Improving the Naturalness of Synthetic Speech by Utilizing the Prosody of Natural Speech

Toshimitsu Minowa, Ryo Mochizuki and Hirofumi Nishimura
Multimedia Solution Laboratories, Matsushita Communication Industrial Co., Ltd.

ABSTRACT

The quality of synthetic speech is greatly improved if a prosody of natural speech is adopted instead of a rule-based prosody. In order to apply this effect to an arbitrary word synthesis, the authors propose a new prosody control method. According to the result of a listening test, it was shown that rhythm could be independently controlled from pitch and power whereas pitch and power should be dependently controlled. Therefore, it seems that pitch and power control method should be derived from the same speech. However, in an embedded types of practical arbitrary word synthesis, the amount of memory is so limited that there is little room for redundant data. So, authors systematically derived the prosody (pitch interval and pitch waveform amplitude) of continuously uttered mono syllable speech which cover most Japanese accent types. Syllables are chosen in accordance with the categories of Japanese consonants which are classified by manner and place of articulation. By this method, the naturalness of the synthetic speech achieved almost the same preference score with the one that copied the prosody of natural speech.

1. INTRODUCTION

The quality of synthetic speech has been improved due to several excellent ideas such as pitch synchronous overlap add technique, template based prosody and so on, so that speech synthesis has become to be expected to convey not only linguistic information but also paralinguistic information such as emotion[1][2]. The authors have tried to improve naturalness of synthetic speech[3]. The naturalness here means a impression which the listener has if the listener perceives some "speaker's atmosphere", that is intention, speaking style or emotion, from the synthetic speech as shown in Figure 1. So, this "naturalness" does not seem to be generated simply by perturbation of pitch interval nor perturbation of pitch waveform amplitude. It seems to be generated by a dynamical feature of prosody. However authors are not trying to do emotion synthesis but trying to reproduce the speaker's atmosphere whose speech data is used to build the speech synthesizer. According to Miyatake's categorization[4], this atmosphere can be categorized as "tone" rather than "speaking style". Hereafter the terminology, "tone", is used in this sense. It is not easy to reproduce the tone by rules. On the other hand, it is well known that the tone can be reproduced by copying the prosody of natural speech[5]. But this approach requires a huge amount of memory. Therefore this approach is not regarded as suitable for embedded applications. However, in case of Japanese word synthesis, the time course of pitch frequency can be determined if number of morae and the place of accent is specified.

Since it can be common for all words that belong to the same accent type, even a nonsense word has the same pitch pattern. So, if we assume that there is a representative syllable in each Japanese accent category, and duration and amplitude can be represented by those of such a syllable, target prosody can be easily generated by concatenating such representative syllable's prosody that is cut out from a word in which the syllable is continuously uttered.

![Figure 1. Generation and perception of a tone.](image)

2. A VECTOR OF PROSODY

In natural speech, pitch frequency, pitch waveform amplitude and rhythm are interrelated so that some tone is formed. Therefore, it seems that they should be controlled dependently in speech synthesis. The authors propose a new method to control the prosody using vectors of these components. First, listening test was executed to analyze the importance of combining these components. Then, some characteristics of the prosody vector are examined. The result leads to a new method of generating the prosody for word synthesis.
2.1 Cues to perceive a tone

In order to study the cues to perceive a tone, a listening test was executed. It was a preference test using synthetic speech generated by various combinations of prosody as shown in Table 1. The number of subjects was 9. They listened pairs of synthetic speech using headphones. The sentence was “Bideo kinoo wa arimasenn (There is no video functionality).” The subjects judged which is more likely to be uttered by human being with some tone. The result of the experiment was shown in Figure 2.

<table>
<thead>
<tr>
<th>stimulus</th>
<th>NP</th>
<th>FA</th>
<th>RPi</th>
<th>RPiPw</th>
<th>RT</th>
</tr>
</thead>
<tbody>
<tr>
<td>pitch</td>
<td>org</td>
<td>org</td>
<td>msy</td>
<td>msy</td>
<td>org</td>
</tr>
<tr>
<td>amplitude</td>
<td>org</td>
<td>flt</td>
<td>org</td>
<td>msy</td>
<td>org</td>
</tr>
<tr>
<td>rhythm</td>
<td>org</td>
<td>org</td>
<td>org</td>
<td>iso</td>
<td></td>
</tr>
</tbody>
</table>

where
org: Prosody derived from the original natural speech.
flt: Constant amplitude.
msy: Prosody derived from the continuously uttered monosyllable (/no/) speech in which the speaker mimicked the original speech.
iso: Isochronous rhythm.

Therefore it can be concluded that cues for perceiving a tone lies in a combination of pitch and rhythm or a combination of pitch and amplitude. From the view point of developing a small foot print speech synthesizer, neither NP nor FA is realistic because they require unlimited amount of memory. Therefore, we should focus on RPiPw, namely pitch and amplitude should be derived from the same speech whereas rhythm does not have to be derived from the speech. Since it was also made clear that subjects were very sensible to rhythm, the rhythm control method should be as elaborate as possible. From these results, the vector of prosody was defined as follows.

\[ V(N) = \left( \frac{F_0(N)}{F_{0}\text{max}}, \frac{A(N)}{A_{\text{max}}} \right) \]  

where

\[ F_0(N) \]: Inverse of pitch interval of Nth pitch waveform.
The pitch interval is defined as the interval between pitch driving points [3].
\[ F_{0}\text{max} \]: Maximum \( F_0(N) \) in the vowel.
\[ A(N) \]: Maximum amplitude of Nth pitch waveform.
\[ A_{\text{max}} \]: Maximum \( A(N) \) in the vowel.

Figure 2. Preference score of test materials which have various combinations of natural and rule based prosody.

From Figure 2, it was made clear that:
(1) In these materials, NP was the bast of all.
(2) If pitch and rhythm are derived from the original natural speech, amplitude can be even flat (FA).
(3) If rhythm is derived from the original natural speech, pitch and amplitude can be derived from other utterance (RPiPw).
From these figures, it is clear that none of them are similar so that we should have V(N) for each vowel.

2.2 The effect of position of syllable

As shown in Figure 3 - Figure 7, there seems few similar vectors. In order to confirm this, correlation coefficients between components of V(N) in each syllable were calculated for /nononono'nono/ and /dadadada'dada/. A female narrator uttered these nonsense words for 3 times. Then mean correlation coefficients were calculated. The result is shown in Table 2. As is expected, except for the first and second vowel of /nononono'nono/, there is no similar vectors. So, it seems that we should have the vector in every syllable position.

<table>
<thead>
<tr>
<th>Vowel</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
<th>5th</th>
<th>6th</th>
</tr>
</thead>
<tbody>
<tr>
<td>/nono</td>
<td>-0.71</td>
<td>-0.76</td>
<td>0.45</td>
<td>0.92</td>
<td>0.86</td>
<td>-0.01</td>
</tr>
<tr>
<td>/dada</td>
<td>0.75</td>
<td>0.18</td>
<td>0.45</td>
<td>0.51</td>
<td>0.48</td>
<td>0.00</td>
</tr>
</tbody>
</table>

2.3 The effect of difference of vowels

In order to investigate the effect of difference of vowels, the same calculation was executed for /nuninono'nena/ and /nanunene'noni/. The result is shown in Table 3. Compared to Table 2, it was made clear that these are close to those of /nononono'nono/, so it can be concluded that vowel difference has little effect on prosody vectors.

<table>
<thead>
<tr>
<th>Vowel</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
<th>5th</th>
<th>6th</th>
</tr>
</thead>
<tbody>
<tr>
<td>/nanu</td>
<td>-0.84</td>
<td>-0.84</td>
<td>0.31</td>
<td>0.69</td>
<td>0.90</td>
<td>0.04</td>
</tr>
<tr>
<td>/nuni</td>
<td>-0.81</td>
<td>-0.60</td>
<td>0.74</td>
<td>0.34</td>
<td>0.82</td>
<td>-0.10</td>
</tr>
</tbody>
</table>

2.4 The effect of difference of consonants

In order to investigate the effect of difference of consonants, the same calculation was executed for /okorosorotonoho/. The result is shown in Table 4.

<table>
<thead>
<tr>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
<th>5th</th>
<th>6th</th>
<th>7th</th>
<th>8th</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.57</td>
<td>0.71</td>
<td>0.94</td>
<td>-0.26</td>
<td>0.77</td>
<td>-0.45</td>
<td>0.96</td>
<td>-0.32</td>
</tr>
</tbody>
</table>

From Table 4, it was made clear that unvoiced consonants have different effects from voiced consonants. In order to evaluate the difference quantitatively, statistical test was executed using the T value defined by expression (2-1).

\[
T = \sqrt{\frac{z_1^2}{M-3} + \frac{z_2^2}{N-3}} (2-1)
\]
where \( z_i \) is a z-transform of correlation coefficient \( \rho 
\]

\[
\frac{I}{2} \log \left( \frac{I + i}{I - i} \right)
\]

\( i \)

(2-2)

Table 5 shows the result between 3rd and other vowels in /okorosorotonoho/. |T| of 1.96 means significance level at 5%.

Table 5. T value defined by expression (2-1) between 3rd vowel /(r)o/ and other vowels.

<table>
<thead>
<tr>
<th>1st /o/</th>
<th>2nd /(k)o/</th>
<th>4th /(s)o/</th>
<th>5th /(r)o/</th>
<th>6th /(t)o/</th>
<th>7th /(n)o/</th>
<th>8th /(h)o/</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.37</td>
<td>-1.07</td>
<td>-2.49</td>
<td>-0.93</td>
<td>-2.76</td>
<td>0.18</td>
<td>-2.57</td>
</tr>
</tbody>
</table>

3. Application to arbitrary word synthesis

Based on the results in section 2, the authors propose a new prosody control method for word synthesis. While the position and the preceding consonants have effect on V(N), the difference of vowel has little effect on V(N). So, in order to build reference V(N) database, representative syllables are chosen as in Table 6, and nonsense words which consist of continuous chain of one of these syllables were recorded. These nonsense words cover all the Japanese accent types up to 9 morae. Vowel was fixed on /o/.

When a target word is input to the synthesizer, each syllable is mapped to the reference V(N) database, then the corresponding representative syllables’ V(N) are concatenated to form the prosody of the target word. For example, prosody of [nodorodo'oto] is generated for synthesizing /midoriga'oka/.

![Figure 8](image)

4. Evaluation

In order to evaluate the quality of synthetic speech, preference test was executed on conventional method, proposed method and natural prosody. The subjects listened to pairs of these synthetic speech and judged which is more likely to be uttered by human being with some tone. The number of subjects was 5. Sampling frequency was 22.05kHz. The result is shown in Figure 8. It is shown that proposed method achieved almost the same score with natural prosody.

Cp: Natural Prosody
Cn: proposed method
Sn: Authors’ conventional Method

Figure 8. Preference score for natural prosody, proposed method and conventional method. Test materials are /midoriga'oka/, /nagarade'Nki/, /haQkeejimapa'radaisu/ and /seQkeezumipo'rideNto/.

5. CONCLUSION

A new prosody control method is proposed. It utilizes the concatenation of syllable prosody of systematically recorded word speech. By this method, the quality of synthetic speech has become almost the same as one that copied the prosody of natural speech.

REFERENCES


