International Journal of Electrical, Electronic and Communication Sciences
ISSN: 2517-9438
Vol:4, No:12, 2010

Applications of Support Vector Machines on Smart Phone Systems for Emotional Speech Recognition

Wernhuar Tarng, Yuan-Yuan Chen, Chien-Lung Li, Kun-Rong Hsie and Mingteh Chen

***Abstract*—**An emotional speech recognition system for the applications on smart phones was proposed in this study to combine with 3G mobile communications and social networks to provide users and their groups with more interaction and care. This study developed a mechanism using the support vector machines (SVM) to recognize the emotions of speech such as happiness, anger, sadness and normal. The mechanism uses a hierarchical classifier to adjust the weights of acoustic features and divides various parameters into the categories of energy and frequency for training. In this study, 28 commonly used acoustic features including pitch and volume were proposed for training. In addition, a time-frequency parameter obtained by continuous wavelet transforms was also used to identify the accent and intonation in a sentence during the recognition process. The Berlin Database of Emotional Speech was used by dividing the speech into male and female data sets for training. According to the experimental results, the accuracies of male and female test sets were increased by 4.6% and 5.2% respectively after using the time-frequency parameter for classifying happy and angry emotions. For the classification of all emotions, the average accuracy, including male and female data, was 63.5% for the test set and 90.9% for the whole data set.

***Keywords*—**Smart phones, emotional speech recognition, social networks, support vector machines, time-frequency parameter, Mel-scale frequency cepstral coefficients (MFCC).

I. INTRODUCTION

WITH the popularity of smart phones and the 3G mobile Internet, mobile communications have already become an integral part of modern life. Mobile Internet is the use of high-speed communication networks to achieve the interactive activities and transactions among people. Skiba *et al.* [1] defined mobile communication technologies as the utilization of wireless terminal equipment for web browsing, email delivery, online trading, news reading, radio listening, voice and video conferencing services, etc. According to the statistics of "Opera Mini", the world's annual growth rate of mobile Internet has increased substantially in recent years due to the development of wireless networks and the use of smart phones, indicating the mobile Internet as a global trend.

Wernhuar Tarng, Yuan-Yuan Chen, and Chien-Long Lee are with the Institute of Computer Science, National Hsinchu University of Education, 521 Nanta Rd., Hsinchu, Taiwan 300, R. O. C. (Phone: 886-3-5213132#5903; Fax: 886-3-5622918; e-mail: wtarng@mail.nhcue.edu.tw).

Kun-Rong Hsie and Mingteh Chen are with the Communication Engineering Department, Chung Hua University, 707, Sec.2, Wu-Fu Rd., Hsinchu, Taiwan 300, R.O.C. (e-mails: kr.hsieh@chu.edu.tw, michael@emulationlab.com).

Internet community is a social phenomenon formed by a group of people communicating with each other through various electronic devices. The social networks are widespread and not restricted by time or space, so people can join together to form an Internet community group based on information sharing and emotional support. As the growth of micro blogging services, lots of people are using mobile phones for the access of wireless Internet and they are more active than the other users. Recently, micro blogging services such as Twitter, Facebook, and Plurk have become fast growing and very popular social networking websites. Therefore, the combination of smart phones and social networking websites can increase the interaction of people and their community groups.

Speech communication is an important way to express one's ideas and emotions. In addition to semantic statements, the accent and intonation in a sentence is also a very important part of showing emotions. In the course of a dialogue, the same words by a different tone of presentation can cause the feeling of different emotions. Thus, it is an important issue in emotional speech recognition to obtain emotional data from the features of speech such as pitch and volume for analysis. For example, if the speech features in a phone call can be obtained and analyzed using machine learning techniques, then one can recognize the speaker's emotional status and make a suitable response.

In this study, the support vector machine was used to develop a mechanism for recognizing the major emotions of human speech, including happiness, anger, sadness and normal. To reduce the computation time and complexity, the mechanism divides various parameters into the categories of energy and frequency and uses a hierarchical classifier to adjust the weights of acoustic features for training. In addition, a time-frequency parameter obtained by the continuous wavelet transform was used to identify the accent and intonation in a sentence during the recognition process to achieve better results.

In this study, the Berlin Database of Emotional Speech was used by dividing the speech into male and female data sets for training and testing. An experiment was conducted to compute the average accuracies of the test set and whole data set. The emotional speech recognition mechanism proposed in this study can be used on smart phones to combine with social networking websites and the functions of micro blogging services, allowing users to integrate more easily with their

community groups so as to increase the interaction and concern among people.

II. Related Researches

In this section, the related researches about emotional speech recognition are described, including classification of emotions, emotional speech databases, speech features and the classifiers used in emotional speech recognition.

1.1 Classification of Emotions

To investigate emotional speech recognition, one must first understand the classification of emotions. Human beings are born with emotional reactions, which have a great impact on their mental and physical health as well as social relationships. Psychologists have conducted a lot of researches in emotion classification, and they tried different models to analyze and classify emotions, for example, (1) Discrete model: the most representative one is the circumplex model of personality and emotions proposed by Plutchik [2] in 1980. He divided the basic human emotions into eight types: trust, joy, anticipation, anger, disgust, sadness, surprise and fear, which were combined into ther more complex emotions, to distinguish and explain the complex human emotions. (2) Dimensional model: in 1980, Russell [3] used the distribution of different arousal and pleasure levels in a 2D coordinate system to distinguish between 28 adjectives of emotions (Figure 1), where each quadrant contains a number of emotional adjectives. The model can be used to explain the performance of different emotions. In this model, arousal expresses the intensity of emotions and its performance ranges from "excited" to "calm"; pleasure stands for the evaluation of emotions and its performance ranges from positive to negative. For example, both "happy" and "angry" have higher levels of arousal but different levels of pleasure, and thus they reveal different emotional results.

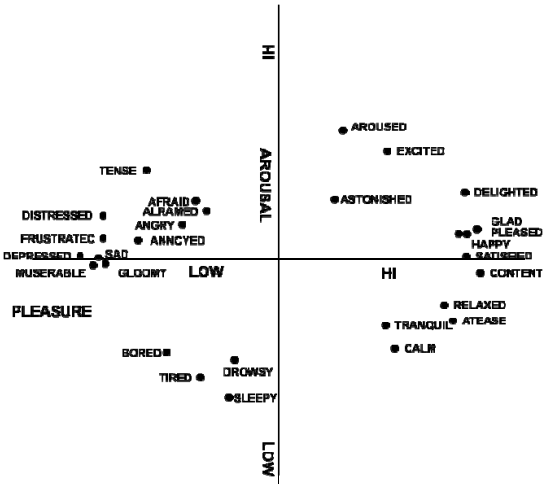

Fig. 1. Russell's arousal-pleasure emotional model

(3) Biological and neural model: Posner *et al*. [4] classified human emotions based on the neurological point of view. They recorded the statuses in different parts of cerebral cortex and analyzed the data based on the reactions of human brains to different emotions. Yang's study [5] pointed out that an emotion is not just a feeling and it is a reaction from human body to help the survival mechanism. Human emotions are generated directly by sending messages from the amygdala to the cerebral cortex or indirectly by hormone secretion through the hypothalamus to the body, causing physiological changes such as muscle contraction and the increase of blood pressure and heart rate. These physiological changes are feedbacked to brain's sensory cortex, which then sends the messages to the frontal lobes. Thus, the physiological changes are often explained as emotions, and the records of physiological responses can also be used as a form of emotional classification.

1.2 Emotional Speech Databases

In general, the emotional speech databases can be divided into three kinds based on the ways they are produced, i.e., natural emotional speech, simulated by actors and actresses, and directed by human-machine interaction, and different kinds of databases may result in different recognition rates. Hun and Zhang [6] found that a higher recognition rate (about 90%) for angry and normal emotions could be achieved if the emotional speech data were simulated by actors and actresses, and the recognition rate was reduced to 75% if the speech data were simulated by a normal person. The recognition rate dropped to 65% when using real emotional speech data. In this study, the Berlin Database of Emotional Speech [7] was used for training and testing. The speech data were produced from 10 irrelevant sentences, with 16-bit audio format and 16 KHz sample rate in WAV files, and the length of each sentence ranges from 3 to 8 seconds. The speech data were simulated by 5 actors and 5 actresses and divided into 7 emotional categories, namely, angry, tired, disgusting, fear, happiness, sadness and normal.

1.3 Speech Features

The purpose of feature selection is to determine the most representative features in emotional speech data as the basis for recognition, so it is very important for increasing the accuracy of the results. In the process of emotional speech recognition, this study tried to identify the features of emotional speech from the dynamic audio signals. In addition to the three elements of human voice [8], i.e., timbre, intensity (or amplitude), and pitch (or frequency), Table 1 also lists the features often used in the related researches. Among them, Mel-scale frequency cepstral coefficients (MFCC) contain several features commonly used in emotional speech recognition. Also, through the conversion of MFCC, the frequencies of perceived audio signals can be closer to the actual physical experience of human hearing.

Table I The features used in the related researches

Features	Researches
Frequency, Amplitude, Duration, Formant	Cai [9]
Frequency, Amplitude, Formant, MFCC	Kwon [10]
Frequency, Amplitude	Schuller [11]
Frequency, Amplitude, MFCC	Vogt [12]
Frequency, Amplitude, Duration, Formant, Bandwidth	Petrushin [13]

1.4 Classifier

In this study, the emotional speech data were analyzed with machine learning methods after feature extraction. There are a few commonly used speech recognition techniques, including Gaussian mixture model (GMM) [14], hidden Markov model (HMM) [15], linear discriminant analysis (LDA) [16], k-nearest neighbor (KNN) algorithm [17, 18], neural network algorithms [19] and support vector machines (SVM) [20]. These classifiers are of different approaches and have different features. After selecting the feature vector, the classifiers can use the speech database for training their feature models, and then add a new data set to perform emotional speech recognition.

III. Research Method

The frequency of human voice ranges from 60Hz to 1000Hz, but the human ear can receive the audio between 10Hz and 20K Hz in which 1.2KHz~1.4KHz is the most sensitive frequency band. In the process of recording, the human voice is converted into analog signals through a microphone, and then transformed into digital signals through the analog to digital converter (ADC) for processing and analysis. The flowchart of emotional speech recognition used in this study is shown in Figure 2, where the pre-processing step extract the features from the speech data to generate representative data (or feature vector), which can be used by the classifier for training the feature model. Then, the new data set can be added to obtain the recognition results.

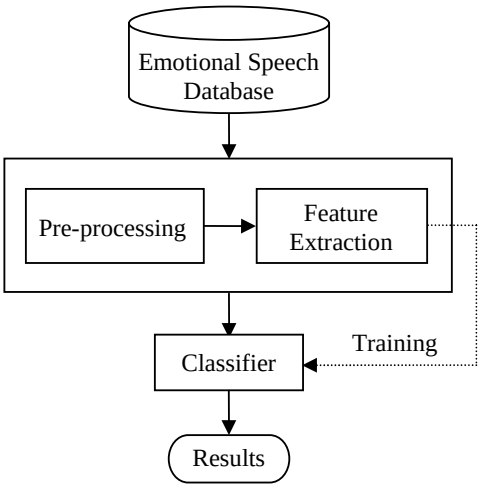

Fig. 2. The flowchart of emotional speech recognition

The emotional speech data used in this study were selected from 4 categories of the database, i.e., angry, happy, sad and normal emotions. The data contained a number of 339 speech samples (189 male samples and 150 female samples), in which 75% of the samples were used as training set, and 25% of the samples were used as test set (Table 2).

Table II The emotional speech data used in this study

	Anger	Happiness	Sadness	Normal	Total
Male	67	45	40	37	189
Female	60	27	38	25	150
Total	127	72	78	62	339

2.1 Pre-processing

Before extracting the features, the original emotional speech data have to go through the pre-processed step (Figure 3), which converts the speech data of various lengths into a simple format containing several frames with the same size through signal compensation and segmentation. Therefore, it is easier for the feature extraction and classification in the follow-up steps.

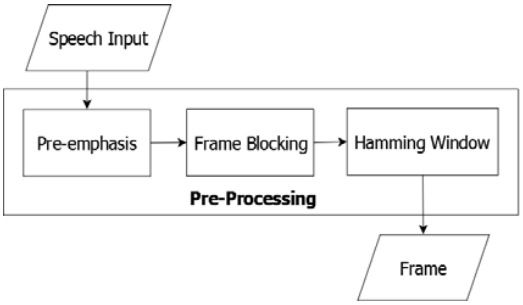

Fig. 3. The pre-processing step to extract speech features

The purpose of pre-emphasis is to compensate for the loss of high-frequency portion due to the conversion from analog to digital signals such that the amended spectrum is closer to the human auditory system. To facilitate the signal analysis and processing, the emotional speech data are first segmented into a number of audio frames of smaller time units. According to Rabiner and Schafer [21], the selection of frame size around 10-20ms is the most appropriate because the frame must contain at least two basic cycles of the audio signals, and there are some overlapped portions between the adjacent frames to prevent the audio signals from changing too much. The final part of pre-processing step is to multiply the frame signals by a fixed Hamming Window function (Figure 4), which aims to amplify the central part of the frame signals and reduce the gaps at the junctions of the frames. After the completion of pre-processing step, speech signals can be converted into concentrated signals in small-size audio frames of the same time unit.

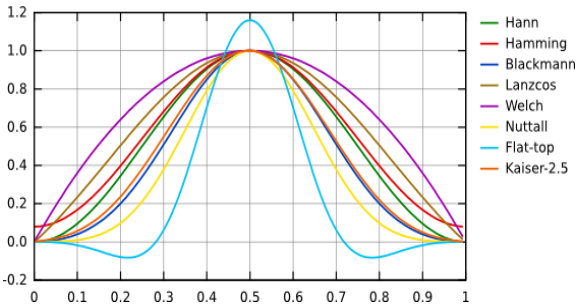

Fig. 4. The window functions used in speech processing

2.2 Time-frequency Feature

In general, the methods of emotional speech recognition first divide the speech data into a number of shorter frames, and then analyze the segmented frames based on the extracted features. Therefore, this study proposed a time-frequency parameter

obtained by continuous wavelet transforms to identify the accent and intonation in a sentence during the recognition process. The wavelet transform is a powerful tool to analyze non-stationary signals in time and frequency domains. After the speech data are processed by continuous wavelet transforms using Morlet mother wavelet, the maximum wavelet transform coefficient at each sampling point is selected among all depth layers. Also, the threshold is set to eliminate the high-frequency part of the signals for reducing the impact of unvoiced sounds. Finally, the time for the maximum instantaneous frequency of normalized audio signals to occur in a sentence is calculated and used as an important parameter of speech features.

The following figures show the results of wavelet transform performed on the emotional speech data, which were produced by the same actors speaking the same words to express different emotions. In these figures, the upper part is the analyzed signals, and a lower scale means a higher frequency and thus a better time resolution; the middle part shows the changes of wavelet transform coefficients in coloration mode; the lower part shows the time when the highest frequency occurs.

In Figure 5, the result of wavelet transform for angry emotion is displayed and it can be seen that the highest frequency occurs at about 0.5 seconds. Figure 6 is the result of happy emotion, and the maximum frequency occurs at 2.7 seconds. Figure 7 is the result of sad emotion and the highest frequency occurs at 1.1 seconds. Figure 8 is the result of normal emotion and the highest frequency occurs at 1.6 seconds. Thus, the time for the highest frequency to occur in emotional speech data varies when the speaker has different emotions. In addition to the 28 features obtained from the segmented frames, the time-frequency parameter is also used as a feature to identify the emotions of the speaker for achieving better recognition results.

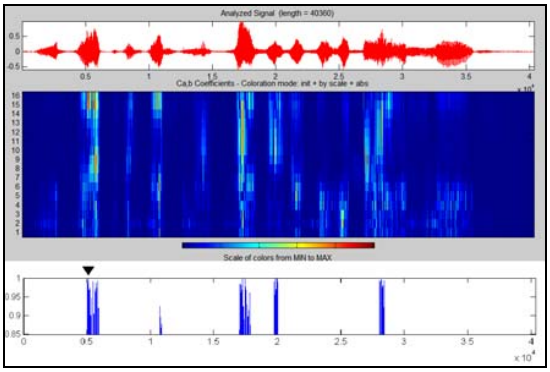

Fig. 5. Result of wavelet transform for angry emotion

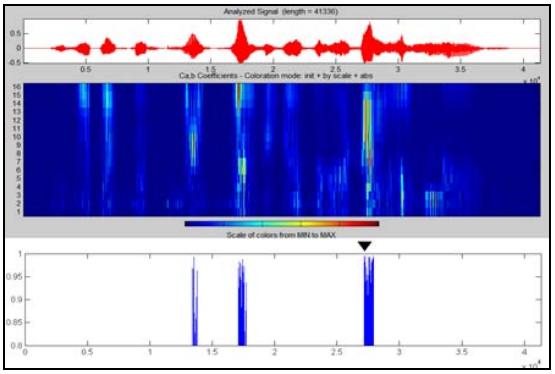

Fig. 6. Result of wavelet transform for happy emotion

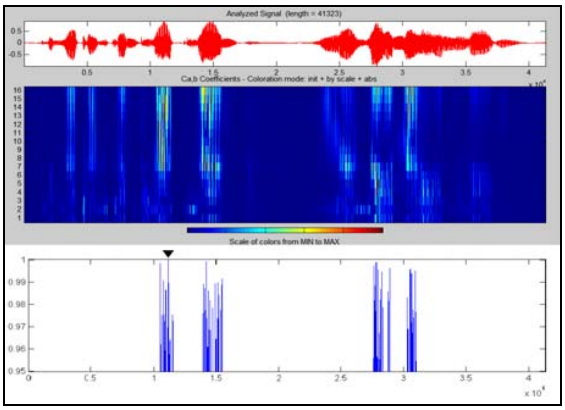

Fig. 7. Result of wavelet transform for sad emotion

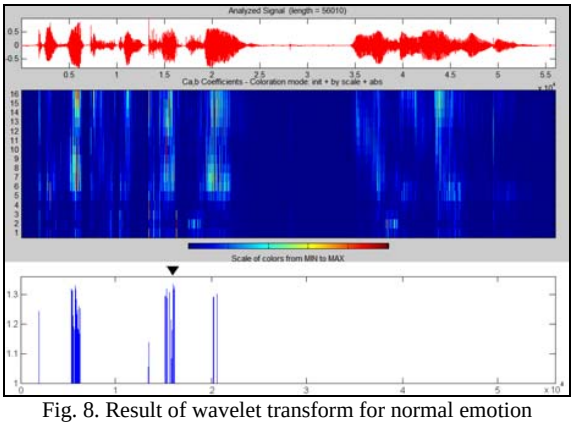

Fig. 8. Result of wavelet transform for normal emotion

2.3 Hierarchical Classification

Considering of the relationship between emotions and their features as well as the computational complexity, this study decided to adopt the hierarchical classification to deal with the features of arousal and pleasure separately (Figure 9).

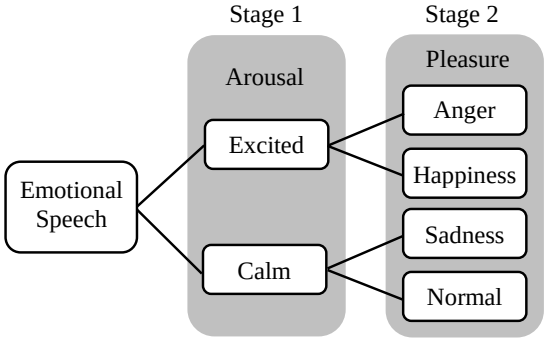

Fig. 9. Two-stage hierarchical classification

The first stage performs classification based on the arousal features, using the energy-related features for training to divide the speech data into excited and calm categories. The second stage performs classification based on the frequency features. The excited category is divided into happy and angry emotions, and the calm category is divided into sad and normal emotions. The objective of using hierarchical architecture is to improve the accuracy of emotional speech recognition while reducing the computation workload of support vector machines during their operation.

The speech features used in the hierarchical classifier are shown in Table 3, consisting of 28 basic feature parameters as well as the time-frequency parameter proposed in this study. The classifier starts with the feature classes of volume and zero-crossing rate in Stage 1, and then the feature classes of pitch and spectrum in Stage 2. The time-frequency parameter is also used in Stage 2 to improve the recognition rate. A total of 3 support vector machines and 29 feature parameters were used in to classify the emotional speech data into four categories, i.e., anger, happiness, sadness and normal.

TABLE III SPEECH FEATURES USED IN THE HIERARCHICAL CLASSIFIER

	Feature Class	Feature Parameters
Stage 1	Volume	volume average, maximum volume, minimum volume, the volume median, standard deviation of the volume
	Zero-crossing Rate	zero average, zero rate of maximum, minimum value of zero rate, zero rate of intermediate values, standard deviation of zero-crossing rate
Stage 2	Pitch	average pitch, maximum pitch, minimum pitch, standard deviation of pitch
	Spectrum	Mel frequency cepstral coefficients 1-13, time-frequency parameter

2.4 Support Vector Machine

The SVM classification algorithm, proposed by Vapnik [20], is a machine learning method based on the statistical learning theory, mainly and often used to perform the classification for a large number of high-dimensional data. The algorithm is often used in pattern recognition, image processing, bio-informatics and other related fields. The concept of SVM is to find a multi-dimensional hyperplane between two sets of data, and use this hyperplane to distinguish these two sets of data (Figure 10).

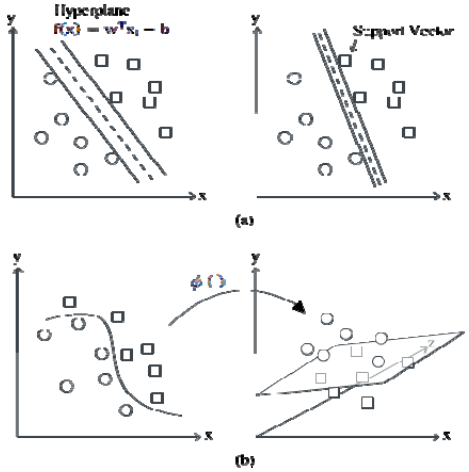

Fig. 10. (a) Using support vector machine for classification
(b) Using a mapping function to distinguish non-linear data sets

However, if the data sets are nonlinear and separable, a mapping function must be used to transform the data into a higher-dimension feature space for classification. By using the kernel function, SVM can calculates the inner product of feature vectors in a higher-dimension space to reduce the computation time and complexity. The SVM developed in this study uses the Radial Basis Function (RBF), an efficient kernel function for the nonlinear problem, and the LibSVM tool developed by Lin [22], which can deal with high dimensional and a large amount of data. Thus, it can be used to obtain the optimal solution for all training data, rather than a local optimal solution.

IV. EXPERIMENTAL RESULTS

In this study, an experiment was set up to test the functions of smart phone user interface and social networking website and the accuracy of emotional speech recognition. Considering the computation time required in processing real-time emotional speeches, 29 features were used for training and testing the SVM mechanism. The hardware of experiment was the smart phone CHT9110 produced by Chunghwa Telecom, Taiwan. The software executed on the smart phone was written by C#.NET programming language and tested on Windows Mobile 6.0 smart phone development platform. The website was set up by PHP dynamic web pages, Apache web server and MySQL database (Figure 11).

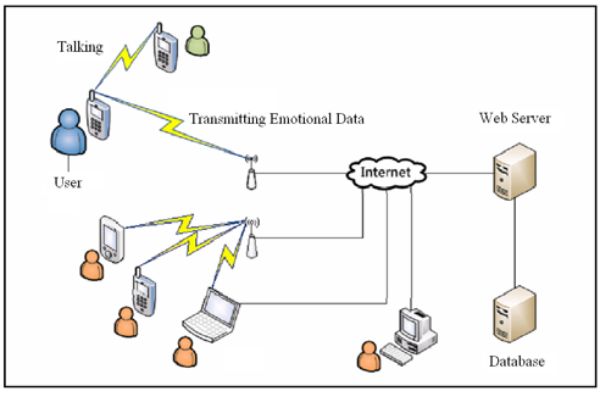

Fig. 11. The experiment conducted on the smart phone system

Users must install the software on their smart phones before they can use the emotional speech recognition system. When they are talking with friends on the mobile phones, the analysis results of emotional speech will be transmitted to the web server immediately. Friends in the same group can see the emotional status of each other through the access of computers or mobile devices at the same time, and make a phone call or send some messages to express their concern.

Generally, the features of emotional speech vary from person to person, but males have lower fundamental frequencies than females. In average, male voice frequency ranges from 62Hz to 523Hz, while female voice is about 110Hz to 1000Hz. In this study, the male and female speech data were trained and tested separately. In the experiment, the first test only used 28 acoustic features for training and testing, while the second test added in the time-frequency parameter to investigate if it is effective for increasing the accuracy of emotional speech recognition.

According to the experimental results (Table 4), the accuracy of male speech data for the test set is decreased from 61.7% to 57.5% (-4.2%) after using the time-frequency parameter, and the accuracy of female speech data for the test set is also decreased from 60.5% to 57.9% (-2.6%) after using the time-frequency parameter. Similarly, the accuracies of male and female speech data for the whole data set are also decreased by 1% and 0.6% respectively when the time-frequency parameter was used. Thus, using the time-frequency parameter in present classification model can not improve and it can even reduce the accuracy of emotional speech recognition.

TABLE IV THE ACCURACY OF THE TEST SET AND WHOLE SET (%)

	Test 1		Test 2	
	Test set	Whole set	Test set	Whole set
Male	61.7	90.4	57.5	89.4
Female	60.5	90.1	57.9	89.5
Average	61.2	90.3	57.7	89.4

To understand the reason for the reduction of accuracy after using the time-frequency parameter, this study further analyzed the precision rates of recognizing various emotions. Based on the results of Test 2 (Table 5), angry emotion has the highest precision rate, with male=84.6% and female=81.8%. On the contrary, normal emotion has the lowest precision rate, with male=40% and female=37.5%. According to our inference, the features of volume and pitch were not so obvious for the speech data of normal emotion and thus caused erroneous classification into excited category in Stage 1 or sad emotion in Stage 2.

TABLE V PRECISION RATES OF EMOTION RECOGNITION IN TEST 2 (%)

Sex / Emotion	Male	Female
Anger	84.6	81.8
Happiness	60.0	40.0
Sadness	77.8	88.9
Normal	40.0	37.5

Further results are provided (Table 6) to analyze the accuracy of using or without using the time-frequency parameter in Stage 2 for the classification of happy and angry emotions. According to these results, the accuracy of male test set is improved from 72.7% to 77.3% (+4.6%) and the accuracy of female test set is also improved from 63.2% to 68.4% (+5.2%), after using the time-frequency parameter for classifying excited category into angry and happy emotions. However, using this parameter can not improve the accuracy in the classification of sad and normal emotions. On the contrary, it could even reduce the accuracy substantially, with the accuracy of male test set decreased from 72.2% to 55.6% (-16.6%) and the accuracy of female test set decreased from 78.6% to 64.3% (-14.3%).

TABLE VI ACCURACY OF THE TEST SET IN STAGE 2 (%)

Experiments	Male		Female	
Emotions	Test 1	Test 2	Test 1	Test 2
Anger & Happiness	72.7	77.3	63.2	68.4
Sadness & Normal	72.2	55.6	78.6	64.3

According to the analysis of test results in Stage 2, the time-frequency parameter can only improve the classification rate between angry and happy emotions. Therefore, this study modified the structure of the hierarchical classifier in Stage 2, where the time-frequency parameter was used for classifying angry and happy emotions only (Figure 12).

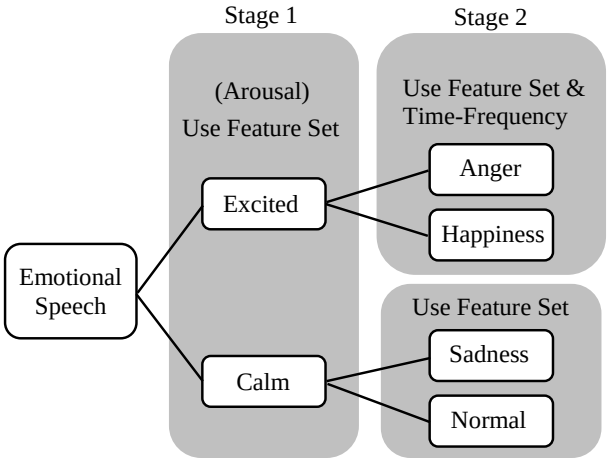

Fig. 12. The modified structure of hierarchical classifier

Using the modified hierarchical classifier, the results of Test 3 (Table 7) show that the accuracy of male test set is increased from 61.7% to 63.8% (+ 2.1%) and the accuracy of the whole

set is increased from 90.4% to 91.0% (+0.6%). Similarly, the accuracy of female test set is increased from 60.5% to 63.2% (+2.7%) and the accuracy of the whole set is increased from 90.1% to 90.8% (+0.7%). For the classification of all emotions, the average accuracy, including male and female data, is 63.5% for the test set and 90.9% for the whole data set. In comparison with the results of Test1, the accuracy for the test set and whole data set is increased by 2.3% and 0.6%, respectively, by using the modified hierarchical classifier.

TABLE VII ACCURACY OF TEST SET AND WHOLE SET AFTER MODIFICATION (%)

	Test 1		Test 3	
	Test Set	Whole Set	Test Set	Whole Set
Male	61.7	90.4	63.8	91.0
Female	60.5	90.1	63.2	90.8
Average	61.2	90.3	63.5	90.9

To test the functions of emotional speech recognition on the smart phones, one can use the web browser to access the social networking website through 3G Internet. The program installed on the smart phone can recognize the emotion of speaker and transmit the results to the web server immediately. Then, a small icon is used to show the emotional status of the speaker (for example, a smiling face means happy emotion), which can be seen by the users in the same community group (Figure 13). The design of convenient graphical user interface (GUI) on the smart phones enables users to login into the website easily and use the functions provided by the system such as editing personal information and albums, sending texts, photos, and voice messages to express their concern.

Fig. 13. The emotional statuses shown on the smart phone

V. CONCLUSIONS

With the portable and exclusive advantages and convenient access to the Internet, there is a promising future for the smart phones. This study combined the functions of smart phones with SVM emotional speech recognition system and social networks, enabling users to know their friends' emotions and send some messages to express their concern. Through a few simple steps, people can connect to the website and be aware of each other's feelings through immediate interaction.

Based on Russell's arousal-pleasure emotional model, this study used the support vector machines to classify speech data according to the emotional features. To identify the accent and intonation in a sentence during the recognition process, a time-frequency parameter obtained by continuous wavelet transforms was used in addition to 28 acoustic features for training and testing. The preliminary results showed that the time-frequency parameter only improves the recognition rates of angry and happy emotions, but it also reduced the recognition rates of sad and neutral emotions. Based on this finding, the hierarchical classification model was modified such that the time-frequency parameter was used for classifying angry and happy emotions only. The modified classification model can further increase the accuracies of male and female test sets by 4.6% and 5.2%, respectively. For the classification of all emotions, the average accuracy of male and female data is 63.5% for the test set and 90.9% for the whole data set.

In this study, the emotional speech recognition mechanism can effectively classify happy and angry emotions, but there is still some room for improvement regarding the other two emotions. In the future, the system can be combined with other classifiers, such as artificial neural networks, or use semantic identification to further improve the recognition rate. This study only divided speech data into four emotions, and the future studies may include more emotions or some other features, such as cry and laugh, to enhance the recognition rate.

In recent years, the market of smart phones has grown quickly. Combining the power of 3G Internet and community groups, the emotional speech recognition mechanism developed in this study can be applied on smart phones to increase the interaction and among people. The mechanism can be applied to interactive games or intelligent toys by providing appropriate feedback for children to achieve better learning results and experiences. The system can also be used as an assistant tool for professional counseling, and it can dynamically monitor patients' emotional changes and then provide appropriate on-line help.

REFERENCES

[1] Skiba, B., Johnson, M., Dillon, M. and Harrison, C., (2000). Moving in mobile media mode, http://www.regisoft.com/articles/lehman.pdf.
[2] Shneiderman, B. (1992). Designing the user interface: strategies for effective human-computer interaction. Reading: Addison-Wesley.
[3] Plutchik, R. (1980). A general psychoevolutionary theory of emotion. San Diego, CA: Academic Press.
[4] Russell, J. A. (1980). A circumplex model of affect. Journal of Personality and Social Psychology, 39, 1161-1178.
[5] Posner, J., Russell, J. A. and Peterson, B. S. (2005). A circumplex model of affect: an integrative approach to affective.
[6] Yen-Kung Yang (2003). Science Development. 367, 70-73.
[7] E. Douglas-Cowie, R. Cowie, and M. Schröder. (2000). Emotional speech: towards a new generation of databases. Speech Communication, a special issue on Speech and Emotion, 40(1–2), 33-60.
[8] Cover, T. M and Hart, P. E. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13, 21-27.
[9] Dimitrios Ververidis and Constantine Kotropoulos. (2006). Emotional speech recognition: Resources, features and methods. Speech Communication, 48 (9) 1162-1181.
[10] Cai, L., Jiang, C., Wang, Z., Zhao, L., and Zou, C. (2003). A method combining the global and time series structure features for emotion recognition in speech. In Proceedings of International Conference on Neural Networks and Signal Processing (ICNNSP'03), 2, 904-907.
[11] Kwon, O. W., Chan, K., Hao, J., and Lee, T. W. (2003). Emotion recognition by speech signal. The Eighth European Conference on Speech Communication and Technology (EUROSPEECH'03), Geneva, Switzerland.
[12] Schuller, B., Rigoll, G., and Lang, M. (2003). Hidden Markov model based

speech emotion recognition. 28th IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP'03).
[13] Vogt, T. and Andr´e, E. (2006). Improving automatic emotion recognition from speech via gender differentiation. Language Resources and Evaluation Conference.
[14] Petrushin, V. A. (2004). Emotion recognition in speech signal: experimental study, development, and application." Sixth International Conference on Spoken Language Processing (ICSLP).
[15] Reynolds, D. A. and Rose, R. C. (1995) .Robust text-independent speaker identification using Gaussian mixture models. In Proceedings of the European Conference on Speech Communication and Technology, 963–966.
[16] Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 77, 257-286.
[17] K. Fukunaga. (1990). Introduction to statistical pattern recognition. San Diego, CA: Academic Press.
[18] Cover, T. M and Hart, P. E. (1967).Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13, 21-27.
[19] E. H. Han, G. Karypis and V. Kumar. (2001). Text categorization using weight adjusted k-nearest neighbor classification. Pacific-Asia Conference on Knowledge Discovery and Data Mining, 53-65.
[20] Rabiner, L. R. and Ronald W. Schafer. (1989). Digital processing of speech signals. Prentice-Hall, Inc., Englewood Cliffs, NJ.
[21] Yao X. (1999). Evolving artificial neural networks. Proceedings of the IEEE , 87(9), 1423–1447.
[22] V. N. Vapnik. (2000).The nature of statistical learning theory. Chapter 5-6, 138-167, Springer-Verlag, New York.
[23] C. C. Chang and C. J. Lin (2001). LIBSVM: a library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm.