recognition

Study Group	Feature Vector
Dellaert F. [19]	pitch, rhythm, voiced parts, slope
Petrushin V. A. [20]	pitch, bandwidth, energy, duration, formant
Amir N. [21]	pitch, intensity, duration
Kwon O. [31]	pitch, log energy, formant, MFCC
Scherer K. R. [30]	pitch, energy
Cai L. [32]	duration, pitch, formant, power,
Litman D. [36]	pitch, energy, temporal, turn
Lee C. M. [22] [37]	pitch, energy, duration, formant

III. Data

We use two databases in this paper: the acted speech corpus and the natural speech corpus. As shown in Table IV, the first database, there are three selected speakers who were asked to read 443 sentences with acted emotions in Chinese. The second database contains 172 natural speech sentences recorded from existing movie products. All the speech corpus are recorded in the format of 16 kHz, 128 kbps, 8 bit and single track.

Three different languages were used to test the content-independence and the cross-cultural robustness of the proposed method. Although both Mandarin and Cantonese are spoken in China, they are quite different in grammar, tone and pronunciation.

Both of the two databases contain three different emotion states: *negative*, *positive* and *neutral*. For the database with acted speech corpus, the speakers were asked to act three negative emotions: *anger*, *fear*, and *sadness*; while another two positive emotions (*happiness* and *excitation*) were also requested.

A. Acted Speech Corpus

There are 443 sentences of emotional speech recorded by three non-professional speakers, one male and two females, in this acted speech corpus. The total number of the text sentences to be read is 53, and each of them is repeated at least 8 times. All the samples are recorded under a noise-free laboratory environment.

The distribution for the three emotional states is: 248 negative examples (56%), 106 positive examples (23%), and

TABLE IV
Databases used in the experiments

DB	Language	Emotion	♯Sp.	♯Sen.	Avg. Sen.
I	Mandarine (Chinese)	Acted	2	337	147
	Shanghai (Chinese)	Acted	1	106	106
II	Am. English	Natural	6	121	20
	Cantonese (Chinese)	Natural	5	51	10

6th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2007)
0-7695-2841-4/07 $25.00 © 2007 IEEE

IEEE
COMPUTER
SOCIETY

53 neutral examples (11%). This data set was used as the training data set in the experiment.

The purpose of this database is used to construct the original training data set to build the classification model and also to provide a comparison with the natural speech corpus.

B. Natural Speech Corpus

The natural speech corpus contains 172 sentences of emotional speech that were recorded. We segmented and labelled the sentences manually. Each sentence was only spoken by one person. We did not remove the background noise in the original recording files, that is, the dialogue represents the real environment of daily life.

There are 72 examples (41%) for the negative state, 45 examples (26%) for the positive state, and 55 examples (32%) for the neutral state in this database.

Natural speech corpus was used as the testing data set to test whether or not the classification method we proposed can also work well in the real dialogue environment.

IV. Pre-processing and Basic Feature Extraction

As mentioned in the previous section, the continuous acoustic features extracted directly from the original speech signal are usually considered as the most reliable features for efficient emotion speech recognition. In this section, the basic acoustic features, such as *pitch*-related, *intensity*-related, and *duration*-related features, are described in detail. In addition, a set of mathematically transformed features by two of the popular transformation methods, Discrete Fourier Transform and Mel-scale Frequency Cepstral Coefficients, are also considered.

A. Basic Acoustic Features Extraction

No matter what features are selected by the various feature selection methods in the previous work done by any research group, they are all derived from a set of basic acoustic features.

The basic acoustic features are those features we can extract directly from raw speech signals, which are in accordance with most studies in this field.

- **Pitch-related features** mean value, maximum, minimum, median, standard deviation, range (difference between maximum and minimum), variance value and change rate. This group of features are usually considered as the most important features that capture monotone speech or highly accented syllables and achieved good results in a wide range of applications.
- **Intensity-related features** normal intensity features (mean value, maximum, minimum, median, standard deviation, range, variance value and change rate) calculated by summing up the absolute values for each data frame; and the relative intensity features (mean value, maximum, minimum, median, standard deviation, range, variance value and change rate). *Intensity*-related features are usually used to capture the voice power in speech, and the normal values (energy) and relative value (powerDb)can be treated separately.
- **Duration-related features** cross zero rate features, such as mean value, maximum, minimum, median, standard deviation, range, variance value and change rate. CZR features are the radios that count the times of the speech wave crossing the zero point, which can capture the temporal characteristics in the prosody.

B. Mathematic Transform on Basic Acoustic Features

- **Discrete Fourier Transform (DFT)**
  Sometimes it is hard to identify the frequency components by only considering the original signals; the signals in time domain are normally converted to the frequency domain by taking the Fast Fourier Transform (FFT). The Discrete Fourier Transform function implements the transform of the feature vectors of length N by:

  $$X(k) = \sum_{j=1}^{N} x(j)\omega_N^{(j-1)(k-1)}$$

  where $\omega_N = e^{(-2\pi i)/N}$ is an Nth root of unity. After applying DFT on the speech signals, the feature vector is enlarged with another 56 features (8 pitch-related, 8 energy-related, 8 powerDb-related, 8 zcr-related, phase-related, powerDb-related.The DFT processed Phase and PowerDb features were obtained to cover all acoustic features obtained, but the pure phase-related features and powerDb-related features were discarded for showing scarcely any effects.
- **Mel-scale Frequency Cepstral Coefficients (MFCC)**
  This is another popular signal transform method that was used for feature extraction in speech recognition work in recent decades [23] [24].

  $$C_m = \sum_{k=1}^{N} cos[m \times (k - 0.5) \times \pi/N], m = 1, 2, \ldots, 12$$

  MFCC generated 12 new features based on pitch, phase and powerDb features.

There are 32 basic acoustic features plus the transformed 52 features, therefore, the total number of the candidate acoustic features for further selection is 84.

V. Ensemble-learning Based Feature Selection Method

A. The Algorithm

In this study, we proposed a new method that combined the random forest ensemble and simple decision tree together. Decision tree is a well-known data mining and machine learning algorithm adapted in 1986 (ID3 algorithm by Quinlan), while the random forest ensemble was developed by Zhou *et al.* in [25].

As a traditional classification algorithm that is easy to understand and is used widely in data mining and machine learning studies and other applications as well, we do not repeat the algorithm of decision tree in this paper. Our focus is centralized on the random forest ensemble learning as described in `RF2TREE` method.

6th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2007)
0-7695-2841-4/07 $25.00 © 2007 IEEE

We use the random forest ensemble learning method used in `RF2TREE` to train our datasets in the experiment to take advantage of its good performance to work with small data sets. The data sets we had contained 615 examples but the number of candidate features was 84. Because of the complex representation of the speech waves, it is impossible to collect a huge number of speech examples that might cover all situations in daily life. Therefore, `RF2TREE` can save much work time and human resources.

Algorithm 1

```
Require: training set S, decision tree algorithm L
Ensure: F = Random Forest, DT = Decision Tree
  $F = RandomForests(S, L)$, {train a random forest F from S via RandomForests}
  $DT = L(S)$, {train a decision tree DT from S}
  $S' = GenVirExp(F)$, {use the random forest to enlarge the training data set by generating the virtual examples as more as possible}
  $DT' = L(S \cup S')$, {grow a decision tree from the enlarged training data set}
  $FS = DT \cup DT'$, {generate the selected feature set by combine the outputs of two decision trees}
```

The generated random forest consists many decision trees, and each tree is grown in the same way: suppose the number of training examples is *n*, each example has *m* variables; and then randomly sample *n* examples with replacement from the original training dataset; specify a number $l \ll m$ so that at each tree node, *l* variables are randomly selected out of the *m* variables and the best one on these *l* variables is used to split the node; therefore, each tree is grown to the largest extent possible, and the new instances are generated in a style that all the possible values of different variables that have not appeared in the original training data set are tried.

Then, each of the selected candidate feature variables was ranked by a voting system, in which the selected features from decision tree and random forest ensemble were sorted on the weights they gained. The bigger the weight value the feature variable has, the higher possibility it will be chosen.

The output of the described algorithm is a set of selected variables, which are considered as the most efficient features for emotion recognition in our study.

B. Justification

Suppose we denote X as the input space and the set of the class labels is Y, then our target is to work out a function: $F : X \rightarrow Y$. Let F_T denote the function implemented by a decision tree trained on a given training data set, and its probability to approach F can be expressed as:

$$P_{F_T} = P_{F=F_T} = 1 - P_{F \neq F_T} = 1 - err_T \quad (1)$$

where err_T denotes the error rate of the decision tree. In the same way described in [25] err_T can be broken into three parts: err_T^c, err_T^n and err_T^s.

$$err_T = err_T^c + err_T^n + err_T^s \quad (2)$$

err_T^c is an error term caused by the limited learning ability of the decision tree; err_T^n is the an error term caused by the noise contained in the training data set; while err_T^s is an error term caused by the limitation of the finite samples.

Since err_T^c can be extremely small, and the noise can be removed by data pre-processing, it is obvious that the performance of a decision tree is usually restricted by the training data set that may not contain a sufficient amount of the data samples to capture the target distribution. That is, err_T can be dominated by err_T^s principally.

On the other hand, we can also obtain the probability to approach the function F_E implemented by a random forest ensemble trained on the same training data set as following equations:

$$P_{F_E} = P_{F=F_E} = 1 - P_{F \neq F_E} = 1 - err_E \quad (3)$$

$$err_E = err_E^c + err_E^n + err_E^s \quad (4)$$

Unlike the simple decision tree, the error term caused by limitation of the finite samples, err_E^s is much smaller than err_T^s. Due to the fact that the original training data set is enlarged by using a random forest ensemble, err_E^s is decreased at any rate. However, the error rate caused by the limited learning ability may be enhanced in the generated training data set, which does not contain all possible feature vectors. That is, assuming the noise error rate can be ignored, err_E is not only dominated by err_E^s, but also by err_E^c.

If we use F_{TE} as the function implemented by the combined model for both RF2TREE and decision tree, then the probability to approach F_{TE} can be expressed as:

$$P_{F_{TE}} = P_{F=F_{TE}} = 1 - P_{F \neq F_{TE}} = 1 - err_{TE} \quad (5)$$

where

$$err_{TE}^{*} = err_{TE}^{c*} + err_{TE}^{s*} \quad (6)$$

According to the above justification, $P_{F_{TE}}$ can be greater than either P_{F_T} or P_{F_E} so far as the following equations are satisfied:

$$err_T^{s*} < err_T^{s*}; \quad err_{TE}^{c*} < err_E^c \quad (7)$$

Therefore, it shows that the performance can be improved if we combine a decision tree with the random forest ensemble method used in `RF2TREE` together. The experiments described in the rest of the paper also verified this fact.

VI. Experiments

A. Feature Selection Process

In the experiment, we extracted 84 features for each instance to develop the training data set. The total number of instances in the data set was 615. To have equal prior probability, the instances in the data set were selected 10 times in a random manner to avoid an imbalance selection situation.

After the data pre-processing, each speech sample in the database was converted into a piece of data record with a high dimension of 84 variables. Then we applied a classifier described in the previous section to the data sets using CrossValidate-10 which set up the training data set (about

IEEE COMPUTER SOCIETY

TABLE V

RECOGNITION ACCURACY RESULTS

Classifier	Overall database		Acted database		Natural database	
	84-feature	29-feature	84-feature	29-feature	84-feature	29-feature
C4.5	74.90	76.51	88.82	88.81	39.60	46.20
Random Forest	78.86	80.58	91.42	92.96	47.71	51.08
RF2Tree	73.17	78.05	85.56	88.67	43.60	46.51

90%) and the testing data set (about 10%). As the output of the classifier, the 29-feature data set (see Table VIII) was generated.

To test the performance of this specifying set of 29 features, a simple `C4.5` classifier and a pure RandomForest classifier were also applied on the generated 29-feature data sets and the original 84-feature data sets. All the classification results listed here were the average values implemented for 100 runs.

Additionally, we specially developed a test data set that only consisted 172 segments taken from the existing movie products, which has a similar dialog environment that is very close to what we had in our daily life. The tested result of this data set shows the performance of the final selected feature to be applied in the practical cases.

B. Results

Table V shows the results classified by 3 algorithms related to our experiment, $C4.5$, *Random Forest* and `RF2TREE`. Each algorithm was applied to both 84-feature data sets, and the selected 29-feature data sets. Learning from the figures in the table, we can find that it is possible to use less well-selected features to achieve the same and even better performance for emotional speech recognition.

Furthermore, the ability to detect an individual emotional state is improved by using the set of the selected 29 features rather than all possible feature vectors. The confusion matrices of the natural speech corpus are shown in Table VI and Table VII, respectively. The recognition rates for both negative emotion and neutral emotion are increased ($55.1\% \rightarrow 57.3\%, 41.9\% \rightarrow 45.9\%$ in average). In contrast, the classification rate of positive emotion performs worse for the 29-feature set due to the higher misclassification of negative and positive (negative → positive: $45.9\% \rightarrow 47.2\%$).

By the comparison of the selected 29-feature set, shown in Table VIII with the original 84-feature set, there are some emotion categories we can follow:

- Although it seems that we can obtain a lot of information from the speech signals, a majority (55 out of 84) are actually irrelevant and can be ignored in practice.
- Pitch-related, Energy-related features and PowerDb-related features are the most important among all the features, which is the same conclusion as mentioned in psychology studies. Nearly two-parts of the selected features are from these three groups.
- Phase-related features have poor performance, which has shown the least effect on.
- the basic acoustic features still work well together with the processed ones. More than half of the selected features are the basic acoustic ones.

TABLE VI

CONFUSION MATRIX FOR **84**-FEATURE TESTING DATA SET

	Negative	Positive	Neutral
Negative	**55.1**	23.6	21.3
Positive	45.9	**28.9**	25.2
Neutral	34.5	23.6	**41.9**

TABLE VII

CONFUSION MATRIX FOR **29**-FEATURE TESTING DATA SET

	Negative	Positive	Neutral
Negative	**57.3**	20.8	21.9
Positive	47.2	**27.3**	25.5
Neutral	34.5	19.6	**45.9**

According to the above analysis of the experiment results, it can be summarized that a set of specifying acoustic features performs well, since it carries most of the useful information we need for emotion recognition.

VII. CONCLUSION

Emotional expression is a normal instinct for human beings, but from the previous work on emotional speech recognition, we can find that it is really difficult for machines to acquire due to the difficulties of patterning vocal cues.

However, according to a century of study in behavioral biology, psychology, the speech and communication sciences, and the data mining and machine learning methods that can be used for emotional speech recognition, we believe that we can find a satisfactory solution to the above problems.

In the study for this paper, we started our work by extracting the basic acoustic features on the two speech corpus we collected, such as *pitch*-related, *intensity*-related, and *duration*-related features. We extracted 32 basic features and processed

TABLE VIII

THE SET OF THE SELECTED SPECIFYING ACOUSTIC FEATURES

pitch	pitch_max	pitch_mean	pitch_min
	pitch_median	pitch_range	pitch_std
	fft_ pitch_changerate	pitch_changerate	
intensity	energy_max	energy_min	energy_median
	fft_energy_changerate	powerDb_max	powerDb_mean
	powerDb_min	powerDb_range	fft_powerDb_max
	fft_powerDb_mean	fft_powerDb_std	
duration	zcr_min	zcr_median	zcr_std
	fft_zcr_mean	fft_zcr_min	fft_zcr_median
phase	fft_phase_min		
MFCC	mfcc2	mfcc3	mfcc4

6th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2007)
0-7695-2841-4/07 $25.00 © 2007 IEEE

DFT and MFCC transforms on them to get the candidate set of 84 features.

Then we designed and developed a classification method, by which we could combine the advantages from a simple decision tree and a random forest ensemble learning. It is justified that the developed method can reduce the error rates caused by both the limitation of the finite data samples of the decision tree method and the restricted learning ability of the random forest ensemble learning. That is, the training data set can be enlarged by generating new virtual data samples while the most important information in the acoustic features needed for classification is also kept after the process.

Unlike the traditional feature selection methods, such as PFS, FS and PCA, our method can be applied on a small database that may contain the finite data samples but with high-dimensional feature vectors. Another advantage of our method is that it can be applied on the data set in which the variables are largely independent of one another.

The experiment results indicate that it is still possible to improve the performance of the emotional speech recognizor. Future work could be done in the following aspects:

- a human listening test may be taken into future experiments to strengthen the reliability of our results.
- explore more emotion states to be recognized, such as including four or six emotional classes.
- continue the work to the next classification stage, test the performance of the specifying feature vectors with the other recognition methods, such as Fuzzy model, Neural Network, or HMM.
- finally, the accuracy results of the acted speech corpus and the natural speech corpus have a big difference, which may be caused by some non-technique problems, such as the limited amount of the speech corpus collected in the experiment. One solution to this problem is to collect more data samples or even to build a public emotional speech database in uniform formats. To achieve this target, we need the cooperation of the other research groups and to share the resources.

REFERENCES

[1] M. Pantic and L. J. M. Rothkrantz. Toward an affect-sensitive multimodal human-computer interaction. *PROCEEDINGS OF THE IEEE*, 91(9), 2003.

[2] J.R. Davitz. Personality, perceptual, and cognitive correlates of emotional sensitivity. *in The Communication of Emotional Meaning*, 1964.

[3] I. Fónagy. A new method of investigating the perception of prosodic features. *Lang. Speech*, 21, 1978.

[4] C.E. Williams and K.N. Stevens. On determining the emotional state of pilots during flight: An exploratory study. *Aerospace Medicine*, 40, 1969.

[5] R. Coleman and R. Williams. Identification of emotional states using perceptual and acoustic analyses. *Care of the Professional Voice*, 1, 1979.

[6] R. Van Bezooijen. *Characteristics and Reognizability of Vocal Expressions of Emotions*. Dordrecht, The Netherlands: Foris, 1984.

[7] G. Fairbanks and W. Pronovost. An experimental study of te pitch characteristics of the voice during the expression of emotion. *Speech Monograph*, 6, 1939.

[8] W. L. Hoffe. On the relation between speech melody and intensity. *Phonetica*, 5, 1960.

[9] Z. Havrdova and M. Moravek. Changes of the voice expression during suggestively influenced states of experiencing. *Activitas Nervosa Superior*, 21, 1979.

[10] I. Fónagy and K. Magdics. Emotional patterns in intonation and music. *A. Phonet. Sprachwiss. Kommunikationsforsch.*, 16, 1963.

[11] K. Sedlacek and A. Syhra. Speech melody as a means of emotional expression. *Folia Phoniatrica*, 15, 1963.

[12] G. Kotlyar and V. Mozorov. Acoustic correlates of the emotional content of vocalized speech. *J. Acoust. Academy of Sciences of the USSR*, 22, 1976.

[13] G.L. Huttar. Relations between prosodic variables and emotions in normal american english utterances. *J. Speech and Hearing Research*, 11, 1968.

[14] D. Crystal. *The English Tone of Voice*. Edward Arnold, London, UK, 1975.

[15] N. Tsapatsoulis G. votsis S. Kollias W. Fellenz R. Cowie, E. Douglas-Cowie and J. Taylor. Emotion recognition in human-computer interaction. *IEEE*, 18(1):32–80, Jan 2001.

[16] C.E. Williams and K.N. Stevens. Emotions and speech: Some acoustical correlates. *J. Acoust. Soc. Am*, 52, 1972.

[17] J. Sulc. Emotional changes in human voice. *Activitas Nervosa Superior*, 19, 1977.

[18] K. Scherer W. Johnson, R. Emde and M. Klinnert. Recognition of emotion from vocal cues. *Arch. Gen. Psych.*, 43, 1986.

[19] T. Polzin F. Dellaert and A. Waibel. Recognizing emotion in speech. *ICSLP*, 1996.

[20] V.A. Petrushin. Emotion recognition in speech signal: Experimental study, development, and application. *ICSLP*, 2000.

[21] N. Amir. Classifying emotions in speech: A comparison of methods. *EUROSPEECH*, 2001.

[22] S. Narayanan C.M. Lee and R. Pieraccini. Recognition of negative emotions from the speech signal. *IEEE Automatic Speech Recognition and Understanding Workshop*, 2001.

[23] C. Chen M. Song, J. Bu and N. Li. Audio-visual based emotion recognition-a new approach. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR0́4)*, (1063-6919/04), 2004.

[24] M. Bulut C.M. Lee, S. Yildirim and A. Kazemzadeh. Emotion recognition based on phoneme classes. *IEEE*, 2005.

[25] Z.H. Zhou. Mining extremely small data set on software reuse. 2004.

[26] S. Tomkins. Affect, imagery, consciousness springer publishing company. 1962.

[27] R. Cowie S. McGilloway and E.D. Cowie. Prosodic signs of emotion in speech: Preliminary results from a new technique for automated statistical ananlysis. *ICPhS 95*, 3, 1995.

[28] K. Takahashi J. Nicholson and R. Nakatsu. Emotion recognition in speech using neural networks. *IEEE*, (0-7803-5871-6/99), 1999.

[29] F.S. Wei T.L. Nwe and L.C. De Silva. Speech based emotion classification. *IEEE*, 2001.

[30] G. Rigoll B. Schuller and M. Lang. Hidden markov model-based speech emotion recognition. *ICASSP, IEEE*, II:1–4, 2003.

[31] J. Hao O. Kwon, K. Chan and T. Lee. Emotion recognition by speech signal. *Eurospeech*, Sep 2003.

[32] Z. Wang L. Zhao L. Cai, C. Jiang and C. Zou. A method combining the global and time series structure features for emotion recognition in speech. *IEEE Int. Conf. Neural Networks and Signal Processing*, (0-7803-7702-8/03), 2003.

[33] C.H. Park and K.B. Sim. Emotion recognition and acoustic analysis from speech signal. *IEEE*, (0-7803-7898-9/03), 2003.

[34] Y. Wang M.W. Bhatti and L. Guan. A neural network approach for human emotion recognition in speech. *IEEE International Symposium on Circuits and Systems*, 2003.

[35] E.H. Kim K.H. Hyun and Y.K. Kwak. Improvement of emotion recognition by bayesian classifier using non-zero-pitch concept. *IEEE International Workshop on Robots and Human Interactive Communication*, (0-7803-9275-2/05), 2005.

[36] D. Litman and K. Forbes-Reley. Recognizing emotion from student speech in tutoring dialogues. *ASRU-IEEE Automatic Speech Recognition and Understanding Workshop*, 2003.

[37] C.M. Lee and S. Narayanan. Emotion recognition using a data-driven fuzzy inference system. 2003.

6th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2007)
0-7695-2841-4/07 $25.00 © 2007 IEEE