Comparison of Recognition Methods for Emotions Involved in Speech

Kazuhiko Takahashi
Faculty of Engineering
Yamaguchi University
Tokiwadai Ube, Yamaguchi, Japan
kylyn@yanaguchi-u.ac.jp

Ryohei Nakatsu
Faculty of Science and Engineering
Kansaigakuin University
Sanda, Hyogo, Japan
nakatsu@ksc.kwansei.ac.jp

Abstract

This paper investigates the characteristics of recognizing emotions contained in human speech. The concept of artificial neural network (ANN) is adopted for a recognition algorithm. An approach based on the support vector machine (SVM) or hidden markov model (HMM) is also investigated as an alternative recognition method. Using a large database of phoneme-balanced Japanese words, three recognition systems are trained and tested. To evaluate the emotion recognition results, emotion recognition testing is carried out with human subjects. The obtained average emotion recognition rates are 51% using ANN, 51% using SVM, 32% using HMM, and 55% with humans. Experimental results confirm that the emotion recognition rate achieved by using the ANN or SVM in the speaker- and context-independent mode is feasible and that ANN and/or SVM are well suited to this task.

1 Introduction

Speech involves a great deal of information: verbal information that conveys the speaker's will, personal information that expresses aspects of the speaker's personality such as age and gender, and emotional information that expresses the speaker's emotion or mood. In speaking text, several different meanings can be expressed depending on how it is said. For example, with the word "really" in English, a speaker can ask a question, express either admiration or disbelief, or make a definitive statement. An understanding of text alone cannot successfully interpret the meaning of such a spoken utterance. Therefore, it is clear that nonverbal information such as emotions plays an important role in communication.

Emotion recognition in speech has many potential applications. One possible use is as an aid to speech understanding. Speech understanding has traditionally treated emotion as noise, however, it is possible that by recognizing the emotions in speech one could subtract them from the speech and improve the performance of speech understanding systems. Another possibility is that an emotion recognition system could serve as a kind of 'emotional translator'. Emotions are often portrayed differently in different cultures and languages (e.g., one type of intonation that indicates admiration in Japanese can indicate disbelief in English). A method of translating emotions between languages can help improve international communication. Moreover, in order to make interactions between humans and machines more natural, the capability of communicating with humans by using both verbal and non-verbal communication channels will be essential on man-machine interfaces [13].

Although the importance of non-verbal aspects of communication has been recognized, most research has involved nonverbal information for images (e.g., facial expression [12] and gesture [16]) while little research has been done in the recognition of emotions involved in human speech [2, 3, 5]. With this and its potential uses in mind, we studied the recognition of emotions involved in speech and developed an emotion recognition system using an artificial neural network (ANN) [10]. A recognition rate of approximately 50% was achieved for speaker- and context-independent mode, however, it was not clear whether the emotion recognition rate of our system is satisfactory or not.

In order to evaluate the emotion recognition system, this paper investigates the characteristics of emotion recognition methods based on ANN, support vector machine (SVM), hidden markov model (HMM), and human subjects. First we explain the recognition of emotion involved in speech. Next, we describe how to design a recognition system using ANN, SVM or HMM. We then carry out recognition experiments with the ANN, SVM, HMM, and human subjects. Finally, we discuss the emotion recognition characteristics.

2 Emotions Involved in Speech

In this study, emotion recognition experiments were carried out with the following assumptions.

- Consciously and purposefully expressed emotions are treated since they are easier for humans to recognize [9] and significantly easier to gather data on.

- Eight emotional states (joy (J.), teasing (T.), fear (F.), sadness (Sa.), disgust (D.), anger (A.), surprise (Su.), and neutral (N.)) are selected by considering several works of emotion classification systems [4, 7, 8, 14].

- By carrying out training with a number of different speakers and a large set of phoneme-balanced words, a recognition system achieves speaker- and context-independence.

- A combination of two kinds of speech features, phonetic features and prosodic features, is considered because it is difficult to express emotions by only controlling prosodic features as these two kinds of features are tightly combined in uttering speech.

Recognizing emotions is a difficult task because people mainly rely on meaning recognition in daily communication. This is why speech recognition research has long treated emotions contained in speech as simply fluctuations or noise. What makes the situation more complicated is that emotional expressions are consciously or unconsciously intertwined with the meaning of speech. In the unconscious state, context plays a more important role than emotional features. As a result, the intensity of emotional expression varies dramatically depending on the situation. Our final target is to recognize emotions in speech even if emotional expression is unconsciously mixed with the meaning of speech. However, for the time being, this is not our research target for the above reasons. Instead, the strategy adopted here is to treat speech intentionally uttered with specific emotional expressions but not speech with unconscious emotion expressions.

2.1 Speech database

In gathering data, we adopted 100 phoneme-balanced Japanese words (e.g., "daidokoro (kitchen)", "ikioi (force)", "jyuuichigatsu (November)", etc.) because our target is context-independent emotion recognition. Since we utter most of these words without any special emotion in our daily life, it is difficult for ordinary people to intentionally utter them with emotions. Therefore, we first recorded the utterance of 100 words with each of the eight emotions by two voice actors (male and female). Then each of our subjects listens to the recordings of the voice actors and tries to imitate them. A total of 100 speakers, 50 male and 50 female native Japanese speakers, served as subjects. Each subject uttered a list of 100 Japanese words eight times, one time for each of the eight emotions.

2.2 Feature Extraction

As the phonetic features, we adopted linear predictive coding (LPC) parameters [6], which are typical speech feature parameters often used for speech recognition. The prosodic features consist of three factors: amplitude structure, temporal structure, and pitch structure. Speech power and pitch parameters are used to express amplitude structure and pitch structure, respectively, and each can be obtained in the LPC analysis. In addition, a delta LPC parameter, which is calculated from the LPC parameters and expresses a time variable feature of the speech spectrum, is adopted since it corresponds to temporal structure.

The speech feature is obtained in the following way. Analog speech is first transformed into digital speech by an A/D converter at an 11 kHz sampling rate and 16 bit accuracy. The digitized speech is then arranged into a series of frames, where each is a set of 256 consecutive sampled data points. LPC analysis is carried out and the following feature parameters are obtained for each of these frames.

$$\boldsymbol{f}^T = \left[\begin{array}{ccccccc} P_w & p & \delta & c_1 & c_2 & \cdots & c_{12} \end{array} \right], \qquad (1)$$

where P_w is speech power, p is pitch, δ is delta LPC parameter, and $c_j (j = 1, 2, ..., 12)$ is LPC parameter. These speech features are extracted from each utterance as shown in Fig. 1. First, the period where speech exists is extracted based on the information of speech power. Speech power is compared with a predetermined threshold value; if the speech power exceeds the threshold value for a few consecutive frames, the speech is determined to be uttered. After the beginning of the speech period, the speech power is also compared with the threshold value; if the speech power is continuously below the threshold value for another few consecutive frames, the speech is determined to no longer exist. Once the period of an utterance has been determined, the utterance is divided into 20 intervals of equal length in time. Let these 20 intervals be expressed as the vectors $\boldsymbol{f}_k (k = 1, 2, ..., 20)$. Each of the vectors has the 15 feature parameters of Eq. (1) for that interval.

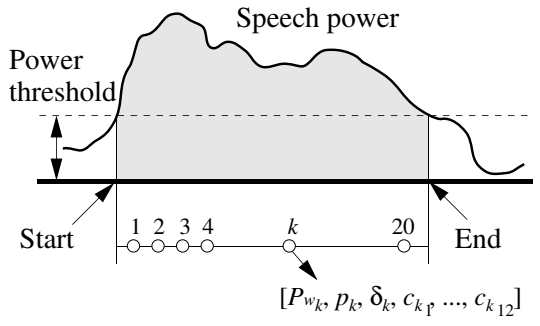

Figure 1. Speech feature extraction.

3 Emotion Recognition Methods

3.1 Emotion recognition by ANN [10]

The configuration of the ANN for emotion recognition is shown in Fig. 2. The network is composed of eight sub-ANNs, with one network for each of the eight emotions that are examined. The feature vector which is composed by the 20 speech vectors of 15 parameters, $\boldsymbol{F}_v^T = [\ \boldsymbol{f}_1^T \quad \boldsymbol{f}_2^T \quad \cdots \quad \boldsymbol{f}_{20}^T\ ]$, is simultaneously fed into all sub-ANNs. The output from each sub-ANN is a value (v_1, v_2, ..., v_8) representing the likelihood that the utterance corresponds to that sub-ANN's emotion. Decision logic selects the best emotion based on these values.

Each sub-ANN is a four-layer PDP model (300-M_1-M_2-1 network) and is trained by a back-propagation algorithm. The number of nodes in the intermediate layers, M_1 and M_2, varies depending on the specific emotion. A sigmoid function was used as an activation function of the nodes.

3.2 Emotion recognition by SVM

The SVM is a learning algorithm based on statistical learning theory [15]. Originally the SVM is designed for two classes classification by finding the optimal hyperplane where the expected classification error of test samples is minimized. There are several approaches to apply the SVM for multiclass classification. In this study, the one-vs-all method [1] is implemented.

Figure 3 shows the processing flow of the emotion recognition using SVM. Eight SVMs that correspond to each of the eight emotions were used. The ith SVM is trained with all of the training data in the ith class with positive labels, and all other training data with negative labels. Given a labeled set of N training data ($\boldsymbol{x}_i$, y_i), where $\boldsymbol{x}_i \in R^n$, $i = 1, 2, ..., N$, and y_i is the class of $\boldsymbol{x}_i$ ($y_i \in 1, 2, ..., 8$), the optimal hyperplane of the ith SVM can be found by solving the following problem:

$$\min_{w^i, b^i, \xi^i} \quad \frac{1}{2}(\boldsymbol{w}^i)^T \boldsymbol{w}^i + C \sum_{j=1}^{N} \xi_j^i (\boldsymbol{w}^i)^T$$

constrained by:

$$(\boldsymbol{w}^i)^T \phi(\boldsymbol{x}_j) + b^i \geq 1 - \xi_j^i \quad (if \quad y_j = i)$$

$$(\boldsymbol{w}^i)^T \phi(\boldsymbol{x}_j) + b^i \leq -1 + \xi_j^i \quad (if \quad y_j \neq i)$$

$$\xi_j^i \geq 0 \quad (j = 1, 2, ..., N)$$

where the training data $\boldsymbol{x}_i$ are projected to high dimensional feature space by the function ϕ, $\boldsymbol{w}^i$ is the weight vector, ξ_j^i is the slack variable that is introduced to account for non-separable data, C is the margin parameter that quantifies the trade-off between training error and system capacity. Solving the dual formulation of this problem, the optimal hyperplane can be defined by decision functions $y_i = (\boldsymbol{w}^i)^T \phi(\boldsymbol{x}) + b^i (i = 1, 2, ..., 8)$. The feature vector $\boldsymbol{F}_v^T$ is simultaneously fed into all SVMs ($\boldsymbol{x} = \boldsymbol{F}_v^T$) and the output from each SVM (y_1, y_2, ..., y_8) represents the likelihood that the utterance corresponds to that SVM's emotion. Decision logic selects the best emotion; the SVM that gives the largest value is chosen, and the class C_s (where $s = \arg\max_i y_i$) indicates the recognition result.

3.3 Emotion recognition by HMM

HMM is an nondeterministic state machine that, given an input, moves from state to state according to various transition probabilities. In each state, HMM generates output symbol probabilistically; this needs to be related to pattern features in an application-dependent manner. HMM, which is particularly suitable if the structure of the object sought is relatively clear, is commonly used in speech recognition

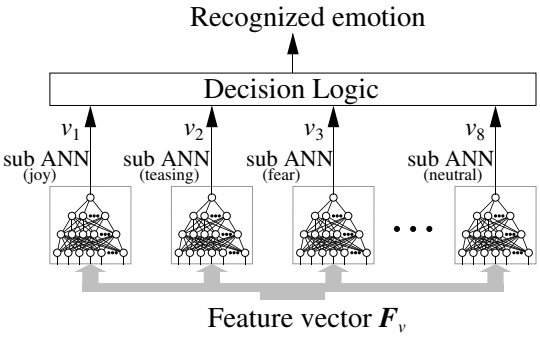

Figure 2. ANN-based emotion recognition system.

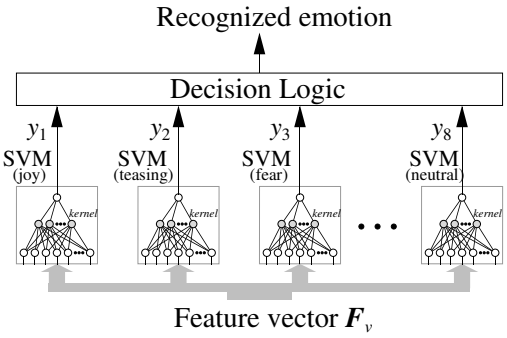

Figure 3. SVM-based emotion recognition system.

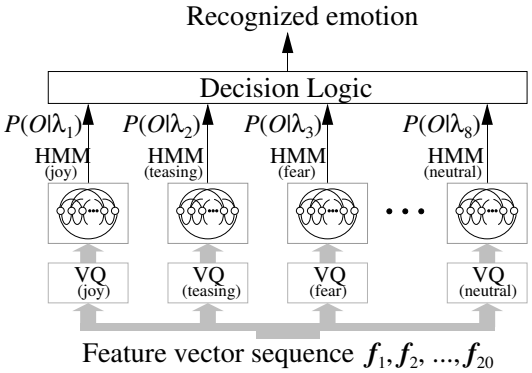

Figure 4. HMM-based emotion recognition system.

[11]. Since phoneme structures are the basis for the content of words or sentences, HMM is appropriate in speech recognition. The structure of emotions is not clear, however, we test HMM as an alternative recognition method.

Figure 4 shows the processing flow of the emotion recognition using HMM. Eight HMMs that correspond to each of the eight emotions were used. In this study, a discrete HMM modeled with an ergodic model was used. The input to the recognition system is the sequence of the feature vector $F = \boldsymbol{f}_1^T, \boldsymbol{f}_2^T, ..., \boldsymbol{f}_{20}^T$. First, a vector quantization (VQ) is carried out to transform the sequence of the feature vector F to a symbol sequence, $O = O_1, O_2, ..., O_{20}$. We used the category-separated VQ [17] in which the codebook B_i corresponds to category C_i and HMM λ_i ($i = 1, 2, ..., 8$). The codebook was created by using the LBG algorithm with the regular Euclidian norm. Next, each symbol sequence was fed into each HMM, which has been trained and re-estimated with the Baum-Welch algorithm using each emotional category's teaching data. The HMM then calculates the probabilities $P(O|\lambda)$ between the input symbol sequences and the teaching data with the Viterbi algorithm. Since category-separated VQ is adopted, a penalty function based on the distance between the input feature vector and the code word is considered in the probability calculation. Finally, the HMM that gives the highest probability is chosen in the decision logic, and the category C_s (where $s = \arg\max_i P(O|\lambda_i)$) indicates the recognition result.

4 Emotion Recognition Experiment

We used the large speech database for training ANN, SVM or HMM and carried out the following two types of recognition experiments to evaluate the performance of the recognition systems. In closed testing, utterances spoken by the speakers included in the training sets (#1-#30) are used for the recognition experiment. In open testing, utterances spoken by the speakers not included in the training sets (#31-#50) are used for the recognition experiment. The ANN, SVM, and HMM were trained and tested separately for male data and for female data.

Table 1. Recognition rates using ANN for male data [%].

	J.	T.	F.	Sa.	D.	A.	Su.	N.
J.	**31.00**	13.00	6.17	5.59	7.68	7.39	11.95	16.62
T.	10.36	**42.03**	8.05	4.99	11.50	5.48	8.78	8.81
F.	5.87	8.62	**40.18**	11.60	13.15	6.06	5.30	9.22
Sa.	2.65	2.17	10.96	**66.13**	4.50	2.65	1.18	9.76
D.	4.56	6.28	8.89	3.75	**63.30**	3.17	2.30	7.75
A.	5.42	2.94	4.11	3.40	2.94	**70.05**	4.11	7.03
Su.	10.31	10.55	4.02	2.90	5.09	9.14	**55.12**	2.87
N.	12.63	8.77	9.17	16.80	9.16	10.86	3.29	**29.32**

Table 2. Recognition rates using ANN for female data [%].

	J.	T.	F.	Sa.	D.	A.	Su.	N.
J.	**30.45**	14.67	4.84	0.34	2.82	24.37	12.78	9.73
T.	9.91	**45.21**	14.19	1.25	8.20	7.55	7.60	6.09
F.	2.06	4.21	**47.79**	29.32	6.73	2.82	0.39	6.68
Sa.	0.42	1.22	17.02	**75.35**	2.83	1.09	0.04	2.03
D.	1.90	6.04	12.53	4.74	**71.60**	1.16	0.58	1.45
A.	5.59	3.60	4.46	2.15	2.35	**72.68**	2.74	6.43
Su.	9.82	14.68	3.74	0	3.03	14.58	**52.60**	1.55
N.	13.88	6.99	17.42	16.56	5.85	11.56	0.81	**26.93**

In the training of the ANN, each ANN's network topology is first optimized using a small database composed of ten speakers. The optimization results were: (300-32-8-1) network topology for joy, teasing, fear and neutral sub-ANN, and (300-16-4-1) network topology for sadness, disgust, anger, and surprise sub-ANN. Then ANN was trained using 30 speakers where the maximum allowable error was 10^{-3} and the training epoch was limited to 1.6×10^4. In the closed testing, the averages of the recognition rates for all eight emotions were 57.4% for male data and 56.9% for female data. In the open testing, the recognition rates were 49.5% for male data and 52.8% for female data. The details of the open recognition results are shown in Table 1 for male data and Table 2 for female data. Here the row and column are the input and recognized class, respectively.

In the training of the SVM, we carried out tests with various kernel functions (Gaussian $\phi(\boldsymbol{x}_1)^T\phi(\boldsymbol{x}_2) = K(\boldsymbol{x}_1, \boldsymbol{x}_2) = exp(-\gamma|\boldsymbol{x}_1 - \boldsymbol{x}_2|^2)$, polynomial $K(\boldsymbol{x}_1, \boldsymbol{x}_2) = exp(-\gamma|\boldsymbol{x}_1 - \boldsymbol{x}_2|^d)$, and sigmoid function $K(\boldsymbol{x}_1, \boldsymbol{x}_2) = tanh(a\boldsymbol{x}_1\boldsymbol{x}_2 - h)$) and parametric studies of the margin parameter C to find the

Table 3. Recognition rates using SVM for male data [%].

	J.	T.	F.	Sa.	D.	A.	Su.	N.
J.	**32.33**	6.34	7.56	3.22	0.80	18.6	6.52	21.62
T.	11.26	**45.73**	11.19	2.80	7.48	5.73	6.43	9.37
F.	1.92	6.29	**60.03**	11.25	8.59	2.44	1.63	7.85
Sa.	1.12	1.12	15.83	**61.62**	2.80	6.72	0.14	10.64
D.	1.05	8.16	15.70	2.16	**67.06**	3.49	0.35	2.02
A.	9.44	1.82	7.41	3.36	1.54	**51.05**	1.61	23.78
Su.	20.57	8.12	3.50	0.70	2.24	6.58	**52.76**	5.53
N.	11.34	5.25	13.45	10.36	1.96	16.60	0.63	**40.41**

Table 4. Recognition rates using SVM for female data [%].

	J.	T.	F.	Sa.	D.	A.	Su.	N.
J.	**29.07**	10.57	3.24	1.76	2.96	19.80	15.47	17.12
T.	9.81	**41.92**	10.67	3.10	10.87	7.57	6.91	9.15
F.	1.57	1.96	**48.48**	33.01	5.68	2.60	1.91	4.80
Sa.	1.12	1.86	18.04	**66.89**	3.57	1.61	2.98	3.91
D.	1.41	5.57	6.93	4.87	**78.25**	1.08	0.60	1.30
A.	7.14	4.57	4.80	2.33	2.61	**59.93**	5.32	13.29
Su.	11.96	12.99	4.79	0.41	3.51	15.05	**44.28**	7.01
N.	9.76	5.50	12.15	19.78	3.86	11.35	2.53	**35.08**

Table 5. Recognition rates using HMM for male data [%].

	J.	T.	F.	Sa.	D.	A.	Su.	N.
J.	**27.07**	12.36	7.00	2.01	4.14	17.92	10.00	19.50
T.	18.00	**25.14**	10.43	3.65	8.93	8.50	5.71	19.64
F.	3.28	10.78	**37.64**	14.14	15.79	4.57	1.09	12.71
Sa.	2.79	3.29	19.43	**51.36**	8.93	6.86	0.41	6.93
D.	6.86	9.78	22.43	14.78	**21.76**	4.68	1.28	18.43
A.	14.00	5.78	8.86	2.57	6.78	**37.50**	2.58	21.93
Su.	27.36	13.50	4.64	1.15	2.64	8.71	**32.50**	9.50
N.	12.43	13.71	15.86	8.36	8.86	15.00	1.14	**24.64**

Table 6. Recognition rates using HMM for female data [%].

	J.	T.	F.	Sa.	D.	A.	Su.	N.
J.	**21.31**	20.68	4.79	2.22	7.63	16.21	10.16	17.00
T.	15.95	**18.63**	11.26	4.42	16.00	9.58	7.00	17.16
F.	1.31	1.58	**44.31**	37.53	7.74	1.21	0.21	6.11
Sa.	1.31	2.27	26.00	**55.74**	6.47	0.79	2.00	5.42
D.	9.37	7.32	18.26	18.89	**28.74**	3.37	1.42	12.63
A.	7.10	10.63	5.47	2.02	6.95	**46.10**	10.26	11.47
Su.	14.79	16.79	4.00	1.57	6.741	20.79	**22.95**	12.37
N.	9.31	12.31	13.74	16.68	18.42	6.21	2.22	**23.11**

optimal SVM. As a result, a maximum rate of the closed testing was reached with the Gaussian kernel function ($\gamma = 0.1$) under the margin parameter of 10. In the closed testing, 100% and 99.6% averages of the recognition rates for all eight emotions were achieved for male and female data, respectively. Table 3 and Table 4 show the details of the open recognition results, where the averaged recognition rates were 51.4% for male data and 50.5% for female data.

In the training of the HMM, we conducted tests with various codebook sizes (16, 32, 48, 64, and 128) for the LBG algorithm and various numbers of HMM states (8, 16, 32, and 48) to obtain the optimal HMM. As a result, a maximum rate of the closed testing was reached with a codebook size of 48 and 16 HMM states. In the closed testing, 49.5% and 44.5% averages of the recognition rates for all eight emotions were achieved for male and female data, respectively. Table 5 and Table 6 show the details of the open recognition results, where the averaged recognition rates were 32.2% for male data and 32.6% for female data.

Emotion recognition testing by human subjects was carried out as follows. A total of 28 subjects (22 males and 6 females native Japanese) were served. In this experiment, each of them first listened to correctly classified examples of speech uttered by the voice actors, where the examples were shown four times from each emotion category. The subjects then listened to 450 voice data samples randomly selected from the large speech database and classified each of them into one of the eight emotion categories. The recognition results for male and female voice data are shown in Table 7 and Table 8, respectively. The averaged recognition rates were 58.2% for male data and 51.7% for female data.

Comparing the results leads to the following observations.

- In each recognition method, the open recognition results are similar between male and female data.
- The recognition results of the HMM are worse than those of the ANN and SVM, however, all methods show similar recognition characteristics: negative emotions, such as anger or sadness, are easy to recognize, while positive emotions, such as joy, are harder to recognize.
- Human subjects easily classify negative emotions but seem to have some difficulty in classifying positive emotions.
- The recognition rate of the 'neutral' emotion is low when using ANN, SVM, or HMM, while human subjects show a good recognition rate for this emotion.
- The emotion recognition capability of either ANN or SVM is almost the same as that of the human subjects.

Table 7. Recognition rates by human subjects for male data [%].

	J.	T.	F.	Sa.	D.	A.	Su.	N.
J.	**18.57**	3.51	0.88	1.46	1.02	5.99	7.46	61.11
T.	7.74	**61.61**	2.43	4.40	4.86	4.70	5.46	8.80
F.	0	1.55	**58.74**	30.25	3.52	1.32	1.03	0.59
Sa.	0	1.52	17.96	**70.62**	4.26	0.91	0.30	4.41
D.	0	7.92	1.47	1.91	**86.66**	0.88	1.03	0.15
A.	1.32	1.76	2.94	7.65	6.18	**53.82**	2.35	23.97
Su.	13.03	8.20	1.90	1.90	0.73	19.33	**45.39**	9.52
N.	1.61	1.46	1.46	15.47	3.07	5.99	0.58	**70.36**

Table 8. Recognition rates by human subjects for female data [%].

	J.	T.	F.	Sa.	D.	A.	Su.	N.
J.	**26.54**	4.11	0.59	0.88	0.44	6.60	3.67	57.18
T.	20.00	**39.86**	1.82	4.76	6.01	4.90	9.23	13.43
F.	0	1.54	**42.42**	48.88	5.06	0.84	0.14	1.12
Sa.	1.97	1.13	9.99	**71.87**	8.16	3.23	0.70	2.95
D.	0.15	7.78	0.15	0.59	**90.16**	0.59	0.44	0.15
A.	6.74	4.78	0.70	3.37	7.30	**37.78**	4.21	35.11
Su.	25.00	6.14	2.19	5.70	3.36	9.50	**30.99**	17.11
N.	3.23	2.67	1.69	8.85	3.09	6.04	0.70	**73.74**

These results indicate that an emotion recognition rate of approximately 51% by using either ANN or SVM in the speaker- and context-independent mode is feasible and it is possible for machines to communicate with people by using nonverbal communication capabilities.

5 Conclusions

This paper investigated the characteristics of recognizing emotions contained in human speech. We adopted ANN for a recognition algorithm. For comparison, a system based on SVM or HMM was tested as an alternative recognition method. Using a large database of phoneme-balanced Japanese words uttered by speakers consciously trying to portray an emotion, we trained and tested the recognition systems. To evaluate the emotion recognition results, we also carried out emotion recognition testing using human subjects. The obtained emotion recognition rates were 51% using ANN, 51% using SVM, 32% using HMM, and 55% using human subjects. Experimental results confirmed that the emotion recognition rate achieved by using ANN in the speaker- and context-independent mode is feasible and the SVM is also well suited to emotion recognition task.

References

[1] O. Chapelle, P. Haffner, and V. Vapnik. Support vector machines for histogram-based image classification. *IEEE Trans. on Neural Networks*, 10(5):1055–1064, 1999.

[2] F. Dellaert, T. Polzin, and A. Waibel. Recognizing emotion in speech. In *Proc. of 4th IEEE Int. Conf. on Spoken Language Processing*, volume 3, pages 1970–1973, 1996.

[3] L. C. DeSilva and P. C. Ng. Bimodal emotion recognition. In *Proc. of FG2000*, pages 332–335, 2000.

[4] G. Klasmeyer and W. F. Sendlmeier. Objective voice parameters to characterize the emotional content in speech. In *Proc. of ICPhS'95*, volume 1, page 182, 1995.

[5] K. Kostov and S. Fukuda. Emotion in user interface, voice interaction system. In *Proc. of SMC2000*, volume 2, pages 798–803, 2000.

[6] J. D. Markel and A. H. Gray. *Linear Prediction of Speech*. Springer-Verlag, 1976.

[7] S. McGilloway, R. Cowie, and E. D. Cowie. Pitch variations and emotions in speech. In *Proc. of ICPhS'95*, volume 1, page 178, 1995.

[8] S. Mozziconacci. Prosodic signs of emotion in speech: Preliminary results from a new technique for automatic statistical analysis. In *Proc. of ICPhS'95*, volume 1, page 250, 1995.

[9] I. R. Murray and J. L. Arnott. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *J. of Acoustical Society of America*, 93(2):1097–1108, 1993.

[10] J. Nicholson, K. Takahashi, and R. Nakatsu. Emotion recognition in speech using neural networks. *Neural Computing and Applications*, 9(4):290–296, December 2000.

[11] L. R. Rabiner and B. H. Juang. An introduction to hidden markov models. *IEEE ASSAP Magazine*, pages 4–16, January 1986.

[12] M. Rosenblum, Y. Yacoob, and L. Davis. Human expression recognition from motion using radial basis function network architecture. *IEEE Trans. on Neural Networks*, 7(5):1121–1120, 1996.

[13] J. Sato and S. Morishima. Emotion modeling in speech production using emotion space. In *Proc. of RO-MAN'96*, pages 472–477, 1996.

[14] K. R. Scherer. How emotion is expressed in speech and singing. In *Proc. of ICPhS'95*, volume 3, page 90, 1995.

[15] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[16] C. R. Wren and A. Pentland. Dynamic modes of human motion. In *Proc. of FG'98*, pages 22–27, 1998.

[17] J. Yamato, S. Kurakake, A. Tomono, and K. Ishii. Human action recognition using hmm with category-separated vector quantization. *IEICE Trans.*, J77-D-II(7):1131–1318, 1994.