Speech Communication 11 (1992) 175-187
North-Holland

Voice transformation using PSOLA technique

H. Valbret, E. Moulines and J.P. Tubach
Télécom Paris, Dept Signal, CNRS-URA 820, 46 rue Barrault, 75634 Paris Cedex 13, France

Received 26 September 1991
Revised 23 January 1992

Abstract. In this contribution, a new system for voice conversion is described. The proposed architecture combines a PSOLA (Pitch Synchronous Overlap and Add)-derived synthesizer and a module for spectral transformation. The synthesizer based on the classical source-filter decomposition allows prosodic and spectral transformations to be performed independently. Prosodic modifications are applied on the excitation signal using the TD-PSOLA scheme; converted speech is then synthesized using the transformed spectral parameters. Two different approaches to derive spectral transformations, borrowed from the speech-recognition domain, are compared: Linear Multivariate Regression (LMR) and Dynamic Frequency Warping (DFW). Vector-quantization is carried out as a preliminary stage to render the spectral transformations dependent of the acoustical realization of sounds. A formal listening test shows that the synthesizer produces a satisfyingly natural "transformed" voice. LMR proves yet to allow a slightly better conversion than DFW. Still there is room for improvement in the spectral transformation stage.

Zusammenfassung. In diesem Artikel stellen wir eine neue Technik zur Sprachtransformation vor. Die vorgeschlagene Struktur verbindet einen, von der Methode PSOLA (Pitch-Synchronous Overlap and Add) abgeleiteten, Synthesizer mit einem Modul zur Transformation der Spektralparameter. Der Synthesizer besteht aus einer klassischen Quell-Filter Zerlegung und einer prosodischen Modifikation des Erregungsignals mit Hilfe des TD-PSOLA (Time Domain PSOLA) Schemas. Zwei Wege der spektralen Transformation werden verglichen, die aus der Anwendung in der Spracherkennung entliehen sind: LMR (Linear Multivariate Regression) und DFW (dynamische Frequenzanpassung, Dynamic Frequency Warping). Eine einleitende Vektorquantisierung erlaubt, daß die Transformation von der akustischen Umsetzung der Töne abhängig wird. Ein formeller Hörtest zeigt, daß der Synthesizer eine zufriedenstellende Qualität der "transformierten" Sprache liefert. Zur Zeit ist die LMR Methode leistungsfähiger als die DFW, aber in der Transformation der Spektralparameter ist noch Verbesserungsspielraum vorhanden.

Résumé. Nous présentons dans cet article une nouvelle technique de transformation de timbre de la voix. Cette technique s'articule autour d'un synthétiseur dérivé de l'approche PSOLA (Pitch-Synchronous Overlap and Add) et d'un module de transformation des paramètres spectraux. Le synthétiseur allie décomposition source-filtre et modification prosodique du signal d'excitation par application de TD-PSOLA (Time Domain PSOLA). Deux approches de transformation spectrale, dérivées de techniques d'adaptation en reconnaissance de parole, sont comparées: la Régression Linéaire Multiple (LMR) et l'Alignement Dynamique en Fréquence (DFW). Une étape préliminaire de quantification vectorielle permet de rendre ces transformations dépendantes des réalisations acoustiques des sons. Un test d'écoute formel démontre que le synthétiseur permet d'obtenir une voix "transformée" d'un naturel satisfaisant. L'étape de transformation des paramètres spectraux est perfectible, la LMR donnant pour l'instant des résultats plus probants que la DFW.

Keywords. PSOLA analysis-synthesis; voice conversion; linear multivariate regression; dynamic frequency warping.

1. Introduction

Any human being is able to recognize the voice of a friend calling on the phone. The human ability to identify persons on the basis of their voice has long intrigued researchers. Much work has been dedicated to human and automatic speaker recognition (Hecker, 1971; Atal, 1976; O'Shaughnessy, 1986). These studies have been motivated by different applications such as security systems (access control) and formal identification (investigation). Another main objective is the ability to distinguish speaker-dependent parameters from message-dependent parameters. This research is closely connected to the automatic speech-recognition domain. The knowledge of

linguistic-dependent parameters which do not reflect the speaker's identity would enable speaker-independent automatic speech recognition (ASR). These parameters have not yet been found and large-vocabulary ASR systems perform better when adjusted to unknown users, thus requiring some normalization and adaptation modules (Lee, 1988). The adaptation problem of ASR systems has raised a great deal of interest, whereas, until recently, little effort has been spent on voice conversion in the context of speech synthesis. The latter, however, offers numerous potential applications: personification of text-to-speech synthesis systems based on acoustical unit concatenation (Carlson, 1991), 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.

Let us first review the sources of interspeaker variability. Schematically, it can be attributed to three main factors.

The first factor is closely related to physiology: the overall dimension of the vocal-tract as well as the relative proportions between the supra-glottal cavities (laryngeal, oral, nasal, ...) present important variations from one speaker to another. As demonstrated by Fant (1966, 1975) in the framework of the acoustic theory of speech production, the modifications of the dimension of the vocal apparatus influences upon vowel formant frequencies leading to an important dispersion of the vocalic triangle among speakers (see also (Nordström and Lindblom, 1975; Wakita, 1977; Matsumoto and Wakita, 1986)).

The second factor is linked to the dynamics of speech production: in learning speech mechanisms, each speaker has developed his own articulatory skill (or strategy) to produce each specific succession of phonemes. Such speaking habits are influenced by dialect and social environment. Strategy differences lead to variations of articulatory gestures which are dependent on each sound. Such differences show up in the temporal variations of speech characteristics of different individuals.

Finally, it is important to note that the acoustical characteristics of the glottal excitation can vary greatly.

A perfect voice transformation system should simulate the modifications of these various factors. This task is clearly beyond the capability of current speech technology and knowledge. Simulations of changes in prosodic strategy - fundamental frequency contour, segmental duration, energy - are difficult to implement and are currently out of the scope of this study. We will mainly stress the modifications of the segmental 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. The underlying hypothesis is that every speaker can be characterized by a "spectral print" in some parameter space. Speaker characteristics will be specified through training. Such an approach does not use any specific underlying model of the speaker.

This kind of approach was first studied for voice conversion by Abe et al. (1988), who propose to use vector quantization and the so-called codebook mapping, a well-known technique in speech recognition (Shikano et al., 1986). A learning step generates, separately, mapping codebooks for spectrum parameters, power values and pitch frequencies. Speech is then modified by applying these mapping codebooks and synthesized using an LPC vocoder. A similar approach was developed by Savic and Nam (1991), the codebook mapping being replaced by a multilayer neural network. The results were encouraging, the effectiveness of voice conversion procedure being clearly assessed by formal tests (Abe et al., 1988). However, the speech quality obtained in these pioneering works was limited, mainly due to the use of a standard LPC vocoder.

Our method differs from these techniques in 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 for deriving 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, as a first step, the

acoustical space of the reference speaker into non-overlapping classes by means of a standard clustering technique. We then obtain 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, based on Dynamic Frequency Warping (DFW), aims at obtaining an optimal non-linear warping function of the frequency axis to simulate changes of speaker characteristics. This approach was initially developed to deal with the problem of formant normalization in the context of speaker independent vowel classification (Ainsworth et al., 1984; Matsumoto and Wakita, 1986; Goncharoff and Chandran, 1988). 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 the sense that changes in vocal-tract length produce a non-linear transformation of formant frequencies.

Formal listening tests are designed to evaluate the efficiency of this new voice conversion system. Experiments are, for the time being, limited to a corpus of CVC logatoms, uttered by four male speakers. The choice of such a vocabulary is motivated by two reasons. First, the learning phase can be controlled more accurately (in particular, dynamic time warping is less subject to errors when performed on logatoms than on sentences). Second, prosodic differences mainly reduce to pitch and time scaling. Prosodic strategy does not appear in such short non-sense words.

In Section 2 we will describe the basic components of the proposed analysis-synthesis system used for voice conversion. Special emphasis will be put on the extraction of spectral envelope parameters. This step is crucial in order to obtain high quality prosodic and timbre transformations. 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 is depicted in Figure 1. It involves the three following stages:

- At the analysis stage, 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 the second stage, 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 as well.
- Finally, the synthesis signal is obtained from the modified excitation source and the modified envelope.

In this section, we will focus on the analysis-synthesis system which includes the analysis stage and the prosodic modifications. Spectral transformations will be detailed in the next section.

2.1. Analysis

The analysis system is illustrated in Figure 2.

At the analysis stage, the signal is split into a global envelope component and a flattened source signal. 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 (henceforth referred to as the analysis filter) modelling the spectral envelope of the speech signal. We use a rather large order ($p = 30$) to get correct formant bandwidths and amplitudes. We then obtain the source component by inverse-filtering the input signal: in order to avoid artifacts at LPC frame boundaries, we interpolate the reflection coefficients between successive models on a sample basis.

To determine the all-pole filter, we could have used standard AR estimation methods, such as the consecrated autocorrelation method. But whereas such techniques perform reasonably well for low-pitch male voices, it is well-known that their performance is poor with high-pitched voices (such as

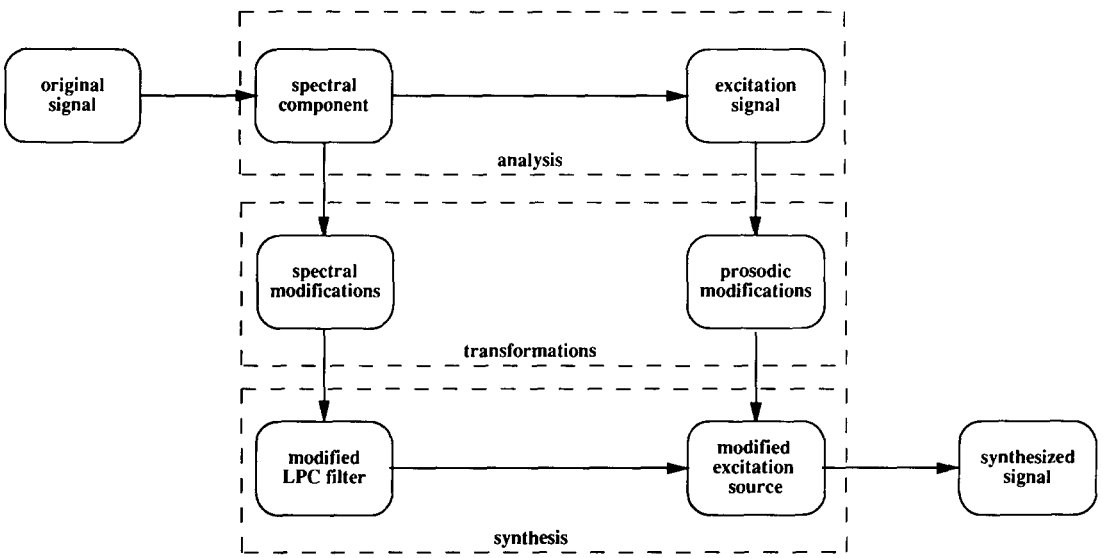

Fig. 1. Voice conversion system overview.

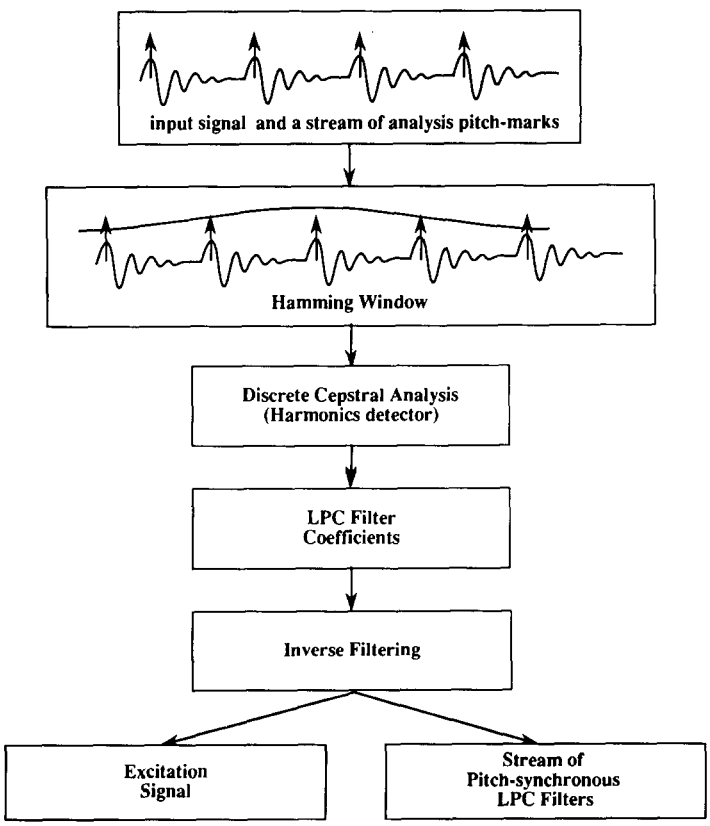

Fig. 2. Analysis of the input signal. A filter modelling the spectrum envelope is estimated at a pitch-synchronous rate by means of the discrete cepstrum technique. This filter is then used to determine an excitation signal by inverse filtering.

female voices), where they tend to model the pitch-harmonics rather than the vocal-tract and glottal characteristics. To alleviate this problem, various methods have been proposed. An attractive approach is to use a modelling technique based on a set of discrete frequencies (El Jaroudi and Makhoul, 1987; Galas and Rodet, 1991). These techniques exploit the fact that the discrete Fourier transform of the output of a linear time-invariant system excited by a periodic excitation is non-zero only for a specific set of frequencies, namely the harmonic frequencies. A reasonable way of estimating the transfer function of a linear system under periodic excitation is thus to find the parameters which approximate, with respect to an appropriate criterion, the harmonic structure of the output signal.

More specifically, the proposed solution proceeds as follows. First the n harmonic frequencies ω_j are determined from an initial pitch estimate f_0; these frequencies are defined as the maxima of the Short-Term Fourier Transform (STFT) amplitudes in the neighbourhood of the integer multiples of f_0. Then the harmonic amplitude $x_j = S(\omega_j)$ associated with this frequency is determined by picking up the amplitude value of the STFT. The window used for STFT estimation is of the Hanning type and its length has been set to 4 times the local pitch period, in order to get a main-lobe narrow enough to resolve the individual pitch-harmonics. The second step consists in finding the spectral envelope $P(\omega)$ that matches x_j at the given frequency points ω_j and minimizes the following criterion, proposed by Galas and Rodet (1990):

$$E = \sum_{j=0}^{n} \iint \mathrm{Pb}_j(\omega, x)(\log P(\omega) - \log x)^2 \, \mathrm{d}\omega \, \mathrm{d}x,$$

where each $\mathrm{Pb}_j(\omega, x)$ is a probability function which takes into account errors in the estimation of both the harmonic frequencies ω_j and the amplitudes x_j.

In our experiments, we have chosen a cepstral representation for $P(\omega)$, defined as

$$P(\omega) = \prod_{k=0}^{m} \mathrm{e}^{c_k \cos(k\omega)},$$

where m (=20) is the number of cepstral coefficients.

The motivation behind this representation is twofold:

- the transformation we first used for spectral mapping, the LMR, is obtained from a cepstral representation for which the "best" distance measure is the Euclidean one;
- preliminary experiments have demonstrated that the cepstral representation gives reasonably good estimates of formant bandwidths, especially in the case of female voices.

Once $P(\omega)$ is determined, we compute the coefficients for the all-pole filter $A(\omega)$ which will be used for inverse filtering operations. The method proceeds as follows: p (=30) autocorrelation coefficients are first estimated by computing the inverse discrete Fourier transform of the spectral envelope $P(\omega)$. The Levinson algorithm is then used to determine the m reflection coefficients solving the Yule–Walker equations associated with this autocorrelation sequence. It is well-known that this set of reflection coefficients define the all-pole approximation $1/A(\omega)$ of $P(\omega)$ which minimizes the Itakura–Saito distance within the set of stable and causal AR models (Markel and Gray, 1976). Other procedures could have been used for this purpose, such as the well-known model reduction method. This approach is much more complex and has not been retained since it does not affect the results in our application.

2.2. *Prosodic modifications*

PSOLA provides a simple framework for performing prosodic modifications (Moulines and Charpentier, 1990; Verhelst and Borger, 1991). In the basic TD-PSOLA system, prosodic modifications are performed directly on the speech waveform; the same approach can be applied as well to the excitation signal resulting from inverse filtering by a filter modelling the spectral envelope, as first shown in (Charpentier, 1988). Here, we briefly review the PSOLA framework and refer the interested reader to (Moulines and Charpentier, 1990) for further details.

The first step of the analysis process consists in decomposing the excitation waveform $x(n)$ into a stream of short-term signals (ST-signals), synchronized with the local pitch-period. These ST-signals are obtained by multiplying the signal by a

sequence of analysis windows. The windows, usually of the Hanning type, are centered around successive instants called pitch-marks, which are set at a pitch-synchronous rate on the voiced portions of the signal and at a constant rate on the unvoiced portions. The window length is proportional to the local pitch period, with the proportionality factor lying between 2 and 3 (2 for low-pitch male voices, 3 for high-pitch female voices).

In the second step, the sequence of analysis ST-signals is converted into a modified stream of synthesis ST-signals synchronized on a new set of time instants, the synthesis pitch-marks. These synthesis pitch-marks (together with a mapping associating the analysis and the synthesis pitch-marks) are determined in order to comply with the desired prosodic modifications. An excitation signal with modified pitch-scale and time-scale is then obtained by overlap-adding the stream of synthesis short-term signals.

The last stage is to synthesize the signal: this is done by filtering the modified excitation signal by synthesis filters synchronized with the synthesis pitch-marks, and derived from the analysis filters through a simple interpolation procedure.

The process is illustrated in Figure 3, 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 artifacts, we simply copy the target prosody of each sentence 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.

In a future work, we plan to apply a histogram-modification technique (Savic and Nam, 1991; Gonzales and Wintz, 1987) to alter the fundamental frequency scale: the average value of the pitch (and its relative dispersion) appears to be the most important clue for speaker identification.

3. Spectral transformations

In this section we will describe how we derive and apply the spectral transformation. Two strategies have been investigated. The first implements a well-known statistical analysis tool: the Linear Multivariate Regression. The second alters the spectral envelope through a combination of frequency warping and amplitude scaling operations.

3.1. Training procedure

The training procedure is presented in Figure 4.

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. These cepstral vectors are obtained using the procedure described in Section 2.1. Note however that, at this stage, the analysis is not synchronized with the fundamental frequency. We rather use a fixed frame rate, set to 10 ms in all our experiments (cf. Section 4).

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 (1978) are applied. Constraints are relaxed at the beginning and end of the words to cope with errors in word end-point setting. The standard euclidean distance between cepstral vectors is used. This operation provides us with a "mapping" between reference and target speaker spaces: each spectral feature frame of the target speaker is associated with one (or possibly more) frame(s) of the reference speaker according to the DTW "optimal" paths.

We could have used this complete mapping to perform voice transformation. However, this solution appears to be quite impractical, since it assumes the storage of the "whole" training set. We have preferred methods that intend to extract the most "significant" information from this mapping.

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). The overall mapping is approximated by a finite set of elementary transforms, each of them

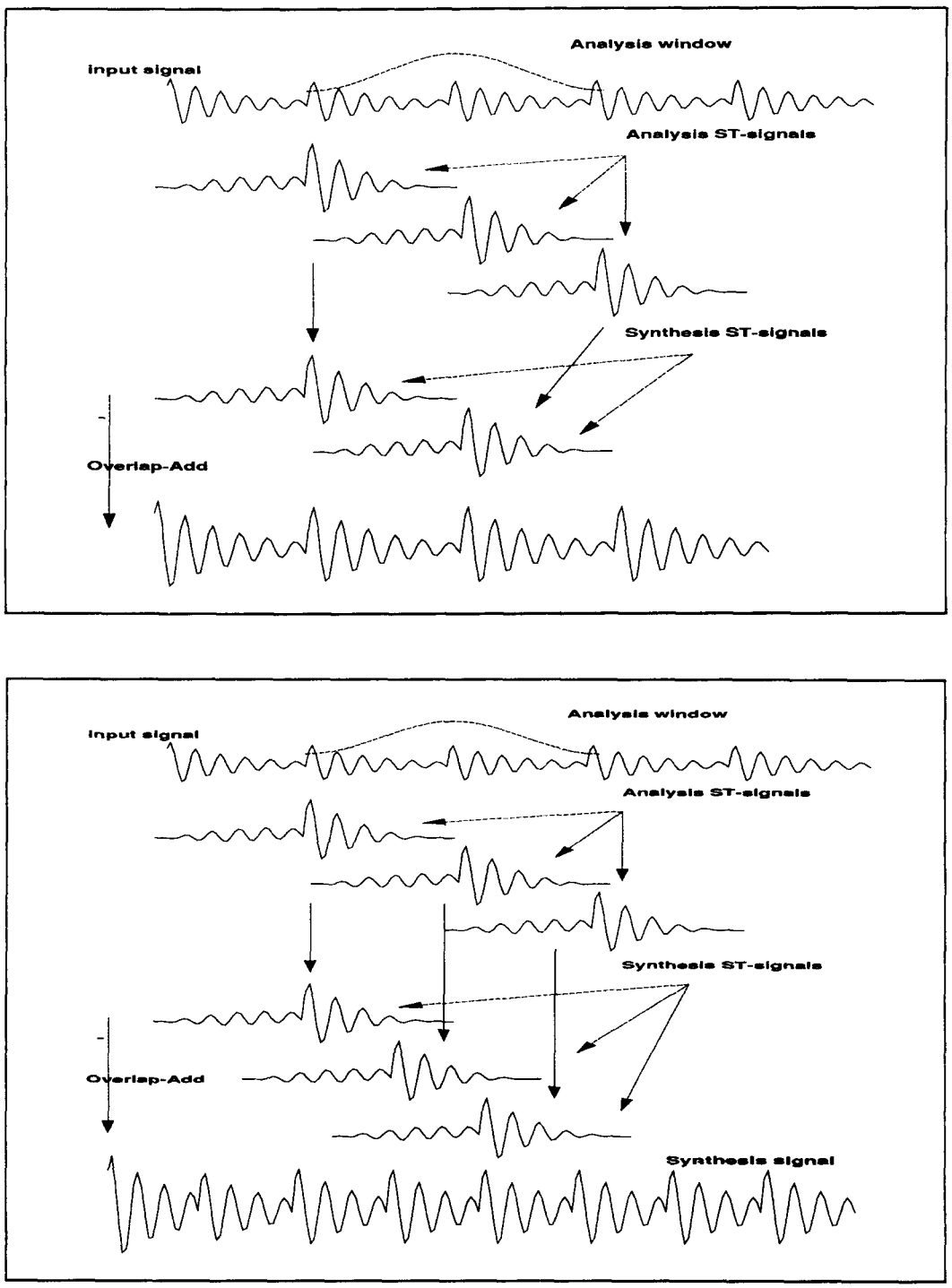

Fig. 3. PSOLA prosodic modification framework. The top panel illustrates a time-scaling operation, aiming at speeding up the speech signal without affecting the pitch contour. In that case, PSOLA boils down to selective elimination of the analysis ST-signals. Similar operations are performed when slowing down the speech signal: synthetic speech would be obtained by duplicating and interpolating several analysis ST-signals. The bottom panel displays a simple pitch-scale modification (without time-scale compensation): the mapping between synthesis and analysis pitch-marks is, in that case, on a one-to-one basis. Pitch-scale modification is achieved by modifying the time delay between pitch-marks.

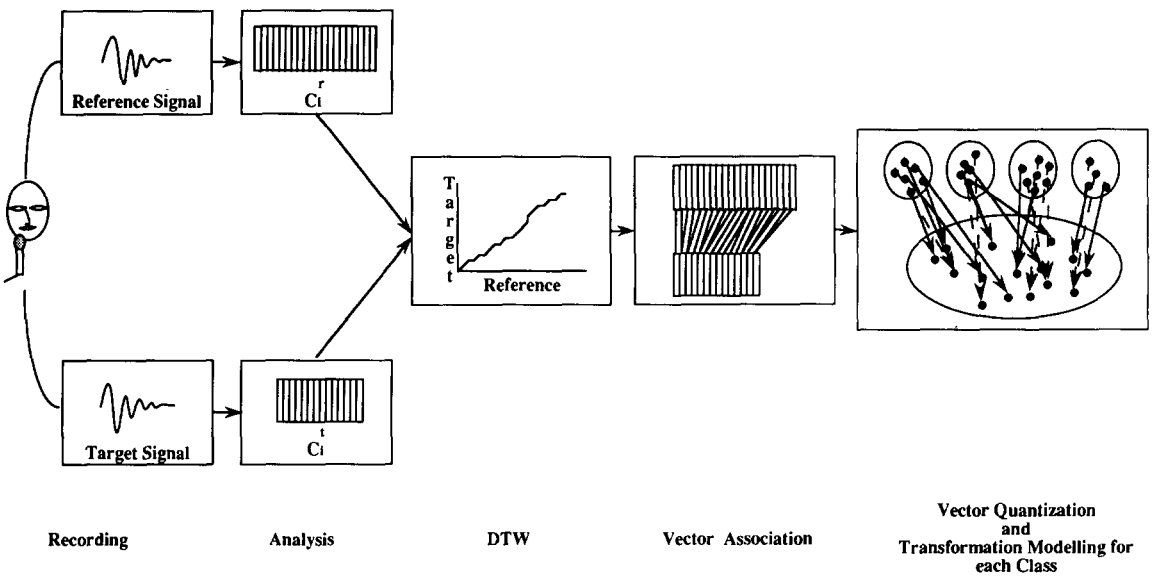

Fig. 4. Training procedure. Sentences uttered by the reference speaker and the target speaker are aligned using a DTW procedure with constraint-relaxation. This procedure allows to define a mapping between the two acoustic spaces of the reference and the target speaker. From this mapping, we learn the spectral transformation by first partitioning the acoustic space of the speaker by means of a VQ procedure and by approximating the transformation within each class.

being associated with a class. This is based on the assumption that a reasonable transformation should depend on the phonetic nature of the sound: it is difficult to imagine that the same spectral transformation is applied to a fricative and a vowel as well. Thus, the number of classes has to be large enough to handle the phonetic contexts of the corpus. Yet it should not be so important as to limit the number of transformations, hence of parameters, to be estimated.

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 $C^{r,q} = \{C_j^{r,q}\}_{j=1}^{M_q}$ denote the set of reference spectral feature vectors belonging to the qth class Q, M_q being the total number of vectors in the qth 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 $C^{t,q} = \{C_j^{t,q}\}_{j=1}^{M_q}$. The LMR consists in finding the optimal linear transformation, that is, the matrix P_q which minimizes 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 jth component of the spectral vectors of the reference and the target sets, respectively:

$$m_j^{r,q} = \frac{1}{M_q} \sum_{k=1}^{M_q} C_k^{r,q}(j).$$

The normalized 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 minimizes the mean-square error between the two sets of normalized vectors and is thus obtained as the solution of the following minimization 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 $\tilde{C}^{t,q} = \{\tilde{C}_k^{t,q}\}_{k=1}^{M_q}$ by the pseudo-inverse of $\tilde{C}^{r,q} = \{\tilde{C}_k^{r,q}\}_{k=1}^{M_q}$:

$$P_q = \tilde{C}^{t,q} \tilde{C}^{r,q\dagger},$$

where † denotes the pseudo-inverse.

Linear multivariate regression can also be interpreted as the search for the conditional expectation of the target spectral vector, knowing the reference spectral vector under the assumption that their distributions are jointly Gaussian.

3.3. Dynamic Frequency Warping (DFW)

The LMR operates on the vector of cepstral coefficients. On the other hand, the DFW operates directly on the spectral envelope. The method's aim is to find an "optimal" warping of the frequency axis in order to approximate the transformation between the target and the reference speaker.

We will first focus on the computation of the DFW path on a pair of log-magnitude spectra. We will then explain how to find a transformation for a given acoustic class.

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

We consider a two-dimensional matching space. At each point (i, j) on the grid, a spectral distortion measure is defined as

$$d^{t,r}(i,j) = |S^t_{(i)} - S^r_{(j)}|.$$

In this plane, a path is given by a set of points $\{C_k\}_{k=1}^P$, where $C_k = (i(k), j(k))$. In constructing this path in the plane, the frequency of the reference spectrum is "warped" depending on the target spectrum. It thus defines a warping function according to

$$j(k) = w(i(k)).$$

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

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

where $\omega(k)$ is a weighting coefficient applied on the kth portion of the path C.

In order to obtain a meaningful warping function, the dynamic procedure for frequency warping follows several constraining conditions. Both functions $i(k)$ and $j(k)$ should increase and respect some continuity conditions. In our implementation, transitions are restricted to

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

The weighting function is defined so as to normalize the distance with respect to the length of the path:

$$\omega(k) = i(k) - i(k-1) + j(k) - j(k-1).$$

In order to suppress unrealistic 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. Matsumoto and Wakita (1986) proposed a rather intricate scheme for that purpose. We have here implemented a simpler algorithm which limits the number of consecutive horizontal or vertical moves, as a function of the frequency. Available results on vowel formant frequencies for different speakers show that the variability of the first two formants is low: we have thus chosen a stricter slope constraint for low frequencies (up to 2000 Hz) than for high frequencies. Whereas $i(0)$ and $j(0)$ are assumed to be both equal to zero, the end point can vary into a predefined interval depending on the type of voice transformation: male to male or male to female. In fact, this interval should be a function of the vocal-tract length of both reference and target speakers. As this knowledge is not easily available, the interval length is set to 1000 Hz for male to male conversion and 2000 Hz for male to female conversion (the difference of male and female formant mean frequencies is about 18%).

Besides the vocal-tract length, the glottal source is another major speaker-dependent characteristic. It has been shown to severely affect the results of DFW. The spectral slope has been found to account for the major part of speaker differences in glottal source spectra. 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 to a given class Q. The preliminary clustering provides us with a set of reference cepstral vectors $\{C_j^{r,q}\}_{j=1}^{M_q}$. The DTW provides us with their corresponding target cepstral vectors $\{C_j^{t,q}\}_{j=1}^{M_q}$. For each analysis frame, a spectral envelope is obtained from the vector of cepstrum coefficients. The spectral tilt is approximated according to the procedure described above. It is then subtracted from the spectrum. For each pair of reference and target spectra, 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 (Figure 5) 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 ensure the consistency of the proposed method. To avoid artifacts, we then work out a median warping function.

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 normalized distance. Hence the new DFW procedure will consist in finding the path C which minimizes the new global distance:

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

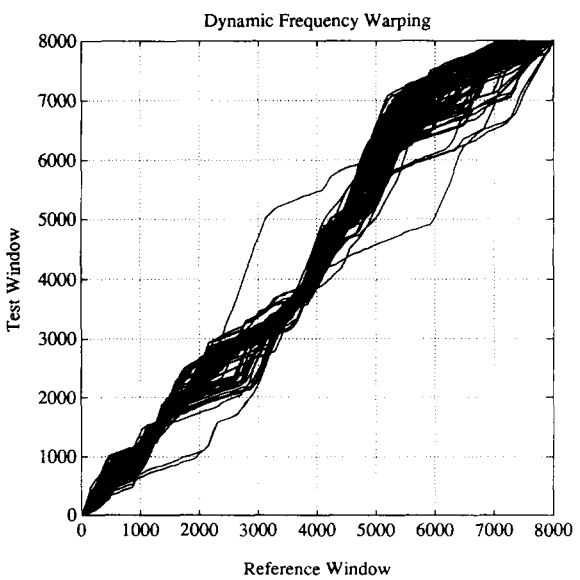

Fig. 5. Consistency of the DFW procedure. A beam of DFW optimal paths for a class is depicted.

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. This is done by finding the nearest code-vector according to the cepstral euclidean distance.
- The transformation related to this subspace is then applied to the vector.

  In the case of LMR, the vector is first normalized: the empirical mean $m_j^{r,q}$ computed during the training phase is subtracted from the analysis vector. The normalized transformed vector is obtained by multiplying the matrix P_q and the original normalized vector. The new vector is then "denormalized" by adding the empirical mean $m_j^{t,q}$. A power spectrum is finally computed.

  In the case of DFW, we first compute the log-magnitude spectrum. We then remove the spectral tilt $t^{r,q}(i) = m^{r,q} i + b^{r,q}$ estimated during the learning procedure. The appropriate warping function is then applied to this spectrum with the tilt removed. We finally add the corresponding target spectral tilt $t^{t,q}(i) = m^{t,q} i + b^{t,q}$ to the warped envelope.
- An 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 9 oral vowels /a,œ,ø,i,o,y,u,e,ɛ/ preceded and followed by the same consonants /b,d,g,p,t,k/ and a set of sustained vowels. Each logatom is repeated 7 times. The first 6 repetitions are used for training the spectral transformation, whereas the seventh is used for testing. For each speaker, the total duration of the corpus is approximately 2 minutes. Material was obtained in a low noise sound isolation booth. The data were digitized at 16 kHz and analysed following the procedure described above. The number of quantization classes is set at 64; this number appears sufficient to allow for the phonetic contexts of the corpus.

Some preliminary experiments were first conducted on another corpus: 100 short sentences – article adjective noun – uttered by both reference and target speakers. But some difficulties appeared as variability in the pronunciation of these sentences (insertion of silence, schwas or epenthetic vowels) leads to an imprecise time-alignment. The "mapping" defined by the DTW procedure was thus erroneous and the quality of the spectral transformation training was degraded.

4.2. Experimental protocol

Once the different transformations are derived using the training corpus, they are applied to the testing signal.

We evaluate the effectiveness of our transformations by conducting some listening tests.

The first experiment consists 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.

The energy contour and the time-scale are normalized by average in order to suppress their respective influence. A preliminary experiment has shown that these prosodic parameters severely affects the listeners' choice.

Listeners are asked to identify the speaker who could have pronounced the third stimulus. They are also asked to select the stimulus which most closely resembles the target speaker stimulus.

A second experiment is designed to evaluate voice quality by a pair-comparison listening test. Each combination of two stimuli among the three stimuli described above and the target stimuli are presented to listeners. Listeners are asked to explain the differences they perceive between both stimuli.

In the near future, we will conduct some further DTW recognition experiments to compare the adaptation capability of such techniques in both speech recognizer adaptation and voice quality conversion.

4.3. Results

The results of Experiments 1 and 2 clearly confirm the results obtained by other authors (Vaissiere, 1974) that the average level of the fundamental frequency is a crucial factor. Even on non-sense words, the average pitch-value seems to be the most important factor for speaker identification: spectral transformation without the correct pitch-modification results in a voice which is not recognized as the target voice; on the other hand, pitch modification without any spectral transformation significantly improves the speaker recognition rate.

It appears that LMR performs slightly better than DFW to modify a speaker's voice. Visual inspection of the spectral envelope shows that the LMR transformed envelope is much closer to the target envelope than the DFW envelope. In particular, DFW can only move formant positions without modifying their amplitudes, whereas LMR is able to cope with both formant frequencies and amplitudes as well. These observations are confirmed by listening tests: LMR speech is most often judged closer to the target speaker than DFW speech. However, LMR speech sometimes displays some audible distortions which have not been fixed yet. On the contrary, DFW speech usually sounds smoother, but creates a kind of "mid-way timbre": the transformed speech is perceived to be "in between" the target and the reference speaker. This means that the DFW does not succeed in removing all the speaker dependent spectral characteristics: this seems to confirm the work on vowel normalization of Ainsworth et al. (1984), where DFW fails to allow speaker independent vowel recognition.

5. Conclusion

In this paper, we propose 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.

Automatic simulation of changes in prosodic strategy are currently out of the scope of this study. Our present work focuses on spectral modifications.

Two methods are proposed and compared to obtain the spectral transformation: first, Linear Multivariate Regression (LMR) which projects the acoustical space of one speaker into the acoustical space of another; second, Dynamic Frequency Warping (DFW) which aims at finding an optimal (and phoneme dependent) non-linear warping of the frequency axis. Special care should be dedicated to the definition, recording and segmentation of the training corpus in order to derive efficient transformations. The performance of this system is assessed on a limited corpus of non-sense CVC (consonant, vowel, consonant) logatoms. Both techniques are shown to succeed reasonably well in modifying speaker identity. LMR appears to perform better than DFW with respect to transforming voice quality but it produces some audible distortions. Further work will be conducted on larger corpora in order to assess the method's robustness.

Acknowledgments

We would like to thank T. Galas for providing the discrete cepstral analysis software and X. Rodet for much fruitful discussion and encouragement. The basic PSOLA analysis-system has been provided by the Centre National d'Etudes des Télécommunications; we are greatly indebted to C. Sorin for allowing us to make use of it. Finally we would like to thank L. Mathan for the many improvements suggested on earlier versions of the article.

References

M. Abe (1991), "A segment-based approach to voice conversion", *Proc. Internat. Conf. Acoust. Speech Signal Process.*, Toronto, 1991, pp. 765–768.

M. Abe, S. Nakamura, K. Shikano and H. Kuwabara (1988), "Voice conversion through vector quantization", *Proc. Internat. Conf. Acoust. Speech Signal Process.*, New York, 1988, pp. 655–658.

W.A. Ainsworth, K.K. Paliwal and H.M. Foster (1984), "Problems with dynamic frequency warping as a technique for speaker-independent vowel classification", *Proc. Inst. Acoust.*, 1984, Vol. 6, No. 4, pp. 303–306.

B.S. Atal (1976), "Automatic recognition of speaker from their voices", *Proc. IEEE*, April 1976, Vol. 64, No. 4, pp. 460–475.

R. Carlson (1991), "Synthesis: Modelling variability and constraints", *Proc. Eurospeech 91*, Genova, Italy, pp. 1043–1048.

F. Charpentier (1988), Traitement de la parole par analyse-synthèse de Fourier: Application à la synthèse par diphones, Doctoral Thesis, Ecole Nationale Supérieure des Télécommunications.

A. El Jaroudi and J. Makhoul (1987), "Discrete all-pole modeling for voiced speech", *Proc. Internat. Conf. Acoust. Speech Signal Process.*, Dallas, 1987, pp. 320–323.

G. Fant (1966), "A note on vocal tract size factors and non-uniform *F*-pattern scalings", *Speech Transmission Lab. Quart. Progress Status Report*, Royal Inst. Techn., Stockholm, No. 4, 1966, pp. 22–30.

G. Fant (1975), "Non-uniform vowel normalisation", *Speech Transmission Lab. Quart. Progress Status Report*, Royal Inst. Techn., Stockholm, Nos. 2–3, 1975, pp. 1–19.

T. Galas and X. Rodet (1990), "An improved cepstral method for deconvolution of source-filter systems with discrete spectra: Application to musical sound signals", *Internat. Conf. Music and Computer*, Glasgow, September 1990.

T. Galas and X. Rodet (1991), "Generalized functional approximation for source-filter system modeling", *Proc. Eurospeech 91*, Genova, Italy, pp. 1085–1088.

V. Goncharoff and S. Chandran (1988), "Adaptative speech modification by spectral warping", *Proc. Internat. Conf. Acoust. Speech Signal Process.*, New York, 1988, pp. 343–346.

R. Gonzales and P. Wintz (1987), *Digital Image Processing* (Addison-Wesley, Reading, MA).

M.H.L. Hecker (1971), *Speaker Recognition, An Interpretive Survey of the Literature*, ASHA Monographs 16 (Amer. Speech and Hearing Assoc., Washington, DC, January 1971).

K.F. Lee (1988), *The Development of the SPHINX System* (Kluwer Academic Publishers, Norwell, MA, USA).

Y. Linde, A. Buzo and R.M. Gray (1980), "An algorithm for vector quantizer design", *IEEE Trans. Comm.*, Vol. COM-28, No. 1, January 1981, pp. 84–94.

J.D. Markel and A.H. Gray (1976), *Linear Prediction of Speech* (Springer, Berlin).

H. Matsumoto and H. Wakita (1986), "Vowel normalisation by frequency warped spectral matching", *Speech Communication*, Vol. 5, No. 2, June 1986, pp. 239–251.

E. Moulines and F. Charpentier (1990), "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", *Speech Communication*, Vol. 9, Nos. 5/6, December 1990, pp. 453–467.

P.E. Nordström and B. Lindblom (1975), "A normalization procedure for vowel formant data", *Internat. Congress Phonetic Sciences*, Leeds, August 1975, Paper 212.

D. O'Shaughnessy (1986), "Speaker recognition", *IEEE ASSP Mag.*, October 1986, pp. 4–17.

S. Sakoe and S. Chiba (1978), "Dynamic programming normalisation for spoken word recognition", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-28, No. 6, pp. 623–635.

M. Savic and I.H. Nam (1991), "Voice personality transformation", *Digital Signal Processing*, Vol. 1, pp. 107–110.

K. Shikano, K.-F. Lee and R. Reddy (1986), "Speaker adaptation through vector quantization", *Proc. Internat. Conf. Acoust. Speech Signal Process.*, Tokyo, 1986, pp. 2643–2646.

J.P. Tubach, G. Chollet, K. Choukri, C. Montacié, C. Mokbel and H. Valbret (1990), "Adaptation au locuteur de systèmes de reconnaissance. Régression linéaire multiple et perceptrons multicouches", *Traitement du Signal*, Vol. 7, No. 4, pp. 285–292.

J. Vaissiere (1974), "On French prosody", *Quart. Progress Report of R.L.E.*, MIT, No. 114, pp. 212–223.

W. Verhelst and M. Borger (1991), "Intra-speaker transplantation of speech characteristics: An application of waveform vocoding techniques and DTW", *Proc. Eurospeech 91*, Genova, Italy, pp. 1319–1322.

H. Wakita (1977), "Normalisation of vowels by vocal-tract length and its application to vowel identification", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-25, No. 2, April 1977, pp. 183–192.