Speech Based Emotion Classification

Tin Lay Nwe, ***Student Member, IEEE,*** **Foo Say Wei,** ***Senior Member, IEEE***

and Liyanage C De Silva, ***Member, IEEE***

***Abstract*-- In this paper, a speech based emotion classification method is presented. Six basic human emotions including anger, dislike, fear, happiness, sadness and surprise are investigated. The recognizer presented in this paper is based on the Discrete Hidden Markov Model and a novel feature vector based on Mel frequency short time speech power coefficients is proposed. A universal codebook is constructed based on emotions under observation for each experiment. The databases consist of 90 emotional utterances each from two speakers. Several experiments including ungrouped emotion classification and grouped emotion classification are conducted. For the ungrouped emotion classification, an average accuracy of 72.22% and 60% are obtained respectively for utterances of the two speakers. For grouped emotion classification, higher accuracy of 94.44% and 70% are achieved.**

Index Terms-- **Emotions of speech, Mel-frequency speech power coefficients, Speech recognition, Hidden Markov Model**

I. INTRODUCTION

HUMANS interact with one another in several ways such as speech, eye contact, gesture etc. Among them, speech communications is the most common in human-to-human interaction. Speech can be said to be the most effective way of communications through which people can readily exchange information without the need for any other tool.

From the signal processing point of view, speech signal includes the linguistic information, speaker's tone and emotion. There are several applications for automatic machine recognition of the type of emotion expressed in a given speech. For example, it may be desirable to include information of emotions to other party in conventional video tele-conferencing and web-based teaching for added effects. In distant teaching, if a student does not understand what the teacher is saying, it may be detected from expression of emotions on his face and in his speech. These responses may have a direct and immediate influence on the teacher who would in turn try to explain the topic again. The emotion-based feedback is especially important in communicating with young children.

In this paper, a system is proposed to classify the emotions including anger, dislike, fear, happiness, sadness and surprise from speech. Several early research works in this area are found in [2], [4], [5], [6], [8] and [17].

In [8], bimodal emotion recognition by human subjects was studied. First, the subjects were asked to listen to sound clips alone. Then, they were shown video clips of the speakers without sound. Finally, video clips and sound track (audio and video data) were presented. In the studies, it was found that anger, happiness and surprise emotions are video dominant. Sadness and fear emotions are audio dominant.

In [6], emotional speeches incorporating happiness, sadness, anger and fear over 1000 utterances from different speakers were classified by human subjects and by computer. Human subjects were asked to recognize the emotion from utterances from one speaker in random order. It was found that human's classification error rate was 18%. For automatic classification by computer, pitch information was extracted from the utterances. Several pattern recognition techniques were used achieving a miss-classification rate of 20.5%.

In [17], classification performance of neural network classifiers on emotions such as sadness, cheerfulness, happiness and anger was investigated. The study made use of neural networks for recognition of basic human emotion and correct recognition rate of 70% is achieved.

Machine recognition of emotions using audiovisual information was conducted by Chan [4]. Six basic emotions of happiness, sadness, anger, dislike, surprise and fear were classified using audio model and video model separately. The recognition rate for audio alone is about 75% and video alone is about 70%. For audio processing, statistics of pitch, energy and the derivatives are extracted. Nearest mean criterion approach was adopted for classification. Joint audiovisual information of facial expression and emotive speech were used. The correct recognition rate is 97%.

For the system proposed in this paper, Mel frequency based short time speech power coefficients are selected as the features to identify the emotional state of the speaker. Subsequently, a universal vector quantizer designed based on LBG algorithm [10] is adopted. Hidden Markov Models are selected as classifiers.

The proposed system is depicted in Fig. 1.

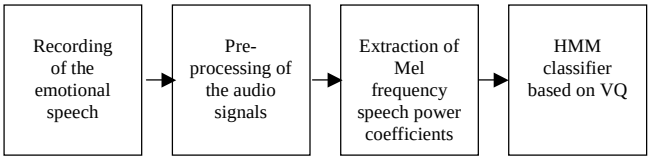

Fig. 1 Block diagram of the proposed system

The authors are with the Department of Electrical and Computer Engineering, National University of Singapore (E.mail: elelcds@nus.edu.sg)

IEEE Catalogue No. 01CH37239
0-7803-7101-1/ 01 / $10.00 © 2001 IEEE.

II. EMOTIONAL SPEECH CORPUS

The experiments conducted in this research were based on two "acted emotion" speech databases. The speech corpus recorded includes six basic emotions such as anger, dislike, fear, happiness, sadness and surprise from two Burmese language speakers.

54 sentences were recorded from the first speaker (Subject A) with 9 sentences for each of the six basic emotions. For the second speaker (Subject B), 90 sentences were recorded with 15 sentences for each emotion. The length of the sentences is between 0.6s to 1.6s. To achieve a very natural sounding of emotional speech corpus, speakers portray the sentences what were experienced in their daily life and performed rehearsals several times. The recording was carried out in a quiet environment with a mouthpiece microphone. The utterances were recorded at audio format of 16-bit PCM (Pulse Code Modulation) and sampling frequency of 22.05kHz.

III. CLASSIFICATION BY HUMAN SUBJECTS

First, an experiment was conducted to investigate how well the human does in classifying the speech corpus. Two human subjects were asked to classify the utterances according to the emotional content. The utterances were played in random order and presented via headphones twice. Human subjects took part in this evaluation stage did not understand the Burmese language and they were asked to classify utterances based on the emotional contents alone.

The performance of the human evaluators is summarized in Table I. Results show that emotions of Subject A can be recognized with higher accuracy compared to those of Subject B since their emotion portraying styles are different from each other. For example, in the case of anger, which might be used to describe anything from seething, Subject A portrayed quite hot anger while Subject B expressed cold-blooded anger. 60% of utterances that were portrayed as anger by Subject B were mis-classified as dislike. We may differentiate these two emotions styles into two main categories of conscious emotions (Subject A) and unconscious emotions (Subject B). Unconscious or concealed emotions are difficult to recognize by accounting emotional content alone since these emotions have less intensity of emotional expression. Our hypothesis regarding low recognition rates of unconscious emotions was that these emotions couldn't bring the variability into acoustical features among different emotion sets and dramatically influences the accuracy of recognition.

TABLE I. AVERAGE ACCURACY OF HUMAN CLASSIFICATION (%)

EMOTIONS	SUBJECT A	SUBJECT B
Anger	100	40
Dislike	66.67	53
Fear	55.56	13.3
Happiness	66.67	13.3
Sadness	55.56	66.67
Surprise	33.33	73.33
Average Perf:	62.96	43.27

IV. SPEECH DATA PROCESSING

A. Choice of suitable features

Many researchers integrated several different techniques in emotion recognition in speech to improve performance of emotional speech recognizer. According to the results of emotion recognition research that has been done to date, the aspect of features that is most suitable for emotion recognition in speech is still in research stage. A possible approach is to apply various different and known feature extraction methods to investigate the way to extract non-textual information to identify emotional state of the utterance.

The feature vector, which is used to represent the emotional speech in our process, aims to preserve the information needed to determine the emotional content of such a signal. Mel based speech power coefficients are chosen as they convey information of short time portions that describe the power of speech signal. These coefficients specially distinguish among various levels of "agitation" or "calm" of a given emotion. Furthermore, these Mel based parameters show the energy migration in frequency domain. It can be shown that for anger and surprise emotions, additional energy was typically moved from low to high frequency bands, in which VQ based discrete HMM has the ability to build an efficient model for these state changes.

B. Speech signal processing front-end

The speech utterance was segmented into 16ms frames, and Hamming window was applied for smoothing and reducing spectral leakage. A frame was overlapped with the next frame by 9ms. Each segmented speech frame is parameterized using 12 Mel-frequency subband energy values. The observed speech frame is transformed to the frequency domain using FFT. Then, a set of 12 Mel scaled filter banks which has frequency span between 200Hz to 3.2kHz, was applied to the FFT power spectrum. Short time energy of each of 12 filter bank output was calculated for each speech frame as expressed mathematically below.

$$S_t(m) = \sum_{k=f_m-b_m/2}^{f_m+b_m/2} \left(X_t(k)\, W_m(k)\right)^2 \quad (1)$$

where

$X_t(k)$ hamming windowed signal in DFT domain,

$S_t(m)$ m^{th} filter bank output

$W_m(k)$ m^{th} subband response,

f_m, b_m center frequency and bandwidth of m^{th} subband.

Details of filter bank implementation may be found in [16]. The above operation results in short time speech power coefficients that present energy distribution among subbands. The feature parameters are then calculated as follows.

$$SE_m^n = \left[10\log 10(S_t(m))\right] / N_m \quad (2)$$

where

n frame number,

N_m number of coefficients in the m^{th} filter bank

m filter bank index

V. FEATURE ENCODING

The vector of 12 speech power coefficients for each speech frame, f_n, was assigned to a cluster by vector quantization. The vector quantizer is composed of a single codebook called universal codebook. The vector quantizer was designed with the LBG algorithm [10]. During training a universal codebook of 64 codeword is constructed by using training speech samples from all observed emotions for each experiment. The vector f_n was assigned the codeword c_n^* according to the best match codebook cluster z_c using (3).

$$c_n^* = \arg \min_{1 \le c \le C} d(f_n, z_c) \quad (3)$$

After encoding all frames $(f_1, f_2, \ldots\ldots f_n)$ by assigning with the best match codebook index, the resultant feature vector Y was obtained for each speech utterance.

$$Y = \left[c_1^* \quad c_2^* \ldots\ldots \quad c_N^*\right] \quad (4)$$

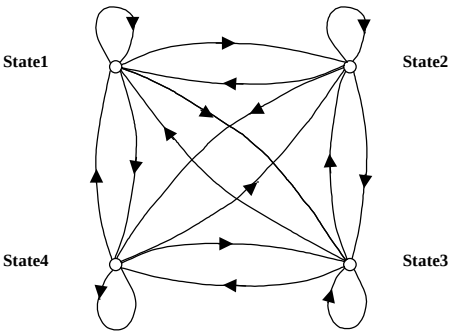

Fig. 2 4-state ergodic model HMM

VI. CLASSIFICATION SYSTEM DESIGN

A. VQ based Discrete HMM emotion recognizer

Once the input has been quantized, then the HMMs were trained. The 4-state ergodic HMM as shown in Fig. 2 was used as this structure achieved the best score compared to left-right structures. For this model, every state can be reached in a single step from every other state of the model. The state transition probabilities and the output symbol probabilities are uniformly initialized. The training is based on the forward and backward algorithm [15]. Separate HMM models for each emotion group are obtained during training phase. The output symbol probabilities are smoothed with the uniform distribution to avoid the presence of too small probabilities or zero probabilities. The utterances were scored using the forward algorithm. 60% of emotion utterances of each of two databases were used to train each emotion model. Recognition tests were conducted on the remaining 40% of each database. Correct recognition rates were listed in Table II for ungrouped emotion classification experiments and Table IV, V and VI for grouped emotion classification experiments. A block schematic diagram of the training and recognition strategy is shown in Fig. 3.

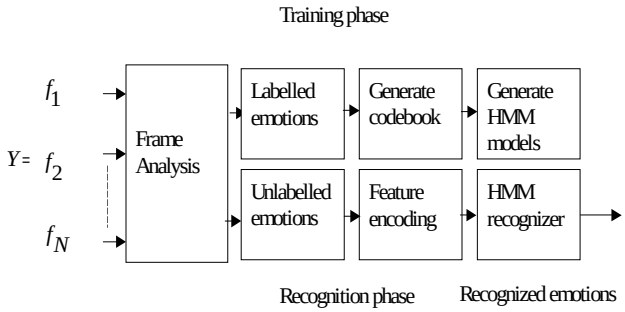

Fig. 3 Training and recognition strategy

VII. RESULTS AND DISCUSSION

As some of the basic emotions such as anger and surprise have some common characteristics, so have dislike and sadness, they may be grouped together and distinguished as a group. This is given the name grouped emotion classification.

Experiments were conducted for ungrouped-emotion and grouped-emotion classification. The recognition rates of ungrouped emotion classification using utterances reserved for testing (unseen) are shown in Table II. It is noted that the classification results are somewhat correlated with human's performance of classification. The anger, dislike and happiness emotions have the highest accuracy for both human and machine classification.

For grouped emotion classification, the grouping shown in Table III is used. Since anger and surprise emotion styles have similar energy distribution, these two emotions are put into one group (G3). Sadness and dislike emotions have similar spectral distributions and are put into one group (G5). The two emotions of fear and happiness have similar level of energy distribution and they are combined into G4.

Three separate experiments were carried out to distinguish the different combinations of groups of emotions. Experiment 1 aims to distinguish between emotion groups G1 and G2. Experiment 2 is carried out to distinguish between G3 and G5 and Experiment 3 to classify G3, G4 and G5. Results of the three experiments are summarized in Tables IV, V and VI respectively.

It is observed that grouping of emotions gives rise to much improved accuracy of classification. Classification score between G1 and G2 is the highest for both speakers since "agitation" level of emotion in G1 differs considerably from of emotion in G2. When three groups, G3, G4 and G5, are to be classified, the accuracy is decreased, as the features of G4 are not very much distinct from the features of G5.

From the results, it can also be observed that the accuracy of classification for utterances of Subject A is consistently higher than that for Subject B. Informal listening tests were made to identify the reasons behind this disparity. It is found that Subject B is of a "lesser emotional person" in that the emotional contents of utterances for the different emotions are not as conspicuous as those of Subject A.

These results demonstrate the potential of VQ based discrete HMM approach based on the features of Mel frequency speech power coefficients. According to the results, there is sufficient information in the Mel frequency speech power coefficients for accurate classification of groups of emotions although it is not sufficient to make finer classification.

TABLE II. UNGROUPED EMOTION CLASSIFICATION SCORES ON UNSEEN DATABASE (%)

EMOTIONS	SUBJECT A	SUBJECT B
Anger	100	40
Dislike	66.67	60
Fear	66.67	60
Happiness	100	60
Sadness	66.67	80
Surprise	33.33	60
Average Perf:	72.22	60

TABLE III. EMOTION GROUPING

EMOTION GROUP	GROUP MEMBER
G1	Anger
G2	Sadness
G3	Anger, Surprise
G4	Fear, Happiness
G5	Dislike, Sadness

TABLE IV. GROUPED EMOTION CLASSIFICATION SCORES ON UNSEEN DATABASE (1ST EXPERIMENT)

EXPERIMENT 1(%)		
Emotion Group	Subject A	Subject B
G1	100	100
G2	100	80
Average Perf:	100	90

TABLE V. GROUPED EMOTION CLASSIFICATION SCORES ON UNSEEN DATABASE (2ND EXPERIMENT)

EXPERIMENT 2(%)		
Emotion Group	Subject A	Subject B
G3	100	80
G5	100	70
Average Perf:	100	75

TABLE VI. GROUPED EMOTION CLASSIFICATION SCORES ON UNSEEN DATABASE (3RD EXPERIMENT)

EXPERIMENT 3(%)		
Emotion Group	Subject A	Subject B
G3	100	90
G4	100	40
G5	83.33	80
Average Perf:	94.44	70

The performance of the proposed system is comparable with that using other approaches such as neural network classifier [17], statistical pattern recognition techniques [6] and nearest mean criterion [4] as shown in Table VII.

TABLE VII COMPARISON OF METHODS

APPROACH	EMOTIONS CLASSIFIED	AVERAGE ACCURACY
Pattern Recognition [6]	Happiness, Sadness, Anger and Fear.	79.5%
Neural Network [17]	Sadness, Cheerfulness, Happiness and Anger	70%
Nearest Mean Criterion [4]	Anger, Dislike, Fear, Happiness, Sadness and Surprise.	75%
Proposed Method	Anger, Dislike, Fear, Happiness, Sadness and Surprise.	72.22%

VIII. CONCLUSION

In this paper, a new approach for recognition of emotion in speech is proposed. The method makes use of Mel based feature parameters and discrete HMM classifier. From the results obtained, it can be concluded that the filter bank analysis coefficients contain information of the features that could significantly reflect the emotional state of a speaker. However, the set of coefficients may be insufficient to represent the fine details of the emotional contents for distinguishing the six basic emotions.

Results also show that using the HMM classifier and Mel based feature parameters, high level of accuracy of classification can be achieved if emotions of similar nature, such as anger and surprise; fear and happiness; dislike and sadness; are grouped together as one for the purpose of classification.

The level of accuracy is also speaker dependent. If the speaker does not reflect his/her emotions in the utterances, there is very little clue for accurate classification.

REFERENCES

[1] C. Becchetti, and L. P. Ricotti, "Speech Recognition Theory and C++ Implementation," John Wiley & Sons, New York, 1998.

[2] S. E. Bou-Ghazale, and J. H. L. Hansen, "HMM-Based Stressed Speech Modeling with Application to Improved Synthesis and Recognition of Isolated Speech Under Stress," IEEE Transactions on Speech and Audio Processing, vol. 6, no. 3, pp. 201-216, 1998.

[3] D. K. Burton, J. E. Shore, and J. T. Buck, "Isolated-Word Speech Recognition Using Multisection Vector Quantization Codebooks," IEEE Transactions on Acoustics, Speech and Signal Processing, vol. ASSP-33, no 4, pp.837-849, 1985.

[4] L. S. Chan, H. Tao, T. S. Huang, T. Miyasato, and R. Nakatsu, "Emotion Recognition From Audiovisual Information," IEEE Second Workshop on Multimedia Signal Processing, pp. 83 –88, 1998.

[5] L. S. Chen, T. S. Huang, T. Miyasato, and R. Nakatsu, "Multimodal Human Emotion / Expression Recognition. Proceedings," Third IEEE International Conference on Automatic Face and Gesture Recognition, pp. 366 –371, 1998.

[6] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing Emotion in Speech," Fourth International Conference on Spoken Language Processing, vol.3, pp.1970-1973, 1996.

[7] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, "Discrete-Time Processing of Speech Signals," Macmillan Pub. Co. ; Toronto, 1993.

[8] L. C. De Silva, T. Miyasato, and R. Nakatsu, "Use of Multimodal Information in Facial Emotion Recognition," IEICE Trans. Inf. & Syst., vol. E81-D, no.1, pp. 105- 14, 1998.

[9] W. H. Equitz, "A New Vector Quantization Clustering Algorithm" IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 37, no.10, pp.1568-1575, 1989.

[10] S. Furui, "Digital Speech Processing, Synthesis and Recognition" Marcel Dekker, New York, pp. 185-204, 1989.

[11] K. W. Law, and C. F. Chan, "Split-Dimension Vector Quantization of Parcor Coefficients for Low Bit Rate Speech Coding," IEEE Transactions on Speech and Audio Processing, vol 2, no. 3, pp. 443-446, 1994.

[12] K. F. Lee, and H. W. Hon, "Speaker-Independent Phone Recognition Using Hidden Markov Models," IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 37, no. 11, pp.1641-1648, 1989.

[13] T. W. Parsons, "Voice and Speech Processing," McGraw-Hill, New York, 1986.

[14] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," Proceedings of the IEEE, vol. 77, no. 2, pp. 257-286, 1989.

[15] L. R. Rabiner, and B. H. Juang, "An Introduction to Hidden Markov Models," IEEE ASSP Magazine, pp. 4-15, 1986.

[16] L. R. Rabiner, and B. H. Juang, "Fundamentals of Speech Recognition," Prentice Hall, Englewood Cliffs, N.J, 1993.

[17] T. Yamada, H. Hashimoto, and N. Tosa, "Pattern Recognition of Emotion with Neural Network," Proceedings of the 1995 IEEE IECON 21st International Conference on Industrial Electronics, Control, and Instrumentation, vol. 1 , pp. 183 –187, 1995.