Prosodic Parameters in Emotional Speech

Kazuhito KOIKE, Hirotaka SUZUKI, Hiroaki SAITO
Dept. of Computer Science
Keio University
{kazuhito, hiro, hxs}@nak.ics.keio.ac.jp

Abstract

Importance of speech prosody is on the increase as spontaneous interaction between human and machines is asked for. This paper examines how prosody contributes emotions to speech. Major elements which determine the emotion are pitch, tempo, and stress of speech. The last two elements correspond to duration and power of syllables, respectively. We choose five emotions to be tested: anger, surprise, sorrow, hate, and joy. To verify our analysis, we have implemented a speech synthesis module which can easily control prosodic parameters of output speech. Responses to the synthesized speech show that the parameters of anger, sorrow and hate are confirmed over 85%. Experimental results also suggest that surprise and joy feelings tend to depend less on its prosody.

2 Synthesis module

Our synthesis module is based on the pitch synchronous overlap add (PSOLA) method. PSOLA is characterized by using original waveforms in human speech as data. For generating speech sound, the waveform in data is added with the delay determined by the pitch. The formant frequency, which specifies the acoustic feature of speech, does not move very much.

There are two fundamental problems in this method. One is the increase of the number of data. It derives from the way PSOLA generates the speech output. The waveform is simply added as it is, thus when a new waveform is needed, a new datum is required. Another problem is deterioration of speech quality accompanied by the variance of the pitch. That is because two waveforms might be overlapped or a blank might appear between the two waves. An overlap of the waveforms happens because the length of the waveform unit does not match to the pitch cycle. To match the both length, the waveform has to be stretched or shrunked.

The waveform is in fact a discrete signal, and cannot be stretched nor shrunked. We introduce a control point concept to make it possible. The control points are set on the peaks and the bottoms of the waveform. They specify the feature of the waveform and have the

<table>
<thead>
<tr>
<th>Table 1: English accent vs. Japanese accent</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Stress</td>
</tr>
<tr>
<td>Pitch</td>
</tr>
<tr>
<td>Duration</td>
</tr>
</tbody>
</table>
relative time, not the absolute time. The waveform is
generated by interpolating the control points, and the
length can be changed.

Although the formant frequency moves if the wave-
form is stretched or shrunken, moving the control
points can suppress that. The control points are also
used for making intermediate waveforms. Since our
current unit of data is a phoneme, a large number of
data is not required.

Our synthesis module is implemented on the idea
above and is an easy-to-use tool*

3 Synthesizing an emotional phrase

We must describe prosody continuous in time because
the synthesized wave is continuous. Thus, when we
use the synthesis module, we must give it a pitch and
an amplitude at every moment. Frequency corre-
sponds to pitch and amplitude corresponds to stress
of the phonemes. It follows from what has been said
that to describe prosody continuous in time can be
regarded as to describe frequency and amplitude con-
tinuous in time. We use Fujisaki model [Fujisaki 84]
to describe continuous frequency.

3.1 Fujisaki model

Fujisaki model is based on an assumption that a con-
tour of a sentence is made up of two kinds of ingredi-
ants. One is a slowly varying component which may
or may not show a slight initial rise and then gradu-
ally decay toward an asymptotic baseline, but may be
resumed or reinforced at certain syntactic boundaries,
at least in the case of Japanese sentences. The other
is local humps (peaks or plateaus) closely correspond-
ing to the accent patterns of words constituting the
sentence. The humps may differ in their height.

The model produces a proportionate change in the
logarithm of frequency by adding these two elements.
In this model, the frequency is expressed by

\[
\ln F_o(t) = \ln F_{\text{min}} + \sum_{i=1}^{I} A_{p_i} G_{p_i}(t - T_{o_i}) + \sum_{j=1}^{J} A_{a_j} \{G_{a_j}(t - T_{1_j}) - G_{a_j}(t - T_{2_j})\} \tag{1}
\]

where

\[
G_{p_i}(t) = \begin{cases} 
\alpha_i^2 t \exp(-\alpha_i t) & (t \geq 0) \\
0 & (t < 0)
\end{cases}
\tag{2}
\]

and

\[
G_{a_j}(t) = \begin{cases} 
\min[1 - (1 + \beta_j t) \exp(-\beta_j t), \theta_j] & (t \geq 0) \\
0 & (t < 0)
\end{cases}
\tag{3}
\]

The symbols in Eqs. (1), (2), and (3) indicate

- \(F_{\text{min}}\): asymptotic value of fundamental frequency in
  the absence of accent components,
- \(I\): number of phrase commands,
- \(J\): number of accent commands,
- \(A_{p_i}\): magnitude of the ith phrase command,
- \(A_{a_j}\): amplitude of the jth accent command,
- \(T_{o_i}\): timing of the ith phrase command,
- \(T_{1_j}\): onset of the jth accent command,
- \(T_{2_j}\): end of the jth accent command,
- \(\alpha_i\): natural angular frequency of the phrase control
  mechanism to the ith phrase command,
- \(\beta_j\): natural angular frequency of the accent control
  mechanism to the jth accent command,
- \(\theta_j\): a parameter to indicate the ceiling level of the
  accent component (generally set equal to 0.9).

3.2 A model of amplitude

We propose a model of amplitude which represents
amplitude continuous in time. The model breaks
phonemes down into four categories of Japanese phonemes.

- **Vowels**
  
  We use the standard normal distribution to de-
  scribe the contour of vowels. We divide vowels
  into two types: vowels which follow a voiced sound
  and vowels which do not. The reason is that we
  can see a distinct difference between those two
  types of vowels.

  We define an amplitude of vowels which follow a
  voiced sound by Eq. (4).

  \[
y = 0.1 + \frac{1.0}{\sqrt{2.0\pi}} \exp\left(\frac{-(\frac{d_p}{2} - 1.0)^2}{2.0}\right) \tag{4}
\]

  where \(T\) is the duration of the vowel.

  We define an amplitude of vowels which do not
  follow a voiced sound by Eq. (5).

  \[
y = 0.1 + \frac{1.0}{\sqrt{2.0\pi}} \exp\left(\frac{-(\frac{d_p}{2} - 3.0)^2}{2.0}\right) \tag{5}
\]

  where \(T\) is the duration of the vowel.

- **Voiced consonants**

  We define an amplitude of voiced consonants us-
  ing the standard normal distribution of Eq. (6).

  \[
y = 0.1 + \frac{0.8}{\sqrt{2.0\pi}} \exp\left(\frac{-(\frac{d_p}{2} - 3.0)^2}{2.0}\right) \tag{6}
\]

* This module will be publicly available. Contact
  hxs6ics.kokoa.uc.jp for distribution information.
where $T$ is the duration of the voiced consonant.

- **Unvoiced consonants**
  We define an amplitude of unvoiced consonants as to take a value between 0.05 and 0.11 at random for each moment.

- **Plosives**
  We define an amplitude of plosives in almost the same way as an amplitude of unvoiced consonants which takes a value between 0.05 and 0.11 at random for each moment. In addition to this we define the sudden change at the beginning by Eq. (7).

$$a = \begin{cases} 
0.05 + 0.005t & (0 < t < 0.2T) \\
0.65 - 0.02t & (0.2T \leq t < 0.4T) 
\end{cases} \quad (7)$$

where $T$ is the duration of a plosive. We use this expression for the first 40% of the plosive, and we adopt the same way as what we do for unvoiced consonants for the rest of the time.

### 3.3 Setting parameters

Here we show how to set parameters to the synthesis module. Current implementation allows the user to input necessary parameter values interactively. The input process is as follows.

1. number of phrases
2. spell each phrase with a space between phrases
3. ratio of each phrase’s duration
4. ratio of silence duration between phrases
5. number of accent commands ($J$)
6. starting time of each accent command ($T_{1j}$)
7. ending time of each accent command ($T_{2j}$)
8. starting time of each phrase command ($T_{0i}$)
9. magnitude of each phrase command ($A_{p_i}$)
10. amplitude of each accent command ($A_{a_j}$)
11. amplitude of each syllable
12. fundamental frequency ($F_{\text{min}}$)

Although it may look difficult to specify the exact timing like $T_{1j}$, some heuristic values are known for synthesizing Japanese.

### 4 Experiments

We synthesized a phrase using our system mentioned in the previous section and had 20 people (male and female at the age in the twenties) listen to them. Five emotion patterns are attached to the phrase: joy, sorrow, anger, hate, and surprise. The synthesized phrase is a Japanese family name (four-syllable word “yamamoto”), so as not to give the listeners any hints of emotion the word implies. Names can also be spoken with various emotions.

#### 4.1 Conditions of experiments

**Experiment I**
First we synthesize a neutral phrase. Then we synthesize an emotional pattern and ask the listener which emotion he/she feels. This experiment is performed three times against randomly chosen three emotions.

**Experiment II**
First we synthesize a neutral pattern. Then we synthesize five emotional patterns (the order is random) and have the listener identify all five emotions.

In short, Experiment I requires listeners’ intuitive response. Experiment II, on the other hand, allows the listeners to examine the sound.

Table 2 shows the parameters set for each emotion (See Eq. (1) for the symbols in the leftmost column).

<table>
<thead>
<tr>
<th>syllable duration</th>
<th>Neutral</th>
<th>Anger</th>
<th>Hate</th>
<th>Sorr.</th>
<th>Surp.</th>
<th>Joy</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.75</td>
<td>1.0</td>
<td>0.3</td>
<td>0.8</td>
<td>0.8</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>0.7</td>
<td>1.0</td>
<td>0.3</td>
<td>0.8</td>
<td>0.8</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>0.5</td>
<td>1.0</td>
<td>0.3</td>
<td>0.8</td>
<td>0.8</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>0.2</td>
<td>1.0</td>
<td>0.3</td>
<td>0.8</td>
<td>0.8</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>0.75</td>
<td>0.7</td>
<td>0.7</td>
<td>0.5</td>
<td>0.75</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>0.7</td>
<td>1.0</td>
<td>2.5</td>
<td>0.75</td>
<td>0.2</td>
<td>1.5</td>
<td>1.0</td>
</tr>
<tr>
<td>0.7</td>
<td>0.7</td>
<td>2.0</td>
<td>1.0</td>
<td>0.88</td>
<td>1.3</td>
<td>1.0</td>
</tr>
<tr>
<td>0.7</td>
<td>1.3</td>
<td>5.0</td>
<td>3.0</td>
<td>0.88</td>
<td>3.0</td>
<td>2.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>syllable amplitude</th>
<th>Neutral</th>
<th>Anger</th>
<th>Hate</th>
<th>Sorr.</th>
<th>Surp.</th>
<th>Joy</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.43</td>
<td>0.43</td>
<td>0.43</td>
<td>0.43</td>
<td>0.43</td>
<td>0.43</td>
<td>0.43</td>
</tr>
<tr>
<td>0.43</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>0.43</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>0.43</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>0.43</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

| $A_{p_i}$ | 0.43 | 0.43 | 0.43 | 0.43 | 0.43 |
| $J$       | 1.0  | 1.0  | 1.0  | 1.0  | 1.0  |

| $A_{a_j}$ | 0.26 | 0.4  | 0.4  | 0.4  | 0.4  |
| $F_{\text{min}}$ | 120 | 100  | 100  | 100  | 135  |

($\alpha_i = 0.003, \beta_i = 0.02, \theta_j = 0.9$)

Although it may look difficult to specify the exact timing like $T_{1j}$, some heuristic values are known for synthesizing Japanese.
4.2 Results

Confusion matrices of Table 3 and 4 show the recognition result. The vertical axis represents correct feelings and the horizontal axis does responded ones.

<table>
<thead>
<tr>
<th>(%)</th>
<th>Anger</th>
<th>Hate</th>
<th>Sorr.</th>
<th>Surp.</th>
<th>Joy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td>75.0</td>
<td>0.0</td>
<td>0.0</td>
<td>16.7</td>
<td>8.3</td>
</tr>
<tr>
<td>Hate</td>
<td>0.0</td>
<td>85.0</td>
<td>0.0</td>
<td>10.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Sorrow</td>
<td>0.0</td>
<td>0.0</td>
<td>100.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Surprise</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>50.0</td>
<td>50.0</td>
</tr>
<tr>
<td>Joy</td>
<td>16.7</td>
<td>0.0</td>
<td>0.0</td>
<td>25.0</td>
<td>38.3</td>
</tr>
</tbody>
</table>

We got high accuracy rate for anger and sorrow both in Experiment I and II, which indicates validness of the parameter values.

In Experiment I, most errors of hate go to surprise and most errors of surprise go to joy. This can be explained as both hate and surprise have a rising pitch pattern at the end of the phrase and both surprise and joy have a high pitch. Remember that comparison only against the neutral emotion is performed in Experiment I, which could lead errors among similar patterns. In Experiment II, where five emotions are relatively compared and the sound characteristics get clear, correct response for anger and hate rises dramatically.

As for hate, attached emotional characteristics need to be refined, because the recognition rate is not as good as anger or sorrow although it also sharply rises in Experiment II.

Unfortunately, recognition rates of surprise and joy are not good both in Experiment I and II. The parameter might not be specific enough for these two emotions. Or from the fact that similar misrecognition is found even in real voice spoken by the humans, it might be difficult to identify such emotion only from its prosody. In Experiment II, we received many responses saying that they can hardly determine which prosody pattern represents joy or surprise after choosing those two emotions out of five. They can tell the difference of the two, but they cannot identify the emotion.

5 Conclusions and Future Work

This paper examined how prosody contributes emotions to speech. We verified that setting the pitch pattern, the amplitude pattern, and duration of each syllable can express a particular emotion. It is safe to say we can find a distinctive parameter value for anger and sorrow.

In synthesizing speech, we need to describe frequency and amplitude continuous in time. The former uses Fujisaki model, and the latter does a new model of amplitude. Our synthesis module uses an improved PSOLA method.

Since the experiment was done only against Japanese phrases, we need to test other languages and against sentences. Preliminary experiments in French show that emotion patterns are somewhat language-dependent.

Acknowledgements

Masataka Suzuki proposed our PSOLA method and implemented our synthesis module.

References


