Auditory Filterbank Improves Voice Morphing

Erika Okamoto†, Toshio Irino†, Ryuichi Nisimura† and Hideki Kawahara†

†Faculty of Systems Engineering, Wakayama University, Japan
{s115016, irino, nisimura, kawahara}@sys.wakayama-u.ac.jp

Abstract

This paper presents a new method for vocal tract length (VTL) estimation and normalization based on a gammachirp auditory filterbank (GCFB) to improve the sound quality in voice morphing. VTL ratios between 28 speakers were estimated based on the spectral distances for all permutations (756 = 28 × 27). The VTL estimation using the mel-frequency filterbank (MFFB), which is a preprocessor for calculating MFCCs commonly used in ASR, was also evaluated for comparison. The results of subjective listening tests of morphed voice sounds with and without VTL normalization are also reported. The objective and subjective results indicate that VTL normalization is essential for voice morphing, and the proposed GCFB-based method outperforms the MFFB-based method.

Index Terms: speech synthesis, voice morphing, auditory filterbank, vocal tract length, VTLN

1. Introduction

Speech sounds convey information about the linguistic contents produced by the shape and the dynamics of the vocal tract and the information about nonlinguistic speaker characteristics related to the vocal tract length (VTL) and the glottal pulse excitation. Humans can easily identify the speaker group as male, female, or as a child from the speech sounds, even with a single word or a mono-syllable. This suggests that the auditory system extracts the VTL information separately from the vocal tract shape information. VTL normalization was modeled as the stabilized wavelet-Mellin transform, which is a successive process that is started with an auditory filterbank [1].

The idea of VTL estimation and normalization has been widely used in speech applications, particularly in automatic speech recognition. In this paper, we propose a method for VTL estimation based on a gammachirp auditory filterbank (GCFB) [2] to improve the sound quality of voice morphing between two speakers [3].

Voice morphing methods are now widely used as research tools for speech perception and singing voice manipulations. A reduced version of morphing, which only requires that temporal reference points be assigned to align two exemplar speech materials, was also proposed to alleviate the time-consuming manual assignment of time-frequency reference points [4]. It was, however, found to be successful only when the two speech samples are within the same gender and age group. The sound quality was largely degraded when the morphing speech samples came from different groups.

The degradation is mainly caused by the difference of the VTLs of the speakers in different groups. In other words, vocal tracts with the same length (i.e., VTL normalization) may improve the quality of the morphed sounds. This idea, which was tested by minimizing the spectral distances between the morphing pairs using the linear transformation of the frequency axes, was found effective [4] when a mel-frequency filterbank (MFFB), which is a preprocessor for calculating MFCCs commonly used in ASR, was used to derive the spectral representation. The question here is whether the MFFB-based method outperforms a method based on more appropriate auditory spectral representations derived by GCFB.

In this paper, the objective estimation accuracy of the VTL ratio was compared between the GCFB and MFFB-based methods as an extension from a previous report [5]. The results of subjective tests to evaluate naturalness are also reported.

2. Vocal Tract Length Estimation

The VTL ratios are estimated from the spectral distance in auditory representations based on GCFB and MFFB. Then regression analysis is applied to get the final estimation, which is used for the voice morphing.

2.1. Auditory filterbank

We used linear and nonlinear versions of GCFBs to introduce the auditory spectral representations for the VTL estimation as:

\[ \text{dcGCFB:} \quad \text{The nonlinear, dynamic compressive GCFB simulates the cochlear characteristics including masking, compression, and two-tone suppression, which are measured by psychophysical experiments [2].} \]

\[ \text{linGCFB: A linear and fast version of GCFB with fixed filter coefficients.} \]

The input speech sound is analyzed into two-dimensional cochlear spectrograms by dcGCFB and linGCFB. The number of channels was 100, and the center frequencies of the filters were equally spaced on the ERBN-number axis between 100 and 6000 Hz. An equal-loudness contour (ELC) filter was also applied to simulate the sensitivity. Sampling rate \( f_s \) was 48 kHz, as in the original design of GCFB. The power of the filterbank outputs was summarized every 5 ms with a 25-ms hammering window to reduce the periodic components that are dependent on the fundamental frequency (F0).

For a detailed comparison, we also calculated the power spectrograms from the MFFBs using a matlab package for MFCC/RASTA [6]. Several kinds of MFFBs were used as an extension from the preliminary report [5], in which only one kind of MFFB was used for the comparison. The mel-frequency function was set to “h mktime”; the pre-emphasis filter was applied to compensate the spectral tilt; a 25-ms hammering window with a frame rate of 5 ms was used for the short-time Fourier transform (STFT). We made four types of MFFB power spectrograms:

\[ \text{MFFB}_{\text{STFT}40}: \quad 40 \text{ channels between 0 and 8000 Hz using the STFT power spectrum} \]
\[ \text{MFFB}_{\text{STFT}120}: \quad 120 \text{ channels using the STFT power spectrum} \]
VTL ratio estimation based on spectral distance

The speech samples of speakers A and B with the same sentence were analyzed by GCCFB or MFFB to derive smoothed spectrograms \( P_A(f, t) \) and \( P_B(f, t) \), where \( f \) is either ERB-frequency \( f_{EB} \) in GCCFB or mel-frequency \( f_{mel} \) in MFFB. Since the phoneme locations in the two spectrograms are different, we deformed the time axis of spectrogram B to align the phoneme boundaries for B with those for A. Deformed spectrogram B is denoted as \( P_{Bn}(f, t) \). To estimate VTL ratio \( r \) between A and B, \( P_{Bn}(f, t) \) is dilated along the frequency axis as \( P_B(r_f, t) \). The spectral distance in the dB scale at time \( t \) is defined as the root mean squared (rms) difference as

\[
D_{dB}(t, r) = \sqrt{\frac{1}{r_H - r_L} \int_{r_L}^{r_H} D_P^2(f, t) df}, \tag{1}
\]

where

\[
D_P = \int_{r_L}^{r_H} \left( \frac{\log_{10} P_A(f, t)}{P_A(t)} - \log_{10} \frac{P_B(r_f, t)}{P_{Bn}(t)} \right)^2 df,
\]

where \( f_L \) and \( f_H \) are the warped versions of the lower and higher limits of the frequency region \( (f_L, f_H) \) and \( P_A(t) \) and \( P_{Bn}(t) \) are the values averaged across the frequencies.

The objective for the VTL estimation is to find the best VTL ratio \( r \) by minimizing distance \( D_{dB}^{\text{total}} \):

\[
r = \arg\min_r D_{dB}^{\text{total}}(r), \tag{2}
\]

where total distance \( D_{dB}^{\text{total}}(r) \) is defined using frame-wise spectral distance \( D_{dB}(t, r) \) in Eq. 1:

\[
D_{dB}^{\text{total}}(r) = \frac{1}{\|V\|} \int_{f \in V} D_{dB}^2(t, r) dt, \tag{3}
\]

where \( V \) represents the set of voiced segments.

2.3. Accuracy measure for VTL estimation

We must provide a measure to represent the accuracy of the VTL estimation. In this report, the distance between the individual VTL ratios estimated using Eq. 2 and the VTL ratios calculated from the whole set of the individual VTL ratios is used as the accuracy measure.

2.3.1. VTL estimation from individual VTL ratios

Three Japanese sentences spoken by 14 male and 14 female speakers were used for the VTL estimation. The speech samples were excerpted from a speech database with simultaneous EGG recording [8]. The sentences consisted of 10, 14, and 20 syllables. The VTL ratios were calculated between two speakers for every pair of speech samples using Eq. 2. We also considered the order (cf. [5]) since \( r \) is applied to one side in Eq. 1. The total number of permutations was 756 (\( \approx 28 P_{27} \)).

When \( m \)-th and \( n \)-th speaker’s VTLs are \( l_m \) and \( l_n \), the VTL ratio is defined as \( r_{m, n} = l_m / l_n \). By introducing logarithmic conversion, this relation is linearized as

\[
\log(r_{m, n}) = \log(l_m) - \log(l_n). \tag{4}
\]

Table 1: Standard deviation (\( \approx \) distance) and best freq. region

<table>
<thead>
<tr>
<th>Method</th>
<th>Standard deviation, ( \sigma )</th>
<th>Best freq. region ( [f_L, f_H] )</th>
</tr>
</thead>
<tbody>
<tr>
<td>dcGCCFB</td>
<td>0.013 ( \leq \sigma \leq 0.076 )</td>
<td>[700, 5000]</td>
</tr>
<tr>
<td>linGCCFB</td>
<td>0.015 ( \leq \sigma \leq 0.054 )</td>
<td>[500, 3000]</td>
</tr>
<tr>
<td>MFFB_STR40</td>
<td>0.020 ( \leq \sigma \leq 0.060 )</td>
<td>[600, 2000]</td>
</tr>
<tr>
<td>MFFB_STR120</td>
<td>0.023 ( \leq \sigma \leq 0.067 )</td>
<td>[600, 2000]</td>
</tr>
<tr>
<td>MFFB_STFT40</td>
<td>0.026 ( \leq \sigma \leq 0.064 )</td>
<td>[800, 3000]</td>
</tr>
<tr>
<td>MFFB_STFT120</td>
<td>0.045 ( \leq \sigma \leq 0.085 )</td>
<td>[800, 3000]</td>
</tr>
</tbody>
</table>

The reverse conditions \( (r_{m, n}) \) were also estimated to check the robustness of the estimation methods. Then the relationship for all permutations is given by the following equation:

\[
\begin{bmatrix}
\log(r_{1,2}) \\
\log(r_{1,3}) \\
\vdots \\
\log(r_{27,28}) \\
\log(r_{2,1}) \\
\log(r_{3,1}) \\
\vdots \\
\log(r_{28,27})
\end{bmatrix} = \begin{bmatrix}
1 & -1 & 0 & \ldots & 0 & 0 \\
1 & 0 & -1 & \ldots & 0 & 0 \\
\vdots & \vdots & \ddots & \ddots & \ddots & \ddots \\
0 & 0 & 0 & \ldots & 1 & -1 \\
0 & 1 & 1 & \ldots & 1 & 1
\end{bmatrix} \begin{bmatrix}
\log(l_1) \\
\log(l_2) \\
\vdots \\
\log(l_{28})
\end{bmatrix}.
\tag{5}
\]

The least squared method is applied to estimate the relative VTLs, \( [l_1, l_2, \ldots, l_{28}] \). Then other VTL ratios \( r \) based on these estimated VTLs are estimated using the following equations:

\[
\hat{l}_{\log} = (H^T H)^{-1} H^T r_{\log},
\]

\[
i = [l_1, l_2, \ldots, l_{28}] = \exp(\hat{l}_{\log}),
\tag{7}
\]

\[
\hat{r} = \exp(H \hat{l}_{\log}).
\]

The estimation accuracy is evaluated in terms of the rms difference or Euclidean norm \( d_{est} \) between individual VTL ratios \( r \) and estimated VTL ratios \( \hat{r} \):

\[
d_{est} = ||r - \hat{r}|| \approx \sigma.
\tag{8}
\]

Note that \( d_{est} \) is virtually identical to standard deviation \( \sigma \) around the identity mapping line \( (\hat{r} = r) \). The accuracy improves when \( \sigma \) is smaller.

2.3.2. Selection of best frequency region

We must select frequency region \( [f_L, f_H] \) in Eq. 1 for the reliable estimation of VTL, because the VTL information in the lower and higher frequency regions are smeared by other speech characteristics. The glottal pulse rate and the shape of the glottal source waveform modify the spectrum at low frequencies. Spectral zeros are produced by the resonances of pyriform fossa [9] in high frequencies around 4 to 5 kHz.

The spectral components in the frequency region midway between these smeared regions are expected to dominantly contain vocal tract resonance information. In other words, individual VTL ratio \( r \) is estimated as a function of the selected frequency region. Consequently, estimation accuracy \( d_{est} (\approx \sigma) \) in Eq. 8 is also a function of the frequency region.
2.4. Results of objective evaluation

The range of standard deviations $\sigma$ and the best frequency region $[f_L, f_H]$ for all methods are summarized in Table 1. The minimum values of standard deviation $\sigma$ are the smallest in dcGCFB, and the greatest in MFFB_STR120. It is clear that GCFBs are generally better than MFFBs. Figures 1, 2, 3, and 4 show the contour maps of the standard deviation for all three sentences when the VTL estimation is performed using dcGCFB, linGCFB, MFFB_STR40, and MFFB_STR120. The abscissa is lower limit frequency $f_L$, and the ordinate is higher limit frequency $f_H$ for the frequency region used in Eq. 1.

It is clear that the standard deviations in dcGCFB and linGFCB are generally smaller than those by the MFFB-based methods. Fig. 1 shows that the standard deviations are relatively small in the higher frequencies between 3000 and 6000 Hz when the lower frequency exceeds 300 Hz. The deterioration above 6000 Hz is simply because the upper limit of the filter center frequency in the GCFBs was 6000 Hz and the estimation was based on the extrapolation beyond the limit. The best frequency regions $[f_L, f_H]$ for dcGCFB and linGCFB are $[700, 5000]$ Hz and $[500, 3000]$ Hz, respectively. In contrast, the best frequency regions for MFFB_STR40 and MFFB_STR120 are $[600, 2000]$ Hz in which the VTL estimation does not seem reliable because important formant information is also distributed between 2000 and 4000 Hz.

Figure 5 shows a scatter plot of individual VTL ratios $r$ and VTL ratios $\hat{r}$ from the regression analysis for dcGCFB and MFFB_STR40 when selecting the best frequency region for each method. The distribution of individual points by dcGCFB is more concentrated. The ratios beyond 1.3 estimated by MFFB_STR40 are unlikely in realistic VTL ratios. The results imply that the proposed GCFB-based method yields better VTL ratio estimates than the MFFB-based method.

3. Subjective Evaluation

We conducted series of subjective tests to evaluate the naturalness of the morphed sound using the VTL normalization methods with the best frequency regions [5]. In this section, we show the preliminary results of moderate sound quality with a sampling rate of 16 kHz.

3.1. Speech stimuli

Three sentences spoken by four male and four female speakers were used to evaluate the quality of the morphed speech sounds. The speech data were the same as in the objective evaluation.
For each sentence, we created the morphing sound between two speakers for every pair of speech samples. The total number of combinations was 56 (= 8 C2). The test stimuli were created by the morphing method based on TANDEM-STRAIGHT [7], and the morphing rate was set to 0.5, which is generally the most difficult condition for maintaining sound quality. The following three sets of stimuli were prepared:

Stimuli N: No VTL normalization
Stimuli G: VTL normalization based on dcGCFB
Stimuli M: VTL normalization based on MFFB

3.2. Experimental conditions

Ten adult listeners (five males and five females) with normal hearing participated. The stimuli were presented over headphones in a soundproof chamber. Thurstone’s paired comparison test was used to evaluate the naturalness of the morphed sounds. A pair of morphed speech sounds was presented sequentially in two intervals. The subjects were instructed to choose the interval that contains the more natural sound.

3.3. Results

The experimental results are shown in Fig. 6 with N, M, and G table. The scores for all three sentences are summarized in Fig. 6(a). The individual scores for the short, middle, and long sentences are shown in Figs. 6(b), (c), and (d). VTL normalization clearly improves the naturalness of the morphed sounds. It also demonstrates that the proposed dcGCFB-based method outperforms the MFFB-based method.

The individual scores for the different gender and the same gender cases are shown in Figs. 6(e) and (f). The improvement by VTL normalization is clearly more salient in morphing utterances spoken by speakers of different genders than the same gender. The proposed dcGCFB-based method is always better than or equal to the MFFB-based method. As a consequence, VTL normalization is essential to improve the sound quality in voice morphing. Subjective tests using high quality, 48 kHz sounds need to be conducted as the next step.

4. Conclusions

This paper proposes a new method using gammachirp auditory filterbanks (GCFBs) for VTL estimation and normalization to improve voice morphing. Objective and subjective evaluations were performed to compare them with the MFFB-based methods. The results clearly indicated that the VTL normalization is essential for voice morphing between different genders, and the dcGCFB-based method outperforms the MFFB-based method.

5. Acknowledgements

This work was supported in part by Grants-in-Aid for Scientific Research (A) 19200017 and (B) 21300069 by JSPS.

6. References