HIGH QUALITY VOICE MANIPULATION METHOD BASED ON THE VOCAL TRACT AREA FUNCTION OBTAINED FROM SUB-BAND LSP OF STRAIGHT SPECTRUM

Ayanori Arakawa, Yoshinori Uchimura, Hideki Banno, Fumitada Itakura\(^1\), Hideki Kawahara\(^2\)

\(^1\)Graduate School of Science and Technology, Meijo University, Siogamaguchi 1-501, Tempaku-ku, Nagoya-shi, Aichi, 468-5802, Japan
\(^2\)Faculty of Systems Engineering, Wakayama University, Sakaedani 930, Wakayama-shi, Wakayama, 640-8510, Japan

\{m0830001, m0951501\}@ccmail.meijo-u.ac.jp, \{banno, itakuraf\}@ccmsf.meijo-u.ac.jp, kawahara@sys.wakayama-u.ac.jp

ABSTRACT

This paper describes a high-quality manipulation method of voice quality based on the vocal tract area function (VTAF) obtained from sub-band LSP of STRAIGHT spectrum. Our research group had developed the manipulation technique of voice quality based on VTAF that can generate natural formant transition. However, it is observed that the generated sound sometimes results in degradation when the input signal has a high sampling frequency. Therefore, we develop a new method that extracts VTAF properly from such input signal. This method firstly divides the input spectral envelope represented by STRAIGHT spectrum into lower and higher frequency bands, secondly extracts the Line spectrum pair (LSP) in each frequency band after spectral flattening that is appropriate for the frequency band, thirdly concatenates a pair of the sub-band LSP, and finally obtains VTAF from PARCOR coefficients converted from the concatenated LSP. A subjective experiment proved that the proposed method is high quality enough.

Index Terms— Speech synthesis, Speech analysis, Vocoders, Vocal system

1. INTRODUCTION

High-quality speech analysis-synthesis systems such as STRAIGHT are used for various manipulations of voice quality such as speech morphing [1] in recent years. The conventional manipulation method based on such speech analysis-synthesis systems can manipulate easily the fundamental frequency (F0) which is associated with a characteristic of the vocal tract. However, the generated sound sometimes results in degradation because the generated spectral envelope can be unnatural. Therefore, we had developed a manipulation method of voice quality using the vocal tract area function (VTAF) [2] which approximates a shape of the human vocal tract [3]. This method enables to generate a natural spectral envelope because modification in VTAF simulates natural change of a shape of the human vocal tract. The method also enables the rule-based modification of voice quality by changing the dimensions of the VTAF.

The VTAF-based method requires in advance the adaptive inverse filter which is adaptively designed to remove characteristics of glottal source and radiation from an input signal. However, the filter is designed for signals with a low sampling frequency such as 8 kHz, it is observed that the generated sound by using the adaptive inverse filter sometimes results in degradation when the input signal has a high sampling frequency. In this paper, we describe a new method that extracts VTAF properly from an input signal with a high sampling frequency.

2. MANIPULATION METHOD OF VOICE QUALITY BASED ON THE VTAF

2.1. Vocal tract area function

Many research groups are investigating a method to estimate a shape of the human vocal tract from X-ray or magnetic resonance imaging (MRI) data [4]. The estimated vocal tract can be useful for applications such as manipulation of voice quality. However, measuring the vocal tract data is not easy because special machinery is required for that. In addition, an exact shape of the vocal tract is not always needed for such applications. Therefore, we decided to employ the VTAF based on the linear predictive coding (LPC), which represents a shape of the human vocal tract simulated by the Kelly’s speech production model, for the applications. VTAF can be easily calculated from the PARCOR coefficients which are obtained in LPC analysis procedure [5].

The speech signal includes characteristics of glottal source and radiation. Therefore, it is required to remove these characteristics from the input speech signal before calculating VTAF. We utilized the adaptive inverse filtering method proposed by Nakajima [6] for removing the characteristics. Nakajima’s method estimates the characteristics and applies the inverse filter of that in each analysis frame. Figure 1 shows the structure of the adaptive inverse filter. The value of \( \epsilon_m \) \((m = 1, 2, 4)\) is determined by the least square method. That is, the following \( L \) is minimized.

\[
L = \sum_{i=0}^{N-1} (x_{m-1,t-i} - \epsilon_m x_{m-1,t-i})^2 + \frac{\epsilon_m^2}{4} x_{m-1,t-2-i})^2, \tag{1}
\]

where \( x_{m,t} \) is a sample value of the output waveform at time \( t \) from the \( m \)-th stage \( (x_{0,t} \) means the original input signal), and \( N \) is the length of window.

The partial derivative equation \( \partial L / \partial \epsilon_m = 0 \) in order to minimize \( L \) leads the following cubic algebraical equation.

\[
C_{22} \epsilon_m^3 - 6C_{21} \epsilon_m^2 + (4C_{02} + 8C_{11}) \epsilon_m - 8C_{01} = 0, \tag{2}
\]
where $C_{jk}$ is the element in the waveform covariance matrix described by

$$
C_{jk} = \sum_{i=0}^{N-1} x_{m-1,t-j-i} x_{m-1,t-k-i}.
$$

(3)

By solving this equation, $\epsilon_m$ is given as a real root satisfying $|\epsilon_m| < 2$. $\epsilon_3$ and $\epsilon_5$ can be easily obtained by

$$
\epsilon_3 = \rho_{2,2}/\rho_{2,0} \quad (4)
$$

$$
\epsilon_5 = \rho_{4,3}/\rho_{4,0} \quad (5)
$$

instead of solving the equation (2), where $\rho_{m,i}$ is the $i$-th correlation coefficient obtained from output of the $m$-th stage. The adaptive inverse filter in Figure 1 can be structured by setting $\epsilon_1, \epsilon_2, ..., \epsilon_5$.

Based on the Levinson-Durbin algorithm, the PARCOR coefficients are calculated from the auto-correlation function of the filtered waveform by the adaptive inverse filter. The Kelly’s speech production model [7] is based on the acoustic tube model. The reflection coefficient $\kappa$ is determined by a cross-sectional area $A$. That is,

$$
\kappa_n = (A_n - A_{n+1})/(A_n + A_{n+1}). \quad (6)
$$

Since the PARCOR coefficient $\kappa_n$ corresponds to the reflection coefficient $\kappa_n$, VTAF can be recursively obtained by

$$
A_n = \begin{cases} 
1 + \kappa_n A_{n+1} & (n = p, p-1, \ldots, 1) \ 
\text{const.} & (n = p+1)
\end{cases} \quad (7)
$$

$A_{p+1}$ corresponds to an area of the glottis, is set to 1.0 cm$^2$ in our method. Then, the obtained $A_n$ is employed for manipulation of voice quality.

### 2.2. STRAIGHT-based VTAF [3]

In our research, the adaptive inverse filter is applied to the STRAIGHT spectrum as follows. In each stage, the auto-correlation function of the previous ($m - 1$-th) stage $\rho_{m-1,i}$ is firstly calculated by

$$
\rho_{m-1,i} = \frac{1}{2\pi} F_{\text{DTFT}}^{-1}[|X_{m-1}(e^{j\omega})|^2] \quad (\omega = 0, \ldots, N - 1), \quad (8)
$$

where $|X_{m-1}(e^{j\omega})|^2$ is the power spectrum obtained from output of the previous stage ($|X_0(e^{j\omega})|^2$ is the original STRAIGHT spectrum), and $F^{-1}_{\text{DTFT}}[\cdot]$ is the inverse DTFT. Secondly, the coefficients of the adaptive inverse filter is obtained by solving

$$
\rho_{m-1,0}\epsilon_m^2 - 6\rho_{m-1,1}\epsilon_m^2 + (4\rho_{m-1,2} + 8\rho_{m-1,0})\epsilon_m - 8\rho_{m-1,1} = 0 \quad (m = 1, 2, 4), \quad (9)
$$

or calculating equations (4) ($m = 3$) and (5) ($m = 5$). In equation (9), $\rho_{m-1,i}$ is an approximation of the elements of the covariance matrix in equation (3). Thirdly, $|X_m(e^{j\omega})|^2$ is updated by applying the filter of $m$-th stage. This procedure is recursively repeated up to $m = 5$.

The resultant amplitude response of the output signal $|X_5(z)|$ is expressed by

$$
|X_5(z)| = |X_0(z)| \prod_{m=1,2,4} |1 - \epsilon_m z^{-1} + \frac{4}{5} z^{-2}|\cdot|1 - \epsilon_3 z^{-2} - |1 - \epsilon_5 z^{-3}|. \quad (10)
$$

By using the output auto-correlation function

$$
\rho_{i,i} = \frac{1}{2\pi} F_{\text{DTFT}}^{-1}[|X_5(e^{j\omega})|^2] \quad (|\omega| = 0, \ldots, N - 1) \quad (11)
$$

as input of the Levinson-Durbin algorithm in the LPC analysis procedure, STRAIGHT-based VTAF can be obtained.

### 3. VTAF BASED ON SUB-BAND LSP

Although the manipulation method of voice quality based on the adaptive inverse filter works fine for signals with a low sampling frequency up to around 16 kHz, it does not work for signals with a high sampling frequency. One reason is that the filter was originally proposed for signals with a low sampling frequency, another reason is that the LPC analysis is not stable in a higher frequency band. Although we have tried resampling the VTAF in the case of a low sampling frequency, we have found that the resampling does not restore properly the spectrum in a higher frequency region.

To solve these two problems, we propose a new method to extract VTAF properly based on sub-band line spectrum pair (LSP). This method firstly divides the STRAIGHT spectrum of an input signal into lower and higher frequency bands, secondly extracts line spectrum frequencies (LSFs) in each frequency band after spectral flattening that is appropriate for the frequency band, thirdly concatenates the sub-band LSFs, and finally obtains VTAF from PARCOR coefficients converted from the concatenated LSFs. The first problem can be solved, because the adaptive inverse filter is applied
only for the lower frequency band. The second problem will also be solved, because LSFs have good interpolation characteristics.

The following is the procedure of the proposed method in each frame. \( f_s \) is the sampling frequency, and \( f_c \) is the cut-off frequency in band-division in the following explanation.

### 3.1. Band division

STRAIGHT spectrum \( |X(e^{j2\pi k/N})| \) is divided into the lower frequency part and the higher frequency part, and the higher frequency part is moved to baseband as follows.

\[
|X^L(e^{j2\pi k/N_L})| = \begin{cases} 
|X(e^{j2\pi k/N_L})| & (0 \leq k < k_c) \\
|X(e^{j2\pi (k-k_c)/N_L})| & (k_c \leq k \leq N_L - k_c) \\
|X(e^{j2\pi (k-N_L+k_c)/N_L})| & (k_c < k < N_L) 
\end{cases}
\]

(12)

where \( k_c = Nf_c/f_s \) is cut-off frequency point, \( N_L \) and \( N_H \) are DFT points of lower and higher frequency bands selected to satisfy the conditions \( 2k_c \leq N_L \) and \( N - 2k_c \leq N_H \). In the equations, the spectra have a frequency region with constant values (\( |X(e^{j2\pi k/N_L})| \) and \( |X(e^{j2\pi k})| \) respectively). Since these regions are only required to prevent from wide spectral dips, another way to prevent from the dips can be chosen instead of using the constant values.

### 3.2. Conversion to auto-correlation function

The obtained spectra are converted into auto-correlation functions \( \rho^L_i \) and \( \rho^H_i \).

\[
\rho^L_i = \frac{1}{N} \mathcal{F}_\text{DFT}^{-1} \left[ |\Psi^L(X^L(e^{j2\pi k/N_L}))|^2 \right] 
\]

\( i = 0, 1, \ldots, N_L - 1, \) (14)

\[
\rho^H_i = \frac{1}{N} \mathcal{F}_\text{DFT}^{-1} \left[ |\Psi^H(X^L(e^{j2\pi k/N_L}))|^2 \right] 
\]

\( i = 0, 1, \ldots, N_H - 1, \) (15)

where \( \Psi^L(\cdot) \) and \( \Psi^H(\cdot) \) are flattening functions for lower and higher bands, and \( \mathcal{F}_\text{DFT}^{-1}[\cdot] \) is the inverse DFT. The flattening functions are independently selected in order to fit with the frequency bands. We selected the adaptive inverse filter as \( \Psi^L(\cdot) \), and cepstral smoothing that set 0th and 1st orders to zero in cepstral domain as \( \Psi^H(\cdot) \).

### 3.3. LSP calculation

The LSFs \( \omega^L_n \) \( (n = 1, \ldots, p_L) \) and \( \omega^H_n \) \( (n = 1, \ldots, p_H) \) are obtained in each frequency bands by a normal LSP analysis procedure. Then, the parameters \( \omega^L_n \) and \( \omega^H_n \) are concatenated together. The concatenated LSFs \( \omega_n \) is obtained by

\[
\omega_n = \begin{cases} 
\omega_n^L & (n = 1, \ldots, p_L) \\
\frac{\omega_n^L + \omega_c + \pi \omega_1^H}{2} & (n = p_L + 1) \\
\omega_c + \frac{\pi \omega_1^H}{\pi - \omega_1^H} & (n = p_L + 2, \ldots, p_L + p_H + 1), 
\end{cases}
\]

(16)

where \( \omega_c \) is a normalized angular frequency of the cut-off frequency expressed by \( \omega_c = 2\pi f_c/f_s \). The resultant analysis order \( p \) equals to \( p_L + p_H + 1 \). In this equation, the LSF where \( n = p_L + 1 \) is the center location between the highest LSF of the lower band and the lowest LSF of the higher band. Though this LSF is not indispensable, yet it contributes to filling up the spectral dip that is sometimes generated between two bands.

If the spectrum of an input signal has a spectral peak around the cut-off frequency \( f_c \), the LSFs cannot represent the peak. However, the peak can be represented by the residual spectrum used in the synthesis stage.

### 3.4. VTAF calculation

The obtained LSFs \( \omega_n \) \( (n = 1, \ldots, p) \) are converted into the PARCOR coefficients \( k_n \) through LPC coefficients \( \alpha_n \). Then, the PARCOR coefficients is converted into VTAF based on equation (7). In the synthesis stage, the residual spectrum that cannot be flattened by \( \alpha_n \) is also used in combination with the obtained VTAF. Note that this residual spectrum includes characteristics of the adaptive inverse filter and the cepstral smoothing.

Figure 2 shows the obtained spectral envelope (upper figure) and log-VTAF (lower figure). You can see that the spectral envelope is properly flattened, and the log-VTAF of the proposed method has a shape that is similar to the conventional method with a low sampling frequency and is made more detailed, while the conventional method with a high sampling frequency takes big values near lips. Figure 3 shows the obtained vocaltractgram which is spectrogram-like display of a VTAF sequence. This figure also indicates that the vocal tract shape of the proposed method is similar to that of the conventional method with a low sampling frequency compared with that of the conventional method with a high sampling frequency.
4. SUBJECTIVE EVALUATION

4.1. Simple voice conversion system

To evaluate our method, we constructed a simple voice conversion system that converts original voice into target voice. This conversion system is based on addition of the log-VTAF difference between original and target into the original log-VTAF sequence. In the following explanation, the frame number \( t \) is introduced. The conversion of VTAF is expressed as

\[
A_{\text{synth}}(t, n) = A_{\text{orig}}(t, n)A_{\text{larg}}(n)/A_{\text{orig}}(n),
\]

(17)

where \( A_{\text{synth}}(t, n) \) and \( A_{\text{orig}}(t, n) \) are the \( n \)-th VTAF of synthetic and original voice in a frame \( n \), and \( A_{\text{larg}}(n) \) and \( A_{\text{orig}}(n) \) are the \( n \)-th VTAF of target and original voice averaged for the long-term. The conversion of F0 is similar to that of the VTAF expressed as

\[
f_{\text{synth}}(t) = f_{\text{orig}}(t)f_{\text{larg}}/f_{\text{orig}}.
\]

(18)

The experiment also includes the spectrum-based method for comparison which adds the log-spectral difference between original and target into the original spectrum sequence described by

\[
|X_{\text{synth}}(t, k)| = |X_{\text{orig}}(t, k)|\bar{X}_{\text{larg}}(k)/|X_{\text{orig}}(k)|.
\]

(19)

Note that this spectrum-based method is only available in the voice conversion system which has both the original and the target voice. While the proposed method can be applicable to rule-based modification of voice quality. The residual spectrum in the proposed method is processed in the same way. We also added the method without STRAIGHT which does not use pitch-synchronous analysis and STRAIGHT spectra while VTAF extraction in each frame is the same as the proposed method. In this case, the frame length and the shift length is 20 ms and 1 ms, respectively.

4.2. Experiment

We have performed a subjective experiment of the opinion method. This experiment includes (1) the spectrum-based method, (2) the proposed method, (3) the proposed method without STRAIGHT, and (4) the conventional adaptive inverse filter method. The sampling frequency \( f_s \) is 44.1 kHz. The total analysis order of LPC \( p \) is 51 in the case of (2)(3)(4). The cut-off frequency \( f_c \) is 5.5 kHz, the analysis orders of the lower band \( p_L \) and the higher band \( p_H \) are 12 and 38, and the DFT points of \( N_s \), \( N_L \), and \( N_H \) are 2048 in the case of (2)(3).

Five vowels of long-tone singing voice are used as input. Singers are two amateur male altoes recorded by us, and two professional male basses and two professional sopranos in the RWC music database [8]. The conversion type includes (a) loudness conversion; low loudness voice to high loudness voice and vice versa, and (b) F0 conversion; low F0 voice to high F0 voice and vice versa. The system converts timbre of input voice so that timbre changes accompanied by these conversions are generated. Four listeners with normal hearing participated in this experiment.

Figure 4 shows the result of the experiment. It is found that the proposed method is better than the conventional method and almost equal to the spectrum-based method. So, it can be concluded that the proposed method is high quality enough. The reason why the spectrum-based method is slightly better is that the proposed method sometimes generates some noises caused by spectral vibration in a higher frequency region. We believe that the problem is fixed by the LSP smoothing in the higher frequency band, and will confirm this in another paper. We also note that the proposed method has a big advantage that it can be used for the rule-based modification of voice quality. It is also found that the proposed without STRAIGHT is really worse than the proposed method only in the F0 conversion. This result proved that STRAIGHT is a powerful tool in F0 conversion.

5. CONCLUSIONS

This paper describes a high-quality manipulation method of voice quality based on the vocal tract area function (VTAF) obtained from sub-band LSP of STRAIGHT spectrum. A subjective experiment proved that the proposed method is high quality enough.

6. REFERENCES


