



PITCH-SYNCHRONOUS WAVEFORM PROCESSING TECHNIQUES FOR TEXT-TO-SPEECH SYNTHESIS USING DIPHONES

Francis CHARPENTIER, Eric MOULINES

Centre National d'Etudes des Télécommunications
22301 LANNION FRANCE

ABSTRACT

We review in a common framework several algorithms that have been proposed recently, in order to improve the voice quality of speech synthesis using diphones [1-3]. These algorithms are based on a pitch-synchronous overlap-add (*PSOLA*) approach for modifying the speech prosody and concatenating diphone waveforms. The modifications of the speech signal are performed either in the frequency domain (*FD-PSOLA*), using the Fast Fourier Transform, or directly in the time domain (*TD-PSOLA*), depending on the length of the window used in the synthesis process. The frequency domain approach is capable of a great flexibility in modifying the spectral characteristics of the speech signal, while the time domain approach provides very efficient solutions for the real time implementation of synthesis systems. We also discuss the different kinds of distortions involved in these different algorithms.

INTRODUCTION

We describe in this paper a family of methods for modifying the prosody of natural speech while retaining a high level of naturalness. These methods are used to improve the voice quality of text-to-speech systems based on the concatenation of elementary speech units, including diphones, demi-syllables, or non-uniform units as proposed in [4]. Such concatenation-based synthesis systems require the use of rather large databases of acoustical units (corresponding for instance to 3 minutes of speech in the case of our French diphone system) and they generally rely on a coding algorithm to compress the size of the database. Therefore the synthesis stage generally involves two different processes:

- a decoding process: the waveform of the acoustical units must be reconstructed from their coded version;
- a concatenation process: the sequence of acoustical units must be concatenated after an appropriate modification of their intrinsic prosody.

The standard linear prediction method (*LPC*) using an excitation signal with a single pulse per pitch period integrates these two processes in a single step, the acoustical units being directly decoded and synthesized with their target prosody. This is possible since the fundamental frequency is an explicit parameter of the *LPC* model. However, such flexibility is counterbalanced by a limited voice quality.

In the methods we propose to raise the voice quality, we explicitly separate the decoding and synthesis processes, thus synthesizing as an intermediate step the original waveforms of the acoustical units. The prosodic modifications are then performed directly on the speech signal, using a *PSOLA* waveform processing scheme. The fundamental frequency is an explicit parameter since the algorithm works at a pitch-synchronous rate

and requires a preliminary pitch period labelling of the input waveforms. In particular, the synthesis process is synchronized with the pitch synthesis period, so that the algorithm controls simultaneously the value of the synthesized pitch and the duration of the synthesized signal.

In this paper, we first present the common *PSOLA* framework and two spectral interpretations of the synthesis process. We describe the time axis warping mechanisms for obtaining time-scale and pitch-scale modifications. Then we analyse successively the frequency-domain and time-domain algorithms, in relation with narrow-band and wide-band conditions of spectral analysis, and we explicit the acoustical distortions involved by each approach.

1. THE PITCH-SYNCHRONOUS OVERLAP-ADD SYNTHESIS FRAMEWORK

The *PSOLA* synthesis scheme involves the three following steps: an analysis of the original speech waveform in order to produce an intermediate non-parametric representation of the signal, modifications brought to this intermediate representation, and finally the synthesis of the modified signal from the modified intermediate representation. We present here these three steps, and then we detail the specific features of the time-scaling and pitch-scaling algorithms.

(1) Pitch-synchronous analysis:

The intermediate representation of the digitized speech waveform $x(n)$ consists of a sequence of short-term signals $x_m(n)$, obtained by multiplying the signal by a sequence of pitch-synchronous analysis windows $h_m(n)$:

$$x_m(n) = h_m(t_m - n)x(n)$$

The windows are centered around the successive instants t_m , called pitch-marks, that are set at a pitch-synchronous rate on the voiced portions of the signal and at a constant rate on the unvoiced portions. The windows $h_m(n)$ are typically of Hanning type and they are always longer than one single pitch period, so that neighbouring ST-signals always involve a certain overlap. Their lengths are usually set to be proportional to the local pitch period, with a proportionality factor μ ranging from $\mu = 2$, for short analysis windows, to $\mu = 4$, for longer ones. These values correspond respectively to 50% and 75% window overlapping factors. Such a window length proportionality rule can be expressed as follows:

$$h_m(n) = h\left(\frac{n}{\mu P}\right)$$

where $h(t)$ denotes the window with length normalized to unity and P denotes the local pitch period.

(2) Pitch-synchronous modifications:

The stream of analysis ST-signals $x_m(n)$ is converted to a modified stream of synthesis ST-signals $\tilde{x}_q(n)$ synchronized on a new set of synthesis pitch-marks f_q . Such a conversion involves three basic operations: a modification of the number of ST-signals, a modification of the delays between the ST-signals, and possibly, a modification of the waveform of each individual ST-signal. The number of synthesis pitch-marks f_q depends on the pitch-scale and time-scale modifications factors, denoted respectively β and γ . The delays $f_q - f_{q-1}$ between two successive pitch-marks must be equal to the local synthesis pitch period. The algorithm works out a mapping $f_q \rightarrow t_m$ between the synthesis and analysis pitch-marks, specifying which analysis ST-signal $x_m(n)$ is to be selected to produce any given synthesis ST-signal $\tilde{x}_q(n)$. In the *Time-Domain PSOLA (TD-PSOLA)* approach, the synthesis ST-signals are obtained by simply copying a version of the corresponding analysis signal, so that the algorithm consists of selecting a certain number of analysis ST-signals $x_m(n)$ and translating them by the sequence of delays $\delta_q = f_q - t_m$:

$$\tilde{x}_q(n) = x_m(n - \delta_q) = x_m(n + t_m - f_q)$$

In the *Frequency-Domain PSOLA (FD-PSOLA)* approach, the synthesis ST-signals are obtained by a frequency-domain transformation of the translated signal $x_m(n - \delta_q)$:

(3) Pitch-synchronous overlap-add synthesis:

Several overlap-add (OLA) synthesis procedures are available to obtain the final synthetic speech. For instance, the synthetic signal $\tilde{x}(n)$ can be obtained by means of the least-square overlap-add synthesis scheme [5]:

$$\tilde{x}(n) = \frac{\sum_q \alpha_q \tilde{x}_q(n) \tilde{h}_q(f_q - n)}{\sum_q \tilde{h}_q^2(f_q - n)}$$

where $\tilde{h}_q(n)$ denotes the sequence of synthesis windows. The additional normalization factor α_q is introduced to compensate for the energy modifications related to the pitch modification procedure. The spectral interpretation of this synthesis scheme is that it minimizes the quadratic error between the spectra of the synthesis ST-signals $\tilde{x}_q(n)$ and the corresponding short-time spectra of the synthetic speech $\tilde{x}(n)$.

An alternative synthesis scheme is the simple overlap-add procedure [6]:

$$\tilde{x}(n) = \frac{\sum_q \alpha_q \tilde{x}_q(n)}{\sum_q \tilde{h}_q(f_q - n)}$$

As in the least-squares synthesis formula, the denominator of this formula plays the role of a time variable normalization

factor: it compensates for the energy modifications due to the variable overlap between the successive windows. Under narrow band conditions, this factor is nearly constant. Under wide band conditions, it can also be kept constant, for particular choice of synthesis windows such as a window length equal to twice the synthesis pitch period. In such cases, and if we assume $\alpha_q = 1$, the synthesis formula reduces to the simplified overlap-add scheme:

$$\tilde{x}(n) = \sum_q \tilde{x}_q(n)$$

In this last formula, the synthetic signal appears as a simple linear combination of windowed and translated version of the original signal. All the operations involved are linear except the windowing operation. Consequently, when combining the *PSOLA* speech modification scheme with a linear filter such as a *LPC*-filter or a low-pass filter, the order of the operations may not be exchanged without modifying the behavior of the overall system.

This simplified overlap-add synthesis scheme leads to a second spectral interpretation of the synthesized signal, in the stationary case. If the original signal is periodic, the resulting synthesis ST-signals are all equal to a single prototype ST-signal $\tilde{x}_q(n) = \tilde{x}_0(n)$. If we assume a constant value for the synthesis pitch, the synthetic signal is obtained by the periodization of $\tilde{x}_0(n)$ at the new synthesis period. Such an operation is equivalent to sampling the spectrum $\tilde{X}_0(\omega)$ of the prototype signal $\tilde{x}_0(n)$ at the new harmonic frequencies of the synthesis signal. In other terms, the amplitude of the synthesis pitch harmonics are given by the amplitude of the short-term spectrum of the prototype signal at the synthesis harmonic frequencies.

(4) Time-scale modifications

Time-scale modifications of speech can be performed either in combination with pitch-scaling, or as a separate transformation in itself. In the latter case, no frequency-domain modifications are required and only the *TD-PSOLA* algorithm is used, independently of the size of the analysis window.

In the simple case where the time-scale modification factor γ is constant, the $f_q \rightarrow t_m$ pitch-mark mapping associates f_q with the analysis pitch mark t_m lying the nearest to the instant γf_q . When slowing down the speech signal, the pitch-mark mapping results into the repetition of several analysis ST-signals (Fig. 1). Inversely, a selective elimination of the analysis ST-signals leads to an acceleration of the speech signal.

In the cases where the speech is speeded up or slowed down by a factor 2, and when the analysis window length is equal to two pitch periods, the *TD-PSOLA* algorithm is analogous to the former *Time-domain Harmonic Scaling (TDHS)* algorithm [7], in which triangular windows are usually utilized. There is a greater difference between these algorithms for other time-scaling factors, since the *TDHS* scheme uses rather a pitch-proportional synthesis rate, while the *PSOLA* schemes use a truly pitch-synchronous rate. The *TD-PSOLA* algorithm must also be compared to the *SOLA* algorithm proposed in [8], which works in an asynchronous way at the analysis stage and uses an autocorrelation technique to resynchronize the synthesis ST-signals with the pitch period.

The acoustical distortions involved by the *TD-PSOLA* scheme are negligible for accelerating the speech rate and for slowing it down by moderate factors. However, when slowing down unvoiced portions by factors in the range of 2 and higher, the regular repetition of unvoiced ST-signals introduces a short-term autocorrelation in the synthesized signal, which is perceived as a tonal noise. A practical solution for a factor 2 is to reverse the time-axis of every repeated version of a ST-signal. Higher factors are not generally required in applications such as diphone synthesis. However, if necessary, it is possible (although costlier) to use a *FD-PSOLA* scheme in order to randomize the phase spectrum of repeated unvoiced ST-signals.

Much slighter tonal noise distortions may also be perceptible for voiced sounds such as voiced fricatives, since their spectrum usually combines voiced and unvoiced frequency regions. Because of the presence of the voiced component, simple solutions as time reversal are not feasible. A proper treatment of such sounds would require a *FD-PSOLA* approach, and a technique for identifying the unvoiced portions of the spectra.

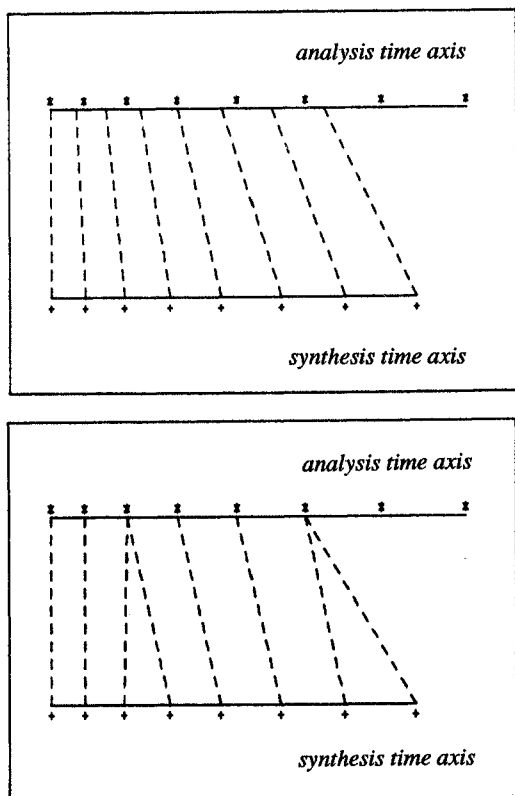


Fig.1 Time-scale modifications using the *TD-PSOLA* algorithm.

On this example, the speech is slowed down. The stars and plus signs represent respectively the analysis and synthesis pitch-marks. On the first figure (above), the dashed lines represent the time-scale warping function between the analysis and synthesis time axes, corresponding to the desired time-scale modification. On the second figure (below), they represent the resulting pitch-mark mapping $f_q \rightarrow t_m$, in this case duplicating two analysis ST-signals out of six.

(5) Pitch-scale modifications

Pitch-scale modifications are slightly more complicated, because they interfere with time-scale modifications. The simpler case is when the signal is to be simultaneously time and pitch-scaled by the same factors $\beta = \gamma$. There is then a one-to-one mapping between the synthesis and analysis pitch-marks: $f_m \rightarrow t_m$. This is illustrated in Fig.2, by the time warping function between the original time-scale and the second one below. But generally, independent time and pitch-scaling factors must be applied, so the mapping is not one-to-one and results into either a duplication or an elimination of some analysis ST-signals. As shown on Fig.2, this case can be seen as the combination of two transformations, the first one modifying the pitch and the time scale by the same factor β , the second one performing a compensatory time-scale modification by the factor γ/β . In fact, the two successive mappings that correspond to these two successive transformations combine into a single overall mapping, so that the time-scale and pitch-scale are performed simultaneously in one single step.

Finally, since pitch modification also involves the time-scaling mechanism, it should be noticed that slight tonal noises on voiced fricatives may not be avoided when simultaneously raising the pitch and slowing down the signal, since the inverse of the time-scaling factor γ/β may then become important.

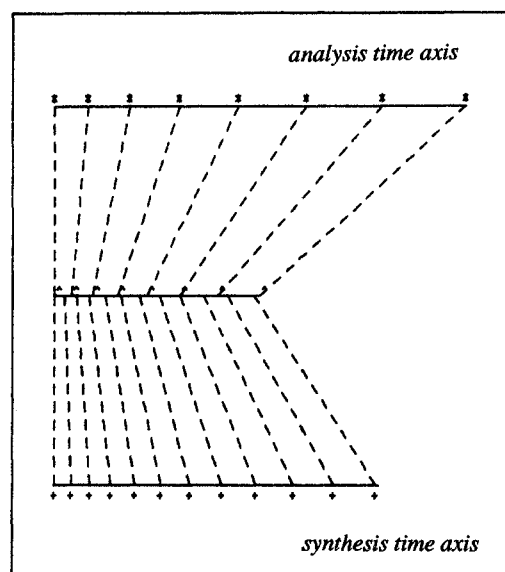


Fig.2 Pitch-scale modifications using *PSOLA* algorithms.

The stars, hat signs and plus signs indicate respectively the analysis pitch-marks, the modified pitch-marks on the virtual intermediate time axis, and the final synthesis ones. The dashed lines represent the time-scale warping functions between the successive time axes. A pitch and time-scale modification by the same factor is performed between the original and the intermediate time axes. A compensatory time-scale modification is performed between the intermediate and the synthesis time axes to achieve an overall independent time-scale modification.

2. PITCH MODIFICATIONS UNDER NARROW BAND CONDITIONS

(1) The TD-PSOLA approach

In the case of a long analysis window, i.e. under narrow-band conditions of spectral analysis, the TD-PSOLA approach leads to reverberant-sounding distortions, due to the discrepancy between the periodicity inherent in the synthesis ST-signal $\tilde{x}_q(n)$ and the synthesized pitch period. In fact, the synthesized signal $\tilde{x}(n)$ can be seen as reverberated version of the desired pitch-modified signal, involving echos with at a time variable delay equal to the original pitch period.

This effect can also be analyzed by using the spectral interpretation corresponding to the simplified OLA scheme. Since no modification is performed on the individual ST-signals, the prototype signal $\tilde{x}_0(n)$ is equal to the original ST-signal $x_0(n)$. Therefore the amplitude of a synthesis pitch harmonic is equal to the amplitude of the original spectrum at that particular frequency. Since the spectrum is obtained through a narrow band analysis, the pitch harmonics involve sharp spectral lobes, related to the main lobe of the analysis window spectral response. The amplitude of a synthesis pitch harmonic will be all the more affected as it departs from the original pitch harmonics, the worse case occurring when the synthesis harmonic lies right in the middle of two neighbouring pitch harmonics. If we denote F_0 the original fundamental frequency and if the pitch-modification factor satisfies the following constraint $1/2 < \beta < 3/2$, the frequency zones of maximum attenuation lie around the multiples of the following frequency:

$$f_{att} = \frac{F_0 \beta}{2 |\beta - 1|}$$

Such attenuations can be observed even for very mild modifications factors such as $\beta = 1.05$, since for an pitch value $F_0 = 100\text{Hz}$ the attenuation zone then lies in the frequency zone around $f_{att} = 1000\text{Hz}$.

(2) The FD-PSOLA approach [1,10]

The FD-PSOLA approach is more appropriate for pitch modifications under narrow-band analysis conditions, because it allows to adjust the pitch harmonics of the ST-signals so that their residual periodicity be consistent with the synthesized pitch value. To do this, spectral modifications are performed on each voiced analysis ST-signal, before feeding it to the PSOLA synthesis scheme. As for the PSOLA general framework, these modifications can also be divided into an analysis, a modification and a synthesis step. We now detail these steps for a given ST-signal.

(a) Frequency-domain analysis

The complex spectrum of the ST-signal is computed using the Fast Fourier Transform, a short-term spectral envelope is estimated and used to derive a flattened version of the complex spectrum. The spectral representation therefore corresponds to the classical source-filter model, consisting of a spectral envelope and of a spectral representation of the source.

(b) Frequency-domain modifications

To obtain a pitch-modified version of the spectrum, the source component of the spectral representation is modified so that the spacing between the pitch harmonics be equal to the new fundamental frequency. We present here two different methods to do this [10].

The first method is the spectral *compression-expansion* technique, illustrated in Fig.3. The frequency axis of the spectrum is linearly warped by the pitch-modifications factor β . To do this, the real and imaginary parts of the original spectrum are linearly resampled to obtained the modified DFT coefficients. As schematized in Fig.3, this method introduces a certain kind of spectral distortion: since it implicitly maintains a one-to-one mapping between the original and synthesis pitch harmonics, it modifies the fine details of the spectrum by carrying its local properties, such as harmonics phases and amplitudes, or the local voiced/unvoiced feature, from one zone of the spectrum to another.

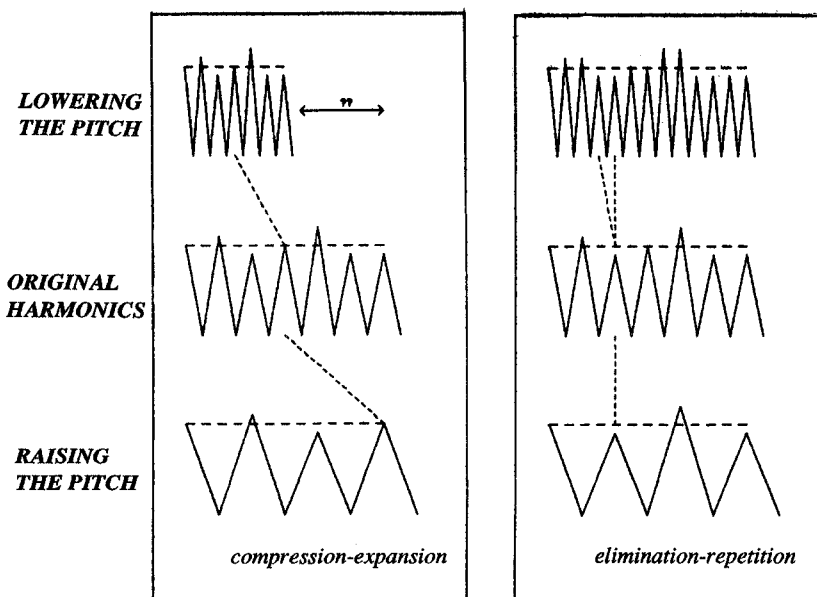


Fig.3 Pitch modification using the FD-PSOLA algorithm and two different spectral resampling methods: the *compression-expansion* method (left) and the *elimination-repetition* method (right). The spectral deviations with respect to an ideally flat source spectrum are enhanced by the horizontal dashed lines. The other dashed lines illustrate the correspondance between the synthesis and original pitch harmonics.

A related problem also occurs when lowering the pitch: an empty spectral zone appears in the high frequencies, requiring techniques for generation of an acceptable spectral distribution in those regions. Hopefully, in spite of these distortions, the method is capable of maintaining very good speech quality.

In fact, the *compression-expansion* pitch modification method works by resampling the original source spectrum at a modified set of frequency points, and it can be seen as a frequency-domain equivalent of a time-domain method developed in the context of *Residual Excited Linear Prediction (RELP)* synthesis: the residual resampling method [9]. The advantage of the *FD-PSOLA* approach is that it allows a time-variable resampling rate, difficult to realize in the time-domain.

An alternative method is the harmonics *elimination-repetition* technique, also illustrated in Fig.3. This method attempts to overcome the distortions of the previous method, by preserving the local spectral properties of the spectrum. The method has a similar organisation in the frequency domain as the *PSOLA* pitch-scaling methods in the time-domain. In case of lowering the pitch, the algorithm inserts new harmonics by repeating original ones, while reducing the spacing between them. In case of raising the pitch, harmonics are eliminated and the inter-harmonics spacing is augmented. This method has some drawbacks: it requires a precise estimation of the pitch value and it modifies the phase coherence between successive harmonics, that is implicitly maintained by the *compression-expansion* algorithm. In order to ensure good quality, the algorithm must compensate for the linear component of the phase spectrum, or equivalently, the original pitch-marks must be set at locations corresponding to the maximal source excitation. In fact, the flexibility of the *FD-PSOLA* approach allows to combine the two methods into a mixed method, using the *compression-expansion* scheme in the low frequencies, in order to preserve the phase coherence of the first harmonics, and the *elimination-repetition* scheme in the high frequencies, in order to avoid the need of high frequency generation.

Finally, we should mention that the *FD-PSOLA* approach can support a wide variety of modifications, such as:

- zero-phasing the low frequencies: to obtain an homogeneous phase distribution, while avoiding the buzziness due to a zero-phase spectrum in the high frequencies;
- modifications of the voice quality: this can be obtained through linear or non-linear modifications of the envelope component, or selective amplification or attenuation of individual pitch harmonics;
- sampling rate modifications: this approach provides an alternative to classical filtering techniques, and is very flexible for modifying the sampling rate by arbitrary ratios. However, the pitch synchronization is not crucial for this application, and an asynchronous *OLA* scheme can be utilized.

(c) Short-term signal synthesis

Finally, the modified spectral representation is converted back to a synthetic complex spectrum, and the synthesis ST-signal is obtained by inverse Fourier Transform.

Both interpretations of the *OLA* synthesis schemes are valid in the case of the *FD-PSOLA* approach. The least-squares synthesis scheme will perform a good match between the modified ST-spectra and the spectrum of the synthesized signal, since they will have a common fine spectral structure. On the other hand,

the simplified *OLA* method will also perform a sound modification, since the amplitude of the modified ST-spectra at the synthesis harmonic frequencies have the right target values, i.e. the values of the spectral envelope.

3. PITCH-MODIFICATIONS UNDER WIDE-BAND CONDITIONS

(1) The TD-PSOLA approach [2,11]

Under wide-band conditions, it is unnecessary to adopt the *FD-PSOLA* approach, since the short-term spectrum no longer has a fine structure and since it can be seen as an acceptable approximation of the spectral envelope. Therefore, when using the least squares synthesis scheme, the synthesized signal will have a spectral envelope similar to the original one. The algorithm is also capable of globally preserving the broad spectral distribution between voiced and unvoiced regions, for sounds such as voiced fricatives, although some modifications will appear at the overlap between voiced and unvoiced regions, due to the wide bandwidth of the analysis window.

If we adopt the spectral interpretation corresponding to the simplified *OLA* scheme, the prototype signal $\tilde{x}_0(n)$ is equal to the original ST-signal $x_0(n)$, as previously under the narrow-band conditions. On the other hand, the short-term spectrum $X_0(\omega)$ of the prototype signal is the convolution of the original pitch harmonics by the frequency response of the analysis window, the bandwidth of which is somewhat greater than the fundamental frequency. Consequently, the amplitude spectrum of the prototype signal is a smeared estimate of the true power spectrum envelope. The amplitude of a synthesis pitch harmonic will therefore involve a certain error with respect to the desired spectral envelope. This distortion can be considered mainly as a widening of the bandwidth of the formant resonances. Hopefully, such a distortion is slightly perceptible, since the differential limen for the perception of formant bandwidths is very high.

Phase distortions are also involved, since experimentally, the best speech quality is obtained when the pitch marks are synchronized with the main excitation of the vocal tract within the pitch period, namely the instant of glottal closure. However, degradations due to a wrong position of the analysis window can also be attributed to formants amplitudes distortions, since the short-term spectrum of the prototype signal may vary a lot with the window position.

(2) The LP-PSOLA approach [3,12]

In fact, it is possible to combine the *TD-PSOLA* approach and *LPC* techniques using a non-parametric or semi-parametric representation of the residual signal (*RELP*, *MPLPC*, *CELP*). This provides a way for compressing the information of the speech databases required by concatenative synthesis, as mentioned in the introduction. A straightforward method is to perform the *TD-PSOLA* modifications of speech after the *LPC* decoding process, thus synthesizing the original waveforms as an intermediate step. However, it is also possible, to perform the *TD-PSOLA* modifications directly on the residual signal, thus interchanging the short-term *LPC* filter and the *TD-PSOLA* synthesis scheme. Such a modification of the synthesis structure defines the *Linear Predictive PSOLA (LP-PSOLA)* approach. The *TD-PSOLA* and *LP-PSOLA* are illustrated in Fig.4, in the case of a multipulse *LPC* coding scheme (*MPLPC*). They are not

strictly equivalent since, as pointed out above, it is not possible to exchange the order of the *LPC*-filtering operation and of the windowing operations inherent in the *TD-PSOLA* process. This leads to somewhat different behaviors in the frequency domain.

A theoretical advantage of the *LP-PSOLA* approach is the better resolution achievable by *LPC* envelope estimation techniques than with the short-term Fourier transform implicitly used in the *TD-PSOLA* approach. This better spectral resolution can be exploited since the *TD-PSOLA* scheme involves less distortions when working on the residual waveform than on the speech itself. Indeed, the spectrum of the residual signal has a roughly flat spectral envelope. The slight variations due to residual resonant and anti-resonant frequency components may be approximated by an amplitude spectrum with broadened spectral peaks. The bandwidths of such broadened resonances are much wider than the main lobe of the synthesis window used in the wide-band *TD-PSOLA* scheme. Consequently, the *LP-PSOLA* algorithm is capable of maintaining the flatness of the modified residual spectrum and of reproducing the original spectral envelope in the synthesized speech. Furthermore, it can preserve the wide-band spectral deviations inherent in the residual signal. In that respect, the method is similar to the *FD-PSOLA* approach using the *elimination-repetition* technique to preserve the spectral deviations. However, it will not be capable of reproducing the sharp spectral deviations due to spectral zeros. Finally, the *LP-PSOLA* method can also be seen as an extension of previous cut-and-splice methods, where the cutting of individual pitch periods of the residual is replaced by a smooth windowing operation, allowing inter-period overlap.

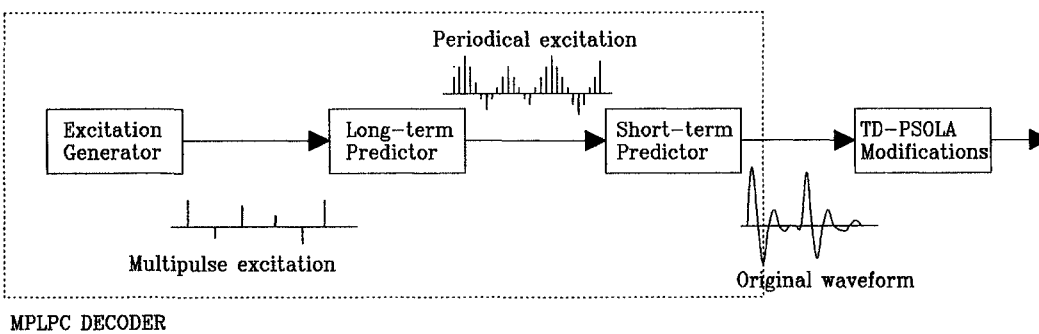
CONCLUDING REMARKS

The *PSOLA* algorithms presented in this paper provide the means for modifying the prosodic parameters of natural speech, such as pitch, duration and energy, while maintaining a very good voice quality in the transformed speech. The *FD-PSOLA* algorithm requires a high computational power (about 5Mflop/s.) and a high memory size for the digitized diphone dictionaries (about 5Mbytes). However, it provides a flexible laboratory tool since it allows a fine control of the speech spectrum. The *TD-PSOLA* algorithm is computationally very efficient and it can be combined in a flexible manner to speech compression techniques in order to reduce memory requirements. The *LP-PSOLA* algorithm can be seen as an optimized combination of the *TD-PSOLA* algorithm, and of specific *LPC* coding technique (such as *MPLPC* or *CELP*).

The speech quality gain brought by these algorithms has been evaluated through a formal test on French synthesized speech using diphones. This test was performed on 16 subjects and on 10 sentences comparing an *LPC* synthesizer with improved excitation, and early versions of the *FD-PSOLA*, *TD-PSOLA*, and *LP-PSOLA* algorithms. The four systems were compared two by two in A-B and B-A pairs for preference. The results have shown that all three algorithms perform much better than *LPC* synthesis, and that they are relatively equivalent among themselves [13].

Current developments include real-time implementation of a diphone synthesis system in two different configurations: using the *TD-PSOLA* algorithm (also called *PSOLA-KDG*) on a personal computer using a 386 processor, and using the *LP-PSOLA* algorithm (also called *PSOLA-MPLPC*, because it uses multipulse excited *LPC*) on a PC-based signal processing board using a TMS320C25 signal processor.

DIRECT APPROACH



LP-PSOLA APPROACH

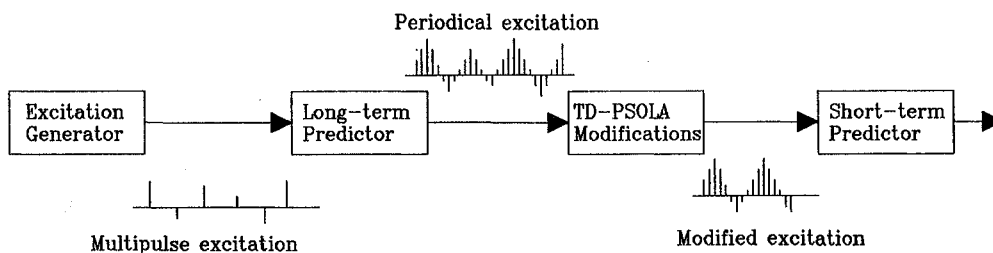


Fig.4 Two combinations of a MPLPC decoder and the TD-PSOLA technique for prosodic modifications

REFERENCES

- [1] F. Charpentier, E. Moulines, "Text-to-speech algorithms based on FFT synthesis", *Proc. Int. Conf. ASSP, New York, 667-670, 1988*
- [2] C. Hamon, E. Moulines, F. Charpentier, "A diphone synthesis system based on time-domain modifications of speech", *Proc. Int. Conf. ASSP, Glasgow, 1989*
- [3] E. Moulines, F. Charpentier, "Diphone synthesis using multipulse linear prediction", *Proc. FASE Int. Conf., Edinburgh, 1988*
- [4] Y. Sagisaka, "Speech synthesis by rule using an optimal selection of non-uniform synthesis units", *IEEE Int. Conf. ASSP, New York, 679-682*
- [5] D.W. Griffin, J.S. Lim, "Signal estimation from modified short-time Fourier transform", *IEEE Trans. ASSP, 32(2), 236-243, 1984*
- [6] J.B. Allen, L.R. Rabiner, "A unified approach to short-time Fourier analysis and synthesis", *Proc. IEEE, 65(11), 1558-1564, 1977*
- [7] D. Malah, "Time-domain algorithms for harmonic bandwidth reduction and time-scaling of speech signals", *IEEE Trans. ASSP, 27(2), 121-133, 1979*
- [8] Roucos S., Wilgus A., "High quality time-scale modification for speech", *IEEE Int. Conf. ASSP, Tampa, 493-496*
- [9] Roucos S., Wilgus A., "The waveform segment vocoder: a new approach for very low rate speech coding", *IEEE Int. Conf. ASSP, Tampa, 236-239*
- [10] F. Charpentier, "Traitement de la parole par analyse-synthèse de Fourier: application à la synthèse par dipphones", *Doctoral thesis, Ecole Nationale Supérieure des Télécommunications, 1988.*
- [11] C. Hamon, "Procédé et dispositif de synthèse de la parole par addition-recouvrement de formes d'ondes", *patent No. 8811517, 1988.*
- [12] E. Moulines, "Approches semi-paramétriques pour les modifications prosodiques et le codage de la parole de haute qualité: application à la synthèse par dipphones", *Doctoral thesis, in preparation, Ecole Nationale Supérieure des Télécommunications, 1989.*
- [13] J.P. Lucas, J.P. Roumiguere, F. Emerard, "Test de quatre algorithmes de traitement de signal pour la synthèse à partir du texte", *CNET internal report, NT/LAA/TSS/375, 1988.*