SPEECH MORPHING BY PROGRESSIVE INTERPOLATION OF SPECTRA

Hideki BANNO\textsuperscript{1}, Shouji KAJITA\textsuperscript{1}, Kazuya TAKEDA\textsuperscript{1}, Kiyohiro SHIKANO\textsuperscript{1} and Fumitada ITAKURA\textsuperscript{2}

\textsuperscript{1} Nara Institute of Science and Technology
Takayama 8916-5, Ikoma, Nara 630-01 JAPAN
\{hideki-b,shikano\}@is.aist-nara.ac.jp
\textsuperscript{2} Graduate School of Engineering, Nagoya University
Fro-cho 1, Chikusa-ku, Nagoya 464-01 JAPAN
\{kaji, takeda, ita\}@nuee.nagoya-u.ac.jp

ABSTRACT

A speech morphing, continuous alternation of speech waveforms of different two speakers, algorithm based on progressive interpolation of spectral envelope and source signal is proposed. The basic scheme of the morphing is 1) find correspondence of unit waveforms of original and target speech. 2) separate speech spectra to envelope and source excitation components. 3) find correspondence of spectral channels of original and target speech for each envelope. 4) interpolate both source signal and envelope. and 5) construct unit waveform and generates morphing speech by PSOLA.

In the objective test, proposed method can reduce the spectral distortion by 1.9 dB compared to the method based on progressive substitution of spectra, when it is used for interpolating two vowels in real speech. The effectiveness of the method is also confirmed by subjective test in which more than 90\% (male to female morphing) or 60\% (female to male morphing) subjects preferred the proposed method.

1. INTRODUCTION

Morphing is a common technology in the field of computer graphics to alter an image of an object to another object continuously. The same kind of processing of speech, speech morphing, has been introduced as continuous alternation of voice individuality from speaker to speaker \cite{1,2,3,4}. Recently, Abe et al. proposed a speech morphing method based on progressive substitution of spectra (PSS) \cite{5}. In the method, higher frequency components than the give frequency boundary is substituted by target speaker’s speech with continuously changing the boundary frequency as time goes on.

In order to realize speech morphing, we have to develop a parametric interpolation of speech characteristics which can generate the intermediate speech which has 1) as smooth changing as possible in time domain and. 2) as natural shape as possible in frequency domain. From these viewpoints, the simple substitution of spectra in PSS method seems to be too simple to generate smooth and natural spectra.

In this paper, we propose a new algorithm of speech morphing which 1) interpolates spectral envelope and source excitation not only independently but also pitch synchronously and 2) associates original and target unit waveforms in both time and frequency domains by means of DTW. The resultant performance of both objective and subjective tests shows that the proposed method can generate better morphing speech than the PSS method.

2. ALGORITHM

The detailed algorithm of the proposed morphing method is described in this section.

First, speech signals of speaker A and B, or to say original and target speaker, for the same utterance are recorded. For each speech signals, time marks are given to the peak amplitude location of an unit waveform corresponding to a pitch period \cite{6}. A pitch period waveform, which we call pitch waveform, is extracted by applying window centered at the pitch mark location and is used for the basic unit of speech morphing. As for unvoiced segment, the same kind of waveform unit is also extracted, whereas the fixed interval windowing is adopted.
Figure 1: Extracting unit pitch waveform by asymmetry variable length window.

2.1 Time warping

To find time domain correspondence between two signals, dynamic time warping is executed by matching two sequences of pitch waveforms of speaker A and B, using the LPC cepstrum distance as the spectral metric. After time warping, the speech signals are associated pitch-waveform-by-pitch-waveform basis. Therefore, interpolating two corresponding pitch waveforms of original and target speech signals will generate a pitch waveform of morphing speech.

Morphing rate, i.e., rate of voice individuality of the target speaker, at each waveform is also decided taking the morphing process period into account as a time domain processing in this stage.

Since interpolation of pitch period is also needed in speech morphing, after finding waveform association, pitch waveform to be processed is re-extracted adaptively to original and target pitch period length. As shown in Figure 1, the lengths of left and right tapered period are adjusted to the shorter period of original and target periods, asymmetrically.

2.2 Separating spectral envelope and source excitation

Separating spectral envelope and source excitation of each pitch waveform is done by liftering FFT cepstrum coefficients. FFT cepstrum coefficients, \( c(n) \), is calculated by

\[
X(k) = \sum_{t=0}^{N-1} x(t) e^{-j \frac{2\pi}{N} kt} \quad (0 \leq k \leq N - 1) \tag{1}
\]

\[
c(n) = \frac{1}{N} \sum_{k=0}^{N-1} \log |X(k)| e^{j \frac{2\pi}{N} kn} \quad (0 \leq n \leq N - 1) \tag{2}
\]

Lower and higher cepstrum coefficients are used as cepstrum representation of spectral envelope, \( e(n) \), and source excitation \( s(n) \), respectively.

\[
e(n) = \begin{cases} c(n) & (0 \leq n < n_\alpha, N - n_\alpha \leq n \leq N - 1) \\ 0 & (n_\alpha \leq n < N - n_\alpha) \end{cases} \tag{3}
\]

\[
s(n) = \begin{cases} 0 & (0 \leq n < n_\alpha, N - n_\alpha \leq n \leq N - 1) \\ c(n) & (n_\alpha \leq n < N - n_\alpha) \end{cases} \tag{4}
\]

2.3 Interpolating spectral envelope

Interpolation of spectral envelopes of two pitch waveforms is done in the log-spectral representation of \( e(n) \), i.e.

\[
E(k) = \sum_{n=0}^{N-1} e(n) e^{-j \frac{2\pi}{N} kn} \quad (0 \leq k \leq N - 1) \tag{5}
\]

Before the interpolation, frequency axis is warped so as to associate the formant frequencies of each speaker. The result of warping is obtained as a mapping defined by
where \( k_A \) and \( k_B \) are the indexes of frequency channel of each speaker. This mapping describes that the frequency \( k \) of speaker A is associated with \( \theta(k) \) of speaker B.

Morphing of spectral envelope can then be realized as the parametric interpolation of two spectral envelopes \( E_A(k) \) and \( E_B(k) \) as follows.

\[
E_M(k) = (1 - r)E_A(k) + rE_B(\theta(k)) \quad (0 \leq k \leq N - 1)
\]

where \( k \) is also interpolated version of frequency defined by frequency mapping function \( \theta(k) \) and morphing rate \( r \)

\[
k = (1 - r)k_A + r\theta(k_A) \quad (0 \leq k_A \leq N - 1).
\]

Finally, the spectral envelope of the morphing speech is reconverted into cepstral representation by

\[
e_M(n) = \frac{1}{N} \sum_{k=0}^{N-1} E_M(k)e^{-j2\pi \frac{k}{N} n} \quad (0 \leq n \leq N - 1).
\]

### 2.4 Interpolating source excitation

The source excitation signal of morphing speech is also calculated in terms of morphing rate and source excitation components of speaker A and B as follows.

\[
s_M(n) = (1 - r)s_A(n) + rs_B(n) \quad (0 \leq n \leq N - 1)
\]

### 2.5 Reconstructing pitch waveform

The cepstrum representation of morphing speech is obtained by superimposing envelope and excitation components of morphing speech by the following equation. (In general, each of \( e_M(n) \) and \( s_M(n) \) consists of lower and higher cepstral coefficients only.)

\[
c_M(n) = e_M(n) + s_M(n) \quad (0 \leq n \leq N - 1).
\]

Then, amplitude spectrum of the morphing speech is obtained by taking FFT of \( c_M(n) \).

\[
|X_M(k)| = \exp \left( \sum_{n=0}^{N-1} c_M(n)e^{-j2\pi \frac{k}{N} n} \right) \quad (0 \leq k \leq N - 1).
\]

On the other hands, the phase of the morphing speech is generated by concatenating the phase spectrum of lower and higher frequency of original and target speakers using

\[
k_d = (1 - r)\frac{N}{2}
\]

as the boundary frequency. A unit pitch waveform of the morphing speech, \( r_M(t) \), can then be obtained by

\[
X_M(k) = \begin{cases} 
\frac{|X_M(k)| |X_A(k)|}{|X_A(k)|} & (0 \leq k < k_d, N - k_d \leq k \leq N - 1) \\
\frac{|X_M(k)| |X_A(k)|}{|X_B(k)|} & (k_d \leq k < N - k_d)
\end{cases}
\]

Whole sentence of morphing speech is, finally, generated by PSOLA[7].

### 3. EVALUATION TEST

In this section, the proposed method is compared with conventional PSS method from both objective and subjective standpoints.

#### 3.1 Illustration of spectral envelopes

First, the spectral envelopes of morphing speech by PSS and proposed method are illustrated for comparison. As for the speech material, a sentential speech (／arayuru geNjitsu o subete jibuN no hore nejima1t\text{-}toda／) of a male (speaker A) and female (speaker B) speakers are used and morphing spectral envelope is calculated for the pitch waveform of ／arayuru／ of 50 % morphing rate. Figure 3 shows the resultant spectra of the proposed method (top) and the PSS method (bottom).

From the figure, it is clear that there is discontinuity between lower and higher frequency regions in the morphing speech of PSS method. This, obviously, is the consequence of the simple replacement of spectral information in PSS method. On the other hand, since the morphing speech is obtained by moving the original formant frequency there is no discontinuity, in the proposed method. The above results clarify that the proposed method can generate more natural morphing speech than PSS.
Table 2: Analysis conditions for measuring spectral distortion.

<table>
<thead>
<tr>
<th>sampling frequency</th>
<th>8kHz</th>
</tr>
</thead>
<tbody>
<tr>
<td>window</td>
<td>Asymmetric Hamming</td>
</tr>
<tr>
<td>window length</td>
<td>~32ms (256point)</td>
</tr>
<tr>
<td>LPC order</td>
<td>12</td>
</tr>
<tr>
<td>Cepstrum order</td>
<td>16</td>
</tr>
</tbody>
</table>

Table 3: Spectral distortion.

<table>
<thead>
<tr>
<th></th>
<th>Proposed Method</th>
<th>PSS Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2.7 dB</td>
<td>4.6 dB</td>
</tr>
</tbody>
</table>

\[ CD = \left( \frac{10}{\ln 10} \right) \sqrt{2 \sum_{n=1}^{\rho} (c_o(n) - c_m(n))^2} \quad (15) \]

Figure 4 shows the spectrogram of (a) coarticulatory transition in natural speech, (b) morphing speech of the proposed method and (c) morphing speech of PSS method, respectively. From the figure, the smooth change in the proposed method can be confirmed.

The measured cepstral distortion is listed in Table 3. As listed in the table, the resultant distortion is smaller in the proposed method than PSS method by about 2.0 dB. This means that the proposed method can generate more natural or to say more similar to the natural coarticulation, morphing speech.

3.2 Objective evaluation by spectral distortion

Objective test of the naturalness of morphing speech is performed by comparing with natural coarticulation between phonemes as follows.

1. Extract pitch waveforms of the initial and the final pitch periods of a natural speech segment of coarticulatory change between phonemes. The transitional portion from /u/ to /o/ of /genjitsuwa/ was used.

2. Generate morphing speech through interpolating the above initial and the final waveforms by continuously changing the morphing rate.

3. Associate each pitch waveform of natural coarticulation and morphing speech.

4. Calculate the below defined cepstrum distances between associated two waveforms

\[ CD = \left( \frac{10}{\ln 10} \right) \sqrt{2 \sum_{n=1}^{\rho} (c_o(n) - c_m(n))^2} \quad (15) \]

Figure 3: Spectral envelopes of original, target and morphing speech. Top: Proposed method. Bottom: PSS method.

3.3 Subjective test

In subjective test, 11 subjects listened to morphing speech of the proposed and PSS methods, to decide which sounds better. The results are listed in Table 4. In both male to female and female to male morphing, the proposed method is preferred to the conventional PSS method. This is due to that there is mismatch in PSS method between lower and higher spectral envelope because of the simple replacement of spectral component.

Table 4: Results of subjective test. Preference score of the proposed method to the PSS method.

<table>
<thead>
<tr>
<th></th>
<th>Preference score</th>
</tr>
</thead>
<tbody>
<tr>
<td>male→female</td>
<td>91 %</td>
</tr>
<tr>
<td>female→male</td>
<td>64 %</td>
</tr>
</tbody>
</table>
4. SUMMARY

In this paper, the speech morphing, continuous alternating of speech signal is developed based on independent manipulation of envelope and excitation components. In the proposed method, both time and frequency warpings are also introduced to associate original and target signals. Thus, the mismatch between original and target speech characteristics, which causes the degradation of the resultant morphing speech, can be reduced from the simple replacement of the spectral components. The effectiveness of the proposed method is confirmed from both subjective and objective tests.

ACKNOWLEDGEMENT

The authors are grateful to Dr. Abe for his comment on this work. The authors are also grateful to Dr. Kawai for pitch marking algorithm and the pitch marked speech materials.

REFERENCES