Speech Emotion Classification with the Combination of Statistic Features and Temporal Features

Dan-Ning Jiang, Lian-Hong Cai
Department of Computer Science and Technology, Tsinghua University, China
jdn00@mails.tsinghua.edu.cn, clh-dcs@tsinghua.edu.cn

Abstract

For classifying speech emotion, most previous systems used either statistic features or temporal features exclusively. However, features of these two kinds appear to be concerned with different aspects of emotion, and should be combined in the task. The paper proposes a classification scheme that enables the combination of both statistic features and temporal features. In the scheme, GMM and HMM are first performed to model the statistic features and temporal features respectively. Then the GMM likelihoods and HMM likelihoods of the speech signal to each class are used as features in further classification. Finally, Weighted Bayesian Classifier and MLP are applied to accomplish the classification. Experiments on Chinese speech corpus demonstrated that the classification scheme could improve the classification accuracy greatly. More detailed analysis indicated that these two kinds of features could compensate each other efficiently in the classification.

1. Introduction

Speech signal conveys not only words and meanings, but also emotions. It would be helpful if a computer could recognize the emotion implied in a given utterance. Till now, most emotion classification systems used either statistic features or temporal features exclusively. The first approach extracts the statistics of acoustic parameters to form one feature vector as the representation of an emotional sentence. Then classifiers dealing with fixed-dimension feature vectors, such as GMM, SVM, MLP, etc, are performed. Much differently, the second approach uses frame-based feature vector sequences and models them by HMM to represent the temporal structure of the speech signal. The two approaches both have achieved pretty good results [1] [2], as consequence of the fact that both statistic features and temporal features convey emotion information. However, these two kinds of features seem to be associated with different aspects of emotion [3] [4]. Statistic features, such as the mean and range of F0, appear to be associated with the arousal dimension of emotion. On the other hand, temporal features are more relevant to the communication of valence, attitude, or intention. For example, disposition to seek information is associated with reports of upward movement in the contour. So, it would be helpful to combine both statistic features and temporal features in the classification.

This paper proposes a classification scheme that enables the combination of statistic features and temporal features. In the scheme, GMM and HMM are first performed to model features of these two kinds respectively. Then the GMM likelihoods and HMM likelihoods of the speech signal to each class are used as features in further classification. Finally, weighted Bayesian Classifier and MLP are applied to accomplish the classification. Experiments showed that the proposed scheme could improve the classification accuracy greatly. More detailed analysis indicated that these two kinds of features could compensate each other efficiently in the classification.

The rest of the paper is organized as follows. Section 2 describes the extraction of features. Section 3 explains the proposed classification scheme. Finally, in section 4, experiment results are shown to evaluate the classification scheme.

2. Feature extraction

In this work, all features are derived from F0, log energy, and duration parameters. Raw feature contours

are first extracted. To extract F0, Yin, a modified autocorrelation algorithm, is applied [5]. The algorithm could estimate ratio of the aperiodic power to total power in speech signal (represented as ap) synchronous with the F0 extraction. F0 with low ap are guaranteed, while those with high ap are supposed to fall in the unvoiced regions and deleted. To get a continuous contour, the spline function is used to interpolate between the guaranteed F0 parameters and smooth the contour. Feature contours of log energy and syllable duration are also extracted. Besides, the raw feature contours also include the first and second order derivatives of the F0 and log energy feature contours.

The statistic features are means, standard deviations, and maximums of the above seven raw feature contours. Thus, each sentence is represented by a 21-dimension feature vector. Features of each dimension are freed of their mean and normalized by the standard deviation throughout the database.

To extract the temporal features, each raw feature contour is normalized throughout the sentence. Generally, the temporal features for HMM are frame-based. However, Chinese is a tonal language, whose syllables have specific F0 patterns. Figure 1 shows the F0 contour of a four-syllable-length sentence which means "you dislike it". The first and third syllable has the fall-rise tone, the second syllable has the fall tone, and the last syllable has the high-level tone. Obviously, the F0 contour shape is highly dependent on the local syllable F0 patterns and much more complicated than that in English. HMM may not model such contours well.

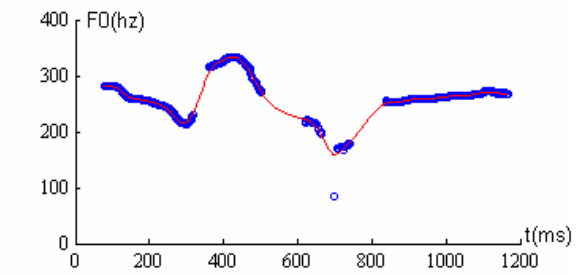

Figure 1. **F0 contour of the Chinese sentence which means "You dislike it".**

The paper uses syllable-based feature vector sequences instead of the frame-based ones. The sequence is syllable-number-length, with each feature vector consists of means, standard deviations, and maximums of the F0 and log energy relevant feature contours in the syllable region, as well as the duration. Thus, variations inside each syllable would be much reduced. Experiment results will show that the syllable-based feature vector sequences achieve a much better performance than the frame-based ones by HMM (in section 4.2).

3. Classification scheme

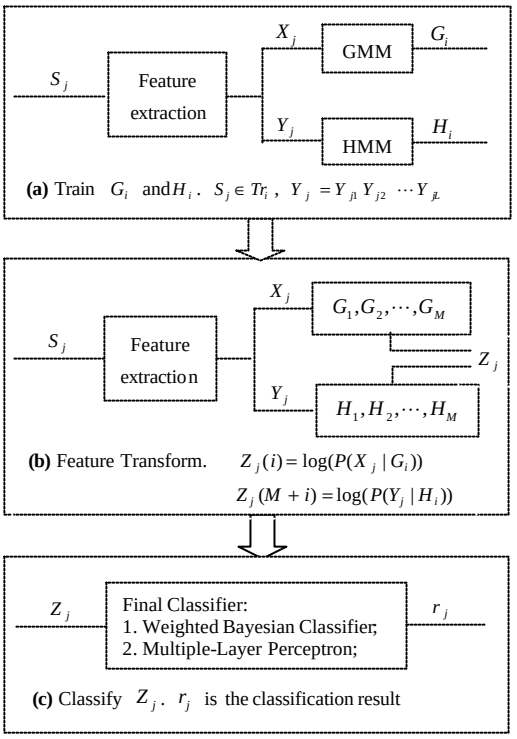

Figure 2. **Paradigm of the classification scheme.**

The classification scheme is composed of three procedures, as illustrated in figure 2. The first procedure (figure 2(a)) is to train the GMM and HMM models for the i-th class, which are represented as G_i and H_i respectively, $0 < i \le M$. S_j is the j-th speech sample in the class-i train set Tr_i; X_j and Y_j are the corresponding statistic feature vector and temporal feature vector sequence respectively.

Figure 2(b) shows the second procedure, which could be interpreted as feature transform. For each speech sample S_j in the whole database, log-likelihoods of X_j to each class are estimated by the relevant G_i, while those of Y_j are estimated by the relevant H_i. Then the GMM likelihoods and HMM likelihoods form the feature vector Z_j for the final classification.

Finally, figure 3(c) shows a typical pattern recognition procedure to get the classification result r_j for each Z_j. In this work, two classifiers are applied. The first one is named as Weighted Bayesian Classifier:

$$r_j = \arg \max_{0<i\le M}[(1-a)Z_j(i)+aZ_j(M+i)] \quad (1)$$

Where a is a weight constant between 0 and 1. It reflects the relative importance of the GMM likelihoods and HMM likelihoods. If a equals 1, then the classification result is entirely determined by the HMM likelihoods. On the contrary, if a equals 0, then it is entirely determined by the GMM likelihoods. When an equaled prior class probability is given, the classifier actually is a Bayesian classifier with the likelihoods to each class as linear combinations of the GMM likelihoods and HMM likelihoods. So it is named as Weighted Bayesian Classifier in the paper.

The second classifier is a one-hidden-layer MLP with 50 hidden neural nodes. MLP could be trained discriminately, and generally has good performance in classification problems.

4. Experiments

4.1. Emotional speech database

The emotional speech database contains six emotion classes: anger, fear, happiness, sadness, surprise, and neutral emotion. There are more than 200 Chinese sentences for each class, which are produced by an amateur actress. Texts for different classes are not all the same, and they all include different sentence types (statements and questions), syllable tones, and accent distributions. All the utterances are recorded in a relative quiet environment, and saved in mono wave files with 16 kHz sample rate and 16 quantitative bits.

In classification, 80 percent of the utterances are used as train data, and the other 20 percent are test data.

4.2. HMM modeling results

This section shows the modeling results by HMM, with both frame-based feature vector sequences and syllable-based feature vector sequences. Table 1 lists the classification accuracy. The HMM state number varies from 2 to 5. It is shown that for state number 2 to 4, the syllable-based feature vector sequences perform much better than the frame-based ones. The accurate rate is at least 17 percent higher. When the state number is 5, HMM fails in modeling some classes with the frame-based feature vector sequences.

Thus, the syllable-based feature vector sequences are used to represent the temporal structure of emotional speech in the below experiments.

Table 1. **Classification results by HMM.** .

Accurate Rate (%)	2-state	3-state	4-state	5-state
Frame-Based	46.1	44.1	57.0	----
Syllable-Based	75.1	73.5	74.4	73.4

4.3. Classification results and comparisons

The classification results by Weighted Bayesian Classifier are illustrated in figure 3. It is shown that with the increase of a, the accurate rate first rises, reaches the maximum when a equals 0.8, and then falls. The highest accurate rate is 80.6%, which is 30.3 percent higher than the accurate rate when a equls 0, and 7.1 percent higher than that when a equals 1. The a value associated with the maximum indicates that the HMM likelihoods are relative more important in the classification.

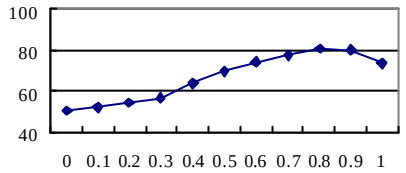

Figure 3. **Classification results by Weighted Bayesian Classifier.**

Table 2 lists the classification results by MLP. The inputs are the GMM likelihoods, HMM likelihoods, and their combination respectively. It shows the similar results as figure 3. By using the combinational features, the accurate rate is 14.5 percent and 10.9 percent higher than those associated with the exclusive use of the GMM likelihoods and HMM likelihoods respectively.

Table 2. **Classification results by MLP.**

Accurate Rate (%)	GMM likelihoods	HMM likelihoods	Both
MLP	68.6	72.2	83.1

More detailed results are further shown to analyze the confusions between different classes. The paper only shows the MLP results, for these two classifiers perform much similarly. Table 3 to 5 are the confusion matrix by using the GMM likelihoods, HMM likelihoods, and their combination respectively. Each emotion class is represented by its first one or several letters.

Table 3. **Confusion matrix by GMM likelihoods.**

(%)	A	F	H	Sad	Sur	N
A	76.7	2.3	7.0	0.0	14.0	0.0
F	4.9	70.7	2.4	0.0	0.0	22.0
H	33.3	0.0	33.3	0.0	28.6	4.8
Sad	0.0	2.4	0.0	95.2	0.0	2.4
Sur	22.7	0.0	18.2	0.0	59.1	0.0
N	3.6	18.2	1.8	0.0	0.0	76.4

Table 4. **Confusion matrix by HMM likelihoods**.

(%)	A	F	H	Sad	Sur	N
A	79.1	7.0	2.3	0.0	7.0	4.7
F	4.9	85.4	2.4	2.4	2.4	2.4
H	7.1	4.8	50.0	0.0	21.4	16.7
Sad	0.0	7.1	0.0	88.1	2.4	2.4
Sur	6.8	2.3	34.1	0.0	54.6	2.3
N	5.5	1.8	5.5	3.6	7.3	76.4

Table 5. **Confusion matrix by combination of GMM likelihoods and HMM likelihoods.**

(%)	A	F	H	Sad	Sur	N
A	81.4	4.7	9.3	0.0	2.3	2.3
F	0.0	100.0	0.0	0.0	0.0	0.0
H	2.4	0.0	83.3	0.0	7.2	7.2
Sad	0.0	4.8	0.0	92.9	0.0	2.4
Sur	9.1	0.0	40.9	0.0	50.0	0.0
N	0.0	5.5	3.6	0.0	0.0	90.9

Table 3 shows that the most confused emotion pairs by using the GMM likelihoods are (anger, happiness), (anger, surprise), (fear, neutral), and (happiness, surprise), while table 4 shows that the most confused emotion pairs by using the HMM likelihoods are (happiness, surprise) and (happiness, neutral). In table 5, except the emotion pair (happiness, surprise), all emotion pairs are separated quite well.

To illustrate the compensation between the statistic features and temporal features more clearly, we defined I_{ij} as measure of the confusion degree between the i-th and j-th class:

$$I_{ij} = P(r = j \mid S \in C_i) + P(r = i \mid S \in C_j) \quad (2)$$

Where S is a test speech sample, and r is the classification result.

Figure 4 compares I_{ij} of the most confused emotion pairs mentioned above. It is shown that the confusion degree associated with the combinational features is much more close to the lower one of those associated with the GMM likelihoods and HMM likelihoods respectively. This indicates that the statistic features and temporal features compensate each other efficiently in the classification.

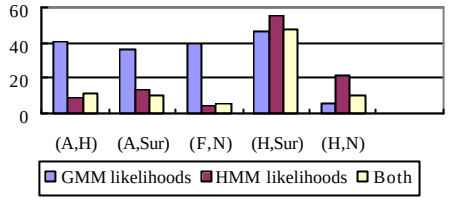

Figure 4. **The comparisons of I_{ij}.**

5. Conclusion

The paper proposes a speech emotion classification scheme that enables the combination of statistic features and temporal features. In the scheme, GMM and HMM are first performed to model the statistic features and temporal features respectively. Then the GMM likelihoods and HMM likelihoods of the speech signal to each class are used as features in further classification. Finally, Weighted Bayesian Classifier and MLP are used to accomplish the classification. Experiments on Chinese speech corpus have demonstrated efficiencies of the classification scheme. By using the combinational features, the accurate rate is at least 14 percent and 7 percent higher than those associated with the exclusive use of the statistic features and temporal features respectively. More detailed analysis indicated that the statistic features and temporal features could compensate each other efficiently in the classification.

6. References

[1] B. Schuller, Gerhard Rigoll, Manfred Lang, "Hidden Markov Model-Based Speech Emotion Recognition", *Proceedings of International Conference on Acoustic, Speech, and Signal Processing*, Vol. II, IEEE Publisher, Jun. 2003, pp.1-4.

[2] O. Kwon, K. Chan, J. Han, etc, "Emotion Recognition by Speech Signals", *Proceedings of Eurospeech*, Sep. 2003, pp. 32-35.

[3] R. Cowie, E. Douglas-Cowie, N. Tsapatoulis, etc, "Emotion Recognition in Human-Computer Interaction", *Signal Processing Magazine*, Vol. 18, No. 1, IEEE Publisher, Jan. 2001, pp. 32-80.

[4] R. W. Picard, *Affective Computing*, MIT Press, Cambridge, Mass. 1997.

[5] A.D. Cheveign, H. Kawahara, "Yin, A Fundamental Frequency Estimator for Speech and Music", *J. Acoust. Soc. Am*. Vol. 111, No. 4, Apr. 2002, pp. 1917-1930.