On a Voice Conversion by using Prosodic Control

Jongkuk Kim, Min-Cheol Hong, Hernsoo Hahn

Department of Information and Telecommunication Engineering, Soongsil University, Seoul, 156-743, Korea
kokjk@hanmail.net

Abstract - Voice conversion is a method that aims to transform the input speech signal such that the output signal will be perceived as produced by another speaker .Speech synthesizers using voice conversion technologies allow developers to create more voices from a single database and users to personalize the synthesizer to speak with any desired voice after a training period. In this paper, we present the method that converts time and pitch scaling using spectral mapping and PSOLA technique with OLA. This new synthesis scheme allows very flexible modifications of the pitch-scale, the time-scale and the spectral envelope characteristics while producing high-quality speech output. This synthesis scheme is thus well suited to voice conversion. Further work will be conducted on a matching method to correspond well with each phonetic information, and larger corpora to assess the robustness of the method.

Index Terms - POSLA, Voice conversion, Prosodic, DTW, Mapping, Pitch, Modification

1. Speech Analysis

Voice conversion is a method that aims to transform the input (source) speech signal such that the output (transformed) signal will be perceived as produced by another(target) speaker. Controlling the synthesized speech is one of the most important issues in extending the application fields of TTS[1] systems. Especially, in amusement and education applications, generating multi-speaker's speech in good quality is strongly required. the parameter controlling and the parameter mapping had been the two major approaches of the voice transformation for speech synthesis.

A perfect voice transformation system should simulate the modifications of vocal-tract characteristics, prosody and glottal excitation. This task is clearly beyond the capability of current speech knowledge and technology. Simulations of changes in prosodic strategy are difficult to implement and are currently out of the scope of this study. We will mainly put the stress on the modifications of the acoustic parameters. In particular, we will focus on a technique which simulates speaker transformation by mapping the acoustic space of one speaker onto the acoustic space of another. Speaker characteristics will be specified through training.

Our method differs from techniques proposed previously by two major aspects: First, we use the PSOLA synthesis framework[4], which has been shown to yield a much more natural output than LPC vocoding does, in applications such as time-scaling or pitch-scaling.

Secondly, we propose and compare two new methods to learn the spectral mapping by DTW and LMR[7]. Dynamic Time Warping (DTW) was the former approach for alignment. It has been shown to be efficient for speaker adaptation in Dynamic Time Warping (DTW)[11] based isolated word recognition experiments. Both methods are based on the simple observation that an optimal transformation should depend on the acoustical characteristics of the sound to be converted. In this paper, we present the method that converts time and pitch scales using PSOLA[1] technique and spectral mapping. In section 3 we will describe the basic system used for the time and pitch conversion system. Section 4 will be dedicated to the experimental procedure and results. Finally, conclusions will be drawn in section 5.

2. Voice Conversion

2.1. Definition

Recent years have witnessed the rapid advances in the speech technology with the increasing number of products which use speech as a means in human-machine interaction. Naturally, speech recognition and TTS7 have been the priorities in research efforts directed at human-machine interaction. The ways to improve naturalness in human-machine interaction is becoming an important matter of concern. It combines the methods of automated knowledge and rule extraction in speech analysis and recognition with the methods of modification and construction in speech synthesis in the light of auditory perception and linguistics.

Voice conversion is a method that aims to transform the input speech signal such that the output signal will be perceived as produced by another speaker. It is hard to determine an optimal method for voice conversion that can achieve success for all possible speaker characteristics and combinations. Different voice conversion systems that employ different methods exist. It is common practice to model the speech waveform as a filter component driven by a source component. The filter corresponds to the vocal tract transfer characteristics which can be estimated using linear prediction (LP) methods. The parameters used in voice conversion that are extracted using LP methods include linear prediction coefficients. A general framework for voice conversion with basic building blocks is shown in Figure 1.

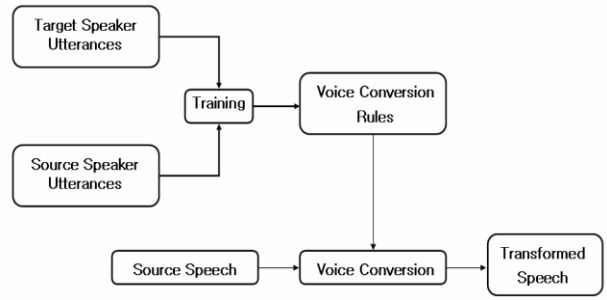

Fig.1. General novice conversion

Modeling and transformation of supra-segmental characteristics such as pitch, duration and energy are well studied and the algorithms developed provide the necessary framework for obtaining high quality output. Time Domain (TD) and Frequency Domain (FD) Pitch Synchronous Overlap-Add (PSOLA)[2] methods are commonly used for pitch and duration scaling. The source component is also referred to as the excitation and source excitation magnitude spectrum is modified to match target speaker characteristics. It is also possible to predict the target excitation from the target training utterances as described.

Principal learning methods were successfully applied for the purpose of training in voice conversion such as Vector Quantization (VQ), Hidden Markov Models (HMMs), Gaussian Mixture Models (GMMs), Artificial Neural Networks (ANNs) and Radial Basis Function Networks (RBFNs)[1,3]. The most straightforward way is to manually mark the corresponding acoustical events on the source and target recordings and to select the acoustical parameters corresponding to the current acoustical event. However, it is much time saving to use automatic methods for alignment before obtaining the mapping function.

Dynamic Time Warping (DTW) was the former approach for alignment. Once training has been completed, the voice conversion system has gathered sufficient information to transform any source speech signal into the target's voice in the transformation stage. This stage employs different methods to modify source speaker characteristics in order to obtain an output that sounds as close to the target speaker's voice as possible. Appropriate modification of source parameters and re-synthesis using the modified parameters produces the output (transformed) speech. The mapping function can be obtained using vector quantization field smoothing.

2.2 Problem

Voice conversion[7] is a fertile field for speech research as the problems of concern are related to almost all of the primary topics in speech processing. The problem of estimating the correspondence between the source and the target acoustical spaces is in fact a learning problem. First, the analysis stage of voice conversion is related to developing appropriate models that capture speaker specific information and estimating the model parameters which are closely related to acoustical modeling, speech coding, and psychoacoustics. Next, the relation between the source and target models must be determined and generalized to unobserved data.

The learning and generalization processes relate voice conversion with speech/pattern recognition, and machine learning. Finally, convenient methods must be employed for processing the source signal with minimized distortion and maximized resemblence of the output to the target speaker. These methods are also addressed in speech synthesis and coding applications. Robustness is perhaps the most important point of concern in voice conversion as the aim is to develop methods that perform well for a wide variety of source-target speaker pairs.

3. Prosodic Conversion System

The system of figure 2 used for time and pitch conversion involves the three following steps. At the analysis step, the target speech waveform is decomposed into two components: a flattened source signal containing much of the prosodic information, and a global envelope component which accounts for the resonant characteristics of the vocal tract transfer function together with the spectral characteristics of the glottal excitation.

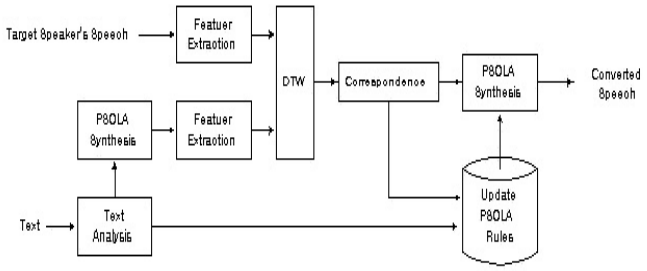

Fig.2. Time and Pitch conversion

In a second step, the two components of the signal are modified: prosodic parameters are altered by applying Time Domain-PSOLA[3] algorithms on the source signal; appropriate modifications of the spectral envelope are applied in the mean time. Each correspondence between the target speech and the synthesized speech by TTS system is extracted using DTW. Finally, the synthesis signal is obtained from the modified excitation source and the modified envelope. The converted speech is obtained by the rescaled time, peak to peak and pitch using PSOLA technique.

3.1 Envelope Feature

The features of target speech and synthesized speech are extracted from the target speech and the synthesized speech. This operation for spectral feature performed by using the all-pole filter. For time and pitch conversion of the synthesized speech[5], we extract time and pitch according to the correspondence. To determine the all-pole filter, we could have used standard AR estimation methods, such as the classic autocorrelation method. But whereas such techniques perform reasonably well for low-pitch male voices, it is well-known that their performances are poor when it comes to high-pitched voices.

To alleviate this problem, various methods have been proposed, among which the so-called "Discrete Cepstrum"[2]. This technique computes the cepstral envelope that matches the analysis short-term spectrum at given frequency points (harmonic frequencies in case of voiced speech). An all-pole filter that best fits (in the least squares sense) the discrete cepstral envelope is then obtained[8,11].

3.2 Prosodic Control

PSOLA provides a simple framework for performing prosodic modifications. In the basic TD-PSOLA system, prosodic modifications are performed directly on the speech

waveform6; the same approach can be applied as well on the excitation signal resulting from inverse filtering by a filter modeling the spectral envelope. Figure 3 is show that analysis by synthesis processing.

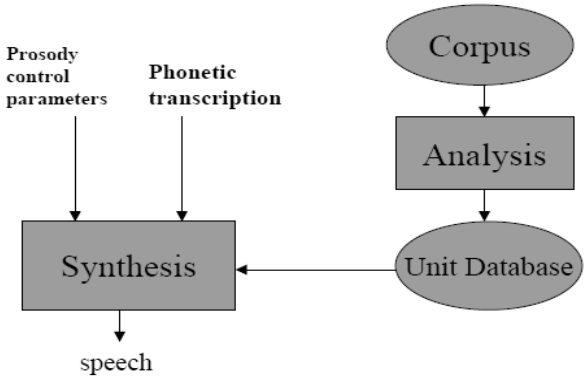

Fig.3. Analysis by synthesis processing

Learning how to modify the prosodic strategy is still an ambitious task. Therefore, to avoid artefacts, we simply copy the target prosody[8,12] of each sentence that is to be converted : the time axis of the reference sentence if first warped in order to align it with the target sentence. Once the evolutive time-scale and pitch-scale transformations are computed, the PSOLA algorithm is performed on the excitation source signal, as described above. Next we will describe how we learn and apply the spectral transformation. Figure 4 is show that synthetic speech by prosodic control.

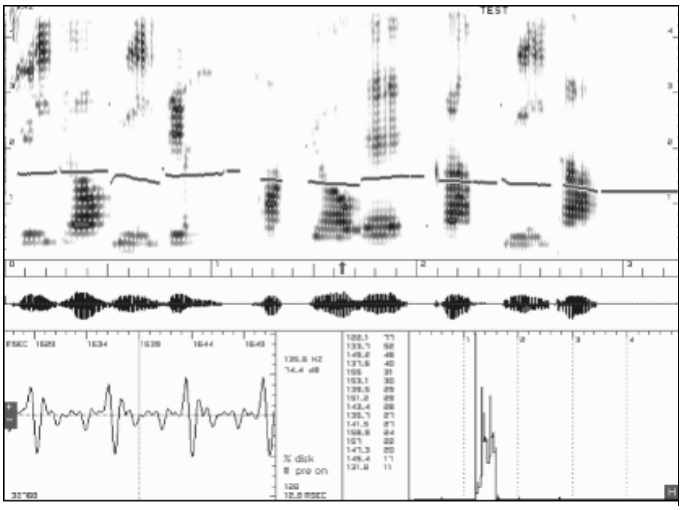

Fig.4. Synthetic speech by prosodic control

3.3 Spectral Conversion

For successive time instants corresponding to the correspondence by DTW[11] and LMR framework, we employ an over-lap adding(OLA)[4,9] reconstruction. Learning how to modify the prosodic strategy is still an ambitious task. Therefore, to avoid artefacts, we simply copy the target pitch of each target speech that is to be converted : the time axis of the source sentence if the first warped in order to align it with the target speech. Once the evolutive time-scale and pitch-scale transformations are computed with the correspondence, the PSOLA with OLA is performed. Figure 5 is show that training process.

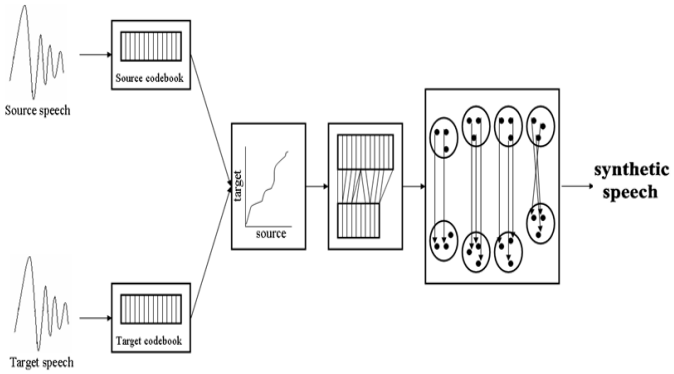

Fig.5. Training process

Two strategies have been investigated. The first one implements a well-known statistical analysis tool : the Linear Multivariate Regression. The second one alters the spectral envelope through a combination of frequency warping and amplitude scaling operations. A training vocabulary uttered by both reference and target speakers is first recorded. This corpus is then analyzed: a stream of cepstral feature vectors[13] is extracted from the speech signal. Note however that, at this stage, the analysis is not synchronised with the fundamental frequency. We rather use a fixed frame rate, set to 10 ms in all our experiments. Each word spoken by the reference speaker is then time-aligned with the corresponding word pronounced by the target speaker, using a standard Dynamic Time Warping technique(DTW).

In order to decrease the mapping complexity, we use a standard unsupervised clustering technique which divides the acoustic space of the reference speaker into non-overlapping classes. In our experiments, the number of classes was set to 64. The overall mapping is approximated by a finite set of elementary transforms, each of them being associated with a class. The second idea is to model the mapping in each class by a simple linear transformation, linear Multivariate Regression(LMR).

The LMR consists in finding the optimal linear transformation, that is the matrix which minimises a "mean square" error between the set of "reference" vectors and the set of "target" vectors. The normalised vectors are obtained through the linear transformation. The linear regression transformation minimises the mean square error between the two set of normalized vectors. In figure 6 is results by mapping processing.

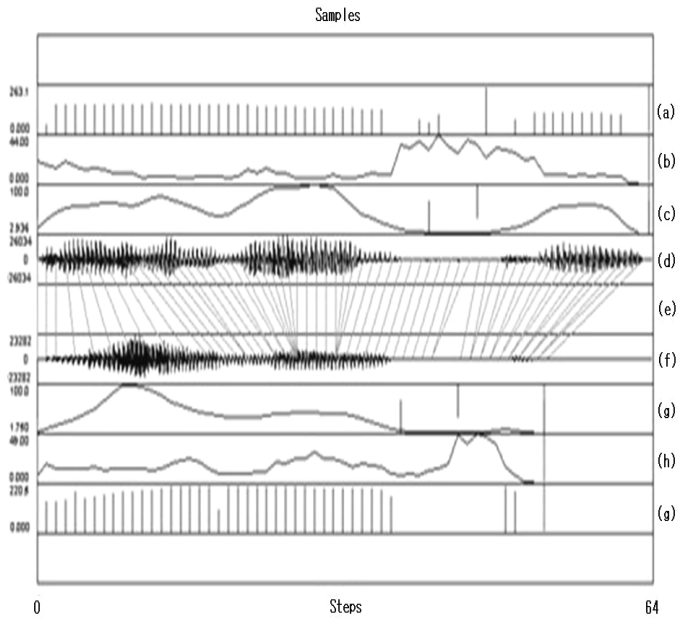

Fig.6. Result by mapping precedure

That is (a) pitch of source speech (b) normalized pitch contour of source using base-line pitch (c) differential pitch contour of (b), (d) source speech (e) Mapping processing source to target (f) target speech (g) normalized pitch contour of source using base-line pitch (h) differential pitch contour of (g), and (i) pitch of target speech.

It appears that the LMR performs better than the LMR speech is most often judged closer to the target speaker than the DTW one. DTW to modify the speaker's voice. However, the LMR transformed speech sounds smoother, but creates a kind of "mid-way timbre": the transformed speech is perceived as being "in between" the target and the reference speaker.

4. Experimental Result

4.1 Processing

This study focuses on various aspects of voice conversion and investigates new methods for implementing robust voice conversion systems that provide high quality output. These characteristics include the based spectral content, vocal tract, pitch, duration, and energy. The training corpus consists of recordings from 3 male speakers. Male to female voice conversion experiments have yet to be conducted. The vocabulary is composed of a symmetrical set of CVC logatoms with 10 oral vowels preceded and followed by the same consonants /b,d,g,p,t,k/, and a set of sustained vowels.

Each logatom is repeated 8 times. The first 6 repetitions are used for training the spectral transformation whereas the seventh and the eighth are used for testing. The total suration of the corpus is approximately 3 minutes. Data are digitized at 11 kHz. We evaluate the effectiveness of our tans formations by conducting listening tests which consist in presenting 3 natural reference and target signals. Listeners are asked to identify the speaker who might have pronounced the third stimulus. They are also asked to select the stimulus which most closely resembles the target speaker stimulus.

4.2 Result

The results clearly point out that the average level of the fundamental frequency is a crucial factor for speaker identification. The training speech consist of recordings from 3 male and female speakers. Male to male, and Female to female voice conversion experiments have to be conducted. Data are digitized at 11 kHz. To obtain correspondence we use DTW and backtracking between the source speech that is synthesized by TTS and the target speech. Distance measure for DTW is likelihood ratio rather than cepstrum.

We evaluate the effectiveness of this conversion system by listening tests which consist of presenting TTS source, target speech and converted speech. Converted speech is not better than the source or the target one, because mapping correspondence is not well matched on phonetic feature. A speaker specific intonational model is developed and evaluated both in terms of accuracy and voice conversion performance. Experimental results show that the proposed algorithm of prosodic control is capable of modifying pitch contours more accurately than non-controlled prosody.

Figure 7-8 is show to voice conversion for pitch contour in sentence unit conversion. That is (a) source, (b) target and (c) conversion speech in figure.

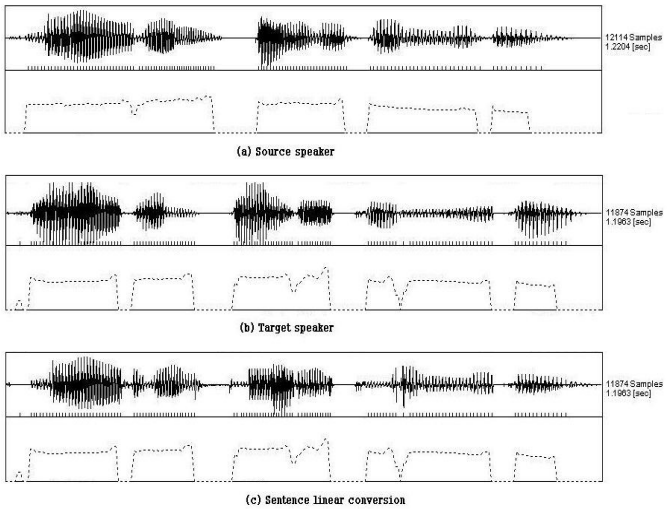

Fig.7. Male to male voice conversion

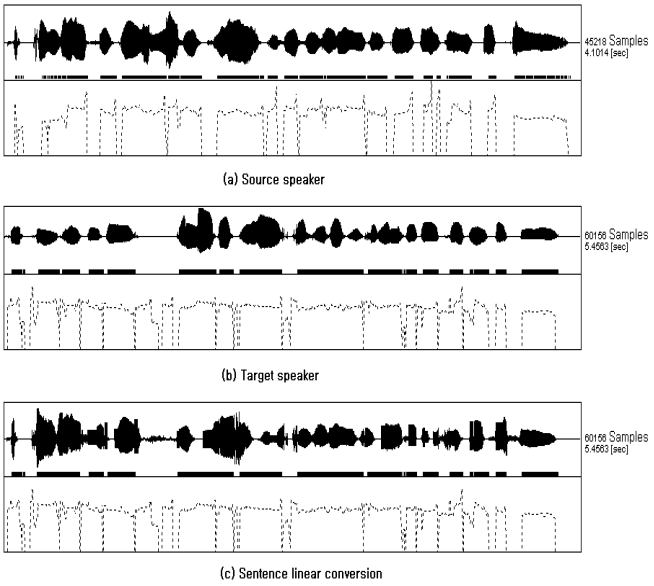

Fig.8. Female to female voice Conversion

5. Conclusions

In this paper, we have presented a time and pitch conversion system which combines the TD-PSOLA technique with OLA for modifying the prosody according to correspondence. This new synthesis scheme allows flexible modifications of the pitch-scale, the time-scale. However this synthesis scheme is not well suited to voice conversion, because correspondences are not well matched between the source phonetic information and the target one. This methods have been proposed and compared to learn the spectral transformation: the first one, the Linear Multivariate Regression projects the acoustical space of one speaker into the acoustical space of another, while the second one, the Dynamic Time Warping, aims at finding an optimal(and speech sound dependent)non-linear warping of the frequency axis.

Both techniques succeed reasonably well in modifying speaker identity, as proven by formal listening tests. Further work will be conducted on a matching method to correspond well with each phonetic information, and larger corpora to assess the robustness of the method. Hereafter we will study new methods for detailed estimation and modification of the pitch contour transformation, vocal tract modeling and we also investigate the design of voice conversion database such as multi-speaker for accurate speech segmentation alignment.

Acknowledgements

This research was supported by the MKE(The Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) support program supervised by the NIPA(National IT Industry Promotion Agency)" (NIPA-2010-(C1090-1021-0010)).

This research was supported by Next-Generation Information Computing Development Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology (No. 2012M3C4A7032182))

References

[1] W. B. Klejin et. al, Speech Coding and Synthesis, Elsevier Science B.V.(1995.)

[2] Wang Rae Jo, JongKuk Kim, and Myung Jin Bae, A Study 0n Pitch Detection in Time-Frequency Hybrid Domain, Springer-Verlag, Vol.-LNCS3406(2005).

[3] Oudeyer PY, The production and recognition of emotions in speech: features and algorithms. Int J Human-Comput Stud 59(1–2)(2003).

[4] Xuejing Sun, Voice Quality Conversion in TD-PSOLA Speech Synthesis, Proc. ICASSP'2000(2000).

[5] Mitra S., Digital Signal Processing, a Computer-based Approach, McGrawHill(2001).

[6] J.K. Kim, W.R, Jo, M.J Bae, A Study on Real Time Prosody Control of Speech, CCCT2003(2003).

[7] Tang, M., C. Wang and S. Seneff, Voice Transformations: From Speech Synthesis to Mammalian Vocalizations, Proc. of the Eurospeech2001 (2001).

[8] Turk, O. and L. M. Arslan, Subband Based Voice Conversion, Proc. of the ICSLP 2002(2002).

[9] Douglas-Cowie E, Campbell N, Cowie R, Roach P, Emotional speech: towards a new generation of databases. Speech Commun 40(2003)

[10] Huber, R., Ramoser, H., Mayer, K., Penz, H., & Rubik, M. Classification of coins using an eigenspace approach. Pattern Recognition Letters, 26(1) (2005).

[11] Schuller B, Seppi D, Batliner A, Maier A, Steidl S, Towards more reality in the recognition of emotional speech. In: Proc. int. conf. on acoustics, speech, and signal processing, (2007).

[12] N. Amir, Classifying emotions in speech: a comparison of methods, Proc. of Eurospeech2001(2001).

[13] M. Pitz, H. Ney, Vocal tract normalization equals linear transformation in cepstral space, IEEE Trans. Speech &Audio Processing(2005).

[14] Kondoz, A. M., Digital speech coding for low bit rate communications systems. John Wiley & Sons(1994).