Statistical Speech Feature Selection for Emotion Recognition

Oh-Wook Kwon*, Kwokleung Chan**, Te-Won Lee**
*Chungbuk National University, **University of California, San Diego

(Received November 16 2005; accepted December 12 2005)

Abstract

We evaluate the performance of emotion recognition via speech signals when a plain speaker talks to an entertainment robot. For each frame of a speech utterance, we extract the frame-based features: pitch, energy, formant, band energies, mel frequency cepstral coefficients (MFCCs), and velocity/acceleration of pitch and MFCCs. For discriminative classifiers, a fixed-length utterance-based feature vector is computed from the statistics of the frame-based features. Using a speaker-independent database, we evaluate the performance of two promising classifiers: support vector machine (SVM) and hidden Markov model (HMM). For angry/bored/happy/neutral/sad emotion classification, the SVM and HMM classifiers yield 42.3% and 40.8% accuracy, respectively. We show that the accuracy is significant compared to the performance by foreign human listeners.

Keywords: *Emotion Recognition, Support Vector Machines, Hidden Markov Models*

I. Introduction

Emotional human-computer interaction is one of emerging research fields in affective computing[1]. In particular, emotional human-robot interaction with an intelligent robot draws much attention because it can make the robot more human-like and more user-friendly. Emotion recognition and synthesis can be done through video signals and/or speech signals. Emotion recognition via a single modality, speech, is favorable in terms of computational complexity and required hardware.

Emotion recognition performance is hard to compare fairly because researchers have used different speech databases in their works. First there is no consensus in the basic emotion set. Most of researchers counted 'angry', 'happy', 'sad', and 'surprise' emotion in the basic emotion set. A few researchers, however, added 'fear' and 'disgusted' emotion[2-3] and the MPEG-4 also defined the same 7 emotions as emotional styles including neutral[4]. Second, the speech databases showed different fluency in expressing emotion. Often, a trained actor played a given situation and hence expressed emotion better than ordinary speakers. In addition, the speech databases often had incompatibility in speaker-dependency, recording environments, or the number of words in an utterance. In spite of the above issues, we describe and compare previous systems to provide a general idea on the state-of-the-art performance level and technologies in feature extraction and emotion classification. The previous studies are summarized into three categories: Emotion recognition based on acoustic information only, combining linguistic information, and audio-visual emotion recognition.

By virtue of its simplicity, most of emotion recognition systems used only acoustic information[5-7]. Some researchers performed stressed/neutral style classification using the Teager energy operator and hidden Markov models for the Speech Under Simulated and Actual Stress (SUSAS) database[8-9]. Recently Ververidis *et al.* reported 51.6% accuracy with 5 emotion categories while human performance was 67.3%[10].

Performance of emotion recognition largely depends on how we can extract relevant features invariant to speaker, language,

Corresponding author: Oh-Wook Kwon (owkwon@chungbuk.ac.kr)
School of Electrical and Computer Engineering, Chungbuk National University, 12 Gaesin-dong, Heungdeok-gu, Cheongju, Chungbuk 361-763, Korea

and contents. There are differences as well as similarities among cultures and languages in representing emotions[11]. With this view, some researchers utilized linguistic information to improve classification accuracy. Polzin et al. used verbal information such as emotion-specific word choice, emotion-specific back-off language models and non-verbal information including prosody and spectral information[12]. Lee and Narayanan improved negative/non-negative detection accuracy by 40.7% for males by combining acoustic and linguistic information[13].

This paper reports emotion recognition performance in real situations where a person interacts with an entertainment robot. While most of previous studies[5-6] use speech data with strong emotion played by actors, this work uses speech data uttered by ordinary persons. We use only acoustic information for emotion recognition. Utilizing linguistic information needs a speech recognizer, which increases system complexity and makes migration to other languages hard. This work also presents some insights into contribution of speech features to emotion recognition and gives comparative evaluation results of support vector machine (SVM)-based classifiers[18] and hidden Markov model (HMM)-based classifiers[15].

The rest of the paper is organized as follows: Section II describes the feature extraction algorithms, feature selection methods, and emotion classifiers adopted in our work. Section III describes the speech database, discusses contribution of each feature, and gives experimental results. The human performance and the factors affecting emotion recognition are discussed in Section IV. Conclusions are given in Section V.

II. Emotion Recognition

Figure 1 shows the block diagram of the emotion recognizer used in this work. In the feature extraction module, frame-based and utterance-based features are extracted. The feature selection module selects feature components to reduce the feature dimension. Finally an emotion classifier decides the emotion based on the feature components.

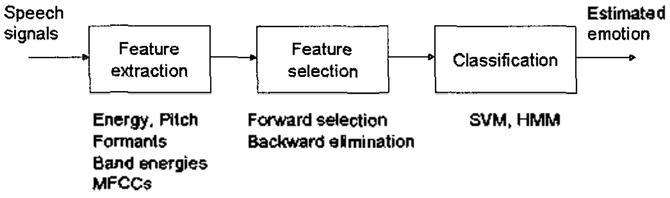

Figure 1. Block diagram of an emotion recognizer.

2.1. Feature Extraction

Figure 2 shows the block diagram of the feature extraction module. The frame shift in feature extraction was 10 ms. We reduce noise by using the Wiener filter[17] and segment only speech parts from an input utterance by using an endpoint detector based on zero crossing rate (ZCR)[16] and frame energy. The frame size to extract all the frame-based features except pitch is 25 ms. In this paper, we regard that the "pitch" and "fundamental frequency" are interchangeable.

2.1.1. Frame-Based Features

In selecting the frame-based features, we have been inspired from the previous studies on the effects of emotion on acoustic-phonetic parameters[5]. Angry speech increases in mean pitch and mean intensity, has higher pitch variability and a wider range of pitch. Angry speech also has larger high frequency energy, a downward directed pitch contour and an increased rate of articulation. Sad speech shows decrease in mean pitch, pitch range, pitch variability. Sad speech has also the pitch contour downward directed; small high frequency energy and rate of articulation. Happy speech increases in pitch, pitch range, pitch variability and mean intensity. It also increases in high frequency energy and in rate of articulation. Spectral information embedded in formants and mel-frequency cepstral coefficients (MFCCs) also reflects the characteristics of speech waveform[14]. Details of the frame-based features are described in the followings.

- Energy

We use the log energy defined as the log of sum of absolute sample values.

- Pitch

Pitch is estimated by finding the time shift that minimizes the average mean difference function (AMDF)[16]. We use the frame size of 60 ms in order to include at least 3 pitch periods in a frame assuming the minimum pitch is 50 Hz. A frame of speech data is windowed by the Hanning window. Every frame is classified into voiced/unvoiced/silence using energy and ZCR. This information is used to derive some statistics from only voiced regions. For example, a continuous contour is needed to compute linear regression coefficients for pitch. The pitch is set to 0 if the frame is not voiced. The time index with the minimum AMDF at the current frame is searched in the plausible range, which is computed by extrapolating the index of the previous last voiced frame based on the maximum admissible change of pitch. For the maximum admissible pitch change, we use 20% at the start of voiced frames and 10% at the inside of voiced frames. If the index is found outside the plausible range, the range is doubled and the same search procedure is repeated. Finally the pitch contour is smoothed with a median filter of length 7 to remove spurious pitch values.

- Formant frequencies

The three formant frequencies (F1, F2, and F3) of a frame are computed from the poles of the all-pole filter which models the vocal tract[16]. The linear predictive coding (LPC) coefficients are computed by using the autocorrelation method[17]. We smooth formant trajectories with a median filter of length 7 and reduce discontinuities as much as possible by post-processing[16]. The formant frequencies are set to 0 for unvoiced or silent frames.

- Band energies

To obtain band energies, we first compute 23 filter-bank coefficients using the feature extraction standard proposed by European Telecommunication Standard Institute (ETSI) for distributed speech recognition[17]. After partitioning the filter bank coefficients into 4 bands, we obtain the band energy by summing all coefficients allocated to a band. The first 3 bands included 5 coefficients sequentially starting from the first coefficient the last band included the remaining 8 coefficients.

- MFCCs

The MFCCs are also computed using the ETSI feature extraction standard[17]. Only the first two MFCCs are used because the two low-order coefficients represent the overall shape of the spectrum while the higher coefficients depend on the phonemic identity of the speech signals.

- Adding Velocity and Acceleration

This module works only for frame-based classification in order to consider the speaking rate. Our preliminary experiments have shown that a faster utterance shows larger pitch variation and spectral change. We add the velocity and acceleration information for pitch and MFCCs, which is widely used for speech recognition to model speech dynamics efficiently[19]. The velocity information vel(t) is obtained by filtering the corresponding pitch and MFCC contours in the temporal direction with a finite impulse response (FIR) filter with the temporal width of 3 frames in our experiments.

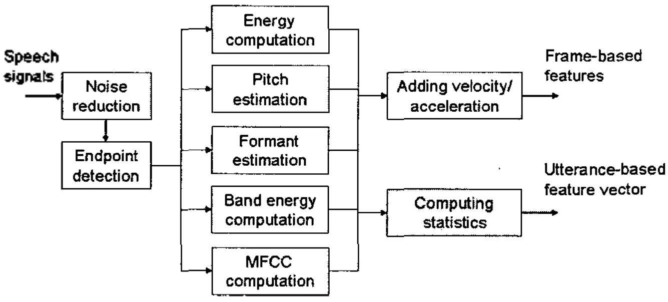

Figure 2. Block diagram of feature extraction.

To reflect the spectral change, the variable spch (t) is defined as the Euclidean norm of the velocity vector for the 12 MFCCs. We compute the spectral change from MFCCs to share computational burden although it may be calculated from the filter-bank coefficients.

To further model the dynamical aspects of the pitch and MFCCs, we use the acceleration information acc(t), which is obtained by filtering the velocity information using the same FIR filter.

Resulting 15 frame-based features constitute a feature vector for HMM-based classifiers.

2.1.2. Utterance-Based Features

While the frame-based features can be used for emotion classification in case of frame-based classifiers such as HMM, we need to convert the frame-based features into a fixed-length feature vector when a back-end classifier works for static pattern recognition, e.g., for SVM. Regarding every frame-based feature as a feature stream, we compute different statistics[24] according to the nature of the frame-based feature and obtain an utterance-based feature vector with the dimension of 59 for each utterance. The pitch and formant frequencies in unvoiced and silence regions are interpolated from adjacent frames so that there are no discontinuities on the contours.

- Computing Statistics

For the pitch stream, we compute 11 statistics the mean, standard deviation, maximum (90th percentile), range between the maximum and minimum (the 10th percentile), skewness, the value of the first frame, the value of the last frame. Another 4 statistics are the mean pitches and linear regression coefficients of the first and last voiced segments. We use the 10th and 90th percentile instead of the minimum and maximum values of the base features to avoid spurious outliers in obtaining the range. The first and the last voiced segments were considered to reflect the fact that energy and pitch of the start and end of an utterance are affected by emotional states.

For the energy stream, we compute 7 statistics: range, low band energy, high band energy, standard deviation, skewness, and the regression coefficients of the first and last voiced segments. We subtract the mean log energy to normalize amplitude according to the speaker's volume.

For the 3 formant frequency streams, we compute 5 statistics each: the mean, standard deviation, maximum, range, and mean distance from the mean pitch.

For the 2 MFCC streams, the maximum, range, mean, and standard deviation were computed, respectively. For the 4 mel-band energies, the mean value was computed for each coefficient. For velocity components of pitch and MFCCs, we obtain the maximum, range, mean absolute value, and standard deviation. For acceleration components of pitch and MFCCs, we use the range and mean absolute value.

We add two duration-dependent components derived from MFCCs. One is the mean MFCC distance between adjacent frames and the other is the duration in frames divided by the mean MFCC distance.

2.2. Feature Selection

Among the many derived features, we want to identify those that contribute more in the classification. This tells us what features and properties of speech are important in distinguishing emotions. Because the feature vector usually has very large dimension, we can improve accuracy as well as reduce the computational complexity by selecting good features.

However, it is forbiddingly time consuming to perform exhaustive search for the subset of features that give best classification. Instead, we used the forward selection and backward elimination methods[20] 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 classification accuracy. Backward elimination starts with the set of all input features and sequentially deletes the next feature that results in least decrease classification errors. Figures 3 and 4 show the pseudocode of the forward selection and backward selection algorithms, respectively.

2.3. Classification

We compared the performance of classifiers based on discriminative and generative models by using SVM[18] and HMM[15], respectively.

```
function ForwardSelection
Select[0] = {};
Remain = {All features};
for n=1 to numFeatures
 for each f in Remain
  Temp = Select[n-1] + f;
    bestAcc = 0; bestFeat = {};
    Train and test using Temp feature set;
    Compute classification accuracy acc;
    if acc〉bestAcc then
      bestAcc = acc;
      bestFeat = Temp;
    end
  end
  Select[n]=bestFeat;
end
```

Figure 3. Forward selection algorithm.

```
function BackwardSelection
Select[numFeatures] = {All features};
Remain = {};
for n=numFeatures-1 to 1
  for each f in Select
    Temp = Select[n+1] - f;
    bestAcc = 0; bestFeat = {};
    Train and test using Temp feature set;
    Compute classification accuracy acc;
    if acc〉bestAcc then
      bestAcc = acc;
      bestFeat = Temp;
    end
  end
  Select[n]=bestFeat;
end
```

Figure 4. Backward selection algorithm.

The SVM is a recently developed technique for solving a variety of binary classification and regression problems. Commonly used kernel functions include the linear, polynomial, Gaussian, and sigmoidal kernels. Hsu and Lin[21] compared various methods proposed to extend the binary SVM to multi-class. They found that the "one-against-one" method is the most suitable for practical use. We used a MATLAB interface version of their LIBSVM[22] for the multi-class problem in emotion recognition.

In HMM-based classification[15], each feature stream is assumed to be generated from a first-order hidden Markov process. In each state of a Markov process, a feature has the observation probability given by a mixture of Gaussian pdfs. We computed the log likelihood of the feature stream and decide the emotion with the maximum likelihood as the final classification result. The HMM-based classifier has the advantages over other static discriminative classifiers that frame length normalization is not necessary. Short-time temporal dynamics is implicitly

Table 1. Confusion matrix (%) of multi-class Gaussian SVM.

	Angry	Bored	Happy	Neutral	Sad
Angry	60.2	8.5	19.5	9.3	2.4
Bored	13.7	37.9	14.7	15.0	18.8
Happy	34.3	9.3	40.7	11.3	4.4
Neutral	25.0	23.0	12.3	29.3	10.3
Sad	8.5	34.9	6.4	13.2	37.0

modeled through the addition of velocity and acceleration components. However, the HMM classifier still has a weakness in modeling long-time dynamics.

III. Experimental Results

3.1. Speech Database

We used the German emotional database recorded with the Sony entertainment robot, AIBO[14]. The sampling frequency was 16 kHz. This database included commands or short greetings with several words. Emotion expression was mostly weak so that even native speakers were often confused about the emotion of the utterances. The set of emotions used in the experiments included 5 emotion classes: angry, bored, happy, neutral and sad. 3534 utterances were used as the training data set and the remaining 1681 utterances were used as the test data set (2:1 data ratio). The training data set included 10 male and 10 female speakers and the test data set included independently another 5 male and 5 female speakers. We used the 2:1 ratio because preliminary experiments with a different organization (4:1 data ratio) had yielded almost similar classification results.

3.2. SVM-Based Classification

We used all 59 features in SVM-based classification experiments. The Gaussian SVM gave the best multi-class classification on the emotion data. Table 1 shows the confusion matrix, and Gaussian SVM achieved an overall accuracy of 42.3%. The easiest to detect was angry emotion and the next easiest ones were happy, bored, and sad emotions. Neutral speech was often misclassified as emotional speech. We note that the confusion matrix is asymmetric. Neutral speech is more often misclassified to angry speech that angry speech to neutral speech.

3.3. HMM-Based Classification

We used the HTK[19] to test the performance of the HMM-based classifier. The covariance of a state was diagonal matrix. 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. Table 2 shows the classification results for 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%. This accuracy improvement was not significant because the 95% confidence interval in the significance test was ±2.4%. Classification accuracy with full covariance matrices, not shown here, was similar to the diagonal matrix case.

Table 2. Confusion matrix (%) of HMM-based classifier with diagonal covariance matrices.

	Angry	Bored	Happy	Neutral	Sad
Angry	61.7	4.1	20.1	11.9	2.2
Bored	11.8	29.0	17.2	26.4	15.6
Happy	30.2	5.2	43.4	19.3	2.0
Neutral	25.0	13.8	19.1	32.9	9.2
Sad	12.4	27.4	7.4	23.4	29.4

3.4. Group Feature Selection Results

We performed group feature selection based on a multi-class classifier. We divided the 59 features into 13 groups: pitch, pitch velocity, pitch acceleration, energy, F1, F2, F3, mel-band energy, MFCC velocity, MFCC acceleration, MFCC1, MFCC2 and duration. We used the same methodology as in the individual feature selection case. But in this case, we considered a feature group at each step of addition or deletion.

The mean true positive (TP) rate over the five emotions was used as the criteria for selecting the next feature to include or delete. We performed feature selection by cross-validation on the training data. The test data were used afterward only to assess the feature selection result. First, we divided the training data into 5 partitions. To achieve speaker independent feature selection, the 5 partitions contain mutually exclusive speakers. In forward selection, at each stage of adding the next feature group, each partition took turns to be the "held-out" set while the classifier was trained on the rest four. The TP rates over the five partitions were then combined to determine which feature group should be added. Backward elimination was done in a similar fashion.

To minimize the effect of random partitioning, the process of forward selection (and backward elimination) was repeated five times, each with a different random partitioning of the training data. A total of five rankings of the features were obtained from forward selection, and another five were obtained from backward elimination. In Figure 5, we plot the cross-validation true positive rate of the training data ("selection TP") against the number of features included from backward elimination. Thin dotted lines represent results from the 5 different partitions, while the thick dashed line is their average in the upper side. Plotted together are the TP rates on the test data ("verify TP"), again the average of 5 random partitions in the thick dashed line and each individual one in thin dotted lines in the lower side.

Similarly in Figure 6 plotted are the results from forward selection. Both plots show that roughly 30 features contribute most to classification and best represent the data. Simple voting was used to combine the five rankings from forward selection into one single ranking. The same was done for backward elimination.

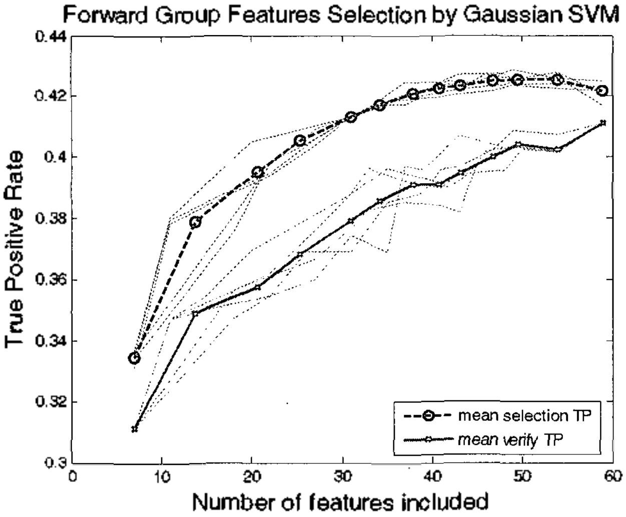

Figure 5. True positive rate as a function of number of features included in backward elimination for group feature selection.

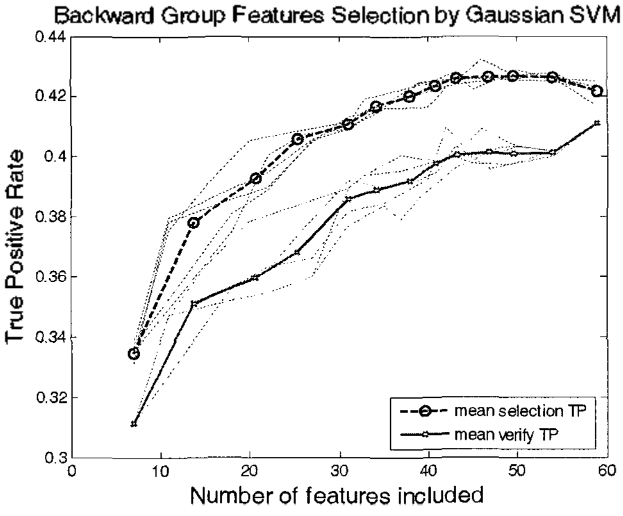

Figure 6. True positive rate as a function of number of features included in forward selection for group feature selection.

Table 3. Classification accuracy as the feature set size increases.

Added feature group (size)	Total number of features	Accuracy (%)
energy (7)	7	33.1
velPitch (4)	11	36.3
pitch (11)	22	37.0
mfcc1 (4)	26	39.7
f1 (5)	31	40.1
All features	59	42.3

The features are then plotted in Figure 7 on a two dimensional space where the x-y coordinate of each feature is its rank by the backward and forward selection. In general the ranking by forward and backward selection agree with each other. Feature groups near origin are considered to be more important in emotion recognition. This figure implies that energy, pitch velocity, pitch, low-order MFCCs, and F1 contribute to emotion recognition while F2, F3, mel-band energy, duration, and acceleration components are less important.

3.5. Performance of Feature Selection

Starting from the empty feature, we added a group of features in the order of the rank given by the forward selection results and evaluated the average classification accuracy. Table 3 shows the results, which implies that only 31 features achieves 40.1% accuracy.

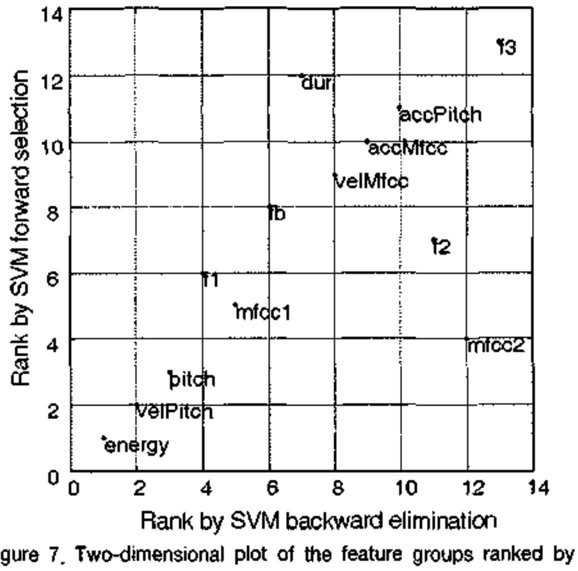

Figure 7. Two-dimensional plot of the feature groups ranked by forward selection (y-axis) and backward elimination (x-axis).

IV. Discussion

4.1. Comparing with the Human Performance

To compare our emotion recognizer with human classifiers, each of 3 graduate listeners classified 479 utterances sampled randomly from the test set. The listeners were totally foreign to the German language. For fair comparison, we trained the listeners with the same amount of training data used for training SVM. This situation is well comparable to our experimental setup. Table 4 is the resulting confusion matrix, which shows that the average human classification accuracy is 40.9%, which is within the confidence interval of 2.4%. For nonnative listeners, humans and machines yield similar performance. Even humans cannot classify emotion with high accuracy without linguistic or visual information. Recent evaluation results by native human listeners also show that humans do not perform significantly better than a machine[23].

In the preliminary experiment, the same 3 graduate students classified the test set without listening to the training set. The classification accuracy was near the chance level of 20%. We also performed similar experiments with 3 graduate students majoring in the German language, who did not listen to the training set. The classification accuracy was also near random selection accuracy because there were no particular association emotion and linguistic contents in the database. These two additional experiments show that nonnative listeners who have the knowledge of the target language still have difficulties in recognizing emotion if the utterance does not have some meanings associated with emotion.

4.2. Factors Affecting Emotion Recognition Accuracy

The performance difference between the SVM and HMM-based classifiers was shown to be similar. This fact implies that classification accuracy does not largely depend on the class of classifiers. Performance difference with different discriminative algorithms is not significantly different. Good feature extraction is a more critical factor in emotion recognition than classifier selection. When discriminative classifiers are used, good feature-length normalization scheme is also important. The final accuracy is largely affected by the performance of core component modules in feature extraction, e.g., pitch tracking and formant tracking.

Table 4. Confusion matrix (%) of foreign listeners after training.

	Angry	Bored	Happy	Neutral	Sad
Angry	43.3	9.4	23.9	15.4	4.9
Bored	8.2	37.7	7.1	13.9	37.5
Happy	25.2	7.7	45.5	22.3	5.9
Neutral	19.1	17.8	16.5	37.7	11.5
Sad	4.3	27.3	7.1	10.6	40.3

Further study is required regarding the exploration of new features better representing prosody and timbre, the improvement of the pitch and formant tracking algorithm, and the development of a more systematic approach to model dynamics of feature streams.

V. Conclusion

Using pitch, energy, formant frequencies, mel-band energies, MFCCs as base features, we analyzed the effects of the features in emotion recognition. Analyzing the factors of candidate features contributing to emotion recognition, we found that pitch and energy are the most significant feature in emotion recognition, which is consistent with the previous theoretical studies. We performed emotion recognition experiments using SVM- and HMM-based classifiers. With the SVM-based classifier, we achieved 42.3% of emotion classification accuracy using 5 emotion classes: angry, bored, happy, neutral and sad. The HMM-based classifier showed similar performance at 40.8%. The machine classifiers performance was shown to be significant when compared with the human performance.

Acknowledgment

This work was supported by Chungbuk National University Grant in 2004. The authors appreciate Sony for permitting them to use the AIBO speech database.

References

1. R.W. Picard, *Affective computing*, (MIT Media Lab Perceptual Computing Section Technical Report No. 321, 1995.)
2. R. Cowie, "Describing the emotional states expressed in speech," ISCA Workshop on Speech and Emotion, Belfast 2000.
3. N. Amir, "Classifying emotions in speech: A comparison of methods," Proc. Eurospeech 2001, Aalborg, Denmark, Sep. 2001.

4. M. Pardas, A. Bonafonte, J.L. Landabaso, "Emotion recognition based on MPEG-4 facial animation parameters," Proc. ICASSP 2002, Orlando, USA, May 2002.
5. K. R. Scherer, "Adding the affective dimension: A new look in speech analysis and synthesis," Proc. ICSLP 96, 1996.
6. S. McGilloway, R. Cowie, E. Douglas-Cowie, "Approaching automatic recognition of emotion from voice: A rough benchmark," ISCA Workshop on Speech and Emotion, Belfast 2000.
7. A. Nogueiras, A. Moreno, A. Bonafonte, J.B. Marino, "Speech emotion recognition using hidden Markov models," Proc. Eurospeech 2001, Aalborg, Denmark, Sep. 2001.
8. G. Zhou, J.H.L. Hansen, and J.K. Kaiser, "Nonlinear feature based classification of speech under stress," IEEE Trans. Speech and Audio Processing, 9 (3), 201-216, Mar. 2001.
9. M. Rahurkar, J.H.L. Hansen, J. Meyerhoff, G. Saviolakis, M. Koenig, "Frequency Distribution Based Weighted Sub-band Approach for Classification of Emotional/Stressful Content in Speech," Proc. Eurospeech-2003, 721-724, Geneva, Switzerland, Sep. 2003.
10. D. Ververidis, C. Kotropoulos, I. Pitas, "Automatic emotional speech classification," Proc. ICASSP 2004, I-593I-596, 2004.
11. A. Tickle, "English and Japanese speakers' emotion vocalization and recognition: A comparison highlighting vowel quality," ISCA Workshop on Speech and Emotion, Belfast, 2000.
12. T. S. Polzin, A. Waibel, "Emotion-sensitive human-computer interfaces," ISCA Workshop on Speech and Emotion, Belfast, 2000.
13. C. M. Lee, S. S. Narayanan, "Toward detecting emotions in spoken dialogs," IEEE Trans. Speech and Audio Processing, 13 (2), 293-303, Mar. 2005.
14. R. Tato, R. Santos, R. Kompe, J.M. Pardo, Emotional space improves emotion recognition," Proc. ICSLP 2002, 2029-2032, Sep. 2002.
15. L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, (Prentice-Hall, 1993.)
16. L. Rabiner and R.W. Schafer, *Digital Processing of Speech Signals*, (Prentice-Hall, 1978.)
17. ETSI Standard, Final Draft ETSI ES 202 050 v1.1.1 (2002-07), Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms.
18. V. Vapnik, *Statistical Learning Theory*, (New York: Wiley, 1998.)
19. S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, P. Woodland, *The HTK Book Version 3.2*, (Cambridge University Engineering Department, 2002.)
20. B.D. Ripley, *Pattern Recognition and Neural Networks*. (Cambridge, U.K.: Cambridge Univ. Press, 1996.)
21. C.-W. Hsu and C.-J. Lin. "A comparison of methods for multi-class support vector machines," IEEE Transactions on Neural Networks, 13, 415-425, 2002.
22. J. Ma, Y. Zhao, and S. Ahalt, OSU SVM Classifier Matlab Toolbox (ver 3.00), http://eewww.eng.ohio-state.edu/~maj/osu_svm/.
23. S. Steidl, M. Levit, A. Batliner, E. Nöth, H. Niemann, "Of all things the measure is man - Automatic classification of emotions and inter-labeler consistency," Proc. ICASSP 2005, pp. I-317I-320, 2005.
24. A.J. Hayter, *Probability and Statistics for Engineers and Scientists*, (PWS Publishing Company, 1995.)

[Profile]

•Oh-Wook Kwon

The Journal of the Acoustical Society of Korea, Vol. 22(3E)

•Kwokleung Chan

1994: The Hong Kong University of Science and Technology, Department of Physics (BS)
1996: University of California, San Diego, Department of Physics (MS)
2002: University of California, San Diego, Department of Physics (PhD)
2002-2003: Postdoctal researcher, Institute for Neural Computation, UCSD
2003-present: Senior Research Scientist at Softmax, Inc
Main Research: Machine learning algorithms, audio signal processing, biological and biomedical data analysis

•Te-Won Lee

The Journal of the Acoustical Society of Korea, Vol. 22(3E)