HIGH-QUALITY PROSODIC MODIFICATIONS OF SPEECH USING TIME-DOMAIN OVERLAP-ADD SYNTHESIS

Eric MOULINES, Christian HAMON, Francis CHARPENTIER
Centre National d'Etudes des Télécommunications
22301 LANNION FRANCE

RESUME

Un algorithme temporel permettant de modifier la prosodie de la parole tout en conservant le naturel du timbre a été proposé récemment dans le contexte de la synthèse par diphones [1]. Cet algorithme est fondé sur le méthode de l'addition-recouvrement synchrone de la fréquence fondamentale (PSOLA) de formes d'onde temporelle. Nous étudions dans cet article les raisons pour lesquelles cet algorithme conceptuellement simple permet d'obtenir des modifications de haute qualité. Nous montrons notamment que deux types de distortion sont possibles, selon la longueur de la fenêtre de synthèse: (1) élargissement de la bande passante des formants pour des fenêtres de synthèse courtes (2) effet de réverbération pour des fenêtres longues. Un compromis satisfaisant consiste à utiliser des fenêtres de longueur égale à deux fois la période fondamentale locale. Dans ce cas, la distortion de l'enveloppe spectrale est la plupart du temps peu importante d'un point de vue perceptif. Cet algorithme peut aussi être appliqué au résidu de prédiction linéaire, ou encore couplé à un codeur prédictif. Les distortions spectrales sont dans ce cas essentiellement liées aux problèmes d'estimation paramétrique de l'enveloppe spectrale du signal de parole.

ABSTRACT

A time-domain algorithm using the pitch-synchronous overlap-add *(PSOLA)* synthesis scheme has been proposed recently in the context of diphone synthesis, and it was shown to provide a very good sound quality [1]. In this paper, we analyse the reasons why the *PSOLA* synthesis scheme can be successfully applied to the speech waveform to produce high quality prosodic modifications of natural speech. The theoretical distortions brought by the algorithm are twofold: (1) a widening of the formant bandwidths for short synthesis windows; (2) a reverberation-like effect for longer windows. A practical tradeoff consists of using a synthesis window twice as long as the local pitch period. In that case, the formants distortion effect is almost not perceptible. Finally, it is shown the *Time-Domain PSOLA* algorithm can also be applied to the residual excitation signal in the context of multipulse linear predictive synthesis. The spectral distortions are slightly different in this case, whereas the resulting speech quality is also judged very good.

INTRODUCTION

A variety of methods based on the pitch-synchronous overlap-add *(PSOLA)* synthesis scheme have been proposed recently to perform high-quality prosodic modifications of natural speech. The main incentive for the development of these methods is the design of a good quality text-to-speech system based on the concatenation of speech segments such as diphones. In such a system, a simultaneous modification of the speech rate and of the pitch level are compulsory to ensure a natural prosody while retaining a natural voice quality.

In the general *PSOLA* framework, the original speech signal is transformed into a sequence of overlapping short-term signals (ST-signals), obtained by pitch-synchronous windowing. This stream of ST-signals is then modified to produce a stream of synthesis ST-signals. The final synthetic is obtained by use of an overlap-add synthesis procedure [2,3].

In the *Frequency-Domain PSOLA (FD-PSOLA)* approach, the synthesis ST-signals are obtained by frequency domain modifications of the analysis ST-signals [4,5]. The pitch harmonics of the short-term spectrum are adjusted so that the periodicity inherent in the synthesis ST-signal be consistent with the synthesized pitch value. The analysis window should be long enough to include at least three pitch periods, or equivalently, the bandwidth of the window must be less than the instantaneous fundamental frequency, leading to a narrow band analysis condition.

In the *Time-Domain PSOLA* approach (called *PSOLA/KDG, KDG* meaning time-domain overlap-add in the breton language), the frequency-domain operations of the *FD-PSOLA* approach are simply by-passed [1,6]. Therefore, the algorithm only involves two types of operations:
(1) elimination or duplication of the ST-signals;
(2) adjustment of the delays between the ST-signals to comply with the specified time-scale and pitch-scale modifications.

The computational complexity of the algorithm is very low since it does not need an explicit extraction the the short-term spectral envelope. Under narrow band conditions such as with the *FD-PSOLA* approach, the *PSOLA/KDG* method will lead to reverberant sounding distortions due to the discrepancy between the periodicity inherent in the synthesis ST-signal and the synthesized pitch value. However, the algorithm is also capable of working under wide band conditions, and the spectral distortions can then be interpreted as a spectral leakage broadening of the formant bandwidths. The optimal synthesis conditions result from a trade-off between these two kinds of distortions. In fact, for intermediate window lengths, ranging between twice to three times the local pitch period, the distorsions are only slightly perceptible and the overall speech quality is judged very good.

Finally, the *Linear Predictive PSOLA (LP-PSOLA)* approach can be seen as the application of the *PSOLA/KDG* approach to a pre-whitened version of the speech signal consisting of the linear predictive excitation signal [7,8]. In other words, *LP-PSOLA* algorithms result from the combination of linear predictive coding of the speech waveform (*RELP,MPLPC* or *CELP)* and *PSOLA/KDG* prosodic modifications, with a permutation of the short-term *LP* filtering and the *PSOLA/KDG* transformation of the speech signal. In fact, the *PSOLA/KDG* transformation can be seen as a linear combination of windowing operations applied to the original signal, and therefore it may not be permuted with a linear filtering operation. Consequently, the distortions involved by the *PSOLA/KDG* and *LP-PSOLA* methods cannot be considered as identical. However, both methods are capable of producing synthetic speech with a natural sounding quality, so that both methods are suitable for text-to-speech applications.

1. DESCRIPTION OF THE PSOLA/KDG ALGORITHM

The digitized speech waveform $x(n)$ is decomposed into a sequence of short-term overlapping signals $x_m(n)$. These ST-signals are obtained by multiplying the signal by a sequence of analysis windows $h_m(n)$, centered around the time origin:

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

The successive instants t_m, called pitch-marks, are set at a pitch-synchronous rate on the voiced portions of the signal. The stream of analysis ST-signals is processed to produce a stream of modified synthesis ST-signals $\tilde{x}_q(n)$ synchronized on a new set of synthesis pitch-marks $\tilde{t}_q$.

The algorithm determines simultaneously the synthesis pitch-marks according to the pitch-scale and time-scale modifications factors, and the mapping $\tilde{t}_q \rightarrow t_m$ between the synthesis and analysis pitch-marks. This mapping specifies which analysis ST-signal $x_m(n)$ is to be copied to obtain any given synthesis ST-signal $\tilde{x}_q(n)$, and the sequence of $\tilde{t}_q$ indicates the delays to be used between the short term synthesis signals:

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

If the signal is to be simultaneously time and pitch-scaled by the same factor β, there will be one-to-one mapping between the $x_m(n)$ and the $\tilde{x}_q(n)$. Generally, independent time and pitch-scaling factors must be applied, so that the mapping is not generally one-to-one, resulting into either a duplication or an elimination of the analysis ST-signals.

The synthetic speech $\tilde{x}(n)$ is then obtained by overlap-adding the stream of synthesis ST-signals, for instance by means of the least-square overlap-add (OLA) synthesis scheme [2]:

$$\tilde{x}(n) = \frac{\sum\limits_q \alpha_q \tilde{x}_q(n)\tilde{h}_q(\tilde{t}_q - n)}{\sum\limits_q \tilde{h}_q^2(\tilde{t}_q - n)}$$

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

Since the ST-signals are not modified, it is reasonable to use the same analysis and synthesis windows: $h_m(n) = \tilde{h}_q(n)$, so that the least-square overlap-add becomes a simple overlap-add procedure [3]. If we change slightly the notations and denote $\tilde{h}_q(m)$ the squared versions of the previous analysis windows, we obtain:

$$\tilde{x}(n) = \frac{\sum\limits_q \alpha_q x(n + t_m - \tilde{t}_q)\tilde{h}_q(\tilde{t}_q - n)}{\sum\limits_q \tilde{h}_q(\tilde{t}_q - n)}$$

The denominator $D(n)$ 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 can be shown to be nearly constant. This can be seen in the simple case where the successive windows are identical and the synthetic pitch-marks are equally spaced. Let T denote the synthesis period and $H(\omega)$ the Fourier transform of analysis windows assumed all identical to a fixed window $h(n)$. The denominator $D(n)$ is obtained by periodization of $h(n)$. It can therefore be recovered from its set of Fourier coefficients $H(2\pi k/T)$:

$$D(n) = 1/T \sum_{k=0}^{T-1} H(2\pi k/T)\exp(j(2\pi kn/T))$$

Under narrow-band conditions, $H(\omega)$ can be approximated as an ideal low pass filter with a cutoff frequency inferior to the spacing of the synthetic signal pitch harmonics, so that:

$$H(2\pi k/T) \approx 0 \quad \text{if } k \neq 0$$

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 [1]. When $D(n)$ can be considered constant, the synthesis formula reduces to the simplified overlap-add scheme:

$$\tilde{x}(n) = \sum_q \alpha_q \tilde{h}_q(\tilde{t}_q - n)x(n + t_m - \tilde{t}_q)$$

In this formula, the synthetic signal appears as the linear combination of windowed version of the original signal. So rigorously, when combining the synthesis scheme with a linear filter such as a *LP*-filter or a low-pass filter, the order of the operations may not be exchanged without modifying the behavior of the overall system. But practically, such modifications of the algorithm structure do not entail great modifications of the speech quality.

2. INTERPRETATION OF THE PSOLA/KDG ALGORITHM

Assume the signal is periodic with period $P_m = P_0$. Therefore, it is reasonable to choose a fixed synthesis windows $h_m(n) = h_0(n)$. The analysis ST-signals are all equal to a single prototype ST-signals:

$$x_m(n) = x_0(n) = h_0(t_0 - n)s(n)$$

According to the simplified overlap-add synthesis scheme, the synthetic signal is obtained by the periodization of $x_0(n)$ at the new synthesis period $\tilde{P}_q = P_0/\beta$ (β denotes the pitch-scale modification factor). The effect of this operation is equivalent to sampling the spectrum $X_0(\omega)$ of the prototype signal $x_0(n)$ at the new harmonic frequencies $\tilde{\Omega}_l = l\beta\Omega$. In other terms, the spectral envelope of the synthetic signal is identical to the short-term spectrum $X_0(\omega)$ of the original signal, with a spectral resolution determined by the analysis-synthesis window $h_0(n)$. Therefore, the amplitude of each harmonic of the synthetic signal is distorted with respect to the ideal spectral envelope. This corresponds to spectral distortions of variable kinds depending on whether wide-band narrow band synthesis conditions are utilised.

Under narrow-band condition, the synthetic signal can be shown to be the output of a comb filter whose frequency response consists of non-overlapping images of the frequency response $H(\omega)$ of the synthesis filter, shifted at the frequency of the original signal pitch-harmonics, and excited by the (assumed known) ideal pitch-modified signal. Since the synthesis signal periodicity does not match with the comb filter frequency response, the synthetic waveform energy can be severely affected: the more the synthetic signal pitch-harmonics depart from their original values, the more attenuated they are, and this may lead to a complete cancellation of the harmonics. It is easy to predict in which frequency zones such attenuation effects will occur. Assume that the pitch-scaling factor satisfies the constraint $1/2 < \beta < 3/2$. The frequency of maximum attenuation effect are integer multiples of the following frequency:

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

where $F_0 = 1/P_0$ denotes the fundamental frequency. Such attenuations can be observed even for very mild modifications factors such as $\beta = 1.05$, since the attenuation zone then lies in the region $f_{att} = 1000Hz$. For these reasons, narrow-band analysis conditions do not seem appropriate.

Under wide-band conditions, the central main lobe of the analysis window is (several times) greater than the fundamental frequency. Consequently, the amplitude spectrum $|X_0(\omega)|$ is a "smoothed" estimate of the true power spectrum envelope $S(\omega)$. When the *PSOLA/KDG* process is directly applied to the speech signal, a major discrepancy between these two spectrums lies in the bandwidth of the formant resonances. As the bandwidth of a formant is usually much less than the bandwidth of the synthesis window, $|H(\omega)|$, the shape of the synthetic signal formants depends primarily on the synthesis window main lobe: the bandwidths of these peaks do not reflect the "true" formant bandwidth. The problem appears to be more severe when the original pitch increases, as the synthesis window length, often chosen to be twice the synthesis period, decreases. In some adverse cases, a fusion of closely spaced formant peaks can even be observed. However, from a perceptual point of view, this mismatch of bandwidth is generally not so severe, as the diffrence limens measured for formant bandwidth is about 40 % for steady vowel (and should be much larger for continuous speech).

The effects of *PSOLA/KDG* modifications on the short-time phase spectrum is difficult to analyse. Even in the idealized stationnary, truly periodic case, the phase of a synthetic signal harmonic cannot be simply deduced from the original signal phase distribution. However, if the simplified "engineering" model for production of voiced speech signal is adopted (a linear time varying filter driven by a periodic train of unit sample), and if the analysis window is assumed symmetric, one can show that the synthetic signal phase distribution depends critically upon the distance D of the center of the window with respect to nearest pitch pulse:

$$\theta(\tilde{\Omega}_l) = \theta_0(\tilde{\Omega}_l) + (\omega_i - \tilde{\Omega}_l)D \qquad 0 \le D < P_0$$

where $\tilde{\Omega}_l$ denotes the synthesis harmonic frequencies, ω_i the frequency of the formant peak closest from the synthetic harmonic, and $\theta_0(\tilde{\Omega}_l)$ is a phase distribution function related to the original phase. This result suggests that minimum phase distortion is achieved when the window is synchronized with the main excitation of the vocal tract within the pitch period, namely the instant of glottal closure. Experimentally, when the window center is shifted with respect to the presumed glottal closure instant, the synthetic speech is first altered and then began to degradate noticiably (when the shift exceeds 30 % of the pitch period). Yet this degradation cannot be only attributed to the synthetic signal phase distribution, but probably also results from formants amplitudes distortions. As the window length is short, the magnitude of the short-term Fourier Transform $X(t, \omega)$ for a given frequency ω, preserves some of the pitch periodicity as function of time. Moreover, the magnitude $|X(t, \omega)|$, is behaves in time as a decaying exponential with a time constant determined by the bandwidth of the nearest formant in the neighborhoud of ω. Consequently, the formant amplitudes reconstructed in the synthesis signal depend on the exact window position within the pitch period. Indeed, we have observed experimentally that an improper synchronization of the synthesis window affects more heavily the amplitude of high-frequency formants, which have larger bandwidths.

3. THE LP-PSOLA SYNTHESIS SCHEME

An possible extension of the *PSOLA/KDG* approach is to apply this time domain pitch modification to a pre-whitened signal. For instance, the prosodic modifications can be combined with a predictive coding system like *MPLPC* or *CELP:* the *PSOLA/KDG* modifications are then applied to the excitation signal driving the LPC synthesis filter. We therefore introduce the *LP-PSOLA* approach in which an *LP*-filtering operation is performed as the final step of the algorithm. The *PSOLA/KDG* and *LP-PSOLA* approach are not strictly equivalent since, as pointed out above, it is not possible to exchange the order of the *LP*-filtering operation and of the windowing operations inherent in the *PSOLA/KDG* algorithm. This leads to somewhat different behaviors in the frequency domain.

A major advantage of the *LP-PSOLA* approach is the better resolution achievable with auto-regressive spectral estimation techniques over that achievable with the classical short-term Fourier transform estimation implicitly used in the *PSOLA/KDG* approach. In fact, the splitting of the speech spectrum into a source and a filter components provides an additional degree of freedom by allowing narrow bandwidth conditions for the spectral envelope estimation and wide bandwidth conditions for the prosodic modifications.

Furthermore, it can be seen that the *LP-PSOLA* algorithm is capable of preserving the wide-band spectral deviations inherent in the residual signal. Indeed, it is well known that the *LPC* analysis filter is not capable of removing speech resonant and anti-resonant frequency components completely so that the prediction residual does not have a flat amplitude spectral envelope. However, the resonants peaks in the residual are nearly equalized, unlike those of the original speech. Thus, the spectrum of the prediction residual may be approximated by an amplitude spectrum with broadened spectral peaks. Under the *PSOLA/KDG* wide-band synthesis condition, the bandwidths of these "broadened" resonances are much wider than the main lobe of the synthesis window. Consequently, the *LP-PSOLA* algorithm can preserve better the global shape of the excitation spectrum. To see this, consider the excitation signal is truly periodical. Its short-term spectrum can be expressed as:

$$X_0(\omega) = \sum_{k=0}^{P-1} D(\Omega_k) H(\Omega_k - \omega)$$

where $D(\omega)$ denotes the envelope of the residual excitation spectrum and Ω_k denotes the original pitch harmonic frequencies. As mentioned above, the *PSOLA/KDG* modification is equivalent to sampling the short-time spectrum at the synthesis pitch harmonic frequencies $\tilde{\Omega}_l$, so the harmonic amplitudes c_l of the synthetic signal are given by:

$$c_l = \sum_{k=0}^{P-1} D(\Omega_k) H(\Omega_k - \tilde{\Omega}_l)$$

As the LP-filter has a very low spectral dynamic, it is reasonable to expand $D(\omega)$ in a Taylor series in the neighborhoud of a pitch-modified signal harmonic frequency $\tilde{\Omega}_l$:

$$D(\omega) = D(\tilde{\Omega}_l) + (\omega - \tilde{\Omega}_l)\dot{D}(\tilde{\Omega}_l) + (\omega - \tilde{\Omega}_l)^2 \ddot{D}(\tilde{\Omega}_l) + \dots$$

By substituting this expansion in the former equation, we obtain:

$$c_l = D(\tilde{\Omega}_l)\left(\sum_{k=0}^{P-1} H(\Omega_k - \tilde{\Omega}_l)\right) + \dot{D}(\tilde{\Omega}_t)\left(\sum_{k=0}^{P-1} (\Omega_k - \tilde{\Omega}_l) H(\Omega_k - \tilde{\Omega}_l)\right) + \dots$$

It is possible to show that the following relation:

$$\sum_{k=0}^{P-1} H(k\Omega - \tilde{\Omega}_t) = 1$$

is a normalization condition that can be achieved by a large class of window. On the other hand, it can be ensured that the remaining terms of the summation are at most of the order of the product of the analysis window bandwidth by a majorant of the spectral derivatives. In conclusion, the *LP-PSOLA* algorithm preserves at the first order the spectral envelope deviation, provided that the dynamic of the spectrum is low. This fact is confirmed experimentally in the sense that the *LP-PSOLA* processed excitation signal is "intelligible".

Finally, the different behavior of the *PSOLA/KDG* and *LP-PSOLA* algorithms is illustrated in Fig.1, for a vowel /i/ pronounced by a female speaker. The resonance broadening effect appears clearly for the second formant both on the synthetic signal obtained by the *PSOLA/KDG* method and on its short-term spectrum. On the other hand, the *LP-PSOLA* method is sensitive to spectral estimation errors such as the locking of the estimated resonances on the pitch harmonics of a high pitch voice. Suprisingly, although the synthesized signals obtained by both methods look quite different, it is very difficult to distinguish them from an auditory point of view.

CONCLUSION

We have presented in this paper the *PSOLA/KDG* algorithm for prosodic modifications of natural speech. The algorithm is based on pitch-synchronous processing of the speech waveform in the time-domain and uses an overlap-add synthesis scheme. It is capable of producing very good sound quality, when the synthesis window is chosen to minimize spectral distortions due to both a comb filtering and a spectral broadening effects. Furthermore, it provides a very flexible solution for speech synthesis, since it can be combined to a variety of coding algorithms. When used in combination with linear predictive coding techniques, it can be modified into a *LP-PSOLA* synthesis sheme, in which the prosodic modifications are performed on the excitation signal rather than the speech waveform. From a theoretical point of view, this is capable of eliminating the spectral broadening effects. However, the practical advantage of such a solution remains unclear, since it relies on an accurate estimation of the short-term spectral envelope, and since human hearing is relatively insensitive to spectral broadening effects.

REFERENCES

[1] C. Hamon, E. Moulines, F. Charpentier, "A diphone synthesis system based on time-domain modifications of speech", *Proc. Int. Conf. ASSP, Glasgow, 1989*

[2] D.W. Griffin, J.S. Lim, "Signal estimation from modified short-time Fourier transform", *IEEE Trans. ASSP, 32(2), 236-243, 1984*

[3] J.B. Allen, L.R. Rabiner, "A unified approach to short-time Fourier analysis and synthesis", *Proc. IEEE, 65(11), 1558-1564,* 1977

[4] F. Charpentier, E. Moulines, "Text-to-speech algorithms based on FFT synthesis", *Proc. Int. Conf. ASSP, New York, 667-670, 1988*

[5] F. Charpentier, "Traitement de la parole par analyse-synthèse de Fourier: application à la synthèse par diphones", *Doctoral thesis, Ecole Nationale Supérieure des Télécommunications, 1988.*

[6] C. Hamon, "Procédé et dispositif de synthèse de la parole par addition-recouvrement de formes d'ondes", *patent No. 8811517, 1988.*

[7] E. Moulines, F. Charpentier, "Diphone synthesis using *multipulse linear prediction", Proc. FASE Int. Conf., Edinburgh, 1988*

[8] E. Moulines, *Doctoral thesis, in preparation.*

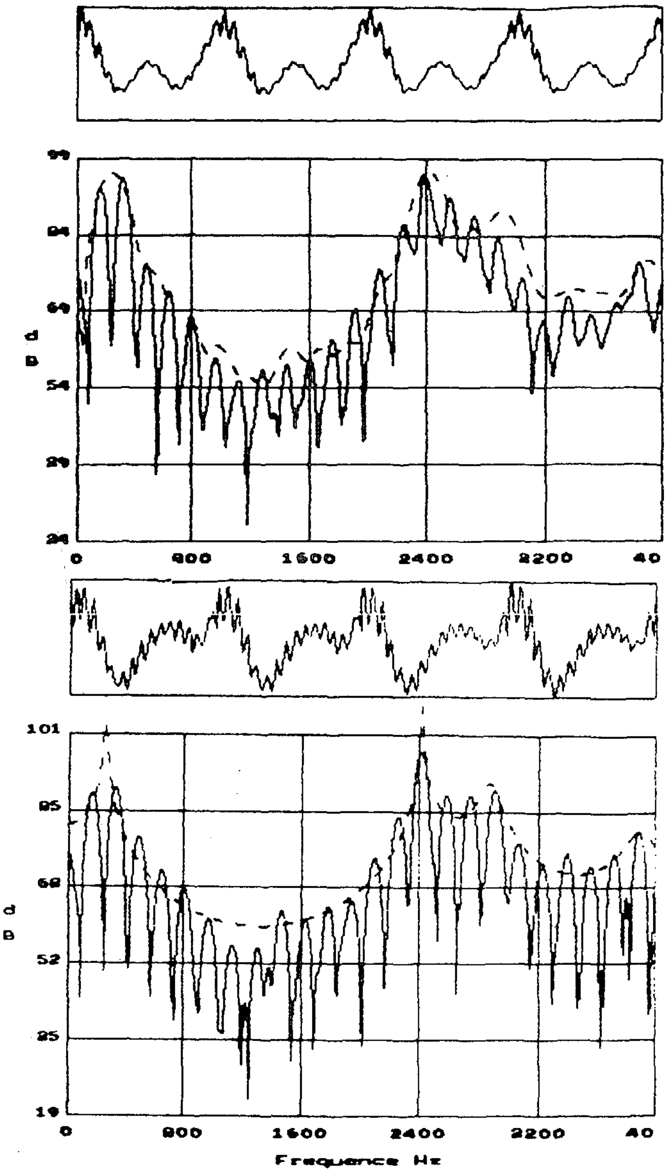

Fig. 1: Pitch modification of a natural vowel /i/ uttered by a female speaker ($\beta = 3/2, f_0 = 260 \quad Hz$). The synthetic signals obtained by the two methods are displayed along with their ST-spectrum using a 25 msec Hanning window.

(a) **PSOLA/KDG** method: the dashed line indicates the magnitude spectrum of the ST-synthesis signal (which corresponds to the implicit spectral envelope)

(b) **LP-PSOLA** method: the dashed line indicates the LP-spectrum envelope (weighted covariance method, 20 coefficients, 14 msec Blackman window)