VOICE TRANFORMATION USING PSOLA TECHNIQUE

H. VALBRET, E. MOULINES, J.P. TUBACH.

Télécom Paris, Dept Signal, CNRS-URA 820
46 rue Barrault, 75634 Paris Cedex 13, FRANCE

ABSTRACT

Whereas speaker adaptation has received much attention for speech recognition, few studies have been devoted to voice transformation for speech synthesis, despite the potential interest of such techniques. We propose in this paper a voice conversion system which combines the TD-PSOLA technique with a source-filter decomposition. The first technique allows prosodic modifications while the second enables spectral envelope transformations. Two approaches to learn spectral alteration are compared : the Linear Multivariate Regression (LMR) and the Dynamic Frequency Warping (DFW).

1. INTRODUCTION

Whereas the adaptation problem of ASR systems has raised a great deal of interest, on the contrary, until recently, little effort has been spent on voice conversion in the context of speech synthesis. Yet, potential applications are numerous : personification of text-to-speech synthesis systems based on acoustical unit concatenation, preservation of speaker characteristics in interpreting systems (Abe, 1991), preprocessing for speech recognition systems. These techniques could also be used to deal with related problems such as enhancement of helium speech signals.

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 (Abe et al, 1988),(Savic and Nam, 1991) by two major aspects :
First, we use the PSOLA synthesis framework (PSOLA stands for Pitch Synchronous Overlap and Add), which has been shown to yield a much more natural output than LPC vocoding does, in applications such as time-scaling or pitch-scaling (Moulines and Charpentier, 1990).
Secondly, we propose and compare two new methods to learn the spectral mapping. Both methods are based on the simple observation that an optimal transformation should depend on the acoustical characteristics of the sound to be converted. We therefore partition, in a first step, the acoustical space of the reference speaker into non-overlapping classes by means of a standard clustering technique. We then learn a transformation for each class. The first technique is the Linear Multivariate Regression (LMR), a well-known statistical analysis tool which aims at projecting the acoustical space of the target speaker onto the reference one. It has been shown to be efficient for speaker adaptation in Dynamic Time Warping (DTW) based isolated word recognition experiments (Tubach et al, 1990). The second method is based on Dynamic Frequency Warping (DFW) (Matsumoto and Wakita, 1986): it aims at learning an optimal non-linear warping function of the frequency axis to simulate changes of speaker characteristics. Contrary to LMR, which does not take into account the specific properties of the speech signal, DFW is more closely related to the acoustic theory of speech production, in that sense that changes in vocal tract length produce a non-linear transformation of formant frequencies.

In section 2 we will describe the basic components of the proposed analysis-synthesis system used for voice conversion. In section 3, we will then describe the two proposed methods for learning the spectral transformations. Section 4 will be dedicated to the experimental protocol. Finally, conclusions will be drawn in section 5.

2. PSOLA BASED ANALYSIS-SYNTHESIS SYSTEM

The system used for voice conversion involves the three following steps :

- At the analysis step, the 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.
- In a second step, the two components of the signal are modified : prosodic parameters are altered by applying Time-Domain-PSOLA (Moulines and Charpentier, 1990) algorithms on the source signal ; appropriate modifications of the spectral envelope are applied in the mean time.

0-7803-0532-9/92 $3.00 © 1992 IEEE

- Finally, the synthesis signal is obtained from the modified excitation source and the modified envelope.

2.1. Envelope extraction (figure 1)

This operation is performed at a pitch-synchronous rate : at successive time instants corresponding to the analysis pitch-marks used in the PSOLA framework, we determine an all-pole filter modelling the spectral envelope of the speech signal. We use a rather large order ($p = 30$, $f_s = 16$ kHz) to get correct formant bandwidths and amplitudes. We then obtain the source component by inverse-filtering the input signal : in order to avoid artefacts at LPC frame boundaries, we interpolate the reflection coefficients between successive models.

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" (Galas and Rodet, 1991). 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.

2.2. Prosodic modifications

PSOLA provides a simple framework for performing prosodic modifications. In the basic TD-PSOLA system, prosodic modifications are performed directly on the speech waveform ; the same approach can be applied as well on the excitation signal resulting from inverse filtering by a filter modelling the spectral enveloppe. We refer the interested reader to (Moulines and Charpentier, 1990) for further details. The process is illustrated in Figure 2, in the two simple cases of time-scaling (upper panel) and pitch-scaling (lower panel).

2.3. Prosodic modification for voice conversion

Learning how to modify the prosodic strategy is still an ambitious task. Therefore, to avoid artefacts, we simply copy the target prosody of each sentence that is to be converted : the time axis of the reference sentence is 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.

3. SPECTRAL TRANSFORMATIONS

In this section we will describe how we learn and apply the spectral transformation. 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.

3.1. Training procedure (figure 3)

A training vocabulary uttered by both reference and target speakers is first recorded. This corpus is then analyzed : a stream of cepstral feature vectors 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. Sakoe and Chiba local constraints (Sakoe and Chiba, 1978) are applied.

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 (Linde et al, 1980). 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 following step consists in modelling the transformations for each class. Our approaches to cope with this problem are presented in the next sections.

3.2. The Linear Multivariate Regression (LMR)

The first idea is to model the mapping in each class by a simple linear transformation.

Let $C^{r,q} = \{C_j^{r,q}\}$ $(j=1,\ldots,Mq)$ denote the set of reference spectral feature vectors belonging to the q^{th} class Q, M_q being the total number of vectors in the q^{th} class. To this set of vectors is associated, through DTW mapping, a set of vectors belonging to the acoustic space of the "target" speaker, denoted $C^{t,q} = \{C_j^{t,q}\}$ $(j=1,\ldots,Mq)$ The LMR consists in finding the optimal linear transformation, that is the matrix P_q which minimises a "mean square" error E between the set of "reference" vectors and the set of "target" vectors.

Let $m_j^{r,q}$ and $m_j^{t,q}$ be the empirical means of the j^{h} component of the spectral vectors of the reference and the target sets respectively. The normalised vectors are obtained through the linear transformation :

$$\tilde{C}_k^{r,q} = C_k^{r,q} - m_k^{r,q}.$$

The linear regression transformation P_q minimises the mean-square error between the two set of normalised vectors and is thus obtained as the solution of the following minimisation problem :

$$\sum_{k=1}^{M_q} \| \tilde{C}_k^{t,q} - P_q \tilde{C}_k^{r,q} \|^2.$$

The solution of this least square problem is straightforward : the matrix P_q is obtained by multiplying $\tilde{C}^{t,q}$ by the pseudo-inverse of $\tilde{C}^{r,q}$.

3.3. Dynamic Frequency Warping (DFW)

Let us draw our attention to the computation of the DFW path on a pair of log-magnitude spectra.

Let $S^r = \{S^r(k)\}$ $(k=1\ldots P)$ and $S^t = \{S^t(k)\}$ $(k=1\ldots P)$ denote the reference spectrum and the target spectrum respectively. $S^r(k)$ and $S^t(k)$ are the k^{th} log-spectral amplitude at the frequency $f_k = kf_s/2P$, f_k being the sampling frequency, P being the total number of frequency lags used to evaluate the Discrete Fourier Transform.

Let us consider a two dimensional matching space. The warping function w can be viewed as a path $C = \{C_k\}$ $(k=1,...P)$ where $C_k = (i(k),w(i(k)))$.

Computing this function consists in choosing among all possible paths in the plane, the path which minimises the frequency normalised distance measured between the target and the reference log spectra, defined as

$$D(S^t,S^r) = \min_C \left\{ \left[\sum_{k=1}^{P} d^{t,r}(i(k),j(k))\ \omega(k) \right] \left[\sum_{k=1}^{P} \omega(k) \right]^{-1} \right\}$$

where $\omega(k)$ is a weighting coefficient applied on the k^{th} portion of the path C defined so as to normalise the distance with respect to the length of the path and d is a spectral distorsion measure :

$$d^{t,r}(i,j) = \left| S^t(i) - S^r(j) \right| \quad .$$

In order to obtain a meaningful warping function, the dynamic procedure for frequency warping follows several constraining conditions.Transitions are restricted to :

$$C_{k-1} = \begin{cases} (i-1,j) \\ (i-1,j-1) \\ (i,j-1) \end{cases} .$$

Whereas i(0) and j(0) are assumed to be both equal to zero, the ending point can vary into a 1000 Hz interval.

In order to suppress irrealistic jumps in the warping function as well as to avoid excursions of the spectral warping function out of reasonable limits, the local slope of the warping path is constrained. The number of consecutive horizontal or vertical moves is limited.

Besides the vocal-tract length, the glottal source is another major speaker dependent characteristic. In attempting to eliminate the glottal effects, the spectral tilt (including both the glottal source spectrum and the vocal-tract transfer function) is estimated by fitting the envelope with a linear function of frequency, using a least-square regression line. Dynamic frequency warping is then performed on the residuals, the log magnitude spectrum minus the spectral tilt.

Let us now focus on the procedure which finds the transformation associated with a given class Q. For each pair of reference and target spectra belonging to Q, the DFW path is computed. We then obtain as many warping functions as pairs of spectral feature vectors within the class. It can be verified that, as expected, the warping functions obtained for vectors belonging to the same class look rather similar. Very few paths deviate from the main "beam", whose width is narrow enough to insure the consistency of the proposed method. To avoid artefacts, we then work out a median warping function. Mean spectral reference and target spectral tilts are also computed.

In further experiments, we propose to compute an optimal DFW path for the whole class. Each pair of reference and target spectra belonging to the given class Q will be taken into account in the frequency normalised distance. Hence the new DFW procedure will consist in finding the path C which minimises the new global distance :

$$D^q = \min_C \left\{ \left[\sum_{k=1}^{P} \omega(k) \sum_{(t,r)} d^{t,r}(i(k),j(k)) \right] \left[\sum_{k=1}^{P} \omega(k) \right]^{-1} \right\}.$$

3.4. Application to voice conversion

During voice conversion, the following procedure is applied to each analysis vector :

- The first step consists in finding the class Q to which the vector belongs.
- The transformation related to this subspace is then applied to the vector.

In the case of LMR, the vector is first normalised. The normalised transformed vector is obtained by multiplying the matrix P_q and the original normalised vector. The new vector is then "denormalised". A power spectrum is finally computed.

In the case of DFW, we first compute the log-magnitude spectrum. The appropriate warping function is then applied to this spectrum with the tilt removed. We finally add the corresponding target spectral tilt to the warped envelope.

- A LPC parameter set is extracted from the transformed spectrum. It is used to synthesize the converted signal.

4. Experimental procedure

4.1. Speech material

The training corpus consists of recordings from 4 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 duration of the corpus is approximately 3 minutes. Data are digitized at 16 kHz.

4.2. Experimental protocol

We evaluate the effectiveness of our transformations by conducting listening tests which consist in presenting 3 stimuli to 3 naive listeners. The first two stimuli are the natural reference and target signals. The third one is chosen randomly among :

- a speech token with modified prosody.
- a speech token with modified prosody and modified LMR spectra.
- a speech token with modified prosody and modified DFW spectra.

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 ressembles the target speaker stimulus.

4.3. Results

The results clearly point out that the average level of the fundamental frequency is a crucial factor for speaker identification. It appears that the LMR performs better than the DFW to modify the speaker's voice. LMR speech is most often judged closer to the target speaker than the DFW one. However, the LMR transformed speech displays some audible distortions. On the contrary, DFW 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. This means that the DFW does not suceed to remove all the speaker dependent spectral characteristics : this seems to confirm the work on vowel normalisation by Ainsworth et al (1984), where DFW fails to allow speaker independent vowel recognition.

5. Conclusion

In this paper, we have proposed a voice conversion system which combines the TD-PSOLA technique for modifying the prosody with a source-filter decomposition which enables spectral envelope transformation. 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.

Two 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 Frequency Warping, aims at finding an optimal (and speech sound dependent) non-linear warping of the frequency axis. Both techniques suceed reasonnably well in modifying speaker identity, as proven by formal listening tests. The LMR performs better than the DFW with respect : transforming voice quality but produces some audible distortions. Further work will be conducted on larger corpora to assess the robustness of the method.

References

M. Abe, S. Nakamura, K. Shikano, H.Kuwabara (1988), "Voice Conversion through Vector Quantization", ICASSP, 1988, pp. 655-658.

M. Abe (1991), "A Segment-Based Approach to Voice Conversion", ICASSP, 1991, pp. 765-768.

W.A. Ainsworth,K.K. Paliwal, H.M. Foster (1984), "Problems with Dynamic Frequency Warping as a Technique for Speaker-Independent Vowel Classification", Proc. Institute of Acoustics, 1984, Vol. 6, Part 4, pp. 303-306.

T. Galas, X. Rodet (1991), "Generalized Functional Approximation for Source-Filter System Modeling", Proc. Eurospeech 91, pp. 1085-1088.

Y. Linde, A. Buzo, R.M. Gray (1980), "An Algorithm for Vector Quantizer Design", IEEE Trans. Comm., Vol. COM-28, N° 1, January 1981, pp. 84-94.

H. Matsumoto, H. Wakita (1986), "Vowel Normalisation by Frequency Warped Spectral Matching", Speech Comm., Vol 5, N° 2, pp. 239-251.

E. Moulines, F. Charpentier (1990), "Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis using Diphones", Speech Comm., Vol. 9, N° 5/6, Dec. 1990, pp. 453-467.

S. Sakoe, S. Chiba (1978), "Dynamic Programming Normalisation for Spoken Word Recognition", IEEE trans. on ASSP., Vol. ASSP-28, N° 6, pp. 623-635.

M. Savic, I.H. Nam (1991), "Voice Personality Transformation", Digital Signal Processing 1, 1991, pp. 107-110.

J.P. Tubach, G. Chollet, K. Choukri, C. Montacié, C. Mokbel, H. Valbret (1990), "Adaptation au Locuteur de Sytèmes de Reconnaissance. Régression Linéaire Multiple et Perceptrons Multicouches", Traitement du Signal, Vol 7, N° 4, pp. 285-292.

Figures

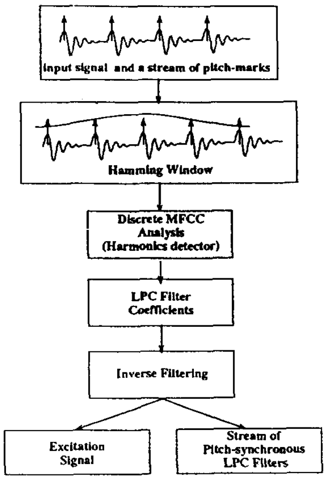
Input signal and a stream of pitch-marks
Hamming Window
Discrete MFCC Analysis (Harmonics detector)
LPC Filter Coefficients
Inverse Filtering
Excitation Signal
Stream of Pitch-synchronous LPC Filters

Figure 1 : Analysis of the input signal

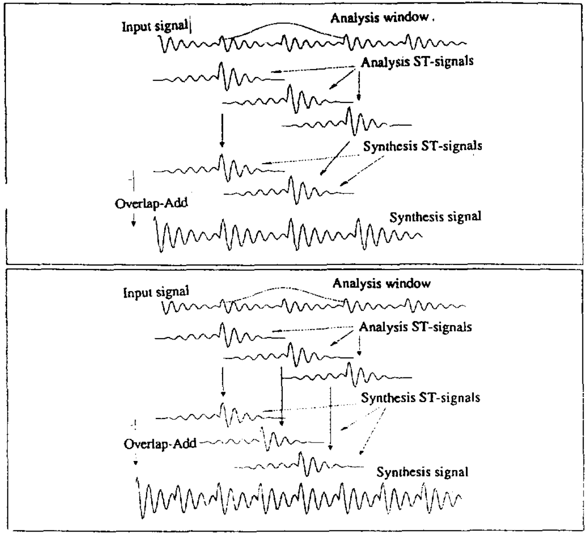

Figure 2 : Prosodic modifications, time and pitch scaling.

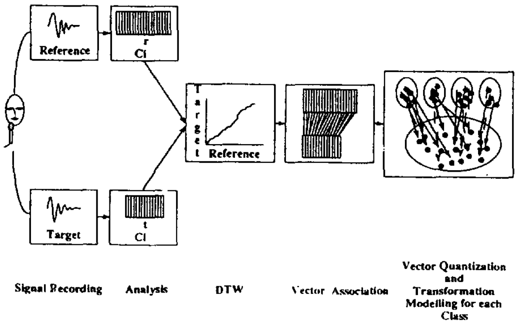

Figure 3 : Spectral Training.