Speech Emotion Recognition System Based on Integrating Feature and Improved HMM

Han Zhiyan , Lun Shuxian , Wang Jian
College of Engineering, Bohai University
Jinzhou, Liaoning, China

***Abstract*-This paper described a new speech emotion recognition system by use of Hidden Markov Model (HMM) aiming at improving speech emotion recognition rate. Seven discrete emotional states (anger, disgust, fear, joy, neutral, sadness, surprise) are classified throughout the work. It integrated different speech features into the system, the system is comprised of three main sections, a pre-processing section, a feature extracting section and a HMM processing section. Results are given on speaker dependent case using the Chinese corpus of emotional speech synthesis database. Recognition experiments show that the method is effective and high speed and accuracy for emotion recognition.**

Keyword-Emotion recognition, Integrating feature, Hidden Markov Model, Speech signal

I. INTRODUCTION

Dealing with the speaker's emotion is one of the latest challenges in speech technologies. Three different aspects can be easily identified: speech recognition in the presence of emotional speech, synthesis of emotional speech, and emotion recognition. In this last case, the objective is to determine the emotional state of the speaker out of the speech samples. Especially in the field of security systems a growing interest can be observed throughout the last year. Besides, the detection of lies, video games and psychiatric aid are often claimed as further scenarios for emotion recognition[1].

Most current emotion recognition systems employs only one type of speech feature, e.g. pitch and their variants, this entails that the system may not be robust. So the incorporation of different speech features into emotion recognition potentially offers a significant degree at a reasonable computational cost.

The recognition platform can be divided into three types. Dynamic Time Warping(DTW), the earliest platform, uses the variation in frame's time for adjustment and further recognition. Later, Artificial Neural Network(ANN) replaced DTW. Finally, Hidden Markov Model was developed to adopt statistics for improved recognition performance.The classical Baum-Welch(B-W) method had traditionally been the first choice for training HMM. However, the B-W method cannot achieve global optimization. In this paper, a genetic algorithm(GA) for training HMM was presented [2-4].

II. FEATURE EXTRACTION

This paper integrates prosody and quality features into system to improve recognization ratio. After a detailed analysis of these parameters, they can provide significant information to system.

1) PROSODY FEATURES

A set of 37 features has been used, 26 of these features model logarithmic F0, energy and durational aspects.

Logarithmic F0: maximum, minimum, maximum and minimum position, mean, standard deviation, regression coefficients, mean square error for regression coefficients, and F0 for the first and last voiced frame.

Energy: maximum, minimum, maximum and minimum positions, mean, regression coefficients, and mean square error for regression coefficients.

Durational aspects: number of voiced and unvoiced regions, number of voiced and unvoiced frames, longest voiced and unvoiced region, ratio of number of voiced vs. unvoiced frames, ratio of number of voiced vs. unvoiced regions, ratio of number of voiced vs. total number of frames, ratio of number of voiced vs. total number of regions.

Additional 11 features, model jitter, tremor and pitch derivative statistics, as suggested by Dellaert[5-6].

2) quality FEATURES

Latest emotion recognition approaches also include information related to articulatory precision or vocal tract properties, e.g. formant structure, as in [7]. There is perceptual evidence, in terms of emotions expression, of the additional importance of phonatory quality parameters, i.e. auditory qualities derived from variations in the glottal excitation.

We chose 16 quality features, describing the first three formants, their bandwidths, harmonic to noise ratio, spectral energy distribution, voice to unvoiced energy ratio, and glottal flow. All the quality features described were obtained using the phonetic analysis software PRAAT.

III. OPTIMIZING AN HMM USING GA

The classical B-W method is an iterative procedure. An initial HMM is composed of some random numbers. Each observation sequence is offered to train this model one by one. A large number of temporary variables have to be calculated for each observation sequence. All the HMM parameters are re-estimated until there is no change. However, only regional optimization solution is obtained. Usually, several initial HMM models are tested, and then the best solution is selected. In this paper, a new algorithm for training HMM based on a GA is presented. An HMM

Published by Atlantis Press, Paris, France.

requires the specification of the three sets of probability measures A, B and π . For convenience, the compact notation $\lambda = (A, B, \pi)$ is used to indicate a complete parameter set of the model. Training an HMM, therefore, involves adjusting the model parameters (A, B, π) to satisfy certain observation sequences more closely. In this paper, the GA is the global optimization tool for training an HMM[8].

Genetic Algorithm: The gene is a biological concept. GA has been used mainly as function optimizers, which have been shown to be effective global optimization tools. For a solution when exploring a large and complex space, the GA is guided by the equivalent biological evolution mechanisms of reproduction, crossover and mutation. The GA is used to train the HMM because the re-estimation of the HMM parameters may be considered as searching for a global optimum in the parameter space.

In nature, the motivating force behind biological evolution stems from inheritance and selection. The main operator of the GA is simulated by using gene crossover, gene mutation and natural selection. The GA can be described as follows:

1) Randomly generate an initial population $p(0) = \{a_1, a_2, \cdots, a_N\}$, where a_i is one chromosome.

2) Compute the fitness $F(a_i)$ of each individual chromosome a_i of the current population $p(t)$, and then select some chromosomes as an intermediate population $p'(t)$.

3) Apply the crossover operation to some chromosomes in $p'(t)$.

4) Carry out a mutation operation at a small probability on some chromosomes.

5) $t = t + 1$; if not end, then go to 2).

In the natural environment, a pair of chromosomes is divided and rebuilt, and then a pair of new chromosomes is produced. The multi-point crossover simulates this process. A crossover probability threshold p_c is defined. The length of a chromosome is L for a random number $0 \le r_j \le 1 (j = 1, 2, \cdots, L)$, if $r_j \ge p_c$, then the next parameter belongs to the other chromosome; otherwise it belongs to the same chromosome as the previous parameter.

The crossover operation leads to regional optimization. However, the mutation operation can help the algorithm to jump out from regional optimization and avoid it being finished too early. We can define a mutation threshold p_m for a random number $0 \le r_j \le 1 (j = 1, 2, \cdots, L)$, if $r_j \le p_m$ then the jth parameter of the chromosome is mutated; otherwise, the jth parameter is copied.

Training HMM: All of the parameters of an HMM are arranged in a line to construct a chromosome. For speech recognition, we use a left-right with one-order jump HMM model structure. The initial state distribution is fixed as $\pi = \{1, 0, 0, \cdots\}$, so the chromosome need not contain them. There are seven states in our HMM. The state-transition probability distribution is $A = \{a_{ij}\}$, in which there are only 12 $a_{ij} > 0$, and the other a_{ij} parameters are always 0. After vector quantization, the codebooks of the speech signal feature vectors have sizes 53. The observation symbol probability distribution B includes $7 \times 53 = 371$ parameters. The HMM $\lambda = (A, B, \pi)$ has $12 + 371 = 383$ parameters. The training of an HMM involves searching for the best setting of these 383 numbers. We encode one chromosome as 383 numbers in which each ranges from 0 to 1.

The GA selects some excellent chromosomes as an intermediate population according to the degree to which they fit the observation sequences. The training of an HMM is based on likelihood maximization. Hence, the fitness function can be defined as $F(a_i) = \sum(\varphi_k)$, in which φ_k is the logarithmic calculation of the viterbi algorithm for the kth observation sequences to the ith chromosome.

However, the HMM has some features. For HMM $\lambda = (A, B, \pi)$, the sum of each row vector of matrix A or B is 1.0. We have to control the gene production. When a new chromosome is generated, we adjust the correspondence numbers of each segment of the row vector to unity.

For training an HMM, we select $p_c = 0.8$ and $p_m = 0.05$.

IV. EXPERIMENT AND RESULT ANALYSIS

Some experiments are conducted to evaluate the emotion recognition system. In our experiment, seven discrete emotional states (anger, disgust, fear, joy, neutral, sadness, surprise) are classified throughout the work. The emotional speech synthesis database was recorded in Chinese language, the recordings were all done in silent rooms and with high quality microphones by seven speakers, data were recorded at a sampling rate of 16kHz, 100 utterances per emotion have been used for the training, while a disjunctive set of 100 utterances per emotion were used testing.

Table 1 shows the confusion matrix of the emotion evaluation using HMM, table 2 shows the confusion matrix of the emotion evaluation using improved HMM. Columns represent the emotion elected in first choice for utterances belonging to the emotion of each row, where A stands for anger, D stands for disgust, F stands for fear, J stands for joy, N stands for neutral, S stands for sadness and P stands for surprise.

The table 1 and table 2 show that the use of integrating feature instead of single feature is much more robust and using improved HMM is more effective than HMM, all emotions are recognized with accuracy higher than 77%.

V. SUMMARIES

The recognition system has been known to improve

emotion recognition rate, especially for speaker dependent case. A method for using genetic algorithms to train HMM has been successfully developed. The main contribution of this study is that it presents the idea of searching for the most optimal HMM. The experiments also show that the approach is superior to the classical method. The proposed system was completely simulated on PC. The simulation results show that the approach is correct and effective. Our effort will be directed toward speaker independent case and muti-modal emotion recognition.

ACKNOWLEDGMENT

The authors wish to deeply thank graduate students who collaborated in the experiments and in the development of the system. The work is also supported by a grant from the National Natural Science Foundation of China (No. 60974071) and a grant from Education Department Excellent Talents of Liaoning Province (No. LR201002).

REFERENCES

[1] R. Cowie; E. Douglas-Cowie; N. Tsapatsoulis; G. Votsis; S. Kollias; W. Fellenz; J.G. Taylor. Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine [J], 2001, PP:32-80

[2] QIN Long; WU Yi-jian; LING Zhen-hua; WANG Ren-hua; DAI Li-rong. Minimum generation error criterion considering global/local variance for HMM-based speech synthesis. IEEE Symp. Acoustics, Speech, and Signal Processing (ICASSP 08) [J], 2008, PP:4621-4624

[3] LIU Peng; LIU Cong; JIANG Hui; F. K. Soong; WANG Ren-hua;. A constrained line search approach to general discriminative HMM training. IEEE Symp. Automatic Speech Recognition & Understanding (ASRU 07) [J], 2007, PP:290-295

[4] E. R. Duni; B. D. Rao. High-rate optimized recursive vector quantization structures using Hidden Markov Models. IEEE Trans. on Audio, Speech and Language Processing [J], 2007, PP:756-769

[5] F. Dellaert; T. Polzin; A. Waibel. Recognizing emotion in speech. IEEE International Conference on Spoken Language Processing (ICSLP 96) [J], 1996, PP:1970-1973

[6] T. Raquel; S. Roc; K. Ralf; J. M. Pardo. Emotional space improves emotion recognition. IEEE International Conference on Spoken Language Processing (ICSLP 2002) [J], 2002, PP:2029-2032

[7] V. A. Petrushin. Emotion recognition in speech signal: experimental study, development, and application. IEEE International Conference on Spoken Language Processing (ICSLP 2000) [J], 2000, PP:222-225

[8] SUN Fang; HU Guang-rui. Speech recognition based on genetic algorithm for training HMM. Electronics Letters [J], 1998, PP:1563-1564

TABLE I. THE CONFUSION MATRIX OF THE EMOTION EVALUATION USING HMM

Labeled emotion	**Recognized emotion**							**Total**
	A	D	F	J	N	S	P	
anger	79	11	4	1	0	0	5	100
disgust	3	91	1	2	1	0	2	100
fear	2	4	78	1	0	0	15	100
joy	6	4	0	72	0	0	18	100
neutral	0	3	1	0	96	0	0	100
sadness	0	6	0	0	0	82	12	100
surprise	2	2	5	7	0	0	84	100
Total	92	121	89	83	97	82	136	700

TABLE II. THE CONFUSION MATRIX OF THE EMOTION EVALUATION USING IMPROVED HMM

Labeled emotion	**Recognized emotion**							**Total**
	A	D	F	J	N	S	P	
anger	83	7	4	1	0	0	5	100
disgust	2	94	0	1	1	0	2	100
fear	2	2	80	1	0	0	15	100
joy	4	4	1	77	0	0	14	100
neutral	0	0	1	0	99	0	0	100
sadness	0	4	0	0	0	86	10	100
surprise	2	2	6	5	0	0	85	100
Total	93	113	92	85	100	86	131	700

Published by Atlantis Press, Paris, France.