HAL
open science

Effectiveness of a periodic and aperiodic decomposition method for analysis of voice sources

Christophe d'Alessandro, Vassilis Darsinos, B. Yegnanarayana

To cite this version:

Christophe d'Alessandro, Vassilis Darsinos, B. Yegnanarayana. Effectiveness of a periodic and aperiodic decomposition method for analysis of voice sources. IEEE Transactions on Speech and Audio Processing, 1998, 6 (1), pp.12-23. hal-02000967

HAL Id: hal-02000967

https://hal.science/hal-02000967v1

Submitted on 6 Feb 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Effectiveness of a Periodic and Aperiodic Decomposition Method for Analysis of Voice Sources

Christophe d'Alessandro, *Member, IEEE,* Vassilis Darsinos, and B. Yegnanarayana, *Senior Member, IEEE*

***Abstract*—Decomposition of speech into periodic and aperiodic components is useful in analyzing and describing the characteristics of voice sources. Such a decomposition is also useful in controlling the excitation source for synthesis. This paper addresses the issue of decomposition of speech into periodic and aperiodic components in the context of speech production. The effectiveness of a recently proposed algorithm for decomposing speech into these components is examined for analysis of voice sources. Synthetic signals are generated using formant synthesis. Different sources of aperiodicity encountered in normal speech production are considered, using a set of parameters to control the synthetic signals. The sources of aperiocity studied are 1) additive pulsed or continuous random noise, and 2) modulation aperiodicities due to variation in the fundamental frequency, jitter, and shimmer. Three types of measures are used to characterize these voices: ratio of energies in the periodic and aperiodic components, perceptual spectral distance, and spectrograms. The results demonstrate the effectiveness of the periodic–aperiodic decomposition algorithm for analyzing aperiodicities for a wide variety of voices, and point out the limitations of the algorithm.**

***Index Terms*— Jitter, periodic and aperiodic decomposition, shimmer, voice quality assessment, voice source analysis.**

I. Introduction

SEPARATION of signals into periodic and aperiodic components is an important issue in signal processing, especially in speech and music. Processing methods for the separation have been explored for speech coding [1], speech synthesis [2]–[5], voice analysis [6]–[10], and musical acoustics [11], [12]. An important question in the context of speech is whether these components represent features of speech production or whether they are merely convenient tools for the representation of the signal. Although much effort has gone into the development of algorithms for analysis of speech in terms of periodic and aperiodic components, few studies performing an actual separation of the components signals have been reported emphasizing the relevance or acoustic significance of such a separation. If these components signals can be associated to acoustic components, then it may be possible to refine the models of the vocal source by incorporating the aperiodic component also.

In this paper, we shall study the effectiveness of a recently proposed method of decomposition for analysis of voice sources [13]. The periodic–aperiodic (PAP) decomposition algorithm is discussed in detail in a companion paper [14]. In our study, we associate the periodic component with the "deterministic" part, and the aperiodic component with the "stochastic" or "random" or "noise" part of the excitation signal. The algorithm attempts to separate the contributions of the individual components of the signal at each sample point in the time and frequency domains.

The following model of speech production for voiced speech is assumed for analysis and synthesis:

$$s(t) = e(t) * v(t) = (p(t) + r(t)) * v(t) \tag{1}$$

$$S(\omega) = |S(\omega)| e^{j\theta_s(\omega)} \tag{2}$$

$$= (|P(\omega)| e^{j\theta_p(\omega)} + |R(\omega)| e^{j\theta_r(\omega)}) |V(\omega)| e^{j\theta_v(\omega)} \tag{3}$$

where

- $s(t)$ is the speech signal, and $S(\omega)$ is the Fourier transform (FT) of $s(t)$;
- $v(t)$ is the impulse response of the vocal tract system, and $V(\omega)$ is the FT of $v(t)$;
- $e(t)$ is the excitation signal;
- $p(t)$ is the quasiperiodic part of the excitation, and $P(\omega)$ is the FT of $p(t)$;
- $r(t)$ is the random part of the excitation, and $R(\omega)$ is the FT of $r(t)$.

The magnitude of the FT is represented by $|\ |$, and the phase as $\theta(\omega)$. The $*$ symbol is used to denote convolution. The aperiodic component can be associated with different situations in speech production. For unvoiced segments, the periodic component reduces to zero. Two sources of aperiodicity in voiced speech signals can be identified (see, for instance, [8]): 1) additive random noise and 2) modulation aperiodicity.

- *Additive random noise*: This source of aperiodicity represents frication or aspiration noise. This type of noise is termed *additive* because the noise source is superimposed onto the voice source. This type of noise is present in segments of voiced fricatives or breathy vowels [9], [15], [16]. Depending on the location of the source of noise in the vocal apparatus, this noise is aspiration noise (generated at the glottis, e.g., vowels, especially when the glottal closure is incomplete), or frication noise (generated at a constriction in the vocal tract, e.g., voiced fricatives).

Manuscript received September 20, 1995; revised February 19, 1997. This work was supported by a grant from the University of Paris, and by the CEC ERASMUS Program in Phonetics and Speech Communication. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Douglas D. O'Shaughnessy.

C. d'Alessandro is with LIMSI-CNRS, F-91403 Orsay, France (e-mail: cda@limsi.fr).

V. Darsinos is with the Wire Communications Laboratory, University of Patras, Patras, Greece.

B. Yegnanarayana is with the Department of Computer Science and Engineering, Indian Institute of Technology, Madras 600036, India.

Publisher Item Identifier S 1063-6676(98)00584-7.

1063–6676/98$10.00 © 1998 IEEE

Because of the different location of the source of noise, aspiration and frication noises have different spectral properties. Typically, frication noise is a highpass noise, and aspiration noise is spread over the whole spectrum. The noise source can be continuous or pulsed (gated noise source).

- *Modulation aperiodicity*: Modulation aperiodicity is a result of variation in the periodicity of the glottal excitation. Aperiodicity may be introduced due to random variations in the duration (jitter), or the peak amplitute of signal periods (shimmer). Aperiodicity may also be introduced due to voluntary changes in the source characteristics as in prosody, and in formant transitions.

In addition, aperiodicity in the signal may also be observed in speech processing due to the effects of finite windows, discretization in amplitude (quantization) and time (sampling), and inadequacy of the speech production model. This type of aperiodicity may be termed *computational noise*.

These sources of aperiodicity are not equally significant. Ideally, a decomposition method should be insensitive to prosodic and spectral variations. The method should also minimize computational noise, and it should be robust. In this paper, we study the performance of a decomposition algorithm proposed in [13] and [14] to analyze various sources of aperiodicity in speech signals. We also discuss the limitations of the algorithm, so that the results of voice source analyses can be interpreted meaningfully.

Section II describes the method used for characterizing voice sources. We first describe the set of test signals used in this study. Different types of voice source signals are simulated through these test signals. The measurement methodology is then presented. We briefly describe the algorithm used for PAP decomposition, and we define the measures used to characterize voices in terms of their periodic and aperiodic components. In Section III, we discuss the results of decomposition on the test signals for characterizing the different types of voice. Section IV concludes the paper.

II. Method

A. Test Signals

Using natural speech for systematic evaluation of source parameters is difficult, because speakers producing the test material typically cannot control the source characteristics over a desired range of variation. This lack of control on production is even more difficult for voice characteristics related to additive noise, jitter, and shimmer [10].

Therefore, synthetic speech is used for evaluation in the present study. It is possible to separately generate the periodic and aperiodic components, and then add them to generate the synthetic speech signal. This will enable us to compare the components derived using the decomposition algorithm with the components in the synthetic signal. Each test signal can be described on the basis of six different components: synthetic aperiodic component (SAC) containing the additive noise, synthetic periodic component (SPC), extracted aperiodic component (EAC), extracted periodic component (EPC), and the signals obtained by summation of the SAC and the SPC, and by summation of the EAC and the EPC. It must be noted that the SAC contains only the additive random noise. Thus, modulation aperiodicity is included in the SPC. Eventually, both sources of aperiodicities are present in the EAC.

Synthetic signals are generated using formant synthesis. Two different computer programs are used for synthesis. The first one, based on the Klatt synthesizer [18], is used for all test signals, except for those involving jitter and intonation. It is called in the following *serial formant synthesizer* because we used only the serial branch for our stimuli. For the jitter and intonation cases, a time-domain parallel formant synthesizer (based on elementary waveforms [19]) is used, because it is necessary to compute the fundamental frequency very accurately. It is called in the following *parallel formant synthesizer*. In the Klatt synthesizer [18, p. 975], quantization of the fundamental period to an integral number of samples is not accurate enough for computing small jitter values, or for obtaining accurate pitch glides.

For the Klatt synthesizer, the synthesis model has two parts, one corresponding to the glottal source and the other to the vocal tract system. The radiation effect is included in the source part, by taking the first derivative of the glottal waveform. The synthetic vowel /*a*/ is chosen in the present study. The serial formant frequencies and bandwidths for the four formants are (700 Hz, 100 Hz), (1200 Hz, 200 Hz), (2300 Hz, 250 Hz), and (3500 Hz, 300 Hz). The glottal source signal is modeled using the Liljencrants–Fant (LF) model [17]. The set of parameters chosen for this model is as follows.

- t_0: fundamental period (varying among test signals).
- t_p: rise time of the flow pulse ($t_p = 0.5 \times t_0$ for all the test signals).
- t_e: time of the maximum peak of the flow derivative ($t_e = 0.7 \times t_0$ for all the test signals).
- E_e: amplitude of the maximum peak of the flow derivative (this parameter defines the amplitude of the synthetic signal E_e is normalized for all the test signals).
- ε: related to the time of the return phase t_a of the flow derivative ($\varepsilon = 0.8 \simeq 1/t_a$ for all the test signals).

The reader is referred to [17] for a detailed description of the LF model. A Gaussian random number generator is used for noise sources. Fig. 1 gives an example of synthetic excitation signal consisting of the glottal flow component, the additive noise component, and the excitation signal.

The parallel formant synthesizer is used for the shimmer and intonation experiments. In this case, the parallel formant frequencies, bandwidths, and amplitudes for the four formants are (650 Hz, 78 Hz, 0 dB), (1100 Hz, 88 Hz, −8 dB), (2900 Hz, 133 Hz, −11 dB), (3300 Hz, 130 Hz, −20 dB). The glottal source is controlled in frequency domain by the spectral tilt, and by an extra low-frequency glottal formant, with formant frequency, bandwidth, and amplitude of (10 Hz, 400 Hz, −4 dB). No additive noise is present in the source with this synthesizer.

Jitter, shimmer, and additive noise are controlled in combination with variations of the fundamental frequency to generate the test signals. The duration of the synthetic signals

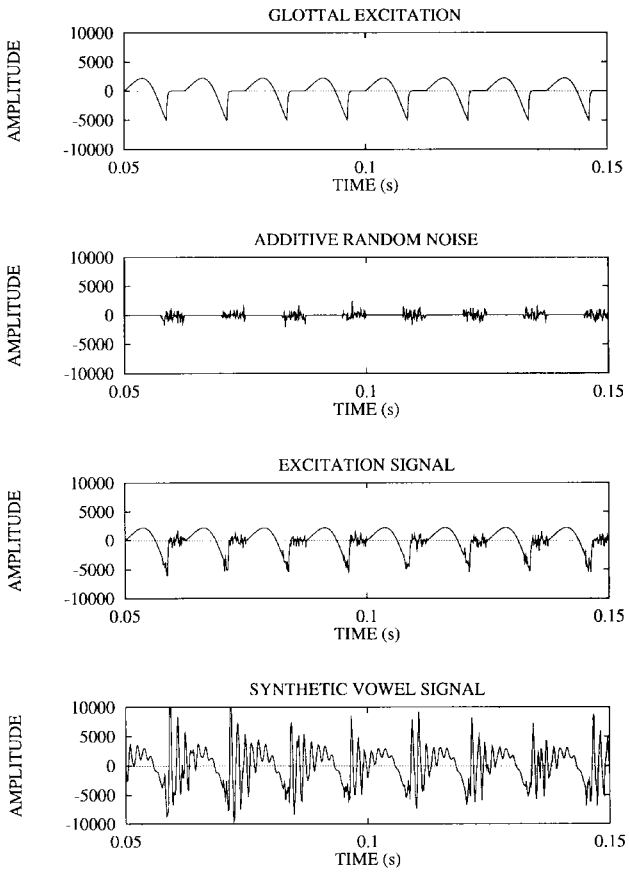

Fig. 1. Illustration of a synthetic signal (serial formant synthesizer). From top to bottom: glottal flow component; additive random noise; glottal flow + additive random noise component (excitation signal); synthetic vowel /*a*/. F0 = 80 Hz, HNR = 5 dB, noise burst duration = 60% of T0.

is 0.4 s. Following are the parameters used to control different types of voices.

1) *Input harmonic-to-noise ratio*: Four input harmonic to noise ratio levels (HNR, i.e., SPC to SAC energy ratio level) are used: ∞ (no noise), 20, 10, and 5 dB. A few trial runs of the speech synthesis program are made in order to obtain the desired HNR level in the signal, independent of the fundamental frequency and the glottal source parameters. The noise amplitude is varied to realize the desired HNR levels. The fundamental frequency (F0) is fixed for a given signal, and is varying among different signals, as it is described in the following. The serial synthesizer is used.
2) *Noise burst duration*: The SAC consists of a burst of noise for a duration of either 60% or 100% of each pitch period. This noise burst is centered around the instant of maximum excitation. The noise burst is a gated white Gaussian noise signal. This choice for the noise source is dictated by previous studies on synthesis and perception experiments [16], [20]. Glottal turbulence noise is commonly assumed to result from a combination of high air flow velocity and imperfect glottis closure, and can be more or less modulated. An example of the synthetic excitation signals is given in Fig. 1. Each of these signals is generated with a fixed fundamental frequency value. Different excitation signals are generated by varying the F0, HNR level, and duration of burst. For each of the three HNR levels and two duration values, 12 values of F0 are used to generate a total of $6 \times 12 = 72$ test signals. In addition, the noise-free test signal (∞ HNR level) is also generated using each of the twelve F0 values. Thus, the total number of test signals in this category is $72 + 12 = 84$. The serial synthesizer is used.
3) *Jitter*: Jitter is defined as the maximum perturbation of F0. Jitter values of 0.00, 0.25, 0.50, 1.00, 1.50, 3.00, and 5.00% are used. They are expressed as a percentage of the duration of the pitch period. Large values for jitter variation are also considered in this study, because they may be encountered in pathological voices. However, jitter in normal voices is generally less than 1% of the pitch period [15], [21]. These signals are generated without additive noise, by varying the average F0 in the range 80–300 Hz. Seven values of jitter are combined with twelve F0 values, yielding to a total of $7 \times 12 = 84$ test signals. The parallel synthesizer is used.
4) *Shimmer*: Shimmer values of 0.5, 1.0, and 1.5 dB are used. These values represent the maximum range of peak amplitude change in the signal (and thus the maximum variation in peak amplitudes of successive signal periods). Large values for shimmer are also considered in this study, because they may be encountered in pathological voices. However, shimmer in normal voices is generally less than about 0.7 dB [22]. These signals are generated without additive noise, by varying the average F0 in the range 80–300 Hz. Three values of shimmer are combined with twelve F0 values, yielding to a total of $3 \times 12 = 36$ test signals. The serial synthesizer is used.
5) *Fundamental frequency*: For each input HNR and burst duration condition, and for each jitter or shimmer condition, fundamental frequency is fixed. F0 was varied among the signals in 12 steps in the range 80–300 Hz (80, 100, 120, 140, 160, 180, 200, 220, 240, 260, 280, 300 Hz).
6) *Fundamental frequency changes*: Variation of F0 during an analysis frame is another source of modulation aperiodicity. This situation is normal in natural speech due to intonation. The effect of these variations is studied using a set of synthetic signals, in which the slope of F0 is varied in the range 0–24 SemiTones per second (ST/s) in seven steps: 0, 1.5, 3, 6, 12, 18, 24 ST/s. Two conditions are used for the baseline frequency, 100 Hz and 200 Hz. These conditions correspond roughly to the pitch variations due to intonation in normal speech. A set of 14 test signals is generated. The parallel synthesizer is used.

Thus, a total of 218 synthetic test signals are generated for studying the influence of additive noise and modulation aperiodicities due to jitter, shimmer and F0 changes, in the F0 range of normal male and female speech. Table I gives a summary of the test signals used in this study. All signals are generated at a sampling rate of 8 kHz.

TABLE I
SUMMARY OF TEST SIGNALS USED FOR VOICE SOURCE CHARACTERIZATION

Effect of additive noise		
	HNR levels:	∞, 20 dB, 10 dB, 5 dB
	burst duration:	60 % and 100 % of the pitch period
	F0 values:	80, 100, 120, 140, 160, 180, 200, 220, 240, 260, 280, 300 Hz
Effect of jitter		
	jitter values:	0%, 0.25%, 0.5%, 1%, 1.5%, 3% and 5%
	average F0 values:	80, 100, 120, 140, 160, 180, 200, 220, 240, 260, 280, 300 Hz
Effect of shimmer		
	shimmer values:	0.5dB, 1.0dB and 1.5dB
	F0 values:	80, 100, 120, 140, 160, 180, 200, 220, 240, 260, 280, 300 Hz
Effect of F0 variation		
	F0 glides:	0, 1.5, 3, 6, 12, 18, 24 ST/s
	baseline F0 :	100 and 200 Hz

B. Algorithm for PAP Decomposition

This section gives a summary of the PAP decomposition algorithm proposed in [13]. Details on the properties of this algorithm can be found in [14]. The complex addition in (3) suggests the importance of both the magnitude and phase of each of the components in the signal. Also, both the periodic and aperiodic components are present at each frequency. The proposed decomposition algorithm used for measurements of EAC and EPC is illustrated in Fig. 2. It contains the following steps.

1) *Extraction of linear prediction residual*: The objective is to separate the components of the source. The method chosen for computing an approximation to the excitation part of the signal is linear predictive (LP) analysis [23] (discrete signals are considered now). The LP residual is obtained by passing the speech signal through a tenth-order inverse filter. The LP residual is associated to the excitation signal, and is denoted by $e(n)$. A frame length of 20 ms and a frame rate of 100 frames/s are used in the LP analysis.

   The LP residual signal is decomposed into short overlapping analysis frames, using a Hamming window. A 255 points ($\simeq$32 ms) window is used, and the frame rate is 200 frames/s for the short-term excitation signal decomposition. A short-time spectrum $E_l(k)$ is computed for each frame, using a 512-point discrete Fourier transform (DFT)

$$E_l(k) = \sum_{n=0}^{N-1} w(n)e(n+lH)\exp\left(-\frac{j2\pi}{N}nk\right) \quad (4)$$

   where e is the discrete-time excitation signal, H is the hop size in number of samples (the spacing between analysis frames), l is the frame index, N is the fast Fourier transform (FFT) size, and w is the analysis window.

2) *Identification of frequency regions of the aperiodic component*: Both periodic and aperiodic components contribute to the DFT coefficients. In the first stage of processing, we identify a subset of the DFT coefficients to form an approximation to the aperiodic component. For this purpose, we determine approximately the frequency regions contributing to the harmonic part and the frequency region contributing to the noise part. This is accomplished by marking the region in the cepstrum corresponding to the vocal tract system, and the regions corresponding to the harmonic and noise parts of the excitation [10]. Because of their distinct nonoverlapping regions in the frequency domain, distribution of energies for each of these components in the frequency domain can be obtained. From these distributions, it is possible to determine the ratio of the harmonic and noise components at each discrete frequency point.

3) *Reconstruction of the aperiodic signal using an iterative procedure*: The knowledge of the ratio of the harmonic to noise parts at each frequency point does not enable us to separate the two components by subtraction, because at each frequency point there is contribution due to both the periodic and aperiodic components. According to (3), it is necessary to use both amplitude and phase, at each frequency, for separating these components in the frequency domain. Obviously, the magnitude and phase are not directly available. We developed an iterative procedure to reconstruct the aperiodic components.

   From the frequency distribution of the harmonic regions in the log magnitude spectrum, we hypothesize, to a first approximation, that the valley regions between two harmonics are mostly due to the aperiodic component. To obtain an approximate aperiodic component, $r_l(n)$, of the residual, we can sum only those DFT coefficients ($k \in F_r$) for which the noise component dominate. That is

$$r_l(n) = \sum_{k \in F_r} E_l(k)\exp\left(\frac{j2\pi}{N}nk\right) \quad (5)$$

   where N is equal to the size of the DFT. Here, F_r is the set of frequency points in the valley regions between two harmonics. The width of the harmonics can be fixed according to the width of the main lobe in the cepstrum.

   Thus, the aperiodic component is set to zero in the harmonic regions, and to the measured DFT values in the regions between the harmonics, i.e., in the noise regions. It is clear that such a comb-filtered noise component cannot represent the aperiodic component in

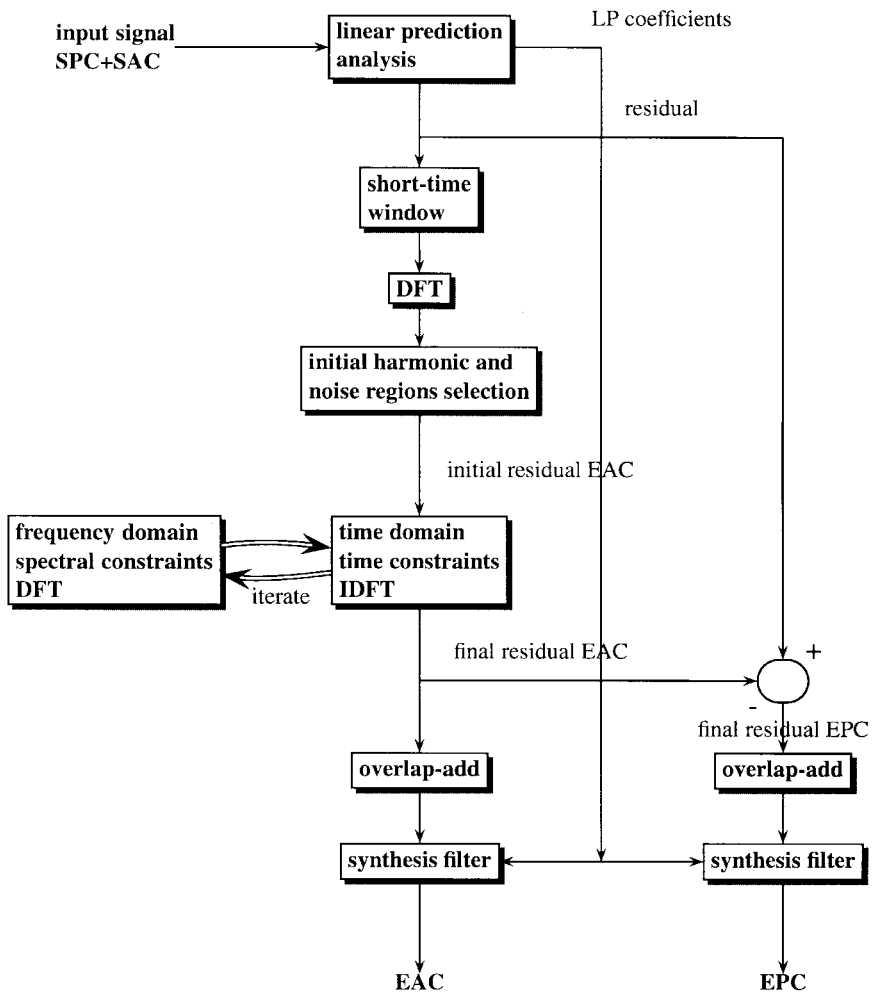

Fig. 2. PAP decomposition algorithm. SAC: synthetic aperiodic component. SPC: synthetic periodic component. EAC: extracted aperiodic component. EPC: extracted periodic component.

the speech signal. It is necessary to estimate the values of the noise component in the harmonic regions. The contribution of the aperiodic component in the harmonic regions is estimated using an iterative algorithm similar to the Papoulis–Gerchberg extrapolation algorithm [24, pp. 244–248]. An estimate of the aperiodic component is obtained by iteratively moving from the frequency domain to the time domain and vice versa, through the inverse DFT (IDFT) and DFT operations.

a) *First iteration*: Suppose we obtained a set of DFT coefficients that form a first approximation $R_l^0(k)$ to the aperiodic component

$$R_l^0(k) = \begin{cases} E_l(k), & \text{for } k \in F_r \text{ (noise regions)} \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

An IDFT is applied to this first approximation, and the corresponding time-domain signal $r_l^0(n)$ is obtained (512 samples in our case). A finite duration constraint is imposed in the time domain, i.e., the signal samples of the aperiodic component in the time domain beyond the analysis frame size are set to zero (as we used a $N/2-1=255$ samples analysis windows, the samples from 255 to 511 are set to zero, numbering the samples from 0 to 511). That is, form a signal

$$\hat{r}_l^0(n) = \begin{cases} r_l^0(n), & \text{for } n < N/2-1 \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

b) mth iteration: Starting with $m=1$, we compute the DFT $\hat{R}_l^{m-1}(k)$ of $\hat{r}_l^{m-1}(n)$, and form the function

$$R_l^m(k) = \begin{cases} E_l(k), & \text{for } k \in F_r \\ \hat{R}_l^{m-1}(k), & \text{otherwise} \end{cases} \tag{8}$$

and compute its IDFT $r_l^m(n)$. The time samples beyond $N/2-1$ are set to zero. That is

$$\hat{r}_l^m(n) = \begin{cases} r_l^m(n), & \text{for } n < N/2-1 \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

The iterative algorithm is continued until the difference (in terms of magnitude of the noise samples) between two successive steps becomes less than a given threshold value, or after a fixed number of iterations. In our experiments, we used $m=10$ iterations. The periodic component is obtained by subtracting the reconstructed aperiodic component noise samples from the residual

signal samples in the time domain. The steps in the iterative algorithm are illustrated in Fig. 2. The convergence of the proposed algorithm is proved in [14].

4) *Synthesis*: The aperiodic component of the residual signal is obtained for each of the overlapping analysis frames. The aperiodic component signal for the entire utterance is derived from these short-time signals, using an overlapp-add procedure as follows:

$$r(n) = \sum_{l=0}^{L-1} r_l^m(n - lH) \tag{10}$$

where L is the number of frame in the utterance.

The periodic component is then obtained by subtraction of the aperiodic component, as follows:

$$p(n) = e(n) - r(n). \tag{11}$$

Finally, the speech signal corresponding to each component is generated by passing the component residual signal through the time varying all-pole synthesis filter.

The algorithm depends on an initial analysis window. Of course, the sidelobes and the main lobe width of the window will influence the regions between the harmonics in the spectrum, and introduce some computational noise. Figures of merit for different window types are available, but it is difficult to theoretically study the effect of the window on the PAP decomposition. For this reason, we designed a set of experiments using synthetic speech. The measurements reported later in this paper give some data on the level of the computational noise introduced by the window. We tried several window types and retained the Hamming window in the light of some preliminary experiments.

In our experiments, we use linear predictive coefficient (LPC) analysis and synthesis because it seemed more relevant to work on an approximation of the source signal. Direct decomposition using the signal rather than the LP residual was also tried. The results show that the decomposition is still good, but a slight loss in accuracy is noticeable. Thus, LPC analysis and synthesis is preferred for evaluation of the algorithm. Also, we fix the same algorithm parameters for all the test signals because we want comparable analysis conditions for all the test signals. In other applications, the algorithm could be used with or without LPC decomposition. Also, it would be better to adapt the analysis parameters (window size, DFT size, frame rate, etc.) to the specific features of the signals processed.

C. Measurement Methodology

In this section, we discuss the measures used in our study to assess the performance of the proposed PAP decomposition algorithm.

As the source/filter decomposition achieved by the LP analysis is not perfect, the measurements are performed on the resynthesized periodic and aperiodic components rather than on the approximation of the excitation source signals. Therefore, the formant structure of the signals is added before the measures are taken. Although the formant structure may play a role in the measures, particularly when harmonics move over the formant structure, the influence of spectral peaks and spectral peaks changes is beyond the scope of this study. Three different measures are used to study various aspects of the voice source characteristics. They are: periodic-to-aperiodic ratio, perceptual spectral distance, and spectrograms.

1) Periodic-to-Aperiodic Ratio: A direct measure of the aperiodic component is the HNR, defined (in decibels) by

$$\mathrm{HNR} = 10 \times \log_{10}\left(\frac{E_p}{E_{ap}}\right) \tag{12}$$

where E_p is the energy in the periodic component and E_{ap} is the energy in the aperiodic component. The energy can be computed both in the time and frequency domains. In either case, the total energy is defined as the sum of the squared amplitudes for all samples. Unfortunately, the periodic and the aperiodic components are not available separately in actual speech. Therefore, to study the performance of the PAP decomposition algorithm, the HNR is computed for synthetic test signals.

HNR's are computed using the periodic and the aperiodic components, before and after decomposition (i.e., using SAC, SPC, EAC, and EPC). The HNR's computed from SAC and SPC correspond to the input HNR (or, simply termed HNR). The HNR computed using EAC and EPC, i.e., after decomposition, can be called the *periodic-to-aperiodic ratio* (PAPR). As mentioned before, HNR is actually a measure of the ratio of the energies in the filtered additive noise (SAC) and in the filtered glottal waveform (SPC). Glottal waveforms (SPC) may contain some modulation aperiodicity. This is in contrast with the PAPR, which is the ratio between the aperiodic and periodic components after decomposition. The aperiodic component (EAC) may contain some modulation aperiodicity, although there is no energy in the SAC. In the present case, the HNR and PAPR for synthetic test signals are computed in the time domain as follows:

$$\mathrm{HNR} = 10 \times \log_{10}\left(\frac{\sum_{k=1}^{N} \mathrm{SPC}^2(k)}{\sum_{k=1}^{N} \mathrm{SAC}^2(k)}\right) \tag{13}$$

$$\mathrm{PAPR} = 10 \times \log_{10}\left(\frac{\sum_{k=1}^{N} \mathrm{EPC}^2(k)}{\sum_{k=1}^{N} \mathrm{EAC}^2(k)}\right) \tag{14}$$

where N is the number of samples in the utterance.

A drawback of the PAPR measure is that the energy due to modulation aperiodicity (jitter and shimmer) is partly merged with the energy of the periodic component and partly with the energy of the aperiodic component. This is because this type of perturbation depends on the way the harmonic signals are generated. One must take the sources of aperiodicity into account when interpreting the results of the decomposition. For natural speech signals, only the PAPR is available. One of the aims of this study is to discuss whether the measured PAPR can be used as an estimate of the actual input HNR for natural speech.

2) Perceptual Spectral Distance: Although comparison of HNR and PAPR may give some indication of the ratio of the periodic and aperiodic components, they do not give an idea of similarity of the generated and extracted signals. Hence, a perceptual spectral distance is used for comparing SPC and EPC. The perceptual distance is a measure of the similarity between the amplitude spectra of the SPC and the EPC, seen on a perceptual frequency scale. This distance is applied to the periodic components only. The SAC is reduced to zero in all the test signals in the "jitter," "shimmer," and "F0 variation" sets of Table I. Therefore, the distance between the SAC and the EAC for these tests signals is not very informative. Contrary to the SAC, all the SPC's have the same amplitude, therefore it is meaningful to compare the distances measured for the different test signals. This is why the distance is applied to the periodic components only. A drawback is that the distance will give only a little insight for comparing the SAC and the EAC.

The perceptual spectral distance is based on critical bands, approximating the frequency selectivity of the ear. The critical band rate is expressed in Barks [25]. For computing the perceptual spectral distance, a total number of 29 1-Bark bandwith bandpass filters are used. The spacing between filters is 0.5 Bark to cover the 15.5 Bark that correspond to the 0-4 kHz frequency range. The filters are implemented in the DFT domain using triangular-shaped spectral windows. Each filter gain is normalized with respect to the first filter, centered at 0.5 Bark. For each analysis frame, the energy in each band of the band-filtered spectrum is computed as follows:

$$e(i) = \frac{1}{B_i + 1} \sum_{k=\omega_i - B_i/2}^{\omega_i + B_i/2} f_i(k) H(k) \tag{15}$$

where $e(i)$ is the energy in ith band, B_i+1 is the bandwidth of the filter expressed in number of DFT coefficients (assuming that B_i is even), $f_i(k)$ is the power gain of the ith filter (varying triangularly with k), $H(\omega)$ is the power spectrum, and ω_i is the center frequency of the ith band expressed in number of DFT coefficients.

The perceptual spectral distance between frames of two signals, is defined as the sum of the energy differences in different bands, as follows:

$$D = \frac{1}{M} \sum_{i=1}^{M} |e_1(i) - e_2(i)| \tag{16}$$

where M is the total number of filters, and $e_1(i)$ and $e_2(i)$ are the energies in the ith band of the first and the second signal, respectively. The perceptual spectral distance is given by the average of the distances for individual frames, and expressed in dB.

It must be noted that the perceptual distance takes all the details of the power spectra into account. This may not be the way human subjects perceive signals, since it is generally acknowledged that more attention is paid to the spectral peaks than to the spectral valleys. Therefore, it is difficult to interpret the distance in terms of audibility of the differences between the SPC and the EPC. Also a drawback of the spectral distance measure is that it can be applied to synthetic signals only, where the input periodic component is known.

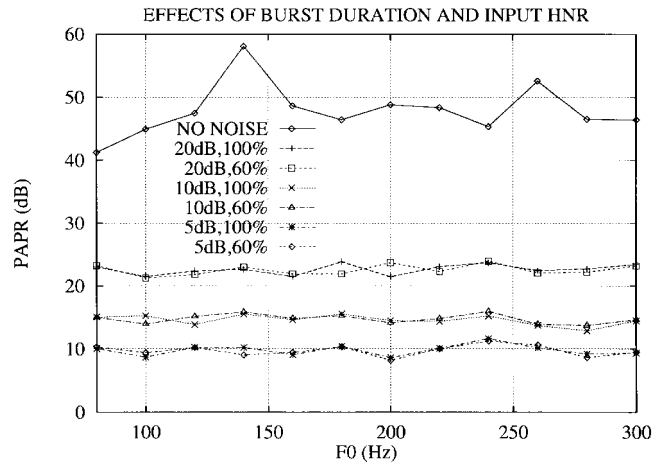

Fig. 3. Effect of input HNR and burst duration on PAPR.

3) Spectrographic Observation: PAPR and spectral distance give some idea on the accuracy of the decomposition. But with synthetic signals, it is possible to examine the time-frequency characteristics of the signals before and after decomposition using wideband spectrograms. The resemblance of SPC and EPC or SAC and EAC can be checked using spectrograms. Spectrograms can be used for natural speech as well, But only a qualitative assessment of differences in signals can be made.

III. Analysis of Results for Various Voice Sources

A. Effect of Additive Random Noise

The effects of HNR, duration of the noise burst, and fundamental frequency on the measured PAPR are shown in Fig. 3. Each line represents one HNR condition for a particular noise burst duration. The x-axis represents F0, and the y-axis represents the measured PAPR obtained after decomposition. It can be seen that the PAPR gives an idea to the input HNR. On average, the PAPR obtained for the 5 dB HNR condition is 10 dB, the PAPR obtained for the 10 dB HNR condition is 14 dB, the PAPR obtained for the 20 dB HNR condition is 23 dB, and the PAPR obtained for the ∞ dB HNR condition is 47 dB. Thus, the PAPR's are higher than the input HNR's. The difference between HNR's and PAPR's seems almost constant for each HNR condition. The noise level is underestimated, which means that there is less energy in the EAC compared to the SAC, except for the no noise condition.

The PAPR is almost constant for all the F0 frequency conditions between 80 and 300 Hz. However, this is not true for the no-noise condition, where an influence of F0 is noticeable. When there is no noise in the SAC, the energy in the EAC is due only to computational noise, which seems sensitive to the F0. This could be explained by the effect of the fixed-size windows used in the algorithm, since the number of signal periods seen in the analysis window depends on F0. We also notice that the duration of the noise burst does not have any significant effect on the measured PAPR.

The algorithm performs well in separating the additive noise and the periodic component. The HNR $= \infty$ condition gives

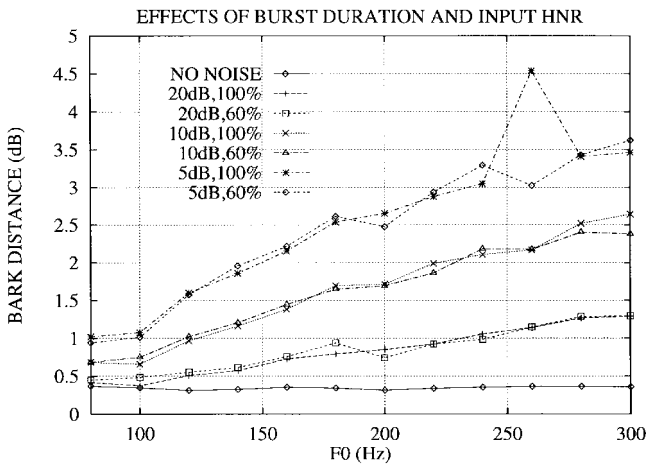

Fig. 4. Effect of input HNR and burst duration on perceptual spectral distance.

an idea of the accuracy of the algorithm. This limiting situation shows the effects of computational noise. When there is no input noise (input $\mathrm{HNR} = \infty$), the separation algorithm gives rise to an aperiodic component which is about 42–58 dB lower than the periodic component.

Quantitative measurements using the spectral distance are shown in Fig. 4. Each line represents one HNR condition for a particular noise burst duration. The y-axis represents the computed distance in dB between the SPC and the EPC. There are no significant differences between the continuous noise and burst noise conditions. The lines for different HNR conditions appear in the opposite order, compared to the PAPR measurements Fig. 3. The distance is small when there is no noise in the original signal, and it increases with decrease in the HNR. This means that the algorithm produces more difference between the original and the extracted periodic components for lower HNR.

Another noticeable effect is that the distance increases with increasing F0. This may be due to the fact that fewer samples are available for specifying the pitch periods for high pitched signals. Therefore, EPC modeling is less accurate for high-pitched signals. The average minimum distance is less than 0.4 dB. The maximum distance is less than 3.5 dB for lower HNR and higher pitch.

Wideband spectrograms for the two noise burst duration conditions are shown in Fig. 5. This figure compares the SAC and the EAC for continuous and pulsed additive noises. Ideally, if the decomposition method were perfect, the two signals in the first and second pairs of signals should be identical. In fact, they appear rather close. The time structure of noise is fairly well captured in the extracted noise signal. Fig. 6 shows the input pulsed noise in the SAC and the EAC. The main time-domain features of the signals are similar. The EAC signal amplitude is lower than the SAC signal amplitude, and some noise is still present in the EPC. This reflects the fact that the additive noise is underestimated in the EAC.

B. Effect of Jitter

Modulation aperiodicities are caused by jitter, shimmer, and F0 variations. In this section, we consider the effect of jitter

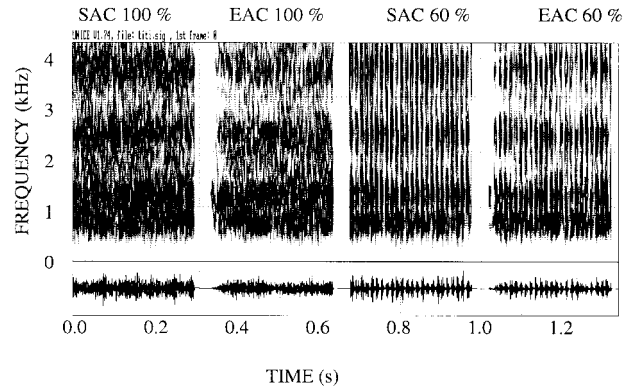

Fig. 5. Spectrograms for the two noise burst duration conditions (serial formant synthesizer). Burst duration: 100 and 60% of T0, F0 = 80 Hz, Input HNR = 5 dB.

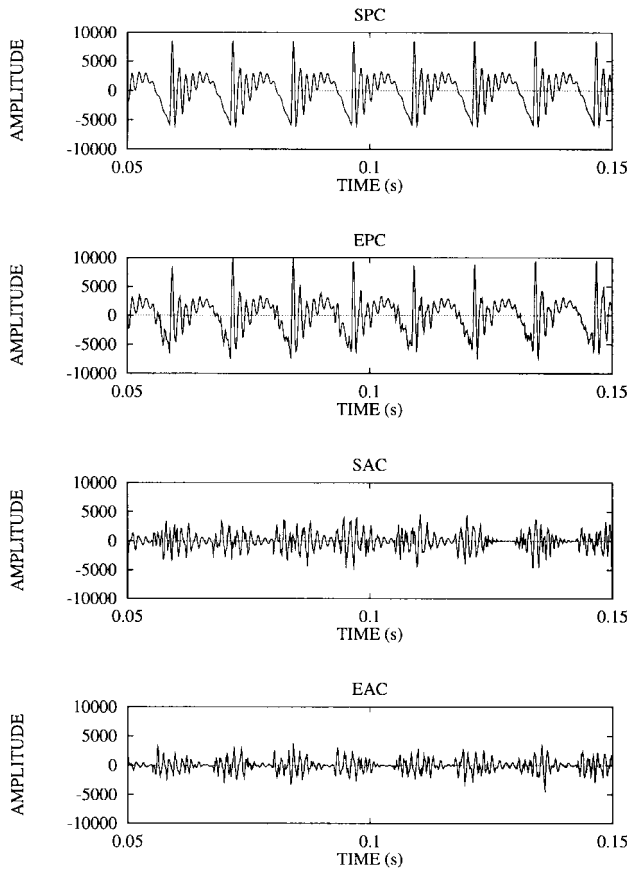

Fig. 6. Comparison of (from top to bottom) SPC, EPC, SAC, and EAC. F0 = 80 Hz, HNR = 5 dB, noise burst duration = 60% of T0. These signals correspond to those in Fig. 1.

on the decomposition algorithm. The effects of the modulation aperiodicities were studied for signals containing no additive noise, and for different fundamental frequency conditions.

PAPR measurements are shown in Fig. 7 for different jitter and fundamental frequency values. Each line of Fig. 7 corresponds to a different jitter value. It must be emphasized here that the test signals were generated using a different speech synthesizer than in the additive noise case. This explains why the 0% jitter condition of Fig. 7 is not exactly same as the HNR $= \infty$ condition in Fig. 3, although the experimental conditions are otherwise identical.

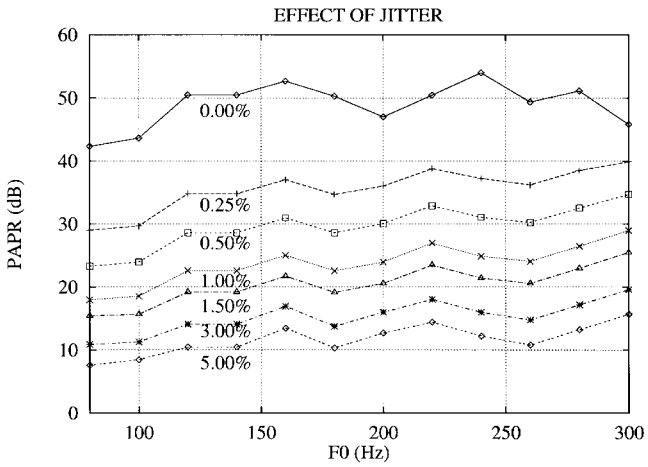

Fig. 7. Effect of jitter on PAPR.

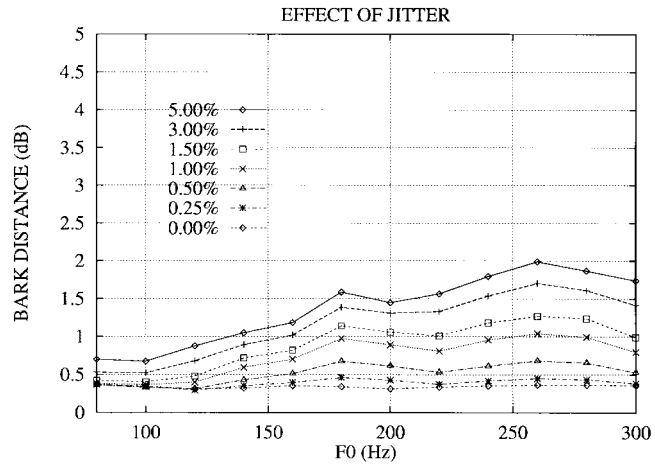

Fig. 8. Effect of jitter on perceptual spectral distance.

It appears that jitter affects the measured PAPR significantly. This is noticeable even for the lower jitter conditions. For the higher jitter condition (5%), the amount of aperiodicity introduced by the jitter seems comparable to the amount of aperiodicity observed in the test signal containing 10 dB HNR additive noise. However, this is just an indication, because the synthesizers used in both experiments are different. Jitter appears a very significant source of aperiodicity in the signal. Moreover, it seems impossible to distinguish between this source of aperiodicity and additive noise using only the PAPR measure.

It is generally known that the effect of jitter on the spectra of voiced speech is to widen the harmonic peaks [8], [20]. This can be explained by the fact that jitter is a random variation of the fundamental frequency, and therefore is a perturbation bounded mostly to the harmonic frequencies. As short-term spectrum computation always introduces some smoothing effect, jitter results in broadening the harmonic peaks. Broad harmonic peaks reduce the accuracy of the algorithm, because the number of initial data points available for extrapolation is reduced. Generally, high-pitched signals produce higher PAPR than low-pitched signals.

The effect of jitter on the spectral distance is shown in Fig. 8. The spectral distance is not very much affected by low jitter conditions. High jitter conditions result in large spectral distances between SPC and EPC. This is particularly true for high-pitched signals. The spectral distances measured for the higher jitter condition are less than to those obtained for the lower HNR additive noise condition in Fig. 4. A high distance could indicate additive random noise rather than modulation aperiodicities. However, it appears difficult to discriminate additive noise and jitter, based on the spectral distance alone.

Spectrograms of EPC and EAC for two jitter conditions are displayed in Fig. 9. In case of little jitter (0.25% jitter), the EAC reduces to zero. That means there is almost no energy in this component. In case of high jitter (5% jitter), the EAC is actually significant. That is, a part of the signal energy is transferred to the aperiodic component. In this case, the EAC might be useful for measuring the jitter effect in a signal.

In summary, high jitter values ($> 1\%$) result in high values for spectral distance and low values for PAPR. This is not usually the case in normal voices, where jitter is generally less than 1%. In this case, the small modulation aperiodicity introduced by low values of jitter is associated with the periodic component. When the jitter increases, the effect of the random F0 variation is reflected in the aperiodic component. In order to show this effect, high jitter values are considered in our experiments, although such values are not common in normal speech. Therefore, for high jitter values, the aperiodic component contains noise coming from both the additive noise as well as from the random variation of the periodicity. If one is looking for a global measure of aperiodicity, which encompasses all sources of noise, the decomposition algorithm is useful. But the algorithm does not give an idea of the different components of aperiodicity.

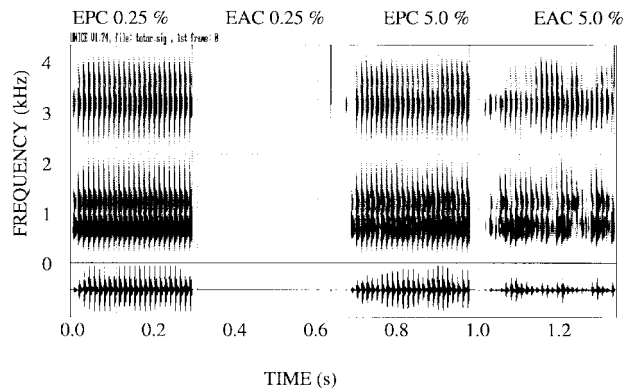

Fig. 9. Spectrograms for two jitter conditions (parallel formant synthesizer). Jitter: 0.25 and 5.0%, F0 = 80 Hz.

C. Effect of Shimmer

Another source of aperiodicity in natural speech signals is shimmer. Shimmer is a random perturbation of the peak amplitude in successive pitch periods. The effect of shimmer on the PAPR is shown in Fig. 10. Each line in this figure represents one shimmer condition. High shimmer values reduce the PAPR. But even for a high shimmer value (1.5 dB), the effect on PAPR appears limited to about 20 dB. Therefore, the effect of shimmer appears less important than the effect of jitter. It has already been noticed [15] that shimmer has less effect than jitter on the spectrum and on the perceived

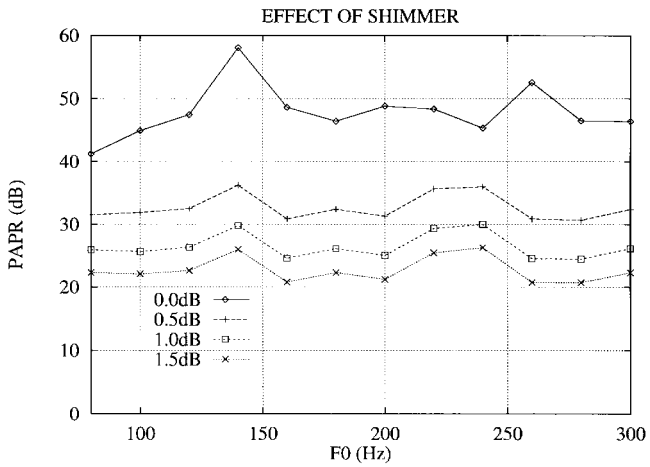

Fig. 10. Effect of shimmer on the PAPR.

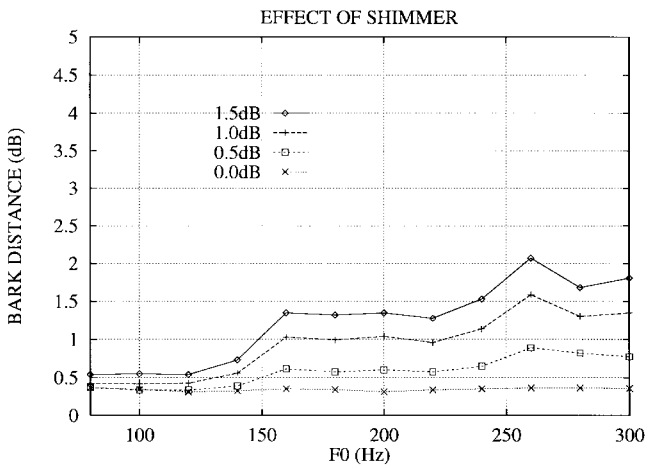

Fig. 11. Effect of shimmer on perceptual spectral distance.

aperiodicity. Thus, we can assume that the shimmer values encountered in normal speech have little effect on the PAPR. This can be explained by the fact that, unlike the effect of jitter, shimmer does not change the locations of the harmonic peaks. It changes rather the amplitudes of the harmonic peaks. The decomposition algorithm relies heavily on the locations of the harmonic peaks.

The effect of shimmer on the spectral distance is shown in Fig. 11. The spectral distance increases with the pitch of the signals. High shimmer values also increase the spectral distance. As in the case of jitter, the spectral distances for high-pitched signals are more informative. The variation in spectral distance is more important at higher F0 than at low F0. This is probably because of the larger spacing between harmonics. The effects of shimmer are comparable to the effects of jitter, as far as the spectral distance is concerned.

Spectrograms of signals containing shimmer are displayed in Fig. 12. The first pair corresponds to EPC and EAC for the 0.5 dB shimmer condition. The second pair corresponds to EPC and EAC for the 1.5 dB condition. For the first pair, little energy is present in the EAC. For the second pair, a small EAC is visible, showing that the effect of shimmer is not very significant.

In summary, shimmer affects both the PAPR and the spectral distance, but only for high values. For shimmer values

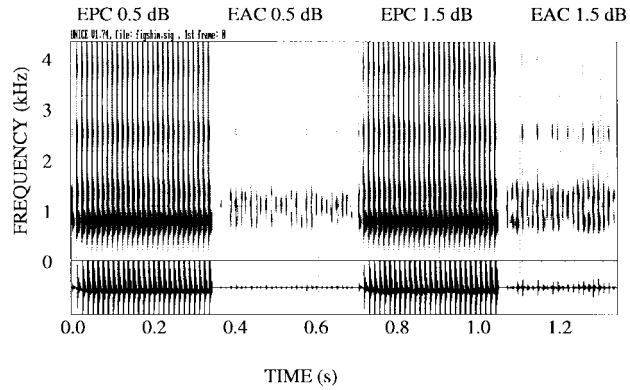

Fig. 12. Spectrograms measured for two shimmer conditions (serial formant synthesizer). Shimmer: 0.5 and 1.5 dB, F0 = 80 Hz.

encountered in normal voices, the EAC is very low compared to the EPC. For high shimmer values, it affects the EAC.

D. Effect of Fundamental Frequency

The effect of F0 can be decomposed into two categories: the effect of the pitch range and the effect of pitch variation. The decomposition algorithm makes use of a fixed analysis window. Therefore, the number of pitch periods in one analysis frame depends on the fundamental frequency. If the pitch range of the specific signals under study is known in advance, it may be possible to adapt the analysis parameters. The issues related to the pitch range have already been examined in the previous sections. It appears that the pitch range has little influence on the PAPR, but more effect on the spectral distance.

The effect of F0 is related to pitch variations. Random variations (jitter) have already been studied. Prosodic variations (intonation) are also responsible for some modulation aperiodicity. The test signals used in this experiment are linear pitch glides. The linear changes are of the same order of magnitude as the pitch glides encountered in normal speech (approximately 0–24 ST/s). In this study, the test signals contain no other form of modulation aperiodicity, no jitter and shimmer, and also no additive noise. The beginning of the pitch glides is located either at 100 Hz or at 200 Hz, corresponding roughly to male and female average pitches, respectively. Therefore, a 12 ST/s glide is either between 100 and 200 Hz, or 200 and 400 Hz, for a tone lasting one second. It must be pointed out that although the extents are expressed in ST/s, linear pitch glides were used.

Fig. 13 shows the PAPR values for F0 glides. Each line represents a particular F0 condition (100 Hz or 200 Hz) for the beginning of the glide. The x-axis represents the rate of F0 change in ST/s. The y-axis represents the measured PAPR. For low pitch glide conditions, the effect of pitch change is not very significant. For very rapid pitch changes, PAPR is rather low. Intonation changes introduce only some variation in the duration of the fundamental period. Jitter introduces the same type of effect, but randomly. A jitter of NJ% means that the maximum possible change in F0 from one period to the following is NJ% of F0. As for intonation variation, if M ST/s is the rate of F0 change, then the percentage NI of F0 change

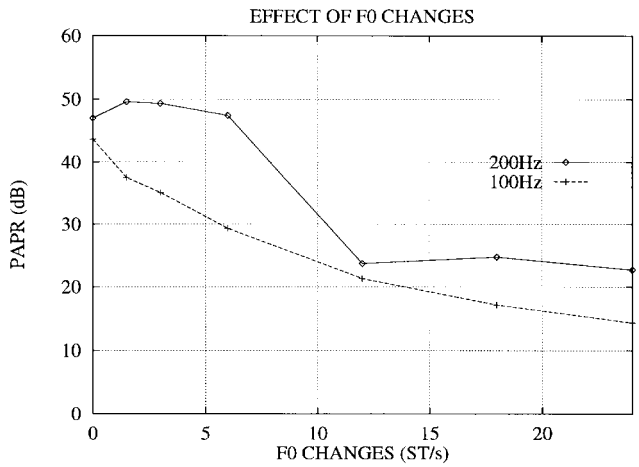

Fig. 13. Effect of F0 changes on PAPR.

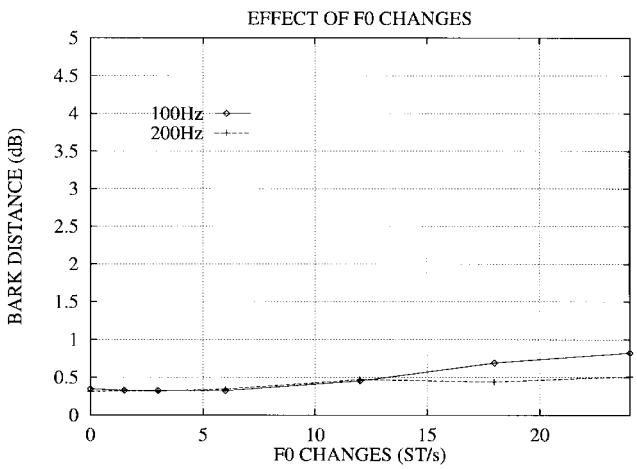

Fig. 14. Effect of F0 changes on perceptual spectral distance.

for one fundamental period T0 is

$$NI = 100 \times (2^{M \times T0/12} - 1). \tag{17}$$

Then, theoretically, a 24 ST/s f0 variation correspond to NI = 1.4% at 100 Hz, and NI = 0.7% at 200 Hz. The corresponding PAPR are close to 14 dB (100 Hz) and, 23 dB (200 Hz) in Fig. 13. The nearest values found in Fig. 7 are obtained for NJ = 1.5% at 100 Hz ($\simeq$ 15.5 dB), and for NJ = 1% at 200 Hz ($\simeq$ 24 dB). The values obtained are relatively close to the values predicted theoretically. Since large intonation variations correspond to large perturbations of the signal periodicity, it is not surprising that the aperiodic component becomes large in these cases. The PAPR is lower for the lower pitch frequency (100 Hz).

The effect of intonation variation on the spectral distance is displayed in Fig. 14. Large pitch glides increase the distance between EPC and SPC. The distances measured are generally quite low, especially compared to the distances obtained for additive random noise.

There is a 20–25 dB difference in PAPR between 0 and 12 ST/s conditions. Therefore, the results obtained for fixed pitch signals are probably too optimistic for real speech, particularly in the case of large F0 variations due to intonation changes. As for jitter and shimmer, the modulation aperiodicity introduced by pitch changes affects the aperiodic component. This is

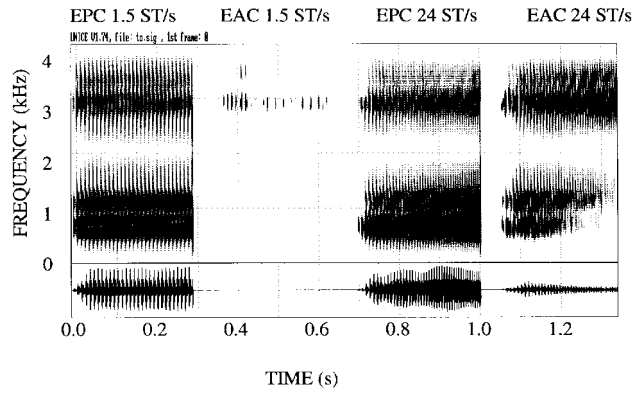

Fig. 15. Spectrograms measured for two F0 changes conditions (parallel formant synthesizer). F0 changes: 1.5 and 24 ST/s, F0 = 100 Hz.

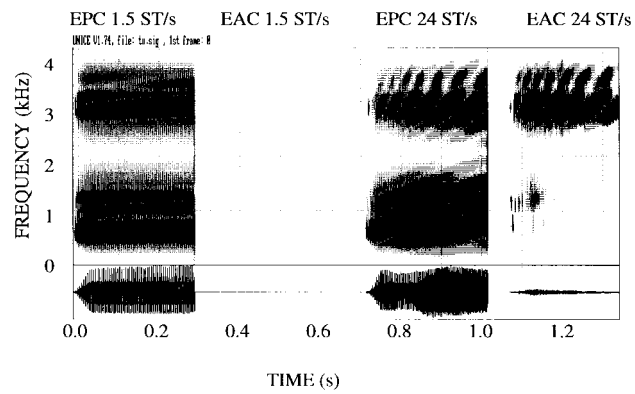

Fig. 16. Spectrograms measured for two F0 changes conditions (parallel formant synthesizer). F0 changes: 1.5 and 24 ST/s, F0 = 200 Hz.

clearly seen in the spectrogram plots in Figs. 15 and 16. Fig. 15 is for the 100 Hz base frequency, and Fig. 16 is for the 200 Hz base frequency. In both figures, the first pair of signals represent the EPC and the EAC for the 1.5 ST/s condition, and the second pair of signals represent the EPC and the EAC for the 24 ST/s condition. When there are few pitch changes, the energy in the EAC is negligible. But for large pitch variations the EAC becomes significant. It seems that pitch changes introduce high-frequency energy in the EAC. However, there is less energy in the EAC for the 200 Hz condition, compared to the 100 Hz condition.

IV. Summary and Conclusion

In this paper, we addressed the issue of the significance of a recently proposed PAP decomposition method for speech signals. The method is based on processing of the LP residual in the spectral domain. An iterative procedure is used for reconstruction of the aperiodic component.

Synthetic voiced speech signals are preferred for assessment, because they allowed easy control of the voice characteristics of the speech signal. Synthetic signals are generated using formant synthesis. The parameters under study represent different sources of aperiodicity encountered in speech production: additive random noise, jitter, shimmer, and pitch changes due to intonation. Three types of measurements are performed for each condition: periodic to aperiodic ratio (PAPR), perceptual spectral distance, and spectrograms.

The PAP decomposition algorithm is able to separate the additive random noise and the periodic component for a wide range of F0 variations. The PAPR is able to give some indication of the input HNR. Therefore, in natural speech also one may use the PAPR as an estimate of the input HNR, especially when jitter and shimmer are low. But for large jitter or shimmer values, the additive random noise and modulation noise are merged in the aperiodic component. While it is still possible to separate the periodic and aperiodic components, it is difficult to separate the different sources of aperiodicity in the aperiodic component. Thus, PAPR may be useful in the analysis of global voice quality, although it cannot be directly interpreted in terms of the underlying speech production parameters, like jitter, shimmer, or pitch changes in the excitation signal or aspiration noise in the voice source, when several different sources of aperiodicity are present in the signal.

The spectral distances are generally low, and reflected the amount of aperiodicity in the signal. The spectral distances are systematically higher for additive random noise than for modulation aperiodicities. This is a possible cue for discrimination between additive random noise and modulation aperiodicities.

The energy level and the time-frequency content of the EAC and the EPC are analyzed with the help of spectrograms. In the case of additive random noise, the spectrograms of the EAC are very close to those of the SAC. The EAC is a random noise showing the same type of modulation than the SAC. In the case of modulation aperiodicities, there is some periodic modulation in the spectrogram of the EAC. In this case, no random noise is present in the SAC, all the modulation aperiodicities are included in the SPC. Therefore, the EAC results only of leakages from the SPC. The amount of leakage in the spectrogram of the EAC is an indication of the amount of modulation aperiodicity in the SPC.

The present studies demonstrate that for normal speech, where modulation aperiodicities are low, the PAP decomposition algorithm can reliably separate additive random noise in the voice source and glottal periodic pulses. Thus, the algorithm can be used for studying the voice characteristics that are linked to aspiration and frication noise in the voice source. As a matter of fact, the output of the decomposition algorithm was successfully used for modification of voice source characteristics in speech synthesis experiments [26]. The algorithm can also be used in other applications such as voice pathology and speech acoustics studies.

ACKNOWLEDGMENT

This work was conducted in part while Prof. B. Yegnanarayana and V. Darsinos were visiting LIMSI. The authors express sincere thanks to the three anonymous reviewers for their valuable suggestions that improved the content and form of the paper.

REFERENCES

[1] D. Griffin and J. S. Lim, "Multiband excitation vocoder," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 1223–1235, 1988.

[2] N. B. Pinto, D. G. Childers, and A. L. Lalwani, "Formant speech synthesis: Improving production quality," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 1870–1886, 1989.

[3] J. Laroche, Y. Stylianou, and E. Moulines, "HNS: Speech modification based on a harmonic + noise model," *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Minneapolis, MN,1993, pp. 550–553.

[4] S. Grau-Grovel, C. d'Alessandro, and G. Richard, "A speech formant synthesizer based on Harmonic + random formant-waveforms representations," *Proc. ESCA-EUROSPEECH-93*, Berlin, Germany, pp. 1697–1700.

[5] T. Dutoit and H. Leich, "MBR-PSOLA: Text-to-speech synthesis based on an MBE re-synthesis of the segments database," *Speech Commun.*, vol. 13, pp. 435–440, 1993.

[6] O. Fujimura, "Approximation to voice aperiodicity." *IEEE Trans. Audio Electroacoust.*, vol. AU-16, pp. 68–73, 1968.

[7] E. Yumoto, W. J. Gould, and T. Baer, "Harmonics-to-noise ratio as an index of the degree of hoarseness," *J. Acoust. Soc. Amer.*, vol. 71, pp. 1544–1550, 1982.

[8] F. Klingholz, "The measurement of the signal-to-noise ratio (SNR) in continuous speech," *Speech Commun.*, vol. 6, pp. 15–26, 1987.

[9] P. Cook, "Aperiodicities in the singer voice source." *J. Acoust. Soc. Amer.*, vol. 91, p. 2434(A), 1992.

[10] G. Krom, "A cepstrum-based technique for determining a harmonics-to noise ratio in speech signals," *J. Speech Hearing Res.*, vol. 36, pp. 254–266, 1993.

[11] X. Serra and J. Smith, "Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Comput. Music J.*, vol. 14, no. 4, 1990.

[12] C. Chafe, "Pulsed noise in self-sustained oscillations of musical instruments," *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Albuquerque, NM, pp. 1157–1160, 1990.

[13] C. d'Alessandro, B. Yegnanarayana, and V. Darsinos, "Decomposition of the speech signal into deterministic and stochastic components," in *Proc. ICASSP-95*, Detroit, MI, pp. 760–763.

[14] B. Yegnanarayana, C. d'Alessandro, and V. Darsinos, "An iterative algorithm for decomposition of the speech signal into periodic and aperiodic components," this issue, pp. 1–11.

[15] J. Hillenbrand, "A methodological study of perturbation and additive noise in synthetically generated voice signals," *J. Speech Hearing Res.*, vol. 30, pp. 448–461, 1987.

[16] D. J. Hermes, "Synthesis of breathy vowels: Some research methods," *Speech Commun.*, vol. 10, pp. 497–502, 1991.

[17] G. Fant, J. Liljencrants, and Q. G. Lin, "A four-parameter model of glottal flow," Quart. Prog. Stat. Rep., Speech Trans. Lab., R. Inst. Technol., Stockholm, Sweden, vol. 4, pp. 1–17, 1985.

[18] D. Klatt D., "Software for a cascade/parallel formant synthesizer," *J. Acoust. Soc. Amer.*, vol. 67, pp. 971–994, 1980.

[19] C. d'Alessandro, "Time-frequency speech transformation based on an elementary waveform representation," *Speech Commun.*, vol. 9, pp. 419–431, 1990.

[20] D. G. Childers, and C. K. Lee, "Vocal quality factors: Analysis, synthesis and perception," *J. Acoust. Soc. Amer.*, vol. 90, pp. 2394–2410, 1991.

[21] H. Holien, J. Michel, E. T. Doherty, "A method for analyzing vocal jitter in sustained phonation," *J. Phonet.*, vol. 1, pp. 85–91, 1973.

[22] Y. Horii, "Vocal shimmer in sustained phonation," *J. Speech Hearing Res.*, vol. 23, pp. 202–209, 1980.

[23] J. D. Markel, A. H. Gray, *Linear Prediction of Speech.* New York: Springer-Verlag, 1976.

[24] A. Papoulis, *Signal Analysis.* New York: McGraw-Hill, pp. 244–248, 1984.

[25] A. Sekey and B. A. Hanson, "Improved 1-bark bandwidth auditory filter," *J. Acoust. Soc. Amer.*, vol. 75, pp. 1902–1904, 1984.

[26] G. Richard and C. d'Alessandro, "Analysis/synthesis and modification of the speech aperiodic component," *Speech Commun.*, vol. 19, pp. 221–244, 1996.

Christophe d'Alessandro (M'95), for a photograph and biography, see this issue, p. 11.

Vassilis Darsinos, for a photograph and biography, see this issue, p. 11.

B. Yegnanarayana (M'78–SM'84), for a photograph and biography, see this issue, p. 11.