ANALYSIS OF APPROACHES TO DESIGN VOICE CONVERSION SYSTEMS

Thai Trung Kien, PhD student
Computer Engineering Department
Belarusian State University of Informatics and Radio electronics
Email: kienthaitrung@yahoo.com

Abstract: *Analysis of approaches to design voice conversion systems is proposed in this paper. Base on analysis we give principles of voice conversion system, speech modes will be used to represent parameters of speech signal and acoustic characteristics will influence the quality of system. Moreover, we compare two voice conversion (VC) systems, which have been used two different speech models (source filter model and harmonic model), to offer general view about them and our point of view design voice conversion systems.*

Keywords: Voice conversion, speech model, acoustic characteristic

1. INTRODUCTION

The voice conversion problem has focused a lot of research effort. For instance, an approach to this problem was speech transformation algorithm using segment codebook (STASC) [3]. The method finds accurate alignments between source and target speaker utterances. Using the alignments, source speaker acoustic characteristics are mapped to target speaker acoustic characteristics. A voice conversion method that improves the quality of the voice conversion output at higher sampling rates is proposed in [4]. This method combines the STASC method with Discrete Wavelet Transform (DWT) to estimate the speech spectrum better with higher resolution. Both these research are combined to use into [16]. The other works suggest a possible way to improve the quality of the converted speech consists of modifying only some specific aspects of the spectral envelope [5], or the location of the formants [7, 8]. Spectral conversion techniques have been also approached by different ways such as [9, 10].

In [10], a different approach is proposed which is based on the TD-PSOLA technique and source filter decomposition. TD-PSOLA technique allows prosodic modification while source filter decomposition enables spectral envelope transformation. Moreover, other method also has been known that is voice conversion system based on the Harmonic plus Noise Model (HNM) [11]. HNM performs a pitch synchronous harmonic plus noise decomposition of the speech signal.

In this paper is analyzed base on methods that were introduced above to give overview also our point of view about approach to design VC systems. The analyses directions are to summarize general stages, the speech models, acoustic characteristic, and to detail analyze typical methods of voice conversion.

2. THE APPROACHES FOR VOICE CONVERSION

A. Voice conversion principle

Voice conversion is the process of automatic transformation of a source speaker's voice to that of a target speaker's. They have general stages in the process that are depicted below:

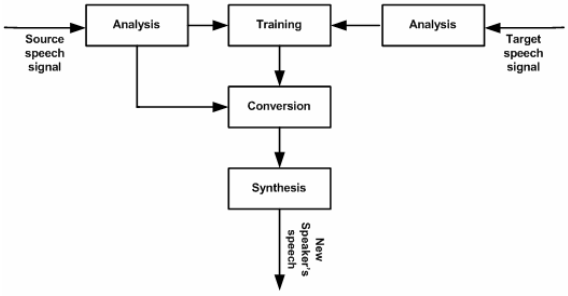

Fig. 1 - The general stages of voice conversion

These stages are:

- Analysis stage: The purpose of this stage is analysis source and target speakers characteristics will be used in the next process stage.
- Training stage: System will be learned the characteristics of source speech and target speech, to create conversion rules.
- Conversion state: In this stage, transformation algorithms are applied that modify characteristics of source speech using the conversion rules obtained in the training stage according to characteristics of target speech.
- Synthesis stage: This inverse process with the analysis process, new speech is created base on parameters obtained from stages above.

B. Speech model for voice conversion

When starting to research about VC system then a speech model is chosen, because it will directly affect quality of VC system. The speech model is a mathematical model and it will be used to represent parameters of speech signal. The speech models of speech processing might be divided into two classes: source-filter modeling with the filter representation of the vocal tract transfer function, and harmonic modeling where the source and the system features are included in the parameters of the harmonic model. In the source-filter model the vocal tract can be modeled either by a filter bank, or a realizable rational transfer function, or

an approximation of a non-realizable exponential function in homomorphic modeling [13].

The source-filter model is represented by the parameters describing the transfer function of the vocal tract model. Two types of the source-filter model are useful for speech processing: the all-pole model known as the autoregressive (AR) model, and the pole-zero models known as the autoregressive moving average (ARMA) model. The AR model of a vocal tract is well known in speech processing as a linear predictive coding (LPC) model [13].

$$P(z) = \frac{G}{1 + \sum_{k=1}^{N_A} a_k z^{-k}}, \qquad (1)$$

where N_A is the order of the AR model, the gain G and the coefficients a_k are the AR parameters or the LPC parameters. AR has the frequency response given by equation:

$$P(e^{j\omega}) = \frac{G}{1 + \sum_{k=1}^{N_A} a_k \exp(-jk\omega)}, \qquad (2)$$

The source-filter model has been used in [3, 4, 5] which will be discussed in method for voice conversion part.

Harmonic model has been shown to be capable of high quality speech processing, particularly in pitch and time scale modification for speech synthesis [11, 12]. The principle of the harmonic speech model is shown in Figure 3.

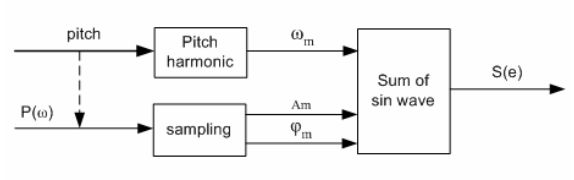

Fig. 2 - Principle of harmonic speech model

Its equation:

$$s(l) = \sum_{m=1}^{M} A_m \cos(\omega_m l + \varphi_m), \qquad (3)$$

$$= \sum_{m=1}^{M} H(m) e^{j(l\omega_m)} \qquad (4)$$

where frequencies ω_m are given by pitch harmonics, and amplitudes A_m and phases φ_m are given by sampling the transfer function of the vocal tract model at these frequencies which is represented in equation (2), H(m) complex amplitude of the m^{th} harmonic, M number of harmonics.

C. Acoustic characteristics for voice conversion

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. The parameters of speech obtained by using speech model in the first stage are parameters, which are called acoustic characteristics. Acoustic characteristics have two types that are the voice source and the vocal tract resonance which are influence on voice individual. They always are very importance for quality of VC system. Parameters of voice source are the average pitch frequency, the time-frequency pattern of pitch (the pitch contour), the pitch frequency fluctuation, the glottal wave shape. And parameters of vocal tract are the shape of spectral envelope and spectral tilt, the absolute values of formant frequencies, the time-frequency pattern of formant frequencies (formant trajectories), the long-term average speech spectrum, the formant bandwidth [2]. On the other hand, acoustic characteristics can be divided into two group are long – term characteristics and sort – term characteristics have been represented in [1]. Almost studies are focused analysis to model and these parameters transformation.

D. Methods for voice conversion

As introduction above, this part will be considered how to approach and implement voice conversion system in [15], [16], which are typical to represent speech models and speaker characteristic transformation. The transformation algorithms are shown in Figures 3 and 4

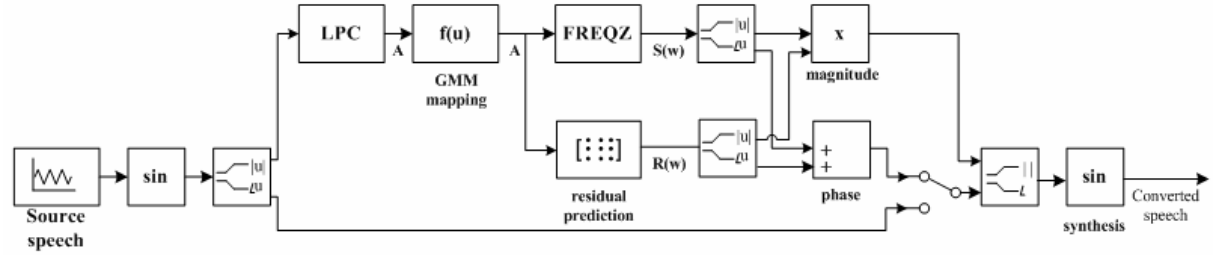

Fig. 3 - Voice transformation algorithm in [15]

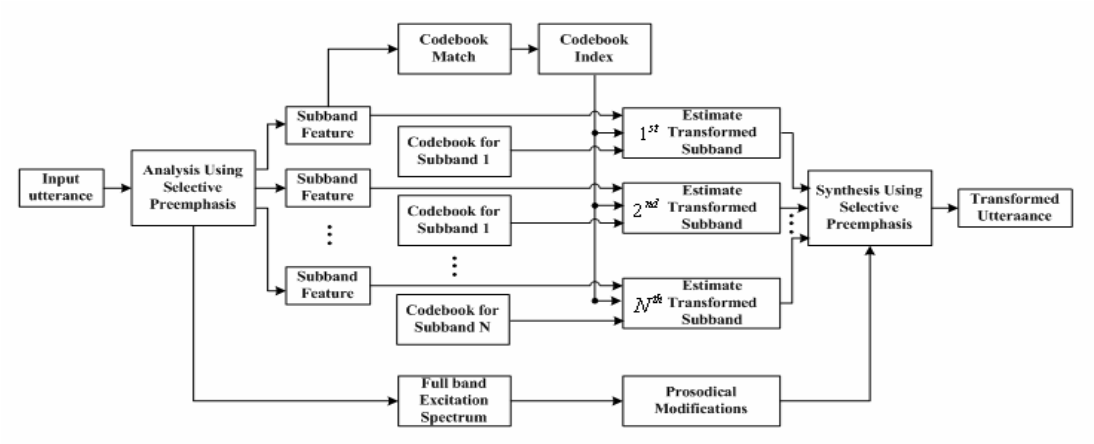

Fig. 4 - Voice transformation algorithm in [16]

In [15], the system is designed to transform the spectral envelope of speech by changing parameters of an all-pole model, using a transformation function implemented by a Gaussian mixture regression model. The author use harmonic speech model to analyze voice while source-filter model is used in [16]. However, the goals of both these models are to obtain line spectral frequencies (LSFs), which will be use to transform spectral characteristics. The reasons that author uses harmonic model are because speech signal consists mostly of harmonics of a fundamental frequency. In fact, the harmonic model parameters correspond to the harmonic samples of the short-term Fourier transform of a perfectly periodic signal. Otherwise, harmonic model has been shown to be capable of high-quality speech processing, particularly in the areas of speech coding and pitch and time-scale modifications in speech synthesis [11, 12]. It is further possible to change the magnitude and the phase of speech spectrum independently and directly by synthesis speech with altered model parameters. An equivalent equation is represented with equation (3) in [15]. At here, author uses minimum phase all-pole model to obtain sinusoidal parameter coding.

In training stage, [15] choose Gaussian mixture model (GMM) approach to implement a locally linear, probabilistic transformation function. The advantage of this model is the fast and accurate fitting of the few model parameters. The residual prediction (RP) system is implemented, which consists of a LPC parameter classifier and a LPC residual codebook. Each class of the classifier is associated with an entry in the codebook. Two data sets are necessary for training the RP system, the set of LPC parameters of voiced frames and the collection of associated LPC residual.

In conversion stage, the system analyzes a source speech and transforms the extracted features to an estimate of target speaker's LSF parameters. For each frame, spectral envelope is transformed by converting the predicted LSF parameter back to LPC filter coefficients. Finally, the transformed sinusoidal spectrum are set by the inverse warped LPC spectrum sampled at the harmonic.

In [16], training is performed by analyzing the source and target utterances using selective pre-emphasis. LSF vectors for each sub-band are obtained and the parameters of the first sub-band are used in the Sentence Hiden Markov Model (HMM), framework for acoustical alignment. The labels generated using the first sub-band is used for the remaining sub-bands. The codebooks are generated for each sub-band for the source and the target speakers. The parameters include LSFs for the vocal tract, instantaneous f_0 values as well as mean and variance of source and target pitch values, durations, and energy values. Two codebooks are generated for the source and the target speaker separately. The codebook entries include average line spectral frequencies, f_0, energy, and duration of each state.

In conversion stage, the sub-band codebooks are used for transforming each sub-band of the vocal tract spectrum separately. This requires the analysis of the input signal using selective pre-emphasis. The full-band excitation spectrum is processed separately for pitch scale modifications. Each sub-band of the vocal tract spectrum is converted separately for each sub-band with the corresponding source and target codebook. Note that the closest codebook entries are estimated using the first sub-band and same indices are used for all sub-bands. The output frame spectrum is obtained by multiplying the modified excitation spectrum with the vocal tract spectrum estimated. Prosodic modifications are used from STASC in [3].

In synthesis stage of [15], after obtained a transformed sinusoidal spectrum, a frame of the speech signal is computed by a weighted summation of harmonic sinusoids. The sinusoidal parameters are treated as constant within one frame of speech, discontinuities are avoided by an overlap-add (OLA) approach that eliminates the need to continuously vary the parameters to interpolate between sine-wave tracks. Beside that, OLA allows for a simple implementation of pitch and time scale modification. After all speech frames are computed, they are weighted, overlapped, and added. While in the synthesis stage of [16], the synthesis LP coefficient vectors are used to reconstruct the vocal tract spectrum. The synthesis LP coefficients can be a modified version of the analysis coefficients depending on the application. They are the target LP

coefficients estimated from codebooks for voice conversion.

E. Evaluation result of methods

The results of methods have been proved that transformed voice of new system has quality than systems use algorithm in [3, 4, and 14]. However, in [15] modeling and prosodic characteristic of the process of voice conversion did not mention, because that the quality could not equal when this implements. The system in [16] the author improve performance to the integration of the LPC residual but the representation of this was limited, thus the result reduce quality. Otherwise VC systems should importance to pitch detection and voiced/unvoiced decision as in [17] because all these parameters are speaker's characteristics.

3. APPLICATIONS

Voice conversion has numerous applications, such as the areas of foreign language training and movie dubbing. It is closely related to the process of speech synthesis, which usually refers to converting text into spoken language, and has many applications, especially relating to assistance for the blind and deaf [1]. Other areas in speech processing, such as speaker verification, have applications in security.

In movie dubbing area, by using the voice conversion, the actors or actresses will be able to speak in another language by their original voice. The movies must be dubbed in foreign countries and the voice characteristics of the original actors/actresses are lost. However, using the VC, the dubber's voice can be converted to the actor's original voice and the voice characteristics of actors/actresses can be preserved. Any actor/actress can speak any language when voice conversion technology is employed. In addition voice conversion can be employed in dubbing applications related to broadcasting, karaoke, Internet voice applications.

4. CONCLUSION

In this paper, we have presented analysis of approaches to design voice conversion systems, including general stages of voice conversion, acoustic characteristics will influence the quality of system and to compare two systems, which used two different speech models. They have been investigated to analyze, from that point, they can be chosen a study direction, combine with algorithm improvements to design voice conversion system. The main targets in our VC system are moved towards as: to improve the speech signal processing algorithms base on analysis above that creating VC system more naturalness, and to develop a real time VC system can be used dubbing area.

5. REFERENCES

[1] T. Toda, "Overview of Voice Conversion" 5th ISCA Speech Synthesis Workshop (SSW5), Tutorial, Pittsburgh, U.S.A., June 2004.

[2] Kuwabara, H., and Sagisaka, Y., "Acoustic characteristics of speaker individuality: Control and conversion", Speech Communication, vol. 16, pp. 165 -173

[3] L.M. Arslan and D. Talkin. "Voice Conversion by Segmental Codebook Mapping of Line Spectral Frequencies and Excitation Spectrum'', EUROSPEECH Proceedings, vol. 3, pp. 1347-1350, Rhodes Greece, September 1997.

[4] Turk, O., and Arslan, L. M., 2002, "Subband Based Voice Conversion", Proceedings of the ICSLP 2002, vol. 1, pp. 289-292, September 2002, Denver, Colorado, USA.

[5] Vích, R., Vondra, M , "Voice conversion based on Spectral Envelope Transformation", www.iiassvietri.it/school2004/School_Materials_1/ oral_contributions/Vondra_short_slides.pdf

[6] P. Zolfaghari and T. Robinson,"Formant Analysis Using Mixtures of Gaussians," in Proc. Int. Conf. of Spoken Language Processing, ICSLP96 (1996 Oct).

[7] Jonathan Malkin, Xiao Li and Jeff Bilmes, "A Graphical Model for Formant Tracking", ICASSP, philadelphia, March, 2005.

[8] *Taoufik En-Najjary, Olivier Rosec, Thierry Chonavel* "A voice conversion method based on joint pitch and spectral transformation", EN-NAJJARY-ICSLP-2004.

[9] Dimitrios Rentzos, Saeed Vaseghi, Qin Yan,"Voice Conversion through Transformation of Spectral and Intonation Features" ICASSP 2004.

[10] Valbret, H, Moulines, E. Tubach, J.P. "Voice Transformation using PSOLA Technique" ICASSP-92, 1992 Page(s): 145 – 148.

[11] Laroche, Y. Stylianou, E. Moulines, "HNS: speech modification based on a harmonic + noise model", ICSLP-1993.

[12] Yannis Stylianou, Applying the Harmonic Plus Noise Model in Concatenative Speech Synthesis, IEEE Transactions on Speech and Audio Processing, VOL. 9, NO. 1, January 2001.

[13] Rabiner, Lawrence R, and Schafer, Ronald W. "Digital Processing of Speech Signals". Bell Laboratories, 1978.

[14] Kain, and Macon – "Design and Evaluation of a Voice Conversion Algorithm Based On Spectral Envelope Mapping And Residual Prediction", In Proceedings of ICASSP '01 (Salt Lake City, UT, May 2001).

[15] A. Kain, "High resolution voice transformation", Ph.D. thesis, Oregon Health and Science University, Portland, OR, Oct. 2001.

[16] O. Turk, "New methods for voice conversion," in PhD. Thesis, Bogaziçi University, Istanbul, Turkey, 2003.

[17] Janer, L.; Bonet, JJ; Lleida-Solano, "Pitch Detection and Voiced/Unvoiced Decision Algorithm based on Wavelet Transforms", In Proceedings ICSLP 96, 1996.