Emotion Recognition by Speech Signals

Oh-Wook Kwon, Kwokleung Chan, Jiucang Hao, Te-Won Lee

Institute for Neural Computation
University of California, San Diego, USA
{owkwon,kwchan,jhao,tewon}@ucsd.edu

Abstract

For emotion recognition, we selected pitch, log energy, formant, mel-band energies, and mel frequency cepstral coefficients (MFCCs) as the base features, and added velocity/acceleration of pitch and MFCCs to form feature streams. We extracted statistics used for discriminative classifiers, assuming that each stream is a one-dimensional signal. Extracted features were analyzed by using quadratic discriminant analysis (QDA) and support vector machine (SVM). Experimental results showed that pitch and energy were the most important factors. Using two different kinds of databases, we compared emotion recognition performance of various classifiers: SVM, linear discriminant analysis (LDA), QDA and hidden Markov model (HMM). With the text-independent SUSAS database, we achieved the best accuracy of 96.3% for stressed/neutral style classification and 70.1% for 4-class speaking style classification using Gaussian SVM, which is superior to the previous results. With the speaker-independent AIBO database, we achieved 42.3% accuracy for 5-class emotion recognition.

1. Introduction

If an entertainment robot recognize emotion, it could respond to its owner differently according to his/her emotional state. Emotion recognition by speech is one of research fields for emotional human-computer interaction or affective computing [1]. Performance in emotion recognition literatures is hard to compare because of the lack in the common databases (DBs). Previous results do not have common basic emotion categories, speaker-dependency, actors or ordinary people, recording environments, or type of utterances (words or sentences) in their databases. Although it is hard to tell which system is better in general, we describe a few previous systems to provide a general idea on current status and approaches. Scherer selected the best 16 features by the jack-knifing procedure for 14 emotion classification and achieved an overall hit rate 40.4% [2]. Polzin *et al.* used emotion-specific word choice, emotion-specific language model and prosody and spectral information [6]. A recent study reported experimental results on stressed/neutral style classification using the Teager energy operator (TEO) and hidden Markov models (HMMs) with the Speech Under Simulation and Actual Stress (SUSAS) database [5].

Performance of emotion recognition largely depends on how we can extract relevant features invariant to speaker, language, and contents. We decided not to use any linguistic knowledge in this work for generality. To reduce the variability in F0, formant frequencies and duration due to speaker and phonetic dependency, it is crucial to normalize the feature. When a back-end classifier working with a fixed-length feature vector is used for classification, it is also important to convert variable-length feature vectors into a representative fixed-length feature vector. Because the resulting feature vector usually has very large dimension, selecting good features is needed to improve accuracy as well as to reduce the feature dimension.

We extracted base features for each frame and formed feature streams by adding velocity/acceleration components. We extracted statistics on the feature streams for discriminative classifiers. Efficient features for emotion recognition were analyzed. By using the text-independent SUSAS database [5] and the speaker-independent AIBO database [3], we compared emotion recognition performance of various classifiers: support vector machine (SVM), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA) and HMM.

2. Emotion Recognition

Emotion recognition in this work has three stages: Feature extraction, feature selection and classification. Base features and statistics were computed in feature extraction. Feature components were analyzed in feature selection. Classification was made by using various classifiers based on dynamic models or discriminative models.

2.1. Feature Extraction

Figure 1 shows the block diagram of feature extraction. We selected the pitch, log energy, formant, band energies, and mel-frequency cepstral coefficients (MFCCs) as the base features based on the previous study results [2] [3] and our preliminary results. The frame shift in feature extraction was 10ms. We first segmented only speech parts from an input utterance by using an endpoint detector based on zero crossing rate (ZCR) and frame energy. For each frame of speech signals, we estimated F0, log energy, three formant frequencies (F1, F2, F3), five mel-band energies, and two MFCCs [8]. The fundamental frequency was estimated by finding the time shift that minimizes the average mean difference function (AMDF). The formant frequencies were estimated by finding the poles of the autoregressive transfer function obtained from the linear predictive coding (LPC) coefficient.

We also added velocity and acceleration information for pitch and MFCCs, respectively, to take the rate of speaking into account and model the dynamics of the corresponding temporal change of pitch and spectrum. Hence we have 15 streams of features including velocity and acceleration components. These streams were used as input vectors of HMM-based classifiers.

For discriminative classifiers, the feature streams were converted into a fixed-length vector for each utterance by computing statistics to represent the streams. Note that conversion to a fixed-length vector was performed only for discriminative classifiers such as SVM, LDA and QDA.

To obtain a fixed-length vector, we computed 11 statis-

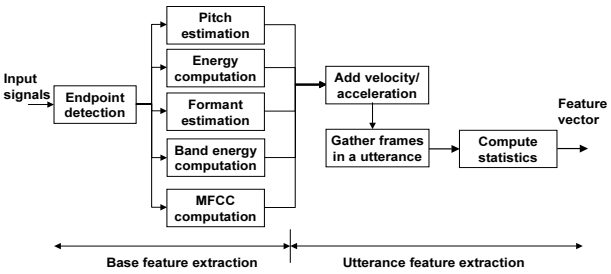

Figure 1: Block diagram of the feature extraction module.

tics from pitch, 7 from energy, 5 for each formant frequency, 4 for each MFCC, and 1 for mel-band energies. The statistics included the 90th percentile, range, mean, standard deviation, skewness, and so on. For pitch we added the pitch of the first frame, the pitch of the last frame, the mean and the regression coefficient of the first/last segments. Regarding log energy, we subtracted the mean log energy to normalize amplitude change according to the speaker volume or the distance of a speaker from a microphone. Two duration-dependent components were added. One is the mean MFCC distance between adjacent frames and the other is the duration in frames divided by the mean MFCC distance. The dimension of one feature vector, the input vector of the feature selection, was 59 for each utterance. The pitch and formants in silence regions were interpolated from adjacent frames so that there are no discontinuities on the contours of the pitch and formants. To obtain the relevant quantities, we used only the voiced region of the input utterance.

2.2. Feature Selection

Out of the many derived features given to the classifiers, we want to identify those that contribute more in the classification. This tells us what features and properties of the speech are important in distinguishing emotions. We could then derive more relevant features accordingly to improve classification accuracy. However, it is forbiddingly time consuming to perform exhaustive search for the subset of features that give best classification. Instead, we used forward selection and backward elimination to rank the features and identify the subset that contributes more in classification. Forward selection sequentially adds one feature at a time, choosing the next one that most increases or least decreases classification accuracy. Backward elimination starts with all input features and sequentially deletes the next feature that most decreases or least increases classification accuracy.

2.3. Classification

We compared the performance of classifiers based on discriminative and generative models by using SVM, LDA, QDA and HMM. Hsu and Lin [9] compared various methods proposed to extend the binary SVM to multi-class and found that the "one-against-one" scheme is more suitable for practical use. We used their scheme for the multi-class problem. For binary classification, we used our MATLAB implementation of the sequential minimal optimization SVM [7].

The HMM-based classifier has the advantages over other static discriminative classifiers that frame length normalization is not necessary and temporal dynamics of the base features can be reflected by using the state transition probability. Short-time temporal dynamics is implicitly modeled through the addition of velocity and acceleration components. However, the HMM classifier is still weak at modeling long-time temporal dynamics.

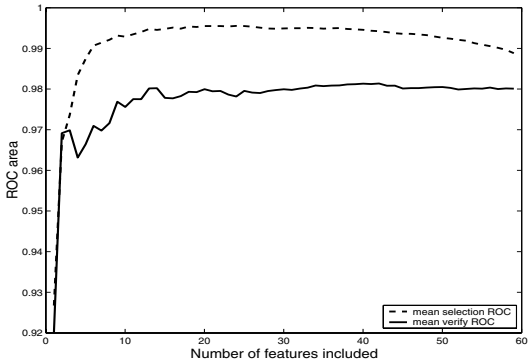

Figure 2: ROC area versus number of features included in backward elimination. The dashed line is the mean cross-validation result of the training data and the solid line is on the test data.

3. Experimental Results Using SUSAS DB

To evaluate performance of the proposed feature extraction method, we chose the SUSAS database [5] designed originally for speech recognition under stress. We used isolated words recorded at 8 kHz sampling rate in various speaking styles. Following the experimental setup of text-independent stress style classification in [5], we chose 3 stressed styles (angry, Lombard and loud) and a neutral style.

3.1. Individual Feature Selection Results

We used the binary Gaussian kernel SVM (GSVM) as our base classifier for feature selection on the task of stressed versus neutral speech classification. The ROC area [10] was used as the criteria for selecting the next feature to include or delete. We performed feature selection by text independent cross-validation on the training data. The test data was used afterward only to assess the feature selection result. First, we divided the training data into six partitions according to the spoken words, utterances of the same word fall into the same partition. In forward selection, at each stage of adding feature, each "word" took turns to be held aside while the classifier was trained on the rest five words. The ROC areas over the six words were then combined to determine which feature should be added. Backward elimination was done in a similar fashion.

In Figure 2, we plotted the cross-validation ROC area, in dashed line, of the training data against the number of features included from backward elimination. Plotted together is the ROC area on the test data, in solid line. The ranked features are plotted in Figure 3 on a two dimensional space where the x-y coordinate of each feature is its rank by the backward and forward selection. The ranks from forward and backward selection roughly agreed with each other.

3.2. Group Feature Selection Results

We divided the features into 13 groups and checked the importance of each group. We used the same methodology as in the individual feature selection case, except we considered the features in group at each step of feature addition/deletion. Backward elimination gave exactly the same ordering as forward selection. The 13 groups were ordered as pitch, log energy, F1, mel-band energy, acceleration of pitch, velocity of pitch, the first MFCC, acceleration of MFCC, duration, F2, velocity of MFCC, F3, and the second MFCC. This showed that pitch and energy are more essential in distinguishing stressed and neutral speech, while MFCCs are less important. This order agreed with that seen in Fig. 3.

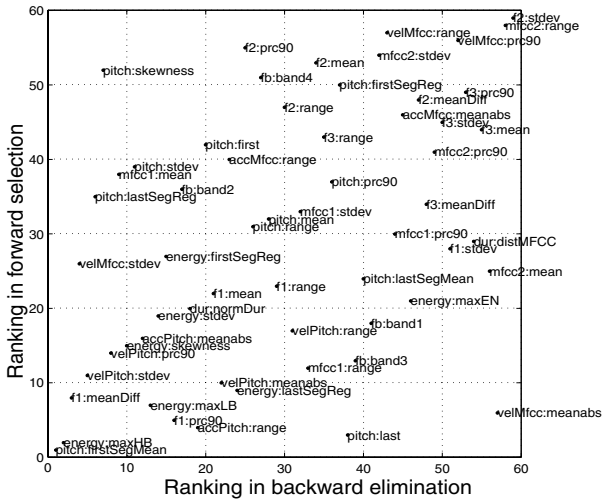

Figure 3: Two-dimensional plot of the features ranked by forward selection and backward elimination. Features near origin are considered to be more important.

3.3. Speaking Style Classification Results

Following the setup of the previous study [5] using the same database, we performed stressed/neutral style classification. We used the GSVM as a classifier. The feature dimension was 59 feature including all statistics obtained from utterances. The detection rates of neutral and stressed utterances were 90.0% and 92.6%, respectively.

By varying a threshold to evaluate the performance with different control detection rate and false alarm rate, we obtained the receiver operating characteristics (ROC) curve as shown in Figure 4. The square mark indicates the performance of the previous study using TEO-based feature and the triangle mark shows the performance with the pitch-based feature. Our feature and classifier performed better than both of the previous results by 3% absolutely.

We performed 4-class style classification using the same configuration as stressed/neutral classification. Table 1 shows the confusion matrix of the classification. The average accuracy was 67.1% and the most confusing pair was loud-angry.

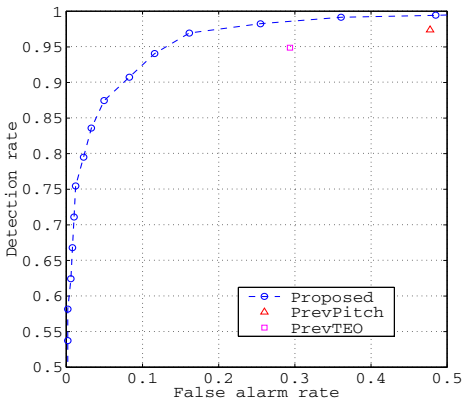

Figure 4: ROC curve of stressed/neutral style classification compared with the previous results.

We used the Hidden Markov Model Toolkit (HTK) [8] to test the performance of the HMM-based classifier. The input was the 15 variable-length feature streams, which consisted of pitch, log energy, F1, F2, F3, 5 filter bank energies, 2 MFCCs, delta pitch, 2 delta MFCCs, acceleration of pitch, and 2 acceleration MFCCs. Each speaking style was modeled by a 5- state HMM with the observation probability distribution of a single full covariance Gaussian distribution. For stressed/neutral style classification, the HMM-based classifier showed 96.3% average accuracy as shown in Table 2. The detection rate was also better than the SVM classifier. The HMM-based classifier showed better performance because the test utterances were mostly short commands and the HMM could mode the temporal change properly. Our experimental results showed that the average accuracy of the HMM-based classifier was 70.1% for 4-class style classification. The tendency of confusion was similar to the case of SVM.

Table 1: Confusion matrix of 4-class style classification using GSVM for the SUSAS database

	Neutral	Angry	Lombard	Loud
Neutral	0.934	0.025	0.037	0.004
Angry	0.093	0.672	0.111	0.124
Lombard	0.164	0.191	0.598	0.048
Loud	0.064	0.294	0.162	0.480

Table 2: Stressed/Neutral style classification using HMM

	Neutral	Stressed
Neutral	0.918	0.082
Stressed	0.023	0.977

4. Experimental Results Using AIBO DB

We used the emotional database in German recorded at 16 kHz by the Sony entertainment robot [3]. This database included short commands or greetings with several words. The emotion expressed in the database was not exaggerated so that even native speakers were often confused about the emotion of the utterances. The basic set of emotions included angry, bored, happy, neutral and sad. The 3534 utterances were used as the training data set and the 1681 as the test data set.

4.1. Pair-wise classification Results

In this experiment, data of one emotion in the first column are mixed with data of another emotion selected from the first row of Table 3. Then we used four classifiers: LDA, QDA, GSVM, linear SVM (LSVM) to separate the mixed two emotions. Note that only the first two characters were used to denote abbreviations of the classifiers. The numbers shown in the table are the true positive value of the test data, i.e. the percentage of the test data over the first column emotion that is correctly classified as that emotion. For example, the first one, 74 means that 74% of angry-emotion test data are correctly classified as angry emotion by using LDA when mixed with bored emotion data. The other three classifiers except QDA gave consistent results, and GSVM was the best most of the time. The happy emotion was most confusing with the angry emotion and they are grouped as high emotion in [3]. The sad emotion was most confusing with the bored emotion and they are grouped as low emotion in [3], which are consistent with the previous results.

A two-dimensional plot for the AIBO database similar to Fig. 3, showed that the 90th percentile of energy in speech and pitch (especially the rate of change) are fundamental in emotion classification, while the skew and kurtosis measure of the various features are least useful in classification. It is believed

Table 3: Pairwise classification accuracy (%) using LDA, QDA, GSVM and LSVM

	Angry		Bored		Happy		Neutral		Sad	
A	LD	GS	74	77	63	65	78	74	82	86
	QD	LS	69	76	34	62	73	76	77	82
B	82	81			73	75	65	66	57	55
	79	80			63	75	58	65	49	55
H	62	66	74	76			76	80	81	83
	86	63	76	75			86	80	86	80
N	68	66	56	61	58	56			69	69
	64	68	60	58	32	57			61	65
S	86	86	59	67	78	80	72	74		
	82	85	65	65	69	80	65	76		

that loudness and change in pitch of speech would separate angry/happy from bored/neutral/sad, it would be interesting to further identify what features to separate out angry from happy, and among bored, neural and sad.

4.2. Multi-class Classification Using GSVM and HMM

The Gaussian SVM gave the best multi-class classification on the emotion data. Shown in Table 4 is the confusion matrix by GSVM. GSVM achieved an overall accuracy of 42.3%. In the HMM, all emotion models had the same number of states and mixtures. The silence model was used in the beginning and ending of an utterance, and its number of states was set to 5. An observation density for each state was modeled by 16 Gaussian mixtures with diagonal covariance matrices. The performance was improved mostly through increasing the number of states, that is, detailed temporal modeling. The average classification accuracy was 40.8%. Classification accuracy with full covariance matrices was inferior by 1-2% to the diagonal matrix case. The confusion matrix showed similar tendency as the SVM case.

Table 4: Confusion matrix of 5-class emotion recognition using GSVM for the AIBO database

	Angry	Bored	Happy	Neutral	Sad
Angry	0.602	0.085	0.195	0.093	0.024
Bored	0.137	0.379	0.147	0.150	0.188
Happy	0.343	0.093	0.407	0.113	0.044
Neutral	0.250	0.230	0.123	0.293	0.103
Sad	0.085	0.349	0.064	0.132	0.370

5. Discussion

We found that performance of the SUSAS database differed from the AIBO database due to speaking style mostly. Speakers in the SUSAS database seem to express their emotion or speaking styles more strongly than the AIBO database. For some utterance in the AIBO database, even native speakers had difficulty identifying the labeled emotion. While the SUSAS database consists of short command words with more constant duration, the AIBO database has larger variation in duration spanning from short commands to long sentences. Another difference is that the SUSAS experiments are for speaker-dependent and text-independent, but the AIBO experiments are speaker-independent and text-dependent where more words were included in the AIBO database. Comparing the AIBO results with the previous results [3], the performance level is quite similar.

The experimental results suggest that we use a Gaussian SVM-based classifier in applications using variable-length utterances and an HMM-based classifier in applications using short utterances. We still have to find a good way to impose speech characteristics in the long time scale when HMM-based classifiers are used. Performance difference with different discriminative algorithms is rather small. Good feature extraction is more critical factor in emotion recognition than classifier selection. Core modules in feature extraction (pitch tracking and formant tracking) largely affect the final accuracy. When discriminative classifiers are used, feature-length normalization is another important factor.

6. Conclusions

Using statistics computed from the base features, we analyzed the effects of the features in emotion recognition. The pitch and energy were shown to play a major role in recognizing emotion, which matches insights. We performed classification using SVM, LDA, QDA and HMM with the SUSAS and AIBO databases. Both SVM and HMM classifiers yielded classification accuracy significantly better than the previous results in the SUSAS database. For the AIBO speech database, we evaluated classification accuracy for 5 emotion classes. Further study is needed to explore new features better representing prosody and timbre, improve the pitch and formant tracking algorithms, and develop a new sophisticated approach to model dynamics of feature streams.

7. References

[1] R.W. Picard, "Affective computing," MIT Media Lab Perceptual Computing Section Tech. Rep., No. 321, 1995.

[2] K.R. Scherer, "Adding the affective dimension: A new look in speech analysis and synthesis," Proc. ICSLP, 1996.

[3] R. Tato *et al.*, "Emotional space improves emotion recognition," Proc. ICSLP, 2002.

[4] S. McGilloway *et al.*, "Approaching automatic recognition of emotion from voice: A rough benchmark," ISCA Workshop, Speech and Emotion, 2000.

[5] G. Zhou *et al.*, "Nonlinear feature based classification of speech under stress," IEEE Trans. Speech, Audio Proc., vol. 9, pp. 201-216, Mar. 2001.

[6] T. S. Polzin, A. Waibel, "Emotion-sensitive human-computer interfaces," ISCA Workshop, Speech and Emotion, 2000.

[7] J.C. Platt, "Fast training of support vector machines using sequential minimal optimization", in Advances in Kernel Methods: Support Vector Machines, B. Scholkopf, C.Burgees, and A. Smola, Eds. Cambridge, MA: MIT Press, 1998, pp. 185-208.

[8] S. Young *et al.*, The HTK Book, 2001.

[9] C.-W. Hsu and C.-J. Lin. "A comparison of methods for multi-class support vector machines," IEEE Transactions on Neural Networks, 13, pp. 415-425, 2002.

[10] D.J. Goodenough *et al.*, "Radiographic applications of receiver operating characteristics ROC curve," Radiology, vol. 110, pp. 89-96, 1974.