Codec Integrated Voice Conversion for Embedded Speech Synthesis

Guntram Strecha, Oliver Jokisch, Matthias Eichner and Rüdiger Hoffmann

Laboratory of Acoustics and Speech Communication
Dresden University of Technology, Germany

Abstract

Voice conversion technologies transform individual characteristics of speech patterns while preserving the original content, and can be widely used in speech processing. Considering limited system resources, in particular, of embedded concatenative speech synthesis, voice conversion may reduce the memory consumption of the acoustic database. Voice conversion enables the intra-gender or cross-gender generation of new voices by using an existing high-quality voice. Usually, voice conversion is based on modification of spectral properties in accord with pitch manipulation. Warping functions in the frequency domain aiming at a reverse vocal tract length normalization (VTLN) is a simplified approach. Consequently, voice conversion itself generates a critical calculation complexity which contradicts the practical constraints of typical embedded and mobile applications.

The authors propose a novel approach for voice conversion by re-using features of a common speech codec. Such a codec is already available in typical mobile applications and the resulting voice quality is widely accepted. The paper investigates the manipulation of the immittance spectral frequencies (ISF) provided by the Adaptive Multi Rate Wideband codec (AMR-WB). This algorithm has been integrated into the embedded speech synthesizer microDRESS.

1. Introduction

Voice conversion used in Text-To-Speech (TTS) systems offers the possibility for personalized speech synthesis. It is of particular relevance to embedded applications with strongly restricted memory resources. TTS systems with very low footprint [1] profits voice conversion techniques by getting additional voices without increasing memory consumption.

Voice conversion is mostly done by modification of the spectral properties in accord with pitch manipulation. Thereby, two different goals of voice conversion have to be discerned – the conversion of a source voice to fit the characteristics of a target voice [2, 3] and the generation of another voice not aiming at specific target voice [4]. The second approach is usually done by manipulating the spectral envelope using a warping function applied in frequency domain.

In this paper we investigate a method for voice conversion, which warps the spectral envelope by manipulating the line spectrum frequencies (LSF). The motivation for this work is the availability of these features provided by the Adaptive Multi Rate Wideband (AMR-WB) codec, which is used for compression of the inventory of the concatenated embedded speech synthesizer microDRESS [1]. Integrating the voice conversion into the speech decoder is the strict continuation of the already integrated acoustic synthesis of the synthesizer into the decoder [5].

2. VTLN based voice conversion

2.1. Targets of voice conversion

In contrast to voice conversion algorithm aiming at fitting the characteristics of a given target voice, the reverse vocal tract length normalization (VTLN) approach generates new voices. The use of VTLN in speech recognition has been widely investigated [6, 7]. VTLN is used in speech recognition to decrease the inter-speaker variability due to the variance of the vocal tract length of different speakers. For our purposes, we used this approach in reverse direction to increase the variability of a given source voice. Usually the normalization is done by warping the frequency axis of the amplitude spectrum.

2.2. Warping functions

Warping the frequency axis of the amplitude spectrum yields in relocating formant frequencies, which is one of the characteristics of a voice. Several warping functions have been proposed. For our approach, the bilinear warping function is used:

\[
\phi_\alpha(\omega) = \omega - 2 \arctan \left( \frac{\alpha \sin \omega}{1 + \alpha \cos \omega} \right)
\]

The warping factor \(\alpha\) controls the level of stretching the frequency axis. Values less than zero results in compression of the spectral envelope at lower frequencies, values greater than zero in stretching towards the higher frequencies (fig. 1). In our TTS system we use this warping function to manipulate the ISF parameters provided by the AMR-WB decoder at run time of the speech synthesis.

Fig. 1: Bilinear warping function for different values of the warping factor \(\alpha\)
3. Voice conversion by manipulation of codec features

As described in [5] in our concatenative TTS system we use the AMR Wideband codec to compress the inventory containing the dipphone segments.

3.1. Adaptive Multi Rate Wideband Codec (AMR-WB)

The AMR-WB is a speech codec based on the code excited linear predictive (CELP) coding model [8]. A $M = 16$ order linear prediction (LP) synthesis filter is used which is given by:

$$H(z) = \frac{1}{A(z)} = \frac{1}{1 + \sum_{m=1}^{M} a_m z^{-m}}$$  \hspace{1cm} (2)

where $a_m, i = 1, \ldots, M$ are the linear prediction (LP) parameters. These LP filter coefficients are converted to the immittance spectral frequencies (ISF). For a 16th order LP filter, the ISF are defined as the roots of the sum $Q$ and difference polynomials $P$:

$$Q(z) = A(z) + z^{-M} A(z^{-1})$$  \hspace{1cm} (3)

$$P(z) = A(z) - z^{-M} A(z^{-1})$$  \hspace{1cm} (4)

The roots of these polynomials are on the unit circle and they alternate with each other. After eliminating the two roots at $z = 1$ and $z = -1$ and with respect to the conjugated roots on the unit circle $e^{lz}$ the polynomials can be written as

$$Q'(z) = (1 + a_M) \prod_{i=0,2,\ldots}^{M-2} (1 - q_i z^{-1} + z^{-2})$$  \hspace{1cm} (5)

$$P'(z) = (1 - a_M) \prod_{i=1,3,\ldots}^{M-3} (1 - q_i z^{-1} + z^{-2})$$  \hspace{1cm} (6)

where $q_i = \cos(\omega_i)$ with $\omega_i$ being the immittance spectral frequencies (ISF). Since the roots of the polynomials alternate with each other on the unit circle the ISF satisfy the property $0 < \omega_0 < \omega_2 < \cdots < \omega_{M-2} < \pi$ [9]. Besides the coded excitation the ISF is transmitted to the decoder, where the LP filter coefficients regained from the ISF by

$$A(z) = \frac{Q(z) + P(z)}{2}$$  \hspace{1cm} (7)

and with

$$S(z) = H(z)E(z),$$  \hspace{1cm} (8)

where $E(z)$ represents the decoded excitation signal, the speech signal $S(z)$ is reconstructed. As the ISF are similar to the line spectrum frequencies (LSF), the voice conversion can be applied to the LSF accordingly.

3.2. Manipulation of line spectrum frequencies (LSF)

The LSF are a representation of the magnitude power spectrum. Since the $\omega_i$ are equally-spaced in the range $\frac{\pi}{M+1}, \ldots, \frac{M\pi}{M+1}$ the corresponding LP coefficients are zero, except $a_0 = 1$, and the spectrum is flat. Therefore, the spacing between the $\omega_i$ forms the spectral envelope. Closely positioned $\omega_i$ parameters correspond to the peaks of the spectrum or the formants, and widely positioned $\omega_i$ parameters correspond to the spectrum valleys. Lower $\omega_i$ corresponds to lower frequency ranges of the power spectrum and higher $\omega_i$ corresponds to higher frequency ranges. The amount of $\omega_i$ commensurates to $i$. These conditions permit the warping of the spectral envelope by manipulating the $\omega_i$.

In our system, we applied the warping function (1) to the differences $\Delta \omega_i$ of the $\omega_i$ to the equally spaced ones:

$$\Delta \omega_i = \Delta \omega(i) = \omega_i - \frac{i + 1}{M + 1} \pi$$  \hspace{1cm} (9)

$$i' = \frac{M}{\pi} \omega_0 \left( \frac{i}{M + 1} \right)$$  \hspace{1cm} (10)

Due to the fact that $i' \in \mathbb{R}$ a quadratic interpolation of $\Delta \omega_i'$ is done for $0 < i' < M - 2$ with

$$\Delta \omega_i' = \Delta \omega(i')$$

$$\Delta \omega_0 = \Delta \omega([i']) + k_1 \Delta_0 - k_2 (\Delta_1 - \Delta_{-1})$$  \hspace{1cm} (11)

$$\Delta_0 = \Delta \omega([i']) - \Delta \omega([i'] - 1)$$

$$\Delta_{-1} = \Delta \omega([i'] + 1) - \Delta \omega([i'])$$

$$\Delta_1 = \Delta \omega([i'] + 2) - \Delta \omega([i'] + 1),$$  \hspace{1cm} (12)

where $k_1 = i' - [i']$, $k_2 = \frac{1}{2} k_1 (1 - k_1)$ and $[i']$ the integer part of $i'$ denotes. The warped $\omega_i$ yielded from the warped differences $\Delta \omega_i'$ by adding the equally spaced LSF:

$$\omega_i' = \Delta \omega_i' + \frac{i + 1}{M + 1} \pi$$  \hspace{1cm} (13)

$\omega_i'$ again follows the ordering property $0 < \omega_0 < \omega_2 < \cdots < \omega_{M-2} < \pi$. The effect of warped LSF with different $\alpha$ is illustrated in figure 2.

![Fig. 2: Effect of different values of the warping factor $\alpha$ on the log-magnitude of $H(n) = \frac{1}{A(n)}$, where $A(n) = \text{FFT}^N((1, a_1, \ldots, a_M))$ and sampling frequency $f_s$.](image_url)

4. Speech synthesis framework and experiments

4.1. Synthesis framework

For testing the warping algorithm described below, it is integrated into the AMR-WB decoder of the speech synthesis sys-
4.2. Voice conversion experiments

To evaluate the warping algorithm, several experiments were done:

1. Re-synthesis of speech signals to avoid quality loss of additional stages of the TTS system.
2. Re-synthesis with \( f_0 \) manipulation without warping to emphasize the effect of warping.
3. Re-synthesis with \( f_0 \) manipulation and warping.
4. Item 1-3, but for synthesis.

The utterances of the re-synthesis experiments were taken from the German PhoneDAT II database. It consists of 3200 signal files recorded at 16 bit, 16 kHz. For the re-synthesis experiments, 3 sentences each for one female and one male speakers were chosen randomly. The synthesis experiments were done with the same sentences of one female and one male speakers. The prosodic features were taken from the re-synthesis sentences and used to control the prosodic manipulation of the acoustic synthesis.

5. Results

The results of the experiments were presented in a listening test to 20 listeners. Additional original utterances from the PhoneDAT II database were included in the listening test. Beside the overall quality evaluation, the listeners were asked to decide whether the presented sentence originated from a female or male speaker. The answer “not sure” were also accepted. In table 1 the experiments were summarized.

The results of the MOS evaluation were shown in fig. 3. Beside the original signals of the PhonDAT database (1), the MOS of the re-synthesis without warping (2), re-synthesis with warping (3), synthesis without warping (4) and synthesis with warping (5) are shown, where the 3 bars each stood for the combination with the manipulated \( f_0 \). (The decrease in MOS when only \( f_0 \) is manipulated and the lowered values when voice is warped showed that more manipulation will induce greater loss in quality.)

The success of the voice conversion is measured by the second listening test. Figures 4 and 5 showed the results of the answers to the question about the gender of the speaker of the presented sentences. The black bars stands for the relative amount of decisions for the opponent gender than the gender of the source voice.

Figure 4 figures out the results of the manipulated female voices. If only the \( f_0 \) is changed to lower values the decisions to the opponent gender increases little. More decisions to the opponent gender achieved by warping the LSF. Already with setting the warping factor to \( \alpha = -0.2 \) and letting \( f_0 \) untouched.
tends to favor male decisions. The tendency is intensified by decreasing $f_0$ as expected.

Figure 5 shows the result of the manipulated male voices. It is seen that the algorithm failed in warping male voices to female ($\alpha > 0$).

One of the reasons for the decreased signal quality due to the warping algorithm is the small amount of LSF used for warping. First, this results in a smoothed magnitude spectrum and therefore in an insufficient splitting of excitation signal and spectral weighting. The spectral envelope of the excitation signal contains characteristics of the vocal tract, which needs to be warped too. And second, the less the amount of LSF parameters the less the precision of the warping algorithm. In spite of the quadratic interpolation (eq. 11) the spacing between the LSF suffers.

Fig. 5: Percentage of answers of the question about the gender of the heard example. The male source speaker were modified by $\alpha$ and $f_0$. Black bars shows the swapped gender

6. Conclusion

The warping of the LSF can be used in speech synthesis to generate different voices from a single voice. If the algorithm is used at the integrated AMR-WB decoder of the synthesis system, the already calculated ISF can be used for warping. The low computational requirements qualifies the method especially for applications in embedded systems. If the amount of LSF/ISF is less, the inter-gender conversion works better for female to male conversion. Increasing the amount of LSF coefficients decreases the distortions of the speech signal due to the warping algorithm.

7. References