Emotion Recognition and Synthesis System on Speech

Tsuyoshi Moriyama, Shinji Ozawa
Faculty of Science and Technology, Keio University
{moriyama,ozawa}@ozawa.ics.keio.ac.jp

Abstract— **In this study, the system that is capable of both recognizing and synthesizing emotional content in speech is developed. At first, the relation information that relates the physical features of emotional speech to the emotional content perceived by listeners is estimated through linear statistical methods, and it is applied to the system. It realizes emotion recognition and synthesis just through easy linear operation using the relation information. In the system, the pitch contour is expressed by the model proposed by Fujisaki (7 parameters) and the power envelope is approximated by 5 line segments (11 parameters), and PSOLA is applied to synthesize the speech. The emotion words among which there is very little correlates were selected from the preliminary statistical experiments. The relation information was verified to be significant and from the result of the experiments, the system was able to recognize and synthesize emotional content in speech as subjects did. Moreover, the emotion recognition system is applied to the emotion measurement module in the cyber shopping system.**

Keywords— **Emotion, Principal component analysis, Multiple regression analysis, Speech processing, PSOLA**

I. Introduction

Recently, HI (Human-machine Interface) grows its importance in accordance with the development of information communication services. The speech interface, which is one of HIs, has the advantage of its simplicity of use, requiring just a small device and the base of the existing communication infrastructures like the telephone network and broadcasting network. So its social needs is considerable at present and also in the future. A speech recognition technology is regarded as the input from human to machine, and in contrast with it, speech synthesis is considered not only as just the machine's output technologies but as what is possible to form a kind of social agents by implemented as an auto-guidance system, warning equipments or welfare utensils like pet robots.

One of speech synthesis methods is the text-based speech synthesis method which has the advantage of being able to synthesize every words from a little text information. It, nowadays, has phoneme information and coarticulation rules in large scale databases and is able to realize high quality synthesis speech. In the preceding studies about speech synthesis, it has investigated its articulation quality primarily, but the control techniques of the prosodic components of speech (e.g. intonation, accent) have not developed enough yet. The prosodic component has been said to communicate much information including human factors like emotion, utterance intention as the so-called non-verbal information. To achieve naturalness in synthesis speech, it is supposed to be necessary to clarify the relation between the prosodic features and communicated non-verbal information.

On the other hand, human factors like emotional information have lately attracted considerable attention. Because for robust speech recognition, such components are the noise and if the relation between the physical distortion and emotional content is clarified, it become possible to remove the components in the prior processing stage. And because for high quality speech synthesis, emotional information is expected to be a kind of breakthrough that the speech quality increases rapidly by letting people have an illusion as if the speech was spoken by man. It means the knowledge of the relation between para-linguistic content and speech parameters is applied to change the speech to include "gentle" or "sorry" atmosphere in speech to raise its naturalness. And also because for the purpose of developing the human-machine interface which is familiar with people, it is necessary of machines to have the ability that can perceive such human factors and imitate them.

So far, such an intelligent speech interface which is able to exchange emotional content with human beings has been investigated. In past studies, emotion words such as "joy" were selected to describe emotional content from the primary emotions reported in psychology, or a few words which are easy to distinguish were chosen provisionally. Also a few prosodic parameters were selected pointed to speech synthesis or based on their simplicity to calculate. It is, however, supposed that there has never been the computation model which is pointed to realize emotion exchange between a human and the machine, and the emotional concept which it is possible to treat as a kind of universal quantity in the engineering systems. Accordingly, the choice of the observable variables such as emotion words and prosodic parameters is supposed to be groundless (mentioned primary emotions were obtained from the experiments using not speech but facial images, so they are not necessarily the same as this case of voices).

In this study, the emotion communication model is proposed at first and after clarification of the definition of emotional information here, the computation model of the observer viewpoints of an emotional expression is shown, which relates the perceived emotion stereotype to analyzed physical parameters of the speech. Then the model is realized by traditional statistical methods, and the system is implemented, which is able to recognize emotional content

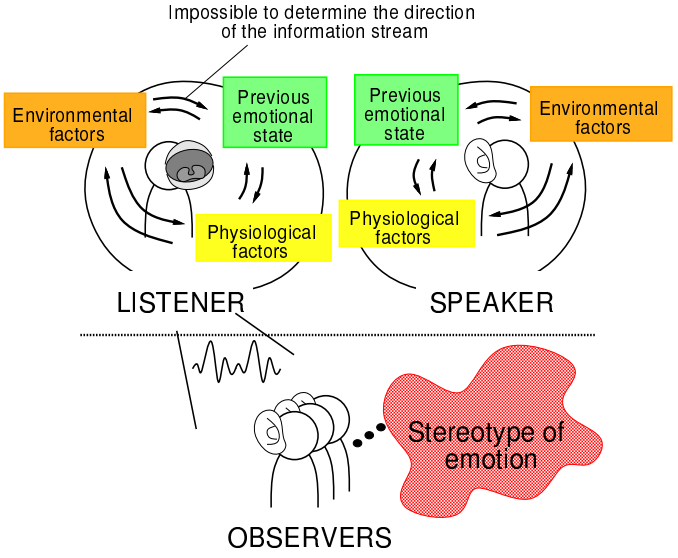

Fig. 1. Emotion communication

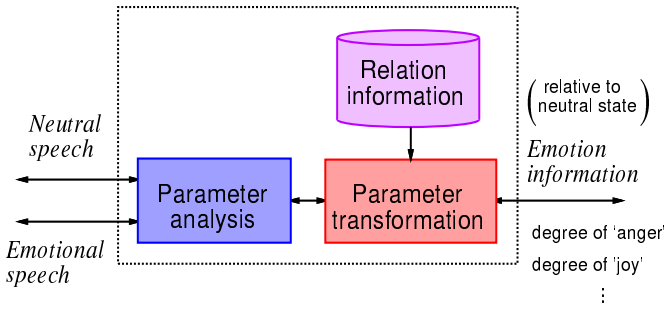

Fig. 2. Proposed relation model

in speech and synthesize an emotional speech from a neutral one based on the estimated relation information in the model. In the next chapter, the model is presented, and how to realize the model is shown in the 3rd chapter. In the 4th, the parameters which represent prosodic components of a speech are explained. The experiments which verify the validity of the model and the effectiveness of the system are shown in the 5th chapter. In the 6th chapter, the application of the system to a speech interface is shown and discussed. The last chapter summarizes and concludes the paper.

II. Relation model

It is supposed that there are two types of vocal emotion. One is the subjective emotion harbored by the speakers or listeners in communication, and the other is the objective emotion by the observer of communication. The former is determined by many factors, for example, the external environment, the inner physiological state and just previous emotional state, and it is impossible to limit the stream direction among those factors and to assume causality(Fig.1). On the other hand, the latter is identified with the stereotype of emotion expression rearranged among observers such as an audience of a play. It seems to be general among people and observable through statistical research. Streeter [4] also noted that the emotion perceived by listeners is measurable by physical aspects of the speech. Therefore, in this study the relation model is based on the latter emotion.

In previous studies, by the way, a few sentences or words were spoken by several speakers for collecting emotion speech which includes personality, verbal meanings, emotion and many kinds of information [5] [6]. And measurement of emotion information has been attempted in such complex information condition. Therefore, it is difficult to separate the aspects caused from the emotional components from others. In this study, meaningless utterances like humming voice spoken by only one person are used for system construction in order to exclude the other factors of personalities and verbal meanings.

On the other hand, the interdependency among observable variables of physical parameters and emotion words has been an important issue in representing their spaces. As the result of statistical experiments by Moriyama [7], strong correlation was found among those observed quantities. So in this study, orthogonal bases are extracted from distribution statistics collected from those observable variables, and using the bases to construct the system which is able to translate physical aspects and included emotional content to each other, it it supposed to be able to avoid the dependency on the choice of observed variables.

In this study, it is assumed that emotion is conveyed by physical variation from neutral speech. Accordingly, the relation between the emotional content relative to neutral speech and physical changes occurred in neutral speech is described in the model proposed in this study. Here neutral speech is defined as the speech which includes all emotions in the same level, and emotional speech is the one that has one emotional content whose level is higher than the other ones. Fig.2 shows the relation model proposed here.

III. Method

A. Obtaining the Relation information

The relation information describes the relation between the emotion information in a speech and physical parameters obtained by analyzing the speech in the shape of a function, and it can be used to transform each other. In this study, the stereotype of emotional impression perceived by the observers is obtained through subjective experiments statistically, and the prosodic changes from the neutral state are observed simultaneously. Then the relation information is estimated as the linear combination between these quantities. The flow chart of obtaining the relation information is shown in Fig.3. It is necessary to prepare many kinds of emotional speech as the teacher in order to cover all over the emotion space. Also, as the condition of the teacher speech, a meaningless word /n/ is uttered with neutral and many kinds of emotion by only one speaker. it is expected that the effects caused from personalities and verbal meanings are avoided. In the subjective experiments, the emotional impression relative to the neutral

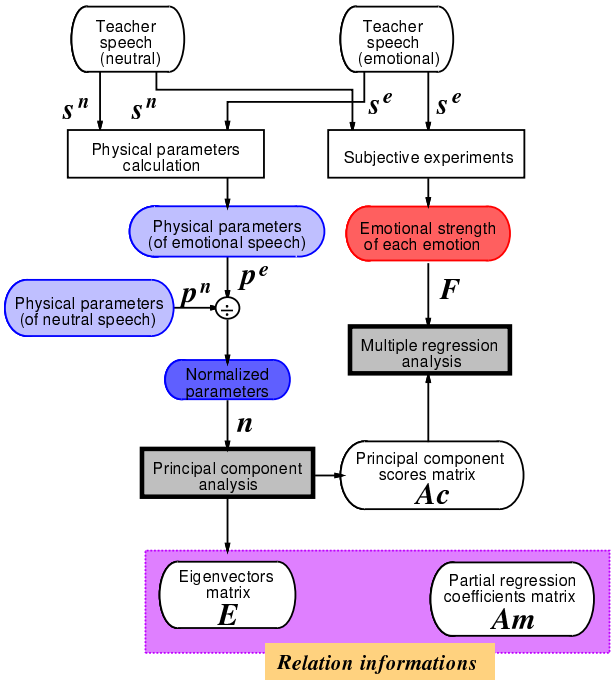

Fig. 3. Emotion communication

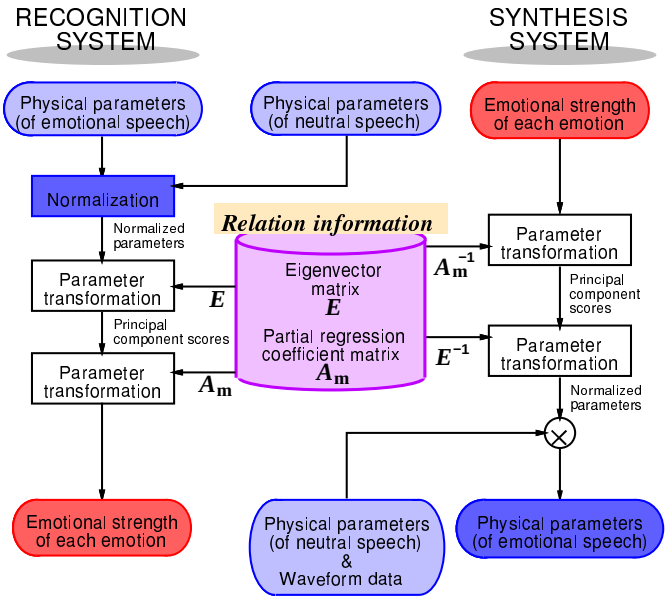

Fig. 4. Proposed system

speech is measured. The subjects evaluate the emotional content perceived from sample speeches in seven steps assessing measurement.

As shown in Fig.3, physical parameter matrix $\boldsymbol{P} = [\boldsymbol{p}_1^e, \boldsymbol{p}_2^e, \cdots, \boldsymbol{p}_n^e] = \{p_{ij} | i = 1, \cdots, m; j = 1, \cdots, n\}$ (m:parameters number, n:sample number) is calculated for all teacher speeches $\boldsymbol{s} = \{s_j | j = 1, \cdots, n\}$. The components of the matrix describe the F_0 contour and the power envelope which is required in the PSOLA (wave synthesis method), and it is necessary that this parameter set is enough to reconstruct the original speech with no distortion of the emotional content. It is explained in detail in the following chapters. The parameter matrix is also obtained from the neutral speech, and the emotional ones $\boldsymbol{p}_j^e = \{p_{ij}^e | i = 1, \cdots, m\}$ are normalized by the neutral one $\boldsymbol{p}^n$. The normalized parameter matrix $\boldsymbol{N} = [\boldsymbol{n}_1, \boldsymbol{n}_2, \cdots, \boldsymbol{n}_n]$ is analyzed by principal component analysis and resolved into orthogonal bases. These orthogonal bases are called the principal component scores and obtained in the shape of a matrix $\boldsymbol{A}_c = [\boldsymbol{a}_1^c, \boldsymbol{a}_2^c, \cdots, \boldsymbol{a}_n^c]$ ($n \times n$ matrix). And the combination coefficients which transforms the parameters into the orthogonal bases is an eigenvector matrix $\boldsymbol{E} = [\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_m]$, which is the relation information between them. Then, the multiple regression analysis is conducted under the condition that evaluation value matrix $\boldsymbol{F} = [\boldsymbol{f}_1, \boldsymbol{f}_2, \cdots, \boldsymbol{f}_n] = \{f_{kj} | k = 1, \cdots, l; j = 1, \cdots, n\}$ (k:the number of terms, evaluation values: the average among the subjects) is the criterion variable and $\boldsymbol{A}_c$ is the predictor variable. As the result of it, the partial regression coefficient matrix $\boldsymbol{A}_m$, which is the relation information between principal components of physical parameters and psychological evaluation values, is estimated.

After obtaining these relation information (the eigenvector $\boldsymbol{E}$, and the partial regression coefficient matrix $\boldsymbol{A}_m$), it is applied to the emotion recognition/synthesis system shown in Fig.fig:fig4. The former system is called the recognition system and the latter the synthesis system in the following chapters, and they are explained in detail.

B. Recognition system

The input to the recognition system is the input speech, of which emotional content is unknown of course, and the neutral speech spoken by the same speaker and the same word, and the output is the emotion information relative to the neutral speech. Physical parameters, discussed in the 4th chapter, of the neutral and input speech ($\boldsymbol{p}^n$, $\boldsymbol{p}^{new}$) are calculated, and the normalized parameter $\boldsymbol{n}^n\boldsymbol{ew}$ is obtained by the Eq.(1).

$$n_i^{new} = p_i^{new} \ / \ p_i^n \qquad (i = 1, \cdots, m) \tag{1}$$

Then, $\boldsymbol{n}^n\boldsymbol{ew}$ is multiplied by the relation information of $\boldsymbol{E}$ and $\boldsymbol{A}_m$ in order, the emotional strength of each emotion words ($\boldsymbol{f}^{new}$) are calculated. In brief,

$$\boldsymbol{c}^{new} = \boldsymbol{E} \cdot \boldsymbol{n}^{new} \tag{2}$$

$$\boldsymbol{f}^{new} = \boldsymbol{A}_m \cdot \boldsymbol{c}^{new} \tag{3}$$

where $\boldsymbol{c}^{new}$ is the orthogonal bases vector (the principal component scores vector) of the physical parameters.

C. Synthesis system

The input to the synthesis system is the neutral speech and the emotional information, and the output is the emotional speech. In this study, the emotional speech is synthesized through waveform editing, PSOLA modified to be suitable for the system is applied. Accordingly the

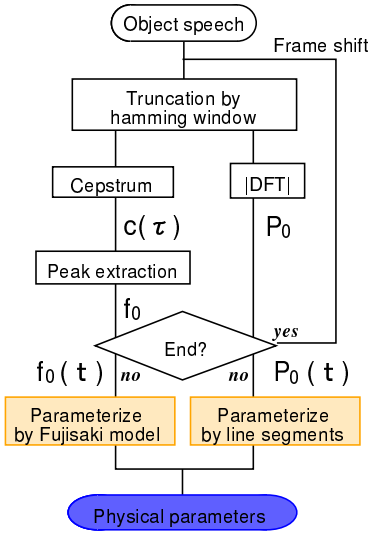

Fig. 5. Flow of calculation of physical parameters

synthesized speech is spoken by the same speaker and the same word as the neutral one, of course. PSOLA (pitch synchronous overlap-add) is capable of transforming the source speech waveform into the speech with given F_0 contour, power envelope and velocity. On the other hand, emotion information to synthesize is given in the shape of the grade how extent each emotion is included.

In the synthesis system, the change of each parameters which add the input emotion information to the neutral speech is obtained by just an inverse operation of the recognition process. Namely, the input emotional strength (grade) of each emotion $\boldsymbol{f}^{neo}$ is multiplied by the relation information of $\boldsymbol{A}_m^{-1}$ and $\boldsymbol{E}^{-1}$ which are the inverse matrix of $\boldsymbol{A}_m$ and $\boldsymbol{E}$ respectively as follows:

$$\boldsymbol{c}^{neo} = \boldsymbol{A}_m^{-1} \cdot \boldsymbol{f}^{neo} \tag{4}$$

$$\boldsymbol{n}^{neo} = \boldsymbol{E}^{-1} \cdot \boldsymbol{c}^{neo} \tag{5}$$

where $\boldsymbol{n}^{neo}$ denotes the difference from the neutral speech's parameter vector. Then it is multiplied by prosodic parameter vector $\boldsymbol{p}^{n}$ of the neutral speech, the emotional speech's one $\boldsymbol{p}^{neo}$ is calculated. PSOLA synthesizes the emotional speech from $\boldsymbol{p}^{neo}$.

IV. Physical parameters

Fig.5 shows the flow of calculation of the parameters. In the following sections, the parameter set is proposed, which describes the temporal contour of F_0 and short-time average power.

A. F_0 contour

In this study, the revision of the F_0 pattern generation model proposed by Fujisaki [1] (called Fujisaki model) is applied to parameterize the pitch contour. In Fujisaki model, it is assumed that F_0 contour in log scale is consist of the phrase component, which descends slowly to the beginning to the end, and the accent component which rises and falls locally. The model equation is following:

$$\begin{aligned} ln F_0(t) &= \ln F_{\min} \\ &+ A_p G_p(t - T_0) \\ &+ A_a[G_a(t - T_1) - G_a(t - T_2)] \end{aligned} \tag{6}$$

where $G_p(t)$ is the impulse response function of the phrase control mechanism, $G_a(t)$ is the step response function of the accent control mechanism. If t > 0,

$$G_p(t) = \alpha^2 t \exp(-\alpha t) \tag{7}$$

$$G_a(t) = \min[1 - (1 + \beta t)\exp(-\beta t), \theta] \tag{8}$$

and $G_p(t) = G_a(t) = 0$ $(t < 0)$. The parameters in the equation is shown in Tab.I.

TABLE I

Proposed model parameters of F_0 contour

T_0	beginning of phrase instruction[ms]
T_1 - T_0	length to accent instruction[ms]
T_2 - T_1	interval of accent instruction[ms]
T_E - T_0	from phrase beginning to speech end[ms]
A_p	amplitude of phrase instruction
A_a	amplitude of accent instruction
$\ln(F_{\min})$	speaker's F_0 standard[ln (Hz)]

B. Short-time average power envelope

The envelope of the short-time average power is approximated by 5 connected line segments which is described by 11 parameters in Tab.II The example of the model is shown in Fig.6.

The procedure of fitting the model to the actual power envelope is as follows: at first, P_{ini} and T_E is observed

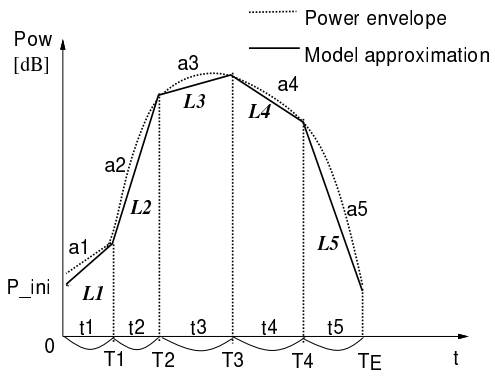

Fig. 6. An example of power envelope model fitting

TABLE II

Model parameters of the power envelope

P_{ini}	Initial power value
a_i	Tendency of each line segments[dB/point]
t_i/T_E	Length ratio of each line segments

from the envelope directly, and $T_{1ini} \sim T_{4ini}$ is determined (by sight now), And then, the Euclid distance between the estimated power envelope and the actual one is minimized by moving $T_1 \sim T_4$ just around the initial values and for each $T_1 \sim T_4$, $a_1 \sim a_4$ varying.

V. Experiments

So far, the relation information is obtained by the mentioned process, and the system shown in Fig.4 has been constructed. Here, the system is evaluated through an emotion recognition experiment and an emotional speech synthesis experiment. However, the ability of the system is dependent on the accuracy of the relation information, so the result of statistical certification of the relation information is indicated in the following chapters at first. Namely, the result of principal component analysis, which relates and summarize the observed prosodic parameters to the orthogonal bases, is shown in A.1, and the result of multiple regression analysis, which estimates linear relation between the physical bases and emotion information, is in A.2. After that, the result of recognition and synthesis experiments is indicated.

A. Accuracy of the relation information

A.1 Result of principal component analysis

As the result of principal component analysis, 5 principal components were extracted shown in Tab.III. The cutoff accumulative ratio of distribution is 80%. The positioning of the prosodic parameters on the basic axes indicates their interpretation. Fig.7 $\sim$ 9, for example, show the positioning of the 1st, 2nd and 3rd principal component, and they seems to be regarded as "steepness", "flatness" and "attenuation" respectively.

A.2 Result of multiple regression analysis

In the subjective experiments, 20 teacher speeches were evaluated by 18 subjects about 9 emotion words. These emotion words shown in Tab.IV had been chosen from 46 emotion words based on the correlation among them,

TABLE III
Result of principal component analysis

Prn.Cmp.	EigenVlue.	Ratio of dist.	Accum.dist.
1	5.522	30.680	30.680
2	3.574	19.854	50.534
3	2.134	11.858	62.392
4	1.734	9.635	72.028
5	1.474	8.188	80.216

TABLE IV
Emotion words used in the subjective experiment

1	anger	4	fear	7	funny
2	joy	5	cynicism	8	flattery
3	sorrow	6	calmness	9	surprise

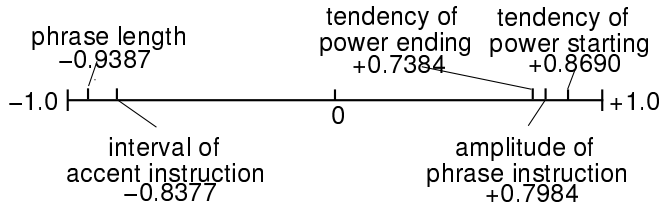

Fig. 7. 1st principal component

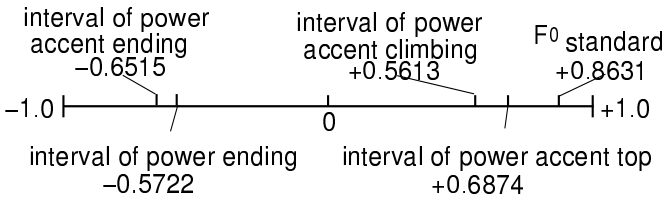

Fig. 8. 2nd principal component

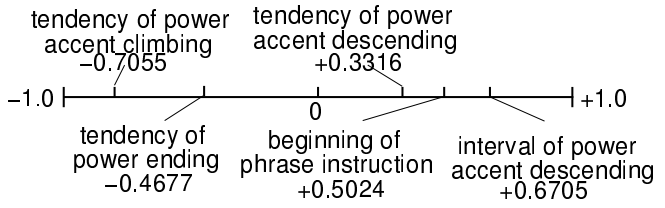

Fig. 9. 3rd principal component

namely if two of them has strong correlation, one party is removed, and if another two of them has large negative correlation, they both are left. The evaluation value vectors were obtained for each emotion words and used as the criterion variables in the multiple regression analysis. They were described as the linear combination of the predictor variables of the physical bases (5 principal components).

Obtained multiple regression equation is following:

$$\begin{aligned} \boldsymbol{f} &= \boldsymbol{A}_m \cdot \boldsymbol{c} \\ &= \begin{bmatrix} 0.495 & -0.012 & -0.691 & 0.274 & 0.315 \\ 0.192 & -0.004 & 0.163 & 0.025 & -0.166 \\ -0.541 & 0.392 & -0.012 & -0.367 & 0.173 \\ -0.168 & 0.340 & -0.082 & -0.264 & -0.033 \\ -0.074 & -0.022 & 0.040 & 0.029 & 0.118 \\ -0.003 & 0.057 & 0.117 & -0.058 & -0.088 \\ -0.111 & -0.205 & 0.249 & 0.033 & 0.110 \\ -0.137 & 0.172 & 0.055 & -0.035 & 0.197 \\ 0.298 & 0.305 & -0.175 & 0.137 & -0.453 \end{bmatrix} \cdot \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \end{bmatrix} \end{aligned} \tag{9}$$

where $\boldsymbol{f} = [f_1, f_2, \cdots, f_9]$ is subjective evaluation value vector, $\boldsymbol{c}$ is the principal component score vector and $\boldsymbol{A}_m$is the partial regression coefficients matrix.

Multiple correlation coefficients and coefficients of determination of superior 5 emotions are shown in Tab.V. As understood from the Tab.V, the coefficients of determination of every emotion has very large value, accordingly it can be said that the relation information is efficient statistically. It also says that it is possible to transform observed prosodic parameters into emotional content perceived by the subjects, and the opposite also consists, too.

TABLE V

Result of multiple regression analysis

emotion order	surprise 1	anger 2	flattery 3	fear 4	sorrow 5
corr.coeff.	0.909	0.907	0.859	0.851	0.843
det.coeff.	0.826	0.822	0.737	0.724	0.710

TABLE VI

Correlation between the system output and the subjects

emotion order	surprise 1	anger 2	sorrow 3	cynicism 4	fear 5
corr.	0.6933	0.5779	0.5215	0.4289	0.2883

B. Emotion recognition experiment

The correlation between subjective evaluation values and the output by the system was calculated. If it is large, it can be said that the system is capable of evaluating emotional content of speech from analyzed prosodic features. 20 sample speeches which is out of the teacher speech and their emotional content are unknown were evaluated by the system as explained in III.B. And simultaneously the subjects who are the same 18 subjects as mentioned in V-A-A.2 evaluated the same sample speeches. The result is shown in Tab.VI. About "surprise", "anger" and "sorrow", the correlation values were large comparatively, so these 3 emotion is well recognized by the prosodic components of the speech.

C. Emotional speech synthesis experiments

The quality of the emotional speech synthesis essentially relies on whether the prosodic parameters proposed in chapter IV are sufficient to reproduce the original speech. So the validity of the parameter set is verified at first, and two types of synthesis experiments are conducted. One is within the condition of system construction, namely /n/ is used for the synthesized word, and the other uses usual words instead. In both cases, emotion information (the intended emotion level is maximum value, and others are minimum value) is transformed into physical parameter changes by the procedure explained in III.C.

C.1 Preliminary experiment

The validity of the parameters used here is verified. 60 speech data spoken with 5 words was parameterized by these 18 parameters, and input to PSOLA to synthesize speech waveforms. As the result of the listening examination by 4 subjects, their naturalness and emotional content were preserved in the synthesized speech although deterioration of naturalness was observed in voiceless or consonants part. Moreover, there were nobody who pointed out lack of emotional content in synthesis speech. Accordingly, it can be thought that the proposed parameter set is sufficient to describe speech, especially about communicating emotional content.

TABLE VII

Result of the synthesis experiment (A)

speech \ terms	anger	joy	sorrow	fear	surprise
anger	**80.4**	42.9	19.6	26.8	62.5
joy	44.6	**42.9**	30.4	44.6	57.1
sorrow	23.2	19.6	**76.8**	60.7	32.1
fear	21.4	28.6	75.0	**62.5**	33.9
surprise	60.7	53.6	33.9	73.2	**91.1**

TABLE VIII

Result of the synthesis experiment (B)

emotion order	surprise 1	anger 2	sorrow 3	fear 4	joy 5
e [%]	87.5	76.8	58.9	57.1	37.5

C.2 Emotional speech synthesis experiment (A)

The neutral speech uttered by the same speaker and the word as the teacher speech and emotion information of a 9 dimensional vector that the maximum value 7 for the intended emotion and minimum value 1 for others are input to the synthesis system. It is investigated whether the intended emotion is included in the synthesized speech by means of comparing the subjective evaluation for the synthesized speech with the input emotional content. The subjective evaluation value vector is transformed into the evalution measurement shown in Eq.(10).

$$e_i = \frac{\sum_{n=1}^{N} v_n^i}{7 \times N} \times 100[\%] \qquad (i = 1, \cdots, m) \qquad (10)$$

where v denotes subjective evaluation values, i is the index to emotion words (evaluation items), n is the index to subjects and N is the number of subjects. In Eq.(10), the denominator (theoretical value) is the total amount for the ith emotion word, and the numerator (actual value) is the sum total of subjective evaluation. The result of the superior 5 emotions are shown in Tab.VII. From Tab.VII, although the other emotion was mainly perceived with the case of "joy" and "fear", the intended emotion is recognized more than the other, especially in the case of "anger", "surprise" and "sorrow". Davits[1959] [11] also reported the difficulty to recognize "joy" and "fear" only by simple prosodic components. It is possible to say from the result that the system is able to synthesize the speech which includes the emotion which is well evaluated only from prosodic components.

C.3 Emotional speech synthesis experiment (B)

In this study, meaningless word /n/ has been used for the purpose of avoiding the effect caused from the verbal meanings of words uttered. It also can be said that the speech condition is limited to 1 mora and a head high accent type. So it is supposed that it is possible to apply the system constructed here to the words which has the

same accent type as the extension because the intonation of them seems to be the same as the case of /n/. In this experiment, /miroyo/ and /owari/ are examined. /miroyo/ has the same accent type, and /owari/ is different. The result of the superior 5 emotion is shown in Tab.VIII. From the result, it can be found that "surprise", "anger" and "sorrow" again got large value.

VI. Application example

In this chapter, the cyber shopping system with the intelligent speech interface is proposed as the example of the application of the system proposed here.

A. Outline of the system

The retrieval process of the proposed cyber shopping system is shown in Fig.10. As shown in Fig.10, it has a simple feedback mechanism. At first, the system presents 4 goods candidates which are chosen from image database at random, and the user evaluates each of them by a speech including the emotional content. Then the system captures the speech and calculates its physical parameters and transforms them into emotion information through the relation information proposed in this paper. On the other hand, since it is necessary to measure how extent the user is pleased with the goods candidate, it is desirable that the emotional content which is based upon to determine the next candidates should be a kind of one-dimensional evaluation measurement. In this study, the estimated emotion information is transformed into psychological basic concepts which are extracted from the 9 emotion words in advance through factor analysis[7]. As the result of the preliminary experiments, it has been clarified that one of the bases is interpreted as "pleased – displeased" axis, so it can be used as the evaluation axis. In brief, the user's utterance is transformed into a one-dimensional evaluation value, and the presented candidates are ranked on "pleased–displeased" emotion axis. Then based on this ranking order, the image retrieval module which will be explained in detail in the following chapter determines the next candidate images. The user repeats this simple evaluation action until the candidates meet his/her own requirement.

If the image retrieval process works under the condition

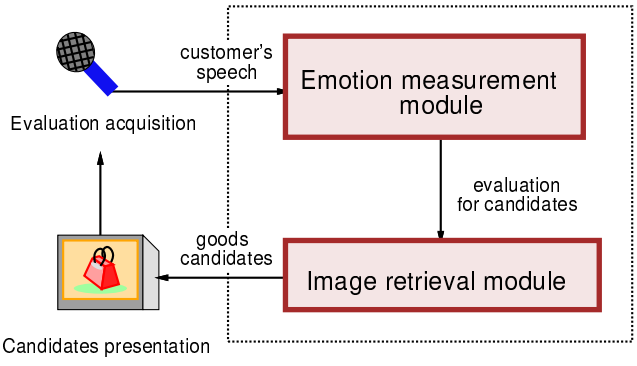

Fig. 10. Retrieving process of the proposed cyber shopping system

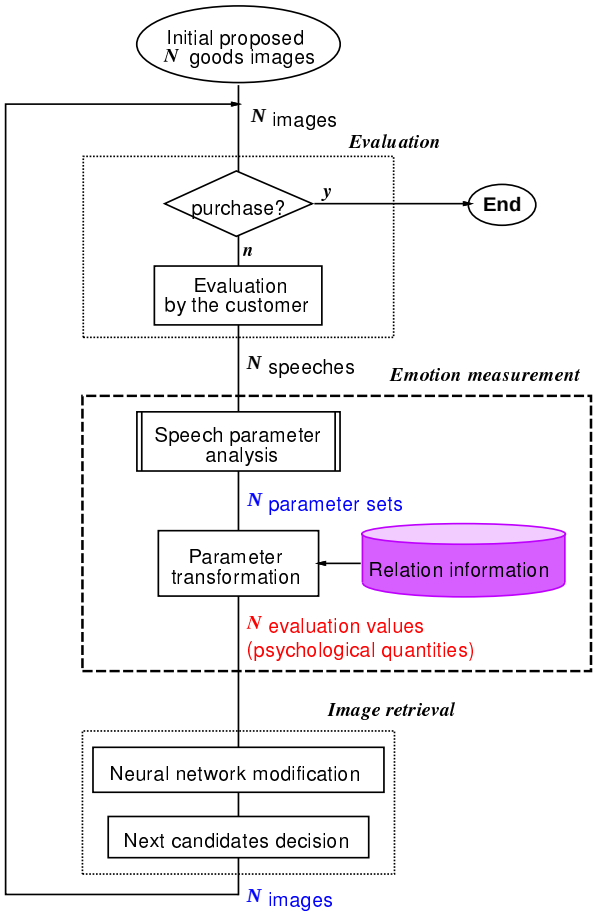

Fig. 11. Image retrieval procedure

that the user gives evaluation in several steps rating by typing a keyboard or clicking a mouse, it is supposed that the user has to translate his/her intuitive and often momentary evaluation to the values on the given scales. It seems burdensome to the users and consequently the system efficiency will decrease. Hence the speech interface is the efficient solution for requirements to reflect user's intuitive and vague request keys to data retrieval on large scale databases.

B. Image retrieval module

The image retrieval procedure is shown in Fig.11. It consists of the neural network[10]. In general, a neural network is trained before being applied, but the method in this system needs no beforehand training steps, because it is performed while it is in use. The cyber shopping system has a goods database which consists of two parts: one is goods images and the other one is physical features of each images. The components of each parts are correspondent one-to-one with the other one. And all images are tagged on with one-dimensional evaluation values (on "pleased" to "displeased" axis in this case). The relation between the evaluation value and physical features of the image is described in the form of weights of the synapses of the neural networks shown in Fig.12 (the former is input to the input

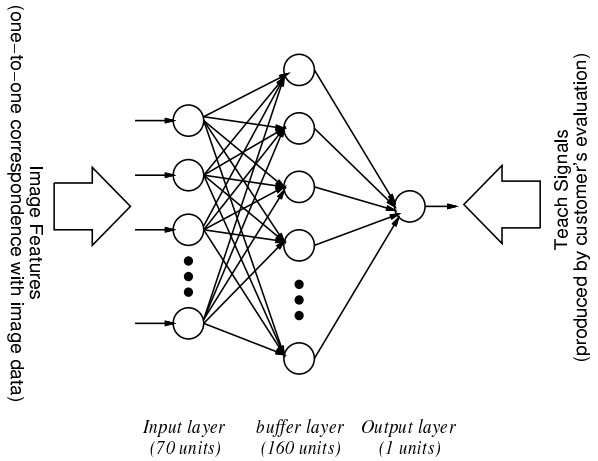

Fig. 12. Structure of neural networks

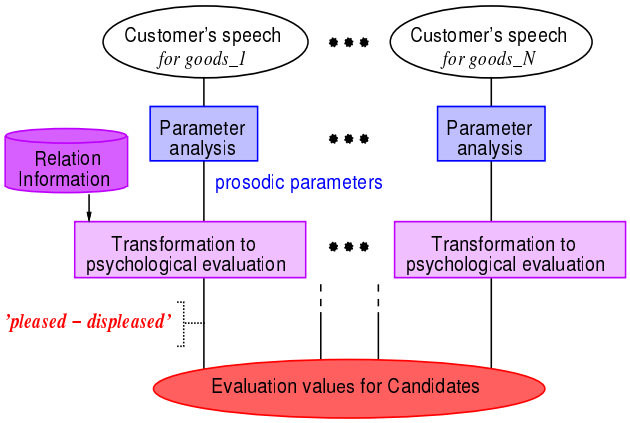

Fig. 13. Emotion measurement module

layer and the latter is input to the output layer). So the images in the database are ranked by one-dimensional axis, and the top image is always what the user is best pleased with. The system presents candidate images at random initially, but from the 2nd time on, it chooses candidates regarding the latest learned weights of synapses. Once the user evaluates a presented image, the network minimizes the root mean square between evaluation values and the output values using current weights of synapses by updating the weights (back–propagation algorithm). After each trial, the system immediately ranks all goods images by given evaluation values, and first 4 goods images are presented as next candidates, which seem to be more desirable for the user than previous ones. Hence, by only repeating one-dimensional evaluation for presented goods, the system can learn the vague retrieval key in the form of the relation between the physical features of images and evaluation values for them, and the user can get a desirable goods fast and efficiently. Retrieval process by the user is in fact training process if it is considered as the teacher signal for the system.

In this system, each image is segmented into parts and the first 10 parts having largest area are selected. Then 7 physical parameters are extracted from each of them. Accordingly there are 70 units in the input layer of the network. These parameters such as mean color, regional area and so on are chosen independent on specific kind of images. So each image can be treated as 70 order physical features vectors. And emotional content in the user's speech is obtained by a procedure as is shown in Fig.13.

C. Experiments

C.1 Condition

300 scenery images are stored in a database with their physical feature parameters mentioned above. The subjects consists of 4 students in the laboratory, who are instructed to retrieve images based on the same sensibility words "refreshing" and "simple and appealing", because it is supposed to be necessary to share the common request image among the subjects. To verify the ability of the system, the subjects are required to evaluate 10 times. If the evaluation values rise every time a time is piled up, it can be said that the system is taking effect. 4 candidate images are presented on the display at a time, and the transition of the average values among evaluation values for each 4 candidates are observed for all trials.

C.2 Results

Fig.14 and Fig.15 shows the transition of the average of the evaluation values for presented candidates. Evaluation values of each subject is indicated in thin lines and the average value is in thick lines. Toward the 5th or 6th trial as is illustrated in Fig.14, evaluation values grows rapidly. It indicates that the system had learned the users' vague requests and presented suitable candidates gradually. From another point of view, it also can be said that it is appropriate to set the trial times to 5 or 6 in this case. By the way, after that, the evaluation values decreased and dispersed. This is because the subjects were required to conduct 10 trials and each image was presented only once, even if it was very desirable. Originally, it is supposed that the subjects would finish the retrieval process at the 5th or 6th trial. In Fig.15, the average evaluation value is gradually increasing, therefore it can be said that presented candidates were getting more and more pleased with them in this case. The distribution of values converges and disperses because the response of the neural networks is very fast, and each image is presented only once.

On the other hand, for the purpose of comparing results, an experiment was conducted under the condition that the subjects evaluated by clicking mouse in 7 steps rating. As the result of it, the case with the emotion measurement module generally presented the best candidate more quickly. Accordingly, it is supposed that the mouse interface forces the subjects to translate their impression for the candidates into values on the given one-dimensional scale unnaturally, while speech interface allows the subjects to express their intuitive evaluation in natural utterance and the system analyzes it and transforms into the general scale.

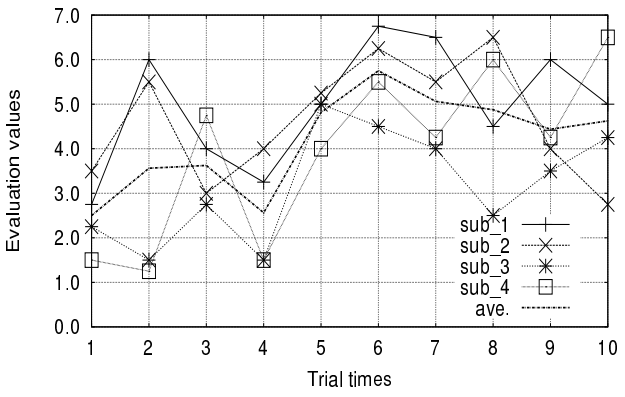

Fig. 14. Transition of evaluation values for "refreshing"

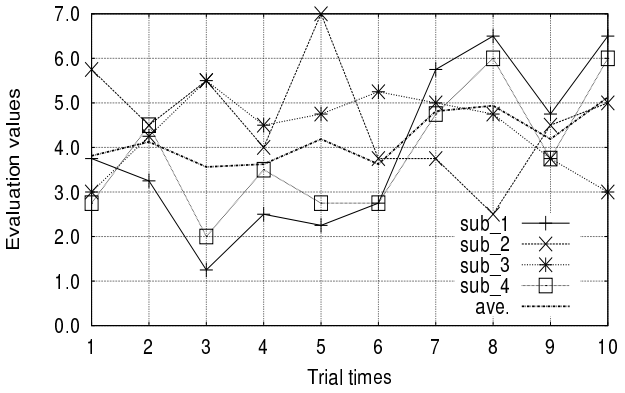

Fig. 15. Transition of evaluation values for "simple and appealing"

VII. Conclusion

In this paper, the relation information which relates emotional impression perceived from speech to physical parameters of the prosodic components of the speech was obtained. Also, the emotion recognition and synthesis system using the relation information is constructed. As the result of the experiments which also verified the validity of the emotion model, it was shown that the relation information has high efficiency statistically, and the system is able to recognize the emotional content in speech and add emotion information to the neutral speech using the relation information. Especially about "anger", "surprise" and "sorrow", both of the recognition system and synthesis system work accurately, it also means that "anger", "surprise" and "sorrow" are well described by the prosodic components of speech. "joy" and "fear" was conversely difficult as reported in the past studies which said it is difficult to measure them only by prosodic components of speech.

The application example was shown, in which the emotion recognition system is applied to the emotion measurement module to capture the user's input to the cyber shopping system. As the result of the experiment of image retrieving, it is clarified that the emotional content in speech is related to an image data of a different media, then the users could find their favorite images naturally and fast. In the near future, the system will synthesize the human-like speech with the emotional expression which is suitable for the presented goods.

The system proposed in this paper is applicable to another applications. It is, for example, possible to encode the multi-modal emotional content of speech and facial expression efficiently in the Internet environment. It contributes to rich communication on the net although it costs just few bits. On the other hand, pet robots which is capable of exchanging human factors like emotion with human beings has been expected to improve welfare services. In such utensils, the system proposed in this paper is thought to be very efficient.

In the future, to extend the knowledge in this study, it is supposed to be necessary to find several patterns of combination of accent types and mora number. Also it is important to clarify the relation between the factors caused from personalities and verbal meanings and emotional impression perceived from the speech.

References

[1] K.Hirose, H.Fujisaki and M.Yamaguchi, *Synthesis by rule of voice fundamental frequency contours of spoken Japanese from linguistic information*, Proc. IEEE ICASSP, 2.13 (March 1984).

[2] P.Lieberman, S.B.Michaels, *Some Aspects of Fundamental Frequency and Envelope Amplitude as Related to the Emotional Content of Speech*, J.Acoust.Soc.Am., Vol.34, Number 7, pp.922-927, July 1962.

[3] R.M.Dawes, E.Kramer, *A proximity analysis of vocally expressed emotion*, Perceptual and Motor Skills, 22, pp.571–574, Southern Universities Press, 1966.

[4] L.A.Streeter, N.H.Macdonald, R.M.Krauss, W.Apple, K.M.Galotti, *Acoustic and perceptual indicators of emotional stress*, J.Acoust.Soc.Am., vol.73, No.4, pp.1354-1360, April 1983.

[5] B.L.Brown, W.J.Strong, A.C.Rencher, *Fifty-four voices from two: the effects of siultaneous manipulations of rate, mean fundamental frequency, and variance of fundamental frequency on ratings of personality from speech*, J.Acoust.Soc.Am., vol.55, No.2, pp.313–318, February 1974.

[6] C.E.Williams, K.N.Stevens, *Emotions and Speech: Some Acoustical Correlates*, J.Acoust.Soc.Am., Vol.52, Number 4, Part 2, pp.1238-1250, 1972.

[7] T.Moriyama, H.Saito, S.Ozawa, *Evaluation of the relation between emotional concepts and emotional parameters on speech*, Proceedings of ICASSP IEEE, SPCH7P(April 22nd 1997).

[8] I.R.Murray, J.L.Arnott, *Toward the simulation of emotion in synthetic speech : A review of the literature on human vocal emotion*, J.Acoust.Soc.Am., vol.93, No.2, pp.1097-1108, February 1993.

[9] H.Levin, W.Lord, *Speech Pitch Frequency as an Emotional State Indicator*, IEEE Trans. Systems, Man, and Cybernetics, vol.SMC-5, No.2, pp.259-273, March 1975.

[10] Y.Kageyama, H.Saito, *Image retrieval system capable of learning the user's sensibility using neural networks*, Proceedings of ICNN IEEE, vol.3, pp.1563–1567, June 1997.

[11] J.R.Davitz, J.R.Davitz, *Correlates of Accuracy in theCommunication of Feelings*, J. Commun. 9, pp.110-117, 1959.