감성과학, Vol. 8, No. 1, pp. 1-7. March 2005

Emotion Detecting Method Based on Various Attributes of Human Voice*

Yutaka MIYAJI**† · Ken TOMIYAMA***

Aoyama Gakuin Women's Junior College**

The Department of Integrated Information Technology, Aoyama Gakuin University***

Abstract : This paper reports several emotion detecting methods based on various attributes of human voice. These methods have been developed at our Engineering Systems Laboratory. It is noted that, in all of the proposed methods, only prosodic information in voice is used for emotion recognition and semantic information in voice is not used. Different types of neural networks(NNs) are used for detection depending on the type of voice parameters. Earlier approaches separately used linear prediction coefficients(LPCs) and time series data of pitch but they were combined in later studies. The proposed methods are explained first and then evaluation experiments of individual methods and their performances in emotion detection are presented and compared.

Key words : Emotion Detection, Prosodic Information, LPC, Pitch, Neural Network

1. Introduction

This paper reports our efforts to develop technologies to detect emotional states of cared persons through voice-sound. These technologies have been developed as part of the Welfare Robotics Project at the Engineering Systems Laboratory at Aoyama Gakuin University, whose aim is to develop care-worker support robots [2, 4].

The general flow of the processing of our detection systems is described first. Then, a series of techniques using individual voice attributes and combinations of those attributes are explained and their performance evaluation experiments are reported. Finally, discussions and subjects for further study are stated.

2. Emotion Detection System Overview

The voice-sounds that we processed in this study include voiced and unvoiced sounds and aspiration. We used only prosodic information here.

The voice parameters that we adopted for our study included linear prediction coefficients (LPCs), time-series data of pitch and voiceprints. Here, attempts using only the first two

* Selected as an excellent paper in 2004 Korea-Japan Joint Symposium on Emotion & Sensibility.

† Corresponding Author : Yutaka MIYAJI(Aoyama Gakuin Women's Junior College)
E-mail : uta@luce.aoyama.ac.jp
TEL : +81-3-3409-7174
FAX : +81-3-3409-7174

parameters are discussed. LPCs roughly represent linear dynamics of the vocal tract and can be found from peaks of spectral density of voice. Pitch, on the other hand, represents the basic frequency of the vibration of vocal chords. The pitch, therefore, does not exist for unvoiced sounds and aspiration. Combining these two voice parameters to realize better detection rates was also attempted.

An index, called comfort level, was defined as a multi-stage indicator of comfort/discomfort level and was introduced to numerically represent the emotional state of speakers. It enables the human emotional state to be processed from an engineering point of view.

Figure 2-1 is an overview of our detection system. It is divided into two major components representing the voice-sound processing unit and the emotional state detection unit. The input of the former unit is the voice-sound and the output is the voice parameters. The latter unit takes voice parameters produced by the former unit to compute the emotional state of the speaker. Processing in the former unit depends on the type of voice parameters produced by that unit but that of the latter unit is done by neural networks (NN) of various sizes and construction.

3. Comfort Level Detection from LPCs

Our first attempt in human emotion detection used LPCs as the voice parameter. This section summarizes the findings of that attempt.

3.1 Voice-sound Processing Unit

Monaural voice samples obtained from a microphone were digitized at 44.1 kHz and 16 bits with DAT as the recording media. The obtained data were divided into frames of 23 [msec] long using a Hamming window. This corresponds to a set of 1024(=210) data points per window. Then, the prepared data set was curve-fitted by an auto-regressive(AR) filter, called a linear predictor, of appropriate dimension. The LPCs are the coefficients of the obtained AR filter.

Preliminary experiments were performed to determine the optimal dimension of the AR filter, namely the number of LPCs. The dimension was varied from 2 to 32 by 2 and was found to yield enough accuracy and little improvement after it reached 8. It was determined to use 8 as the dimension of the filter from this result.

3.2 Comfort Level Identification Unit

Two types of NN were used in this part; a simple three layer network and a recurrent NN[3]. The reason for the adoption of a recurrent NN was our claim that emotion transition involves dynamics. In other words, we believed a dynamic system was needed to generate and to model human emotion.

In the former cases, the input, hidden and output layers had 8, 20 and 5 nodes, respectively. In the latter network, the input layer had 13 nodes with 5 nodes for the output feedback. The number of hidden layer nodes was determined

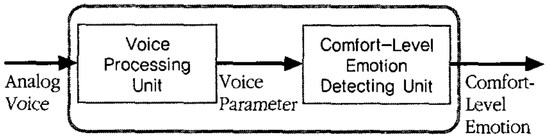

Figure 1. System Pverview

from a series of preliminary experiments with a varied number of nodes. The comfort level of a node with highest output was chosen as the output comfort level if the value was not less than 0.5. The indication of "undetermined" was outputted when the maximum output value did not exceed 0.5. The back propagation method was used for learning. Convergence in learning for each NN was declared when the RMS error becomes less than 0.01.

3.3 Evaluation Experiments

The sentence 'so-nandesuka, e, sorejaa'(I got it, but, then if …) was chosen as the articulation sample because this sentence can be used in scenarios with different emotional context. Several sets of sample voice data corresponding to three comfort levels ("comfort," "neutral" and "discomfort") were obtained from a group of subjects by asking them to show those comfort levels while articulating. The collected data were judged by friends of subjects for their comfort levels and only those data that matched the intended comfort levels were used for experiments.

Recognition experiments were performed with data sets that were not used for learning experiments. The results are listed in Table 3-1 as Type-A and Type-B for regular and recurrent NN cases, respectively. The recognition rate of 46% for the regular NN versus 82% for the recurrent NN prove our claim that there is a dynamics in the transition of the emotional state.

In order to evaluate feasibility in future real-time operation, several levels of downsampling were tested for reduction of data without sacrificing recognition rate. A motivation for this experiment was that the combination of the sampling rate (at 44.1 kHz) and the frame length (23 [msec]) requirements yields a data set too large to be processed in real time. As the results in Table 3-2 show, up to 1/4 down-sampling was possible before significant degradation in recognition rate occurred.

Table 1. Recognition Rates for Two Types of NNs

	Type-A	Type-B
Collect Ratio	46% (633/1395)	82% (1157/1405)

Table 2. Effect of Down Sampling

Down Sampling	Type-D
1/1	94%(1314/1405)
1/2	81%(1142/1405)
1/4	82%(1157/1405)
1/8	66%(923/1405)
1/16	51%(710/1405)

4. Comfort Level Detection from Pitch Time Series

Comfort level detection from pitch time series using a regular NN with input nodes for delayed inputs are reported in this section. The voice processing part is not elaborated on here because the pitch extraction technique is well-known and routines for pitch computation are widely available.

4.1 Emotion Detection Unit for Pitch Experiments

Since pitch time series data do not always exist and since the recurrent NN is susceptible to data loss, a regular three layer NN with extra input nodes for delayed inputs was adopted for the emotion detection unit.

Through a series of preliminary experiments, it was decided to use a total of 5 inputs, from the current pitch p(k) to the pitch p(k-4) at step k-4. This makes the number of nodes in the input layer 5. Expanding the number of the comfort levels from 3 to 5, namely, adding strong and weak levels on both comfort and discomfort levels, was tried but neither the learning nor the recognition results were satisfactory and, therefore, 3 comfort levels were maintained in this experiment. Thus, only 3 or 4 nodes in the output layer were used. The 3 output nodes correspond to the 3 comfort levels and the fourth node is designated as a "no-detection" node that corresponds to the lack of pitch data. The input value of 0 was used when the pitch did not exist. The number of nodes in the hidden layer was chosen to be 10.

4.2 Evaluation Experiments and their Results

The phrase 'doushite so naruno desuka?' (why does it become like that?) was used as the test signal in this experiment. This phrase was chosen because it contains all five voiced sounds, some un-voiced sounds and a no sound interval. It is also easy to articulate this phrase with different levels of emotion.

Four subjects(A, B, C and D), were asked to articulate this phrase with the three comfort levels; comfort, discomfort and neutral. The pitch time series data were extracted from the collected voice samples and then half of them were used for the learning of the NN and the other half for recognition experiments. Note that the total length of the pitch time series data was expected to be and actually was found to be dependent on the subject.

A good overall recognition rate of 72% was achieved. Individual recognition results, listed in Table 4-1, however, indicate that there is a large variation in recognition rates among the four subjects as well as the comfort levels. High recognition rates are observed for the "comfort" and "normal" levels of subject A, the "discomfort" level of subject B and the "normal" and "discomfort" levels of subject C. However, no significant difference is observed with subject D. This indicates that recognition rate is dependent on the subject. This trend is expected, however, because the length and quality of the pitch time series data vary widely from one subject to another.

Table 3. Recognition from Pitch Time Series

	Collect Ratio			
	A	B	C	D
Comfort	25/25	25/42	6/18	13/20
Normal	27/27	11/28	13/13	15/19
Discomfort	10/21	38/38	13/13	11/22

4.3 The Processing of Missing Pitch Data

This subsection reports the result of an experiment to investigate whether the NN is capable of learning no data condition in pitch time series data. As discussed before, the teaching data here included intervals with pitch data of 0, signaling periods where pitch data was missing. The correct output for the pitch data missing period consisted of 1.0 at the "no-detection" node and 0 for all other nodes. The voice data of subject A was used as the actual voice data.

The summary of the recognition results is shown in Table 4-2 where correct answers are given in the left-most column and recognized comfort levels are given in the top row. The right-most column shows the rates of correct answers. As it can easily be seen in the bottom row, the pitch data missing condition is accurately recognized.

Table 4. Detection of no Pitch Data Situation

	Comfort	Normal	Dis-comfort	No-Detect	Collect Rate
Comfort	10	0	15	0	10/26
Normal	0	28	0	0	28/28
Dis-comfort	3	0	22	1	22/26
No-Detect	1	0	0	246	246/247

4.4 Discussions on Pitch Experiments

It was experimentally confirmed that the pitch time series could be used for emotional state recognition. It was found that an NN with time-delayed inputs could be used for recognizing conditions where pitch data was missing as well as the correct comfort level. It was also found that the rate of correct recognition depends on the voice sample. This implies that comfort level recognition may need to be customized to individual cared persons and that an on-the-spot learning session may be needed.

5. Emotional State Detection using both LPCs and Pitch

This section reports several attempts on combining the LPCs and the pitch time series in emotion state detection[1].

5.1 Voice Parameter Data Preparation

A set of voice data from a single subject, similar to those used in Section 4, was also collected here, except that the time series data of the LPC parameters were also obtained. Since the pitch data over five consecutive time instants, p(k) ... p(k-4), are needed and since the pitch calculation was done at every 50 [msec], the LPC data was obtained at every 250(=5X50) [msec] to synchronize the two types of inputs to the NN of the emotion detection unit.

5.2 Combination Experiments

The three different structures listed in Figure 5-1 were tried for the combinations of LPC and pitch. The idea of Structure 1 is to combine comfort levels that are obtained separately from LPC and pitch. It, therefore, uses an NN without input nodes specifically for combining the two comfort levels. The numbers of input / hidden / output nodes of the NNs were: 5 / 10 / 6 (upper left for pitch), 8 / 10 / 5 (lower left for LPC) and 0 / 15 /5 (right for combination). Structure 2 simply combines both LPC and pitch time series inputs into a single NN with the numbers of nodes 13 / 16 / 5. Structure 3 attempts to combine intermediate data in hidden layers of the LPC and pitch NNs rather than the individual comfort levels. The idea for this structure is the following: intermediate data at hidden layers may have subdivided information more suitable to construct combined comfort levels than individual comfort levels from two inputs. Structure 3 had 5 / 10 / 0 (top left for pitch), 8 / 10 / 0 (bottom left for LPC) and 0 / 30 / 5 (right for combination) nodes. In all three structures, the 5 (or 6) nodes in the output layer is for the possibility of future expansion of the detection system to 5 comfort levels.

In the learning experiments, Structure 1 failed to converge. Detection experiments were conducted, therefore, with Structures 2 and 3 only. The results are summarized in Table 5-1. The results indicate that overall detection did not improve much by combining the LPC and Pitch time series data using structured NNs such as Structure 1 and 3. A comparatively better result for Structure 2 indicates that it is better to combine information from LPC and pitch in an early stage since there is no direct mixing of information from LPC and pitch in the first hidden layers in Structures 1 and 3. However, Structure 2 has the largest parameter space and, as such, takes more time to obtain convergence in the learning phase.

From the experiments, it is apparent that a study for a better structure in combining LPC and pitch data is needed. This is now underway and the results will be presented in the future.

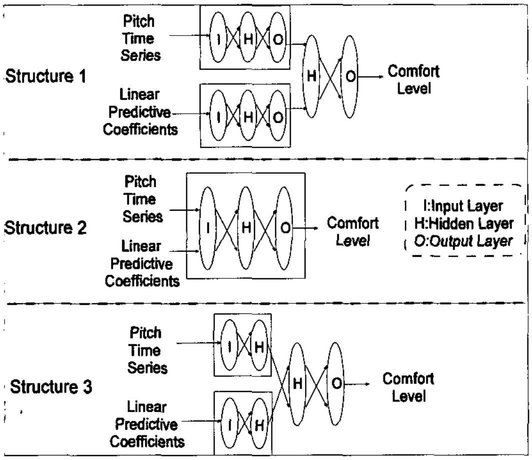

Figure 2. NN Structure 1, 2 and 3

Table 5. Recognition by Structures 2 and 3

	Structure 2	Structure 3
Comfort	71/76	18/65
Normal	65/86	20/71
Discomfort	81/92	43/80
Total	217/254	81/216

6. Conclusions

A series of attempts were made to detect the emotion state of humans from their voice-sounds. Linear prediction coefficients (LPCs) and pitch time series data were tried as primary voice parameters. A set of comfort levels was defined as an indicator of the emotion state of humans and was used with LPCs and pitch. Although reasonable recognition rates were obtained from LPCs and pitch attributes individually, an attempt at combining the LPC and pitch attributes to supplement each other and improve the recognition rate was also made. However, the results of the evaluation experiments indicated that the structure for combining the two needs to be improved.

Studies on evaluating characteristics of emotion detection from voiceprints and effects of various parameters in detection methods reported here are underway. It would also be an interesting undertaking to combine this with emotion detection from face images [5] and to implement the system in a care-worker support robot, a prototype of which was recently developed in our laboratory [6].

The subjects who provided voice sound data in this study gave their consent to participate in the experiments after being fully explained the objectives.

Acknowledgements

The authors would like to express their sincere appreciation for the significant help given by three of their students, Kazutoyo TAKATA, Yoshinori ISHII and Takehiro MINATO. They also express their gratitude to those who provided voice samples. A part of this study was supported by a grant from the Research Institute of Aoyama Gakuin University.

References

[1] Minato, T., Tomiyama, K., Miyaji, Y., & Takata, K., (2001). Integrated Comfort Level from LPC and Pitch Time Series of Voice Sound, The 19th Annual Conference of Robotics Society of Japan, 2H13 (CD-ROM), JAPAN.

[2] Miyaji, Y., & Tomiyama, K. (2003). Towards Realization of Helper-Supportive Robotic System, Journal of Healthcare Engineering Association of Japan, 45(1), 31 - 36, JAPAN.

[3] Takata, K., Tomiyama, K., & Furuta, T. (2000). Discomfort Level Recognition from Voice Sound, 2000 JSME Annual Conference on Robotics and Mechatronics (ROBOMEC '00), A1-80-115 (CD-ROM), JAPAN.

[4] Tomiyama, K., & Miyaji, Y. (2001). Towards Realization of Care Worker Support Robot. In : Ohara, S. & Kaminaga, I. (Ed.), Current Status of Welfare in Japan, IBUNSYA, JAPAN.

[5] Yajima, J., Miyaji, Y., & Tomiyama, K., (2003). Facial Expression Detection from Frequency Domain Data of Facial Images, 2003 JSME Annual Conference on Robotics and Mechatronics (ROBOMEC '03), 2A1-2F-D6 (CD-ROM), JAPAN.

[6] Yanadori, T., Kiuchi, N., Takeuchi, H., Miyaji, Y., &Tomiyama, K., (2004). Development of a Test Bed Care-Worker Support Robot, 2004 JSME Annual Conference on Robotics and Mechatronics (ROBOMEC '04), 1A1-H-63(CD-ROM), JAPAN.

Received : 1 Feb. 2005
Revised : 27 Feb. 2005
Accepted : 28 Feb. 2005