ABSTRACT

A new voice conversion method that improves the quality of the voice conversion output at higher sampling rates is proposed. Speaker Transformation Algorithm Using Segmental Codebooks (STASC) is modified to process source and target speech spectra in different subbands. The new method ensures better conversion at sampling rates above 16KHz. Discrete Wavelet Transform (DWT) is employed for subband decomposition to estimate the speech spectrum better with higher resolution. Faster voice conversion is achieved since the computational complexity decreases at a lower sampling rate. A Voice Conversion System (VCS) is implemented using the proposed algorithm with necessary tools. The performance of the proposed method is demonstrated by both subjective listening tests and applications to film dubbing and looping. In ABX listening tests, the listeners preferred the subband based output by 92.1% as compared to the full-band based output.

1. INTRODUCTION

Voice conversion is a method that aims to transform the characteristics of an input (source) speech signal such that the output (transformed) signal is perceived to be produced by another (target) speaker. A Voice Conversion System (VCS) is a software tool that incorporates the voice conversion algorithm with necessary tools for pre- and post-processing as well as recording, analysis and testing modules in a single interface. The processing tools include waveform editing, pitch & duration scaling, sound effects, enhancement, etc. The VCS serves as an “all-in-one” system that enables fast and high quality voice conversion with maximum capabilities of pre- and post-intervention. Such a system can be used in many applications. Our current research is focused on the applications in film industry. The VCS can be employed for dubbing and translation to a different language by preserving speaker characteristics. Another application is looping (i.e. replacing undesired utterances with the desired ones). Speech by a famous personality can be generated using conversion and/or TTS technology in commercials and news. Another field of interest is the Karaoke applications in which an ordinary voice can be transformed into a famous singer’s voice. Personification of characteristics. Another application is looping (i.e. replacing undesired utterances with the desired ones). Speech by a famous personality can be generated using conversion and/or TTS technology in commercials and news. Another field of interest is the Karaoke applications in which an ordinary voice can be transformed into a famous singer’s voice. Personification of

The acoustic spaces of the source and target speaker are kept without the need to generate and process a separate TTS database for each speaker. Codebook mapping is a general strategy for obtaining one-to-one correspondence between the source and the target acoustical space. Previous research on voice conversion employing codebook based methods include [1], [2], and [3]. The acoustic spaces of the source and target speaker are represented using acoustical features such as formant frequencies [4], LPC cepstrum coefficients, Line Spectral Frequencies (LSFs) [5] and harmonic plus noise model parameters [6]. In this study, we investigate the use of Discrete Wavelet Transform (DWT) and subband processing as a modification to the Speaker Transformation Algorithm Using Segmental Codebooks (STASC) [1] as well as implementing a VCS system using both STASC and the proposed method. In STASC, voice conversion is performed in two steps: Training and Transformation. The training data consist of the same sentences uttered by both source and target speakers. The sentences must be phonetically balanced in order to obtain successful transformation results. The source sentence is aligned first and then the target sentence is force-aligned with it using the “Sentence HMM” method. A Hidden Markov Model (HMM) is generated for each sentence as the name implies. In the transformation stage, the modifications are carried out separately on the vocal tract spectrum and the excitation spectrum. The vocal tract is modified using the source and target line spectral frequencies (LSFs). The excitation spectrum is modified using time-domain / frequency-domain pitch scaling algorithms. Appropriate energy and duration scaling can also be applied. The vocal tract spectrum is estimated from the closest codebook entry LSFs by interpolation. The interpolation creates formant shift problems at higher sampling rates as discussed later. This paper presents a new training approach using subband based acoustical features to obtain better results. The transformation procedure is modified by transforming lower subbands of the original signal while keeping the higher frequency subbands untouched and then reconstructing the transformed signal from the transformed subband signals and the non-transformed subbands. This method also ensures that the regions of the spectrum that contain non-speech components are not distorted by voice conversion.
The next section (Section 2) starts with the description of the proposed subband based voice conversion method including brief information on the Discrete Wavelet Transform (DWT). Section 3 presents the experimental results obtained via informal listening tests (ABX tests) as well as a description of the test database and the testing method. Finally in Section 4, the results are discussed.

2. METHOD

Employing the subband based approach for voice conversion at higher sampling rates has proven to be more successful than the full-band approach because of the following reasons:

- Modeling the source and target spectra is more difficult as the sampling rate increases. The number of parameters (i.e. LSFs) should be high enough to represent the spectrum accurately. It is common practice to use 16-18 LSF coefficients to represent the spectrum at 16kHz. When the sampling rate is higher than 16kHz, the number of sufficient LSF coefficients increase for an accurate spectral representation. However, when we increase the order beyond 18, we get significant distortion in transformation quality. In the case that we use 18th order LPC to represent the spectrum at higher sampling rates, the transformed speech has a child-like voice quality. This is due to shifts in formants towards higher frequencies by interpolation. The interpolation is performed using the equations (1) and (2):

\[ v_i = \frac{-\gamma d_i}{\sum_{i=1}^{L} e^{-\gamma d_i}}, \quad i = 1, \ldots, L \]  

where \( d_i \) is the distance of the incoming source LSF vector to the \( i \)th source codebook LSF vector and \( v_i \) is the weight corresponding to that codebook entry. \( L \) is the codebook size.

- An estimate of the target LSF vector can be calculated as follows:

\[ \tilde{w}_k^i = \sum_{l=1}^{L} v_i T_{lk}, \quad k = 1, \ldots, P \]  

where \( T_{lk} \) represents \( k \)th LSF in the \( l \)th target codebook entry. \( P \) is the prediction order. When the LSFs are interpolated in the baseband range (i.e. 0-5512.5Hz), the amount of shift in the formants will be less when compared to the full-band range (i.e. 0-22050Hz).

- The distortion generated by voice conversion at non-speech subbands is prevented.

- Training is computationally the most intensive process in voice conversion especially when the source and target acoustical spaces are well covered using appropriate amount of training data. The time needed for subband based training is much lower than the full-band case.

- As the number of parameters in the codebook increases, the search from the codebook (which contains typically 1000 - 5000 speech units) to find the entry (or entries) to be used in the transformation takes more time. Transformation takes less time as sampling rate is decreased.

The subband and full-band based voice conversion methods are incorporated into a VCS system including tools and user interfaces for recording, waveform and signal processing, analysis, testing, training and transformation. In the following subsections, we describe the subband processing procedure and the modifications on the STASC training and transformation algorithms.

2.1. DWT and subband processing

Subband decomposition can be implemented efficiently using the DWT. Following characteristics of the DWT make it an attractive tool for designing filterbanks:

\[ \text{Figure 1: Full-band vs. Subband Based Voice Conversion at 44.1KHz. Source is a male speaker and target is a female speaker. The frames above are taken from the utterance “It’s h/a/rd to t/e/ll an original from a forgery.” The dotted spectra (full-band conversion) has formant shifts below 5KHz which create the child-like voice quality. (Whole spectrum is not displayed).} \]
• Perfect reconstruction is guaranteed if appropriate filter pairs are used.
• FIR filters can be used which are guaranteed to be stable and to have linear phase.
• Subband decomposition (reconstruction) can be realized fully in time-domain using convolution and decimation (interpolation).
• Aliasing can be prevented since appropriate lowpass and highpass filters are used before decimation and after interpolation.

DWT can be implemented either using a pyramid structure or using a lattice filtering approach. In this study the pyramid structure described in Fig. 2.a is implemented. A cascade of this basic structure can be used to implement DWT of any order [7].

DWT is employed in both training and transformation for voice conversion. The impulse and phase responses of the FIR filter pairs used in subband processing are given in Fig. 2.b.

2.2. Subband based training

The flowchart for the subband based training algorithm is given in Fig. 3. First, source and target training utterances are decomposed. Using the subband signals for automatic alignment, the codebooks are generated which contain the acoustical parameter mapping between the source and target speakers. It is observed that using only the lower subbands in the training procedure provides satisfactory performance. This is natural because human speech is restricted to lower spectral regions due to the physical properties of the human voice production mechanism. For 44100 Hz recordings, we have used 4 subbands each covering 5.5 KHz frequency range. The first subband covers approximately all the speech components to be used in the generation of the codebooks. As the sampling rate is decreased to 11025 Hz, training takes much shorter time. Subband codebooks for each speaker are generated. The codebooks contain LSFs, f0, energy and duration as acoustical information. Although extra processing must be carried out for subband decomposition, the time needed for the overall training process is reduced considerably. This is due to the fact that DWT based decomposition can be realized fully in time domain using FIR filters.

2.3. Subband based transformation

The flowchart in Fig. 4 presents the subband based transformation algorithm. The input speech from the source speaker is first decomposed using the DWT filterbank into subbands. The subband signals are processed separately for transforming the characteristics from the source speaker to the target speaker in the subband domain. Lower subband signals are transformed in the current system using the corresponding subband codebooks. The transformation is a frequency domain filtering operation in which the source vocal tract spectrum is transformed to the target vocal tract spectrum using the codebook entries. Post-filtering may be applied optionally for removing audible noise in different frequency bands. In most of the cases, attenuating high frequencies produce better results. Another approach (which we preferred) leaves high frequency subbands as they are without presenting extra distortion due to transformation. We have noticed that transforming higher frequencies does not add much to the quality of the transformation but increases distortion as non-speech components are modelled and transformed as speech. In this case, the subbands that will not be transformed are filtered with appropriate bandpass filters and kept for the reconstruction step. The output signal is generated by reconstructing the transformed signals using the subband reconstruction filterbank. Non-transformed subbands are just added to the output to get the full-band output signal. Prosodic modifications including pitch scaling, time scaling, and energy scaling are performed on the full-band output signal.

3. EVALUATIONS

ABX listening tests were performed to evaluate the subjective performance of the proposed method. We have used different combinations of 5 male and 5 female speakers as the source and the target. 4 speakers (1 female + 3 male) were native American-English speakers and the remaining were native Turkish speakers. Four types of voice conversion is performed as far as the gender of the source and the target is concerned (female-to-female, female-to-male, male-to-female, and male-to-male). First, training and test utterances were recorded at
44.1KHz. Full-band and subband based codebooks are generated by two separate training sessions for each conversion. In the subband case, 4 subbands were employed and only the first subband is converted. Each of the 20 subjects listened to 10 sentences and 10 words that were transformed using the subband and full-band codebooks. The subjects were provided with three recordings each time they were asked to make a decision: (A) Full-band based conversion output, (B) Subband based conversion output, and (X) Target recording. The conversion outputs are presented in random order and the listener is asked to judge whether (A) or (B) sounds more like the target speech (X). The order in which the full-band and the subband based output is presented is also changed randomly. The subband based voice conversion output was preferred over the full-band based output by 92.9% for sentences, 91.3% for words, resulting in an overall preference rate of 92.1%.

The proposed voice conversion method is applied for looping using a movie segment of 70 seconds. Some swearwords were present in the original audio track making it unsuitable for TV broadcasting. As new recordings of the original actor are not available, voice conversion is applied on a dubber recording who replaces undesired words with desired ones. Most of the listeners did not recognize the modification and they reported that the audio track sounded natural. The subband based voice conversion strategy explored in this study becomes an appropriate tool for dubbing and looping applications as at least 44.1KHz sampling rate is necessary.

Figure 5: Preliminary perceptual test results on the importance of subbands in identifying speakers.

We are also conducting a set of perceptual experiments to determine the importance of subbands in perceiving the speaker identity [8]. For this purpose, a preliminary test is performed. First, 10 words uttered by 4 male and 4 female speakers are recorded. The signals are then bandpass filtered to obtain 10 subband signals covering the frequency range from 0 – 22050 Hz. The center frequencies and the bandwidths are adopted from psycho-acoustical studies on perceptual coding. In fact, much more subbands were used in these studies, but we have used less subbands as most of the energy of speech signals are concentrated in frequencies lower than 4KHz. Each subject listened to 50 pairs of sound files. The first sound file in each pair is the full-band version of a word uttered by a speaker and the second file is a subband of the same word uttered by the same speaker or a different speaker. Each subject made a decision on each pair by giving a score on the similarity of the speakers in the first and the second sound file. The mean and variances of the scores given at each subband in the preliminary tests are presented in Fig. 5. The highest mean score is obtained at the second subband (1034 Hz – 1895 Hz range) indicating that this subband is the most important subband for perceiving the speaker identity. The standard deviations of the test scores for each subband were between 5.0 and 6.2 with a mean of 5.9.

4. CONCLUSIONS

In this study, we have proposed a new method based on subband processing for voice conversion. The performance of the subband based method is compared with the fullband based algorithm employed in STASC. STASC training and transformation algorithms were modified for better voice conversion at higher sampling rates. The full-band and subband based methods are incorporated into a VCS system that is easy to use and equipped with necessary tools for recording, analysis, testing and modification. Listening tests are performed to demonstrate the validity of the new approach. The enhancement at higher sampling rates is particularly important for film dubbing and looping applications. Our current research is concentrated on the importance of different subbands for voice conversion. The subband based algorithm will be modified according to the results of the perceptual tests. We will explore the possibility of employing more subbands, perhaps at the lower frequency range to obtain higher resolution.

5. REFERENCES