Speech Emotion Recognition Using Gaussian Mixture Model

Xianglin Cheng
Computer Engineer Department
Zhongshan Polytechnic
Zhongshan,China
286863096@qq.com

Qiong Duan
Basic Department,
Zhongshan Polytechnic
Zhongshan,China

Abstract—**The importance of automatically recognizing emotions in human speech has grown with the increasing role of spoken language interfaces in human-computer interaction applications. In this paper, a emotion classification method base on GMM is presented. Five primary human emotions, including anger, surprise, happiness, neutral and sadness, are investigated. For speech emotion recognition, we combined 60 basic features to form the feature vector. Finally, the features of the speech were extracted by PCA were sent into the improved GMM for classification and recognition. Results show that the selected features are robust and effective for the emotion recognition .**

Keywords-Speech Emotion Recognition, Wavelet transform, MFCC, PCA, GMM

I. INTRODUCTION

The study of human centered emotion and cognition, including affective computing and emotion recognition, is a hot research topic in artificial intelligence.Speech emotion recognition is a pivotal question of affective computing[1,2].Generally speaking, the speech emotion recognition ,including ,getting original speech,pretreatment of voice signal, pitching features, pattern matching and get result to the course of four.

In this paper, discuss the category and modeling of emotion firstly; then, give the total frame of the emotional speech recognition system. And explain every parts.The main goal of this thesis is to search the most useful features with analyzing the features related with emotions, and find a recognition model to make use of these features.First extract the feature vector from the speech signal with the principal component analytic (PCA) method, and then use GMM training and testing.

II. EXTRACTING PITCH PERIOD

Pitch period is an important parameter of speech signal[3], and it has been applied in many domains.In this paper, a new pitch detection method based on the wavelet analysis and the analysis of variance is proposed..For the low signal-to-noise ratio , the performance of the method is better than other methods which had been presented.

The method of Pitch period base on the wavelet analysis and the analysis of variance comprises the following steps.xw(n) n=1,2,...,N is voice data frame in the follow.

Step 1:Firstly xw(n) was decomposed by wavelet transformation until we got wavelet coefficients.Fundamental frequency is the reciprocal of pitch period.

Step 2: Analysis of variance were carried out for wavelet coefficients.we got function VSDF(l),l=fs/pmax,fs/pmax+1,...,fs/pmin. The fundamental frequency of speech is between pmin and pmax.fs is sampling frequency .

Step 3:Get maximum of VSDF(l) as max.

Step 4:Set max as threshold.if max is less then th(th is given threshold in advance) ,the frame is unvoiced frame . Go to Step 7 . Pitch period is zero.if max is equal or greater than th,the frame is voiced frame .Go to Step 5.

Step 5:look up the location of a local maximum which is the first greater then threshold in VSDF(l) from l=fs/pmax to l=fs/pmin and set it to lp.

Step 6:Get Pitch period in the form P=lp/fs;

Step 7:The whole processing ends.

III. .EMOTION RECOGNITION OF SPEECH

A GMM[4] classifier is used to recognize emotion in this paper.Through the training and testing of lots of studied samples, the error of the model is limited in a preconcerted range. First extract the feature vector from the speech signal with the principal component analytic (PCA) method, and then use GMM training and testing.

A. Pretreatment of voice signal

Normally, voice signal need to perform preprocessing. The pre-treatment process in a certain extent affect system recognition rate.what we need to do is pre-weighting, en-framing, window adding in preprocessing stage.

The purpose of pre-weighting is to boost the high frequencies of a signal and get the flat frequency spectrum of signals and frequency characteristics.Using window function w(n),we get speech frame with setting spots which is outside of processing regions to zero . Right now, commonly used window function is rectangular window and hamming window.

B. Pitch features

In order to improve the recognition rate, we choose speech features based on the fundamental frequency、time、amplitude、cepstrum. After feature extraction, we get 60 dimension vector as parameters of emotional features .

Published by Atlantis Press, Paris, France.

1) characteristic parameters from fundamental frequency

The fundamental frequency is the vibration frequency of vocal cords.Due to the fundamental frequency is only significant for voiced sound, so the above characteristics are from voiced frame.

Characteristic parameters of fundamental frequency as follows:Maximum of fundamental frequency(F0_max)、Minimum of fundamental frequency(F0_min)、Mean of fundamental frequency(F0_mean)、Standard deviation of fundamental frequency(F0_std)、Range of fundamental frequency(F0_range)、Maximum of the absolute value of difference of Pitch Contour(F0_diff_max)、Mean of the absolute value of difference of Pitch Contour(F0_diff_max)、Standard deviation of the absolute value of difference of Pitch Contour(F0_diff_max) .

2) characteristic parameters from time

We choose short-time average zero-crossing rate ,because speech emotion parameters from time are different . Amplitude of the signals from positive to negative or from negative to positive, we regard this as Zero-crossing. Zero- crossing rate is the number of Zero-crossing in unit time. Speech is a short-term stationary signal,we obtain the average zero-crossing rate according to the frame in the form

$$\mathrm{Z_n} = \sum_{\mathrm{m}=-\infty}^{\infty} |\,\mathrm{sgn[x(m)] - sgn[x(m-1)]}\,|\,\mathrm{w(n-m)}$$

W(n)is rectangular window function.sgn[] is sign function in the form

$$\mathrm{sgn}[x(n)] = \begin{cases} 1 & x(n) \geq 0 \\ -1 & x(n) < 0 \end{cases}$$

We get eight statistical characteristics parameters about short-time average zero-crossing rate using same procedure as before.

3) characteristic parameters from amplitude

The emotional changes have a different effect on amplitude of speech signal.for example,when one is excited,his volume is high and the amplitude is large.when one is sad,his volume is low and the amplitude is small.amplitude is important for Analysis of emotion.

We get short-time energy to measure amplitude in the form

$$\mathrm{E_n} = \sum_{m=-\infty}^{\infty} [x(m)w(n-m)]^2$$

We get eight statistical characteristics parameters about short-time energy using same procedure as before.

4) *characteristic parameters from cepstrum*

Due to fully considered the human auditory characteristics, Mel Frequency Cepstrum coefficient(MFCC)[5] can perfect obtain feature parameters of the speech emotion recognition.

It works using MFCC as follows:

Step 1: Speech signal become short speech frame by the window, then we get the spectrum by short-time Fourier transform.

Step 2: to get power spectrum, we use m MEL band-pass filter to filter,then add energy of each filter.Output power spectrum of the Kth filter is x'(k).

Step 3:we get logarithm of output from filter and the power spectrum of corresponding frequency band,then get L parameters of MFCC in IDCT in the form

$$\mathrm{Cn} = \sum_{k=1}^{M} \log x'(k) \cos(\pi(k-0.5)\frac{n}{M}), n = 1,2,\ldots,L$$

We choose 36 characteristic parameters using the procedure as before with 12 dimension MFCC.Each dimension with three statistical characteristics as follows:Maximum of Short-time MFCC(MFFCC1-12_max)、Mean of Short-time MFCC(MFFCC1-12_mean)、Standard deviation of Short-time MFCC(MFFCC1-12_std).

C. Dimension reduction

In order to find a suitable combination of extracted features[6] ,we used the PCA method to determine beneficial features from among more than 60 speech features.Physical meaning and scope of various characteristics in the original eigenvector are different, A unitary processing for the feature vector was made. The Gaussian model is used to normalize the characters of different speech .

After normalization , we do the dimension reduction. We make feature reduction for N-dimensional vectors{xt}(t=1,2,...T) using PCA as follows:

Step 1:By making use of transformation of coordinate translation,we set average vector Xm as the origin of new coordinate system in the form x*t=xt-xm

$$\mathrm{x}_m = \frac{1}{T}\sum_{t=1}^{T} x_i$$

Step 2:Find out overall covariance matrix R

$$\mathrm{R} = \frac{1}{T}\sum_{t=1}^{T} \overset{*}{x_t}\, \overset{*T}{x_t}$$

Step 3:Find out eigenvalues (ג1,ג2,.....גN) and related eigenvectors (q1,q2,...qN）.

Step 4:The sort order for each eigenvalue is descending order, eg. ג1≥ג2,.....≥גN.we get a transformation matrix A=(q1,q2,...,qM) (M<N) form eigenvectors of in the M larger eigenvalues.

Step 5:Transform N-dimensional original vector to M-dimensional new vector in the form $\mathrm{y}_t = A^t X_t$.

D. Emotion recognition of speech based on GMM

we use GMM to do Pattern recognition in this paper.The Maximum Likelihood(ML) method now is the most popular parametric estimation method of GMM. The purpose of maximum likelihood estimates is to get model parameters in given the training data.In the experiment of emotional speech recognition[7,8] base on GMM , there were two sessions, train and test. During the training, we use EM to get GMM from the features related with emotions. During the testing,we use GMM to get probability of Testing voice features .the emotion with the biggest probability is the result.

1) Training

In emotion recognition of speech system, features extracted from each emotional speech can be expressed as an array F={f1,f2,..,ft,...,fT},ft is eigenvector of speech frame,T is the number of frames in each speech.we build models for each emotion using GMM in the Fig.1 below.

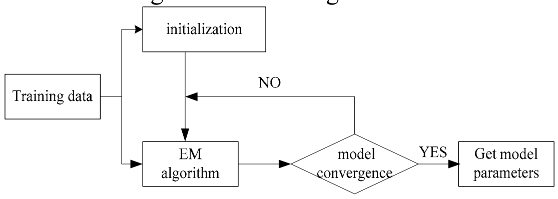

Figure 1. speech enotion recognition system based on HMM

The specific methods to achieve the following :

Step 1 : Initialize parameters of model. It includes the number of Gaussian component and parameters of each Gaussian component.

Step 2: Originally, we are hypothesizing the value of λ in initial GMM model.we get new value of λ using EM until $p(F|\tilde{\lambda}) > p(F|\lambda)$

Step 3 : Repeat Step 2 until $p(F|\tilde{\lambda})$ converges,we get GMM model parameters .

2) Testing

A given vocal signal sequence $\tilde{F} = \{\tilde{f}_1, \tilde{f}_2, ..., \tilde{f}_t, ..., \tilde{f}_T\}$ for test.Probability output of λi in GMM model is $p(\lambda_i|\tilde{F})$. λi is the ith possible emotion .we get it according to Bayesian rule in the form

$$p(\lambda_i \ |\tilde{F}) = \frac{p(\tilde{F}|\lambda i)p(\lambda_i)}{p(\tilde{F})}$$

$p(\tilde{F})$ stays the same when matching .when p(λi) is a certain threshold,what we need to do is to compute $p(\tilde{F}|\lambda_i)$ .Finally,we get the recognition results using Maximum output probability rule. $i^* = \arg\max\ P(F|\lambda_i) \quad 1 \le i \le N$, i* is the emotion we need.

E. Emotion Recognition Results in Mandarin Speech

To test the validity of the algorithms,this paper contrives a few experiments in matlab7.5.We choose 50 speech from mandarin emotional speech database as training data and test data.25 speech signals were randomly selected as training data,others as test data.In experiment,we set pre-emphasis coefficient as 0.9375、the frame length as 32ms、overlap of Frame as 16ms and data window using Hamming window. The dimensions of eigenvectors from MFCC is 12.

We do the experiment related to sex. Training speech signals from persons of the same sex were selected as training set.This system improves measurement precision by averaging the testing values of repeated measurements. The experiment is carried out repeatedly 5 times ,we get the average as the result .

In experiment, the number of Gaussian component (M) is set to 16、32、64、128、256.The accurate identification rate in different genders and the number of Gaussian component is provided in Fig.2 .The emotional recognition ratio of each emotion when M=128 as Tab.1 show.The results show that recognition ratio is greater in men than women.

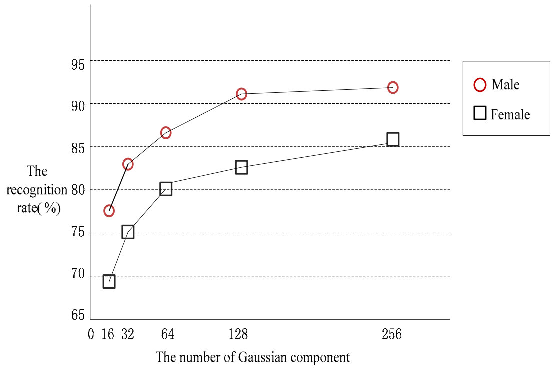

Figure 2. The accurate identification rate of different genders

IV. Conclusion

Automatic speech emotion recognition is a new research area with a wide range of applications in human-machine interactions.It is signification for artificial intelligence research.The feature parameters distillate accuracy influences recognition-rate directly.Two kinds of speech features, long-term and short-term features are studied, to classify five emotional states: anger, happiness, sadness, surprise and a neutral state.To reduce the halving and the doubling errors in pitch tracking, an algorithm based on The wavelet analysis is proposed .we get short-term features of speech .A novel speech emotion recognition, that is speech emotion recognition based on GMM. Experiments show that the recognition is very successful.we will try to improve the method to improve recognition rate in the near future.

References

[1] Chen K,Yue G,Yu F,et al.Reaserch on speech emotion recognition system in e-learning[C].Lecture Note in Computer Science,2007,4489(1):555-558

[2] Razak AA,Yusof MHM,Komiya R.Towards automatic recognition of emtion in speech[C].Proceeding of IEEE International Symposium on signal Processing and Information Technology,2003,548-551

[3] Yacoub, S., S. Simske, X. Lin and J. Burns, "Recognition of Emotions in Interactive Voice Response Systems", In Proceedings of Eurospeech, 2003, Geneva, Switzerland,pp.729-732.

[4] Martin V,Robert V.Evaluation of speech emotion classification based on GMM and data fusion[C].Lecture Note in Artificial Intelligence,2009,5641:98-105

[5] Ramaohan S,Dandapat S.Sinusoidal model-based analysis and classification of stressed speech[J].IEEE Transaction on Audio,Speech and Langusage Processing,2006,14(3):737-746

[6] Park, C.H., K.S. Heo, D.W. Lee, Y.H. Joo and K.B. Sim, "Emotion Recognition based on Frequency Analysis of Speech Signal" ,International Journal of Fuzzy Logic and Intelligent Systems, 2(2), 2002, pp.122-126.

[7] Martin V,Robert V.Recognition of emotion in German speech using Gaussian mixure models[C].Lecture Note in Artificial Intelligence,2009,5641:98-105

[8] Mishra HK,Sekhar CC.Variational Gaussian mixture models for speech emotion recognition[C].Proceedings of International conference on Advance in Pattern Recognition,2009,183-186

TABLE I. EXPERIMENTAL DATA OF DIFFERENT GENDERS (M=128)

	emotion	**neutral**	**happiness**	**anger**	**sadness**	**surprise**
probability using test data of female	neutral	76.6	4.6	3.8	12.6	2.4
	happiness	6.6	83.6	5.6	0.0	4.2
	anger	6.4	13.1	77.2	2.3	1.0
	sadness	6.8	1.2	3.7	87.0	1.3
	surprise	5.0	9.5	7.2	3.2	75.1
probability using test data of male	neutral	94.4	3.6	0.0	0.8	1.2
	happiness	7.0	81.8	8.6	0.0	2.6
	anger	0.0	4.0	94.3	0.0	1.7
	sadness	1.4	3.4	0.0	94.5	0.7
	surprise	1.7	8.2	8.4	1.6	80.1

Published by Atlantis Press, Paris, France.