Speech Emotional Recognition Using Global and Time Sequence Structure Features with MMD

Li Zhao[1,2], Yujia Cao[1,2], Zhiping Wang[2], and Cairong Zou[2]

[1] Research Center of Learning Science, Southeast University,
Nanjing, 210096, China
[2] Department of Radio Engineering, Southeast University,
Nanjing, 210096, China
zhaoli@seu.edu.cn

Abstract. In this paper, combined features of global and time-sequence were used as the characteristic parameters for speech emotional recognition. A new method based on formula of MMD (Modified Mahalanobis Distance) was proposed to decrease the estimated errors and simplify the calculation. Four emotions including happiness, anger, surprise and sadness are considered in the paper. 1000 recognizing sentences collected from 10 speakers were used to demonstrate the effectiveness of the new method. The average emotion recognition rate reached at 95%. Comparison with method of MQDF [1] (Modified quadratic discriminant function), Data analysis also displayed that the MMD is better than MQDF.

1 Introduction

Speech emotion recognition is one of the important parts in emotion processing. The research of the speech emotion can be applied in great many fields, such as Virtual Reality and military affairs. Emotions can be recognized by speech prosody when all word meanings are filtered out [3, 4, 5, and 6]. In addition, many methods are being developed. The multivariable regression and principle component analysis methods have already achieved an average recognition rate above 87.1% [10], while neural net cannot achieve a satisfied result due to the problem of constringency [9]. The global characteristics, which make the whole sentence as a unit, have once been utilized in our research, and the result was good [7]. However, the structure features of time-sequence, which reflects the dynamic characteristics of the emotional change, also have great influences on the recognition of speech emotion. But now, little research focuses on the time sequence, key words and phrases [8]. In this paper, a new emotion recognition method MMD (modified Mahalanobis Distance) based on the combined features of global and time-sequence such as the time, amplitude, pitch and formant construction were presented. And the MQDF [1] method was introduced to compare with MMD. The speech signals, which represent the emotion of happiness, anger, surprise and sadness, have been compared with the neutral speech to find the different emotion

J. Tao, T. Tan, and R.W. Picard (Eds.): ACII 2005, LNCS 3784, pp. 311 – 318, 2005.

features' distribution. The result showed that the new method has recognition rate of 95% among 1000 recognizing sentences collected from 10 speakers. The data analysis also showed that MMD is better than MQDF.

2 Data Collection

Two aspects were taken into account during selecting speech materials in this paper. Firstly, the sentences don't have any emotional tendency; secondly, the sentences can involve all kinds of emotions for analysis and comparison. In addition, the length of the sentence, the construction of consonants, auxiliary words structure and sexual difference were also been taken into consideration. Here, 60 sentences were selected as the speech materials [11]. Neutral, happiness, anger, surprise, and sadness were involved. 10 (males) healthy students (age 20 to 35) were selected in the experiments. The speech signals sampled at a rate of 12KHZ with 16-bit. 3000 sentences (10 speakers×5 emotions×60 sentences) were collected. The data's validity was tested before the experiments, all materials were played randomly, and five listeners (none of the speakers) decided the type of emotions involved in each sentence by their perception subjectively. After repeated comparison and tested by Mcnemar analytical rule [12], the materials, which were ambiguous in emotion, were discarded and they were recorded again until the materials met the need. Among 3000 sentences, 2000 sentences were used for training and 1000 ones were used for test.

3 The Analysis and Extraction of Emotional Characteristics

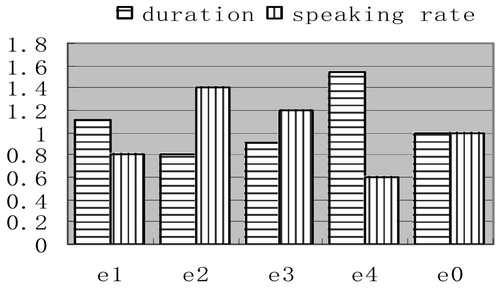
The global features used in this paper include the ratio of duration, changing rate of F_0 (pitch), and F_1 (the first formant), the difference of average amplitude power, mean pitch, maximum pitch, average frequency of the first formant, dynamic range of F_1 and amplitude power between the speech. And the time-sequence features were extracted from the part of pitch in the whole sentence. In MMD the pitch was divided into M parts according to the time and then selected the mean, maximum value and dynamic range of pitch, amplitude power and the frequency of F_1 in each part as the characteristics of time series. While in MQDF, the division is not needed.

3.1 Time

Two parameters were analyzed in time construction: duration of emotional sentence T and the average rate of speaking (syllable/s). T included the parts of silence because these parts contribute to the emotion. An analytic result is shown in figure (1).

3.2 Amplitude

The amplitude of signals closely relates with all kinds of emotional information [8]. Here, the average amplitude energy (A) and the dynamic range (A_{range}) were analyzed. The result is shown in figure (2).



Each word in figure (1) means:

e1	happiness
e2	anger
e3	surprise
e4	sadness
e0	neutral speech

Fig. 1. Time construction

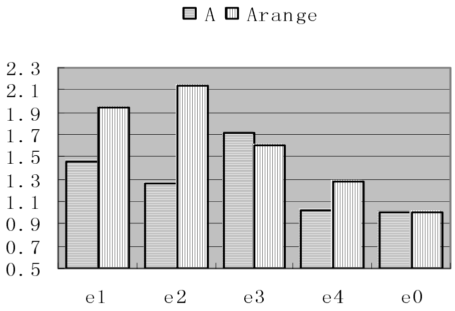


Fig. 2. Amplitude construction

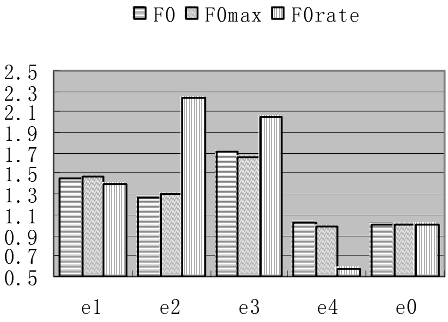


Fig. 3. Pitch construction

3.3 Pitch

The pitch is also an important feature that reflects emotional information [8]. The analyzing characteristics include the average pitch, maximum pitch and the changing rate (F_0, F_{0max}, F_{0rate}) of different emotional speech signals. Here F_{0rate} referred to the mean absolute value of the difference of pitch in each speech signal's frame. Analytic result is shown in figure (3). In addition, because the envelope curve of surprising speech signals is inclined to rise at the end of the sentence, we can distinguish surprise from other emotions.

3.4 Formant

The formant is an important parameter reflecting the features of track, so we researched the parameters of average, dynamic range and changing rate (F_1, F_{1range}, F_{1rate}) of the first formant in our paper. The analytic result is presented in figure (4).

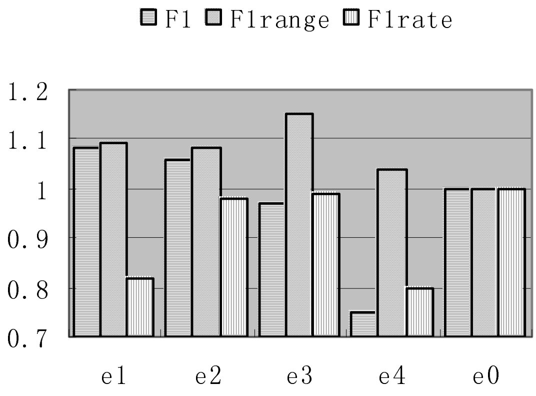


Fig. 4. Formant construction

3.5 Conclusion

According to the above discussion, we got Tab 1 as follow:

Table 1. The variability of the parameters in emotional speech

	T	F_0	$F_{0\ max}$	$F_{0\ rate}$	A	$A_{\ range}$	F_1	$F_{1\ range}$	$F_{1\ rate}$
Happiness	+	+	+	+	+	++	+	+	_
Anger	_	+	+	++	+	++	+	+	_
Surprise	_	++	++	++	++	++	_	+	_
Sadness	++	-	-	_ _	-	+	_ _	+	_ _

(+: increase, ++: increase more, _: decrease, _ _: decrease more, -: no change)

4 Method of Recognition

In this part, MMD is presented in detail firstly, and then MQDF [1] is introduced simply to compare with MMD.

4.1 MMD

The original formula of Mahalanobis Distance is shown in formula (1).

$$d^2(X) = (X - \mu)^t \Sigma^{-1} (X - \mu) \tag{1}$$

Where $X(x_1, x_2, \cdots, x_p)$ is a P-dimension original characteristic vector, μ is mean vector of training samples, Σ is the covariance matrix of the training samples. It is known that the necessary multiplication times of calculating Σ are $p^2 + p$. So the complexity of calculation, EMS memory capacity and the estimation errors of Σ will increase with the increasing of the dimensions of X . Then the recognition percent will decrease. To avoid the above problems, we modified the formula (1). Suppose λ_k is the ith eigenvalue of Σ , ϕ_k is the eigenvector corresponding to λ_k , then we have

$$\Sigma = \sum_{i=1}^{p} \lambda_i \cdot \vec{\phi}_i \cdot \vec{\phi}_i^{\,t} \tag{2}$$

Based on formula (1) and (2), formula (3) could be got.

$$d^2(X) = \sum_{k=1}^{P} \frac{1}{\lambda_k + b} (X - \mu, \phi_k)^2 \tag{3}$$

In order to reduce the effect of the estimated error of eigenvalues on recognizing rate, the offset b was introduced and its value was chosen based on the experimental result. However, because of the application of combined features, the dimension of features space was very big. Moreover, the training data were small relatively, so formula (3) can't simplify calculation to the full extent and raise the recognizing rate. We modified the formula (3) supposing that the covariance matrix of normal Mahalanobis Distance could be segmented as the following form (shown in formula (4)).

$$\Sigma = \begin{bmatrix} \Sigma_1 \ldots\ldots\ldots\ldots\ldots\ldots\ldots 0 \\ \Sigma_2 \cdot\cdot \\ \cdot \\ 0 \ldots\ldots\ldots\ldots\ldots\ldots\ldots \Sigma_M \end{bmatrix} \tag{4}$$

Where $\Sigma_1, \Sigma_2 \ldots \Sigma_M$ is the phalanx of $k \times k$ ($K \times M = P$) . Then the normal *M*ahalanobis *D*istance had the following form (shown in formula (5)).

$$d^2(X) = (X-\mu)^t \Sigma^{-1} (X-\mu) = \sum_{i=1}^{M} (X_i - \mu_i)^t \Sigma_i^{-1} (X_i - \mu_i) = \sum_{i=1}^{M} \sum_{k=1}^{K} \frac{1}{\lambda_{ik} + b} (X_i - \mu_i, \phi_{ik})^2 \tag{5}$$

Where X_i and μ_i are k-dimension vectors and composed by the data from $K_{i-1}+1$ to K_i of vector X and μ respectively. λ_{ik} is the kth eigenvalue of Σ_i , ϕ_{ik} is the eigenvector corresponding to the λ_{ik} . Obviously, the multiplication of formula (5) ($O(K^2M) = O(PK)$) is the $1/M$ times to that of formula (1). Moreover, because the P-dimension vector is segmented to M k-dimension vectors, the ratio of dimension and learning data increase M times for the same amount of learning data when we calculate the covariance matrix of $k \times k$, and the reliability of calculating covariance matrix is more high.

4.2 MQDF

The discriminant criterion based on QDF (quadratic discriminant function) is given as follows:

$$P(\vec{X}\mid k)=\tfrac{1}{2}\ln|\Sigma_k|+\tfrac{1}{2}\left[(\vec{X}-\vec{\mu}_k)^t\,\Sigma_k^{-1}(\vec{X}-\vec{\mu}_k)\right]-\ln p(k) \tag{6}$$

Introduced (2) into (6), we got the modified criterion which is defined as formula (7)

$$P(\vec{X}\mid k)=\sum_{i=1}^{p} {}^{1}\!/_{\lambda_{ik}}\left[\vec{\phi}_{ik}^{\,t}(\vec{X}-\vec{\mu}_k)\right]^2+\ln\prod_{i=1}^{p}\lambda_{ik}-2\ln p(k) \tag{7}$$

Because the estimative errors of eigenvalues and eigenvectors for high-order Σ were higher than them of low-order Σ, a constant $N_0\sigma^2/(N+N_0)$ was used to replace each eigenvalue in $(\lambda_{m+1},\lambda_{m+2},\cdots,\lambda_p)$ which exceeds the threshold of m . Where N represents the total numbers of learning samples, N_0 is a constant, σ^2 is a value which has relation with the eigenvalues. If $N_0/(N+N_0)=\alpha$ $(0<\alpha<1)$, $\alpha\sigma^2=V^2$, introduced the formula (8) into (7),

$$\sum_{i=1}^{p}\left[\vec{\phi}_{ik}\,(\vec{X}-\vec{\mu}_k)^t\right]^2=\|\vec{X}-\vec{\mu}_k\|^2 \tag{8}$$

We got the final calculation formula (9) of MQDF

$$P(\vec{X}\mid k)=\frac{1}{V^2}\left[\|\vec{X}-\vec{\mu}_k\|^2-\sum_{i=1}^{m}\frac{V^2-\lambda_{ik}}{\lambda_{ik}}\left(\vec{\phi}_{ik}^{\,t}(\vec{X}-\vec{\mu}_k)\right)^2\right]+\ln\left(V^{2(p-m)}\prod_{i=1}^{m}\lambda_{ik}\right)-2\ln P(k) \tag{9}$$

5 Results

1000 sentences have been used to the experiment. Let b as 1.3, let M as 4, 6, 8, 10 respectively, and the difference of results was compared. The result is listed in Tab 2.

From Tab 2, it can be see that the recognition result is not always being improved with the raising of *M*. Because of the increase of *M* in each fixed length sentence, the

Table 2. Recognition Result (%)

M Emotion	4	6	8	10
Happiness	89	90	94	91
Anger	91	92	95	92
Surprise	87	88	92	90
Sadness	98	100	100	98
Average	91	93	95	93

amount of data in each part will decrease. Consequently, the analyzing precision will decrease and all these will have influence on the recognition results. So the choice of *M* should be appropriate.

The following list the compared results performed by four methods. In method 1, the global characteristics and MD (Mahalanobis Distance (shown in formula 1)) were used. In method 2, the combined features of global and time-sequence and MD (*M* is 8) were applied. While in method 3, besides the combined features, the MMD presented in this paper was involved to recognize the emotional speech (*M* is 8). And in method 4, the combined features and MQDF were used. The compared results were shown in Tab 3:

Table 3. Recognition Rates [%]

Emotion	Method 1	Method 2	Method 3	Method 4
Happiness	84	82	94	92
Anger	85	83	95	94
Surprise	83	79	92	87
Sadness	95	92	100	100
Average	87	84	95	93

From the Tab 3, it could be seen that the method 3 is excelled to the method 1, 2 and 4. Moreover, the recognizing rate of sadness is higher than any other methods, while the recognizing rate of surprise is lowest among four emotions. And the table also shows that the difference of combined features between the sadness and other emotions (happiness, anger, surprise) is bigger expect for surprise.

6 Conclusion

To find out the distributing rules of different emotional signals, the emotional recognizing method based on the MMD is presented in this paper. The speech signals comprise of four emotions: happiness, anger, surprise and sadness. And MQDF is introduced to compare with MMD. The simulation with 1000 emotional speech sentences was done by MATLAB. We got the result which is closed to the usual behaviors of human. But there is still some work need to do to improve the recognition rate. The more efficient parameters and other modified method will be investigated further.

Acknowledgments

The paper has got the financial support of Education Revitalization Program Oriented to the 21st Century under the Chinese Ministry of Education.

References

[1] *Lili Cai, Chunhui Jiang, Zhiping Wang, Li Zhao, Cairong Zou*, “A Method Combining The Global And Time Series Structure Features For Emotion Recognition In Speech”, IEEE Int. Conf. Neural Networks & Signal Processing, 2003.

[2] *Akemi Iida, Nick Campbell, Soichiro Iga, Fumito Higuchi, Michiaki Yasumura*, “Acoustic Nature and perceptual testing of corpora of emotional speech.”

[3] *Banse, R. & Scherer, K. R.*, “Acoustic profiles in vocal emotion expression,” Journal of Personality and Social Psychology, 70(3), 1996.

[4] *Mozziconacci. S.*, “Speech Variability and Emotion: Production and Perception.” Eindhoven, Netherlands, Technische Universiteit Eindhoven, 1998.

[5] *Scherer, K. R.*, “Speech and Emotional States.” In Darby, J. K. (Ed.) Speech Evaluation in Psychiatry. New York, Grune and Stratton, 1981.

[6] *Soskin W. F. & Kauffman, P. E.*, “Judgements of Emotions in Word-free Voice Samples.” Journal of Communication, 1961.

[7] *Zhao Li, Qian Xiangmin, Zou Cairong, Wu Zhenyang,* “A Study on Emotional Recognition in Speech Signal,” Journal of Software, Vol.12, No.7 2001.7.

[8] *Cowie.R.* “Emotion Recognition in Human-Computer Interaction,” IEEE Signal Processing Magazine, 18(1): 32-80, 2001.

[9] *S.Muraka,* “Emotional Constituents in Text and Emotional Components in Speech,”[Ph. D. Theis] Kyoto, Kyoto Institute of Technology, Japan, 1998.

[10] *M.Shigenaga,* “Features of Emotionally Uttered Speech Revealed by Discriminant Analysis (VI),” The preprint of the acoustical society of Japan, 2-p-18 1999.9.

[11] *Zhao Li, Qian Xiangmin, Zou cairong, Wu Zhenyang,* “A Study on Emotional Feature Analysis and Recognition in Speech Signal,” Journal of China Institute of Communications, Vol.21, No.1, pp18-25 2000.

[12] *Zhao Li, Qian Xiangmin, Zou cairong, Wu Zhenyang,* “A Study on Emotional Feature Extract in Speech signal,” Data Collection and Process, Vol.15, No.1, pp120-123 (2000).