Emotion Recognition and Synthesis System on Speech

Tsuyoshi Moriyama and 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.

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.

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 human. 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 from the primary emotions reported in psychology (mainly investigated about facial expressions), or several emotions were chosen just provisionally. Also about physical aspects, just limitation from the speech synthesis systems in fact determined the parameter sets. However the emotional concept that is possible to be treated as a kind of universal quantity in the engineering systems and the relation model between emotional content and physical aspects is supposed to be necessary.

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 in speech and synthesize an emotional speech from a neutral one based on the estimated relation information in the model.

II. Relation model

To clarify the relation between the physical aspect and the emotional content of the speech, it is necessary to measure those quantities at first. Fig.1 imitates the appearance of speech communication with various emotional expression. As shown in Fig.1, two emotions are observable also in the experiments: The one is intended by the speaker and the other is shared by the observers. The former is consequently filtered by many factors such as the external environment, the inner physiological state and the

0-7695-0253-9/99 $10.00 © 1999 IEEE

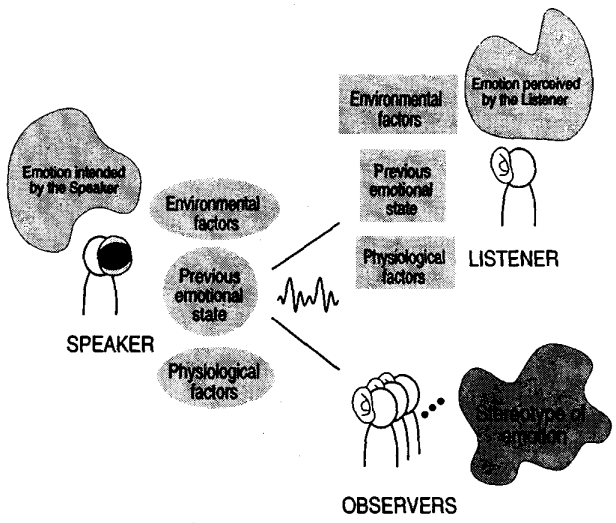

Fig. 1. Emotion communication

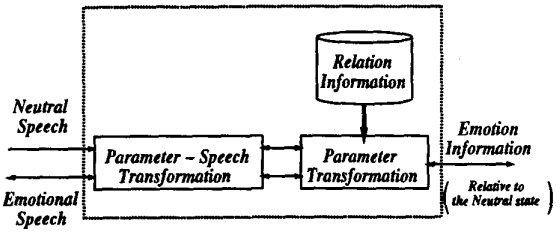

Fig. 2. Proposed relation model

previous emotional state, and it is impossible to decide the causality among them statically. On the other hand, the latter is identified as the stereotype of emotion expression rearranged among observers such as an audience of a play. And 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, the relation model in this study is based on the latter emotion.

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 the 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. Moreover, as the result of statistical experiments by Moriyama [7], strong correlation was found among the observed quantities of physical parameters and emotion words. So in this study, orthogonal bases are extracted from distribution statistics collected from those observable variables, and using the bases to construct the

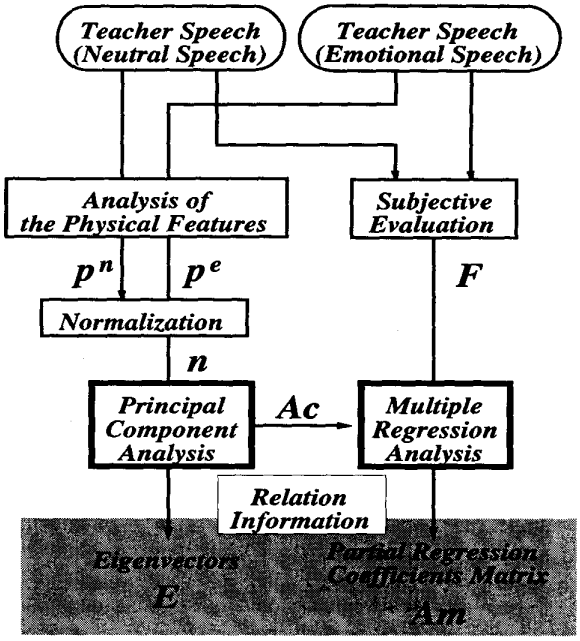

Fig. 3. Flow of obtaining the relation information

system which is able to translate physical aspects and included emotional content to each other, it is supposed to be able to avoid the dependency on the choice of observed variables.

III. Method

A. Obtaining the Relation information

The relation information (Fig.2) describes the relation between the emotion information in a speech and the 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 method of obtaining the relation information is shown in Fig.3 (Moriyama[7] in detail). 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. The obtained relation information (the eigenvector $\boldsymbol{E}$, and the partial regression coefficient matrix $\boldsymbol{A_m}$) are applied to the emotion recognition/synthesis system shown in Fig.4.

B. Recognition system

Physical parameters, discussed in the IVth chapter, of

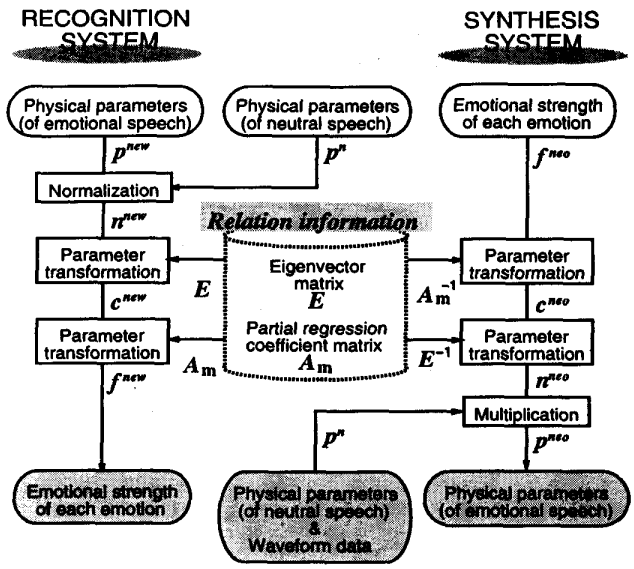

Fig. 4. Proposed system

the neutral and input speech ($\boldsymbol{p}^n$, $\boldsymbol{p}^{new}$) are calculated, and the normalized parameter $\boldsymbol{n}^{new}$ is obtained by dividing $\boldsymbol{p}^{new}$ by $\boldsymbol{p}^n$. Then, the emotional strength of each emotion words ($\boldsymbol{f}^{new}$) are calculated through simple linear operations as shown in Fig.4.

C. Synthesis system

Contrary to the recognition system, the physical changes given to the neutral speech ($\boldsymbol{n}^{neo}$) is obtained through the inverse linear operation to the input emotional strength (grade) $\boldsymbol{f}^{neo}$. $\boldsymbol{n}^{neo}$ is transformed into $\boldsymbol{p}^{neo}$ and input to PSOLA (wave synthesis system) as shown in Fig.4.

IV. Physical parameters

In this study, PSOLA (Pitch Synchronous OverLAp-add) is applied to synthesize speech, which is capable of transforming the source speech waveform into the speech with given F_0 contour, power envelope and utterance velocity. In the following sections, how to calculate those parameters is explained briefly.

A. F_0 contour

In this study, the revision of the F_0 pattern generation model proposed by Fujisaki [1] is applied to parameterize the pitch contour, which is called PFM (Pseudo Fujisaki Model) hereafter. In the PFM, 7 parameters determine the temporal contour of F_0 in TABLE I.

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 Fig.5

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)]

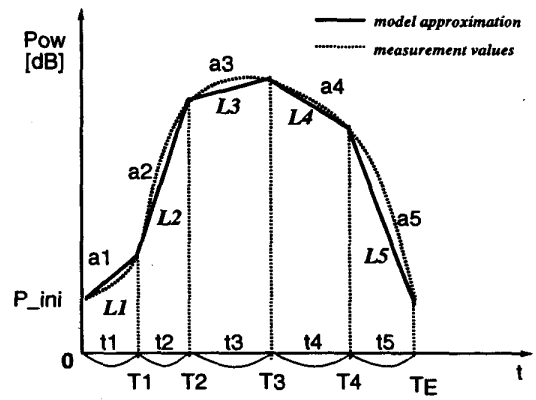

Fig. 5. An example of power envelope model fitting

V. Experiments

So far, the relation information is obtained by the mentioned process, and the system shown in Fig.4 has been constructed. In this section, the system is evaluated through an emotion recognition experiment and an emotional speech synthesis experiment. 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 conducted at first. Then, the result of recognition and synthesis experiments is indicated in order. Furthermore, in the preliminary experiments, it is verified that the 18 prosodic parameters (in PSOLA) are sufficient to reproduce the speech preserving naturalness and the emotional content of it.

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 TABLE II (the cutoff accumulative ratio of distribution is 80%). From the positioning of the prosodic parameters on the basic axes, they can be interpreted. The 1st, 2nd and 3rd principal component is, for example, regarded as "steepness", "flatness" and "attenuation" axis respectively.

TABLE II

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 III

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

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 TABLE III had been chosen from 46 emotion words based on the correlation among them. 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:

$$\boldsymbol{f} = \boldsymbol{A}_m \cdot \boldsymbol{c}$$

$$= \begin{bmatrix} 0.495 & -0.012 & -0.691 & 0.274 & 0.315 \\ 0.192 & -0.004 & 0.163 & 0.025 & -0.166 \\ -0.541 & 0.392 & -0.012 & -0.367 & 0.173 \\ -0.168 & 0.340 & -0.082 & -0.264 & -0.033 \\ -0.074 & -0.022 & 0.040 & 0.029 & 0.118 \\ -0.003 & 0.057 & 0.117 & -0.058 & -0.088 \\ -0.111 & -0.205 & 0.249 & 0.033 & 0.110 \\ -0.137 & 0.172 & 0.055 & -0.035 & 0.197 \\ 0.298 & 0.305 & -0.175 & 0.137 & -0.453 \end{bmatrix} \cdot \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \end{bmatrix} \quad (1)$$

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 TABLE IV. TABLE IV indicates that the relation information is efficient statistically because the coefficients of determination of every emotion has very large value. It also says that it is possible to transform between observed prosodic parameters and emotional content observed by the observers mutually.

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.

TABLE IV

RESULT OF MULTIPLE REGRESSION ANALYSIS

emotion	surprise	anger	flattery	fear	sorrow
order	1	2	3	4	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 V

CORRELATION BETWEEN THE SYSTEM OUTPUT AND THE SUBJECTS

emotion	surprise	anger	sorrow	cynicism	fear
corr.	0.6933	0.5779	0.5215	0.4289	0.2883

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 TABLE V. 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

In this section, 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.

C.1 Emotional speech synthesis experiment (A)

The neutral speech uttered by the same speaker and the same word as the teacher speech (/n/) and emotion information of a 9 dimensional vector (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

TABLE VI

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 VII
RESULT OF THE SYNTHESIS EXPERIMENT (B)

emotion	surprise	anger	sorrow	fear	joy
e [%]	87.5	76.8	58.9	57.1	37.5

evalution measurement shown in Eq.(2).

$$e_i = \frac{\sum_{n=1}^{N} v_n^i}{7 \times N} \times 100[\%] \qquad (i = 1, \cdots, m) \qquad (2)$$

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. The result of the superior 5 emotions are shown in TABLE VI. From TABLE VI, 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] [10] also reported the difficulty to recognize "joy" and "fear" only by simple prosodic components.

C.2 Emotional speech synthesis experiment (B)

Although, the speech condition is limited to the only 1 mora and a head high accent type so far, it supposed to be necessary to treat the voice condition of a variety of accent type in plural morae. It seems to be able to apply the system to those which have the same speech condition as the extension. 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 TABLE VII. From the result, it can be found that "surprise", "anger" and "sorrow" again got large value.

VI. Conclusion

In this paper, the relation information which relates between the emotional content and the prosodic features 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 communication model mentioned in II, the system is able to recognize the emotion from physical features of the speech and add some emotion to the neutral speech. Especially about "anger", "surprise" and "sorrow", both of the recognition system and synthesis system works well, it also means that these emotions are well described by the prosodic components of speech.

In the future, it is necessary to find typical patterns of the speech condition such as 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.

VII. References

[1] K.Hirose, H.Fujisaki and M.Yamaguchi, "Synthesis by rule of voice fundamental frequency contours of spoken Japanese from linguistic information", *Proc. of 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, No.7*, July 1962, pp.922-927.

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

[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*, April 1983, pp.1354-1360.

[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*, February 1974, pp.313-318.

[6] C.E.Williams, K.N.Stevens, "Emotions and Speech: Some Acoustical Correlates", *J.Acoust.Soc.Am., Vol.52, No.4, Part 2*, 1972, pp.1238-1250.

[7] T.Moriyama, H.Saito, S.Ozawa, "Evaluation of the relation between emotional concepts and emotional parameters on speech", *Proc. of IEEE ICASSP, SPCH7P*, April 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*, February 1993, pp.1097-1108.

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

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

[11] G.Fairbanks, "Recent Experimental Investigation of Vocal Pitch in Speech", *J.Acounst.Soc.Am, Vol.11*, 1940, pp.457-466.

[12] H.Schlosberg, "Three Dimensions of Emotion", *Psychol. Rev., Vol.61, No.2*, March 1954, pp.81-88.

[13] M.Cowan, "Pitch and Intensity Characteristics of Stage Speech", *Arch. Speech, Suppl. to December issue.*, 1936.

[14] J.Laver, R.Hanson, "Describing the Normal Voice", *Speech Evaluation in Psychiatry*, edited by J.K.Darby Grune and Stratton, New York, 1981.

[15] E.H.Mamdani, "Application of Fuzzy Algorithms for Control of Simple Dynamic Plant", *Proc. IEE, vol.121, No.12*, 1974, pp.1585-1588.

[16] R.Plutchik, *Emotion, A psychoevolutionary synthesis*, Harper&Row, 1980.

[17] I.Fonagy, "Emotions, Voice and Music", *Research aspects on Singing*, edited by J.Sundberg (Royal Swedish Academy of Music No.33), 1981, pp.51-pp.79.

[18] C.E.Shannon, W.Weaver, *The Mathematical Theory of Communication*, University of Illinois, Urbana, 1949.

[19] E.Kramer, "Judgement of Personal Characteristics and Emotions from Nonverbal Properties of Speech", *Psychol. Bull.60*, 1963, pp.408-pp.420.