Complex emotion recognition system for a specific user using SOM based on prosodic features

Atsushi Iwai*, Yoshikazu Yano**, Shigeru Okuma*

*Department of Electrical Engineering, University of Nagoya, Japan
**School of Economics, University of Nagoya, Japan
{iwai,yoshi,okuma}@okuma.nuee.nagoya-u.ac.jp

Abstract
This paper proposes complex emotion recognition system for a specific user, where complex emotion has mingled emotions. In order to show the differences between individuals, we use Self-Organizing Feature Map (SOM) for proposed system. Additionally, in order that emotion recognition system expresses complex emotion, we propose new method for labeling. We verify proposed system using emotional speech database we recorded. As a result of verifying, this system could express fuzziness emotions such as anger with sadness, and this paper showed the effectiveness for emotion recognition to specify the user.

1. Introduction
It is desired for machines to recognize user’s emotion [1]. For example, machines should reply different action according to user’s emotion. Emotion mingles with the signals human outputs, such as facial expression, speech, and behavior. These signals are effective to recognize user’s emotion. In this paper, we focus on emotion recognition in speech.

A number of studies have been made on emotion recognition in speech, and have shown the effectiveness of the prosodic features in speech [2] [3]. However, because of some differences in prosodic features between individuals, the accuracy of emotion recognition goes down [4]. Therefore, we assumed that emotion recognition system should be constructed for a specific user.

Additionally, each emotion has similar features with each other. It is difficult to distinguish between each emotion[5]. Human emotion is not only basic emotions, such as angry, sad, and neutral. We assumed that emotion recognition system should express complex emotion.

2. Proposed system
2.1. System overview
Each speaker has a habit in speech, so the emotion recognition model for general speaker is not suitable for them. Fig.1 indicates the differences between emotional relationships of each speaker. Human emotion has a mingled areas of the emotions, such as angry with a little sad. Though emotion can be classified crisply in traditional system, they cannot represent the fuzziness of the emotions.

Figure 1: Differences between Speakers

This paper proposes complex emotion recognition system for a specific user, where complex emotion has mingled emotions. In order to show the differences between individuals, we use Self-Organizing Feature Map (SOM) for proposed system. The overview of proposed system is show in Fig. 2.

Emotional speech database consists of speech data and its emotion data of a specific speaker. In the training stage, speech data mingled with known emotion are learned by SOM using the prosodic features extracted from speech data. Then, SOM is labeled by corresponding emotion data. Traditional system used to label nodes using the nearest input vector of corresponding node. Proposed system, however, expresses complex emotion which mingles with some basic emotions like natural, joy, anger, and sadness, by labeling with some input vectors.

In the recognition stage, prosodic features of input
speech data are inputted to trained SOM. Complex emotion is obtained for input speech data.

2.2. Emotional speech database

Speech data are uttered by non-actor people. We deal with four emotions (neutral, anger, joy, and sadness) applied in a number of studies [8]. Emotion data in speech data are given by listeners. They listen to each utterance, and recognize the feelings mingled with the utterance.

The utterances are classified to one of four emotions. The feelings given by listeners are regarded as emotion for the utterances.

2.3. Speech processing

Prosodic features are extracted from one word. In this paper, 5 prosodic features are applied: maximum of pitch, range(simple max-min) of pitch, mean of pitch, maximum of power, and utterance time. Effectiveness of this prosodic features has been shown in a number of studies [6] [7]. These prosodic features are used as a feature vector. The feature vector $x$ is input into the emotion recognition part.

2.4. Training

2.4.1. Learning

SOM is learned by prosodic features. The set $X \in \{x | x_1, x_2, \ldots\}$ is defined as a set of the feature vectors for training. SOM has an output layer consisting of $N \times N$ nodes, each of which represents a weight vector that has the same dimension as the prosodic features. SOM is learned by the set of the feature vectors using competitive learning algorithm. Similar characteristics of prosodic features are mapped closer in the feature map [9].

In first step, we decide the node $c$, which is in the best conformity with a feature vector $x$ in the output layer. In this paper, node $c$ is defined as follows:

$$\|x - m_c\| = \min_i \|x - m_i\|$$  \hspace{1cm} (1)

where $m_i$ is a weight vector of a node $i$ in the output layer. Each feature vector $x(\in X)$ is measured with $m_i$ of all nodes.

In next step, the neighborhood of the node $c$ is learned. The learning rule is defined as follows:

$$m_i(t + 1) = m_i(t) + \alpha(t) \cdot h_{ci}(t)\{x - m_i(t)\}$$  \hspace{1cm} (2)

where $m_i(t)$ is the weight vector update performed during $t$th iteration through of the main loop, $\alpha(t)$ is the learning rate, and $h_{ci}(t)$ is the neighborhood function.

The learning rate and the neighborhood function are defined as follows:

$$\alpha(t) = \alpha_0 \left(1 - \frac{t}{\text{step}}\right)$$  \hspace{1cm} (3)

$$h_{ci}(t) = \exp\left(-\frac{r^2_{ci}}{2\sigma^2(t)}\right)$$  \hspace{1cm} (4)

$$\sigma(t) = 1 + (\sigma_0 - 1) \left(1 - \frac{t}{M}\right)$$  \hspace{1cm} (5)

where $r_{ci}$ is distance of node $i$ and node $c$.

2.4.2. Labeling

Each node of the learned SOM is labeled using emotion data. In this paper, complex emotion model is proposed. Each node in the output layer has likelihood of each emotion, such as $L_{\text{neutral}}, L_{\text{anger}}, L_{\text{joy}}, L_{\text{sadness}}$, in order to express complex emotion. For example, $L_{\text{neutral}}(i)$, likelihood of neutral in node $i$, is defined as follows:

$$L_{\text{neutral}}(i) = \sum_{a \in X_{\text{neutral}}} \frac{1}{\sqrt{2\pi}\sigma_{\text{emotion}}} e^{-\frac{(a-m_i)^2}{2\sigma^2_{\text{emotion}}}}$$  \hspace{1cm} (6)

, where $\sigma_{\text{emotion}}$ is mean of distance between each node, and $X_{\text{neutral}}$ is defined as the set of the feature vectors corresponding to neutral. The same is said of the other emotions.

Neutral ratio expressed in node $i$ is defined as follows:

$$\frac{L_{\text{neutral}}(i)}{L_{\text{neutral}}(i) + L_{\text{joy}}(i) + L_{\text{anger}}(i) + L_{\text{sadness}}(i)}$$  \hspace{1cm} (7)

2.5. Recognizing

The feature vector $x$ extracted from an input utterance is applied to the well-trained SOM. Complex emotion in the input utterance is recognized. We decide the node $c$, which is in the best conformity with a feature vector $x$ in the output layer. In this paper, node $c$ is defined as follows:

$$\|x - m_c\| = \min_i \|x - m_i\|$$  \hspace{1cm} (8)

Emotion in the input utterance is expressed on the node $c$.
3. Experiment environments and results

3.1. Emotional speech database

We have obtained an emotional speech database for training SOM and verifying proposed system. Speech data are recorded from 3 male students at a soundproof room. The uttered word is “okaasan”, means “mother” in Japanese, which each of the four emotions mingled with. The total number of utterance is shown in Tab.1.

Listeners, 3 male students, recognize the mingled emotions. The utterances are classified to one of four emotions. As they are familiar with the speakers, they know speakers’ habit in speech.

In this paper, the nodes in the output layer are organized in a planar configuration and consist of $10 \times 10$ nodes. As we considered a mean of whole distance between each node, $\sigma_{\text{emotion}}$ is decided to 0.02.

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Speaker A</th>
<th>Speaker B</th>
<th>Speaker C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number</td>
<td>220</td>
<td>232</td>
<td>314</td>
</tr>
</tbody>
</table>

Table 1: Emotional speech database

3.2. Experimental result

3.2.1. Complex emotion model

The trained SOMs are shown in Fig.3. All SOM of these figures are learned by the set of speaker A’s utterances. Simple emotion representation is labeled with traditional emotion model, and complex emotion representation is labeled with proposed emotion model. In simple emotion representation, each node is covered with one color as simple emotion. In complex emotion representation, each node is covered with several colors as proportion of four emotions.

The first row in Fig.3 indicates the representation of emotion relationships on speaker A. Though the result of feature map used to change by initial map, similar relations are obtained in several experimental results. Each listener has their own emotion recognition mode, so the relationship of each emotion has differences between them. In simple emotion representation, there are some isolated points of emotion representation. These points only show the emotion data of the nearest input vector. The database contains the word which is hard to classify and is consisted of a hint of emotions. SOM should not represent simple emotion. On the other hand, proposed system can show the varying of mingled emotions.

The second row and the third row in Fig.3 show the intensity of relationship between input vectors and node. Black node has weak connection to input vectors, and could be called as a dead node. This figure indicates that the proposed emotion model has more continual borders on the continual emotion than the traditional one. Because the border between two emotions is vague, the border is expressed faithfully by using proposed emotion model. Additionally, when we treat the stronger connected nodes as dead nodes, where dead node generates is different. In case of traditional emotion model, dead nodes are scattered. In case of proposed emotion model, dead nodes aren’t scattered, and unities of emotion are kept. Emotional combination become strong using proposed emotion model.
3.2.2. Individual differences

The feature map of trained SOM for the specific speaker is shown in Fig. 4. Each row figures are generated from the input data of speaker A, speaker B, speaker C. As SOM is trained by competitive learning, similar prosodic features are mapped closer. In these figures, the pattern of emotion map is different between speakers. For example, in the map of speaker A, neutral area is close to joy and sadness area, but is not close to anger area. These results indicate that there are the differences in prosodic features between speakers. Therefore, it is relevance for the emotion recognition system to specify each speaker.

<table>
<thead>
<tr>
<th>Listener1</th>
<th>Listener2</th>
<th>Listener3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speaker A</td>
<td>[Image]</td>
<td>[Image]</td>
</tr>
<tr>
<td>Speaker B</td>
<td>[Image]</td>
<td>[Image]</td>
</tr>
<tr>
<td>Speaker C</td>
<td>[Image]</td>
<td>[Image]</td>
</tr>
</tbody>
</table>

Figure 4: Differences among individuals

3.2.3. Accuracy of emotion recognition

The accuracy of emotion recognition is shown in Tab. 2. These SOMs are trained with each speaker, and are used to recognize the speaker’s emotion. The average accuracy is 53.6%. Comparing with the other studies [8], this accuracy seems reasonable.

Tab. 3 shows the advantage of a user specification system. When SOM specified for speaker A is applied to speaker B, the accuracy goes down. This result indicates that SOM reflects the characteristic of each speaker adequately.

<table>
<thead>
<tr>
<th>Data for training</th>
<th>Speaker A</th>
<th>Speaker B</th>
<th>Speaker C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutral</td>
<td>14.2%</td>
<td>14.3%</td>
<td>14.4%</td>
</tr>
<tr>
<td>Joy</td>
<td>57.1%</td>
<td>28.6%</td>
<td>28.6%</td>
</tr>
<tr>
<td>Anger</td>
<td>21.4%</td>
<td>46.3%</td>
<td>46.3%</td>
</tr>
<tr>
<td>Sadness</td>
<td>75.0%</td>
<td>6.3%</td>
<td>6.3%</td>
</tr>
<tr>
<td>All emotions</td>
<td>45.5%</td>
<td>29.5%</td>
<td>29.5%</td>
</tr>
</tbody>
</table>

Table 3: Recognition result (listener A model)

4. Conclusion

In this paper, we proposed complex emotion recognition system specified for a specific user in speech using SOM.

Proposed system treat a complex emotion model which is mingled with some basic emotions such as neutral, joy, anger, and sadness. As introduced this model, connections between each node of SOM is much stronger. And this system can express fuzziness emotions such as anger with sadness.

Furthermore, this paper shows the effectiveness for emotion recognition to specify the user.

5. References