An Iterative Algorithm for Decomposition of Speech Signals into Periodic and Aperiodic Components

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

***Abstract*—The speech signal may be considered as the output of a time-varying vocal tract system excited with quasiperiodic and/or random sequences of pulses. The quasiperiodic part may be considered as the deterministic or periodic component and the random part as the stochastic or aperiodic component of the excitation. In this paper, we discuss issues involved in identifying and separating the periodic and aperiodic components of the source. The decomposition is performed on an approximation to the excitation signal, instead of decomposing the speech signal directly. The linear prediction residual signal is used as an approximation to the excitation signal of the vocal tract system. Speech is first analyzed to determine the voiced and unvoiced parts of the signal. Decomposition of the voiced part into periodic and aperiodic components is then accomplished by first identifying the frequency regions of harmonic and noise components in the spectral domain. The signal corresponding to the noise regions is used as a first approximation to the aperiodic component. An iterative algorithm is proposed which reconstructs the aperiodic component in the harmonic regions. The periodic component is obtained by subtracting the reconstructed aperiodic component signal from the residual signal. The individual components of the residual are then used to excite the derived all-pole model of the vocal tract system to obtain the corresponding components of the speech signal. Experiments were conducted using synthetic speech. They demonstrated the ability of the algorithm for decomposition of a synthetic speech signal made of a mixture of periodic and aperiodic components. Application to natural speech is also discussed.**

***Index Terms*—Periodic and aperiodic decomposition, spectral extrapolation, spectral modeling, speech analysis/synthesis, voice source analysis.**

I. Introduction

According to linear acoustic theory, the speech production process can be viewed as consisting of a source component and a filter (or system) component [1]. One of the objectives in speech analysis is to study the characteristics of the source and system by processing the speech signal. Normally, the source is modeled as either voiced or unvoiced, and the voice source as quasiperiodic sequences of glottal pulses. But in real speech even the voiced part consists of some random component [2], [3]. This is particularly obvious in voiced fricative (e.g., /v/, /z/), in breathy vowels (e.g., high vowels in unvoiced consonantal contexts), or in speech produced with a weak phonatory effort. The random component is also present in normal vowels due to turbulence of air around the instant of glottal closure, which gives rise to aspiration noise [4], [5].

In this paper, the harmonic (or deterministic) component is termed as the periodic component and the random (or stochastic) component as the aperiodic component. The study of the aperiodic component of excitation is important in speech analysis. The aperiodic component may help in characterizing voice quality attributes such as breathiness or roughness. *Breathiness* is associated to the impression of glottal air leakage and to turbulence noise during phonation. *Roughness* is defined by the presence of a low-frequency noise component [6]. Moreover, including the aperiodic component in voiced excitation may help to produce a natural sounding synthetic speech [7]. Detailed characterization of the source may also help in generating synthetic speech with desired voice characteristics [8]. Several methods have been proposed to separate a speech signal into periodic and aperiodic components: they are based on sinusoidal modeling, on harmonic plus noise modeling, or on the multiband excitation vocoder.

Serra and Smith [9] proposed a method for analysis and synthesis of musical sounds in terms of a deterministic and a stochastic component. The method is based on sinusoidal modeling, initialy presented by McAulay and Quatieri [10], [11], and further refined by George and Smith [12] . All the energy present at harmonic frequencies is associated to the deterministic component. This is not actually the case in speech, where broadband noise is spread out over the whole spectrum. Thus, the stochastic component cannot entirely be attributed to the random part of the source. Continuity of the frequencies of the sinusoids are avoided in the proposal based on harmonic + noise model [13], where a pitch synchronous analysis is made for the harmonic decomposition. Typically, speech is lowpass filtered, and a fixed number of sinusoids are fitted to the lowpass component. The highpass (above 3–4 kHz) component only is associated with the noise component. Thus, the noise component cannot represent accurately the random part of the source. This model is merely a representation of speech signal suitable for speech synthesis by concatenation, but it can hardly be used for measurement of the aperiodic component in speech. In the multiband excitation vocoder [14], an initial source/filter decomposition is performed. The excitation part is then processed using the short-time Fourier transform (STFT).

Manuscript received August 3, 1995; revised February 20, 1997. This work was supported in part by a grant of the University Paris XI, 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.

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

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.

Publisher Item Identifier S 1063-6676(98)00588-4.

For each frame of the residual, the frequency domain is segmented into voiced and unvoiced regions. These regions are defined using a decision criterion for each category. This method has been applied to speech synthesis by concatenation [15] and to musical sounds [16]. Again, this method takes a binary decision on the frequency regions that are either noisy or voiced. this assumption does not reflect the speech production mechanism, where broadband noise is mixed with quasiperiodic voiced excitation.

None of the above methods seems able to extract the degree of aperiodicity (i.e., proportion of the periodic and aperiodic components of excitation) in a given segment, because they make a binary decision at each frequency sample. But in actual speech, it is likely that a frequency sample may consist of both the periodic and the aperiodic components simultaneously. Therefore, a decomposition method should be able to separate the relative proportion of these components at each frequency sample. This can be done by processing both components in the complex frequency domain, rather than taking decision on magnitude spectra only.

In this paper, the periodic component is associated with the periodic part of the excitation, and the aperiodic component with the random part of the excitation. A key point of the work is the introduction of an iterative algorithm for estimating the aperiodic component of the excitation [17] . The idea is to derive a first approximation of the periodic and aperiodic components using a harmonicity criterion. Each frequency sample for each frame is associated to one or the other of the components. But this approximation must be refined, because both components are physically present at each frequency sample. The iterative algorithm, based on discrete Fourier transform (DFT)/inverse discrete Fourier transform (IDFT) pairs, is used for reconstruction in the complex domain of the aperiodic component in the region labeled as periodic at the first stage of processing. Finally, the periodic component of the excitation is obtained by subtracting the estimated aperiodic component from the excitation signal, and synthesis is performed using a simple overlap-add scheme. Contrary to other methods, no binary decision is taken in the spectral domain, but complex spectra of the periodic and aperiodic components are computed for the full band and for each frame.

In Section II, the model that forms the basis for the method proposed in this paper is described. The actual method for decomposition of the excitation signal into the two components is described in Section III. Some assessment using synthetic speech and illustrations are given in Section IV to demonstrate the utility of the proposed method. Conclusions are given in Section V.

II. Basis for the Proposed Decomposition Method

A. Speech Model

We assume the following model for speech production:

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

where $s(t)$ is the speech signal, $v(t)$ is the impulse response of the vocal tract system, $e(t)$ is the excitation signal, $p(t)$ is the quasiperiodic part of the excitation, and $r(t)$ is the aperiodic part of the excitation.

In the spectral domain we can write

$$\begin{aligned} S(w) &= |S(\omega)|e^{j\theta(\omega)} \\ &= [P(\omega) + R(\omega)]V(\omega) \\ &= (|P(\omega)|e^{j\theta_p(\omega)} + |R(\omega)|e^{j\theta_r(\omega)})|V(\omega)|e^{j\theta_v(\omega)} \end{aligned} \tag{2}$$

where $S(\omega), P(\omega), R(\omega)$ and $V(\omega)$ are the Fourier transforms of $s(t), p(t), r(t)$ and $v(t)$, respectively.

If the periodic and aperiodic parts of the excitation are uncorrelated, then

$$|S(\omega)|^2 = [|P(\omega)|^2 + |R(\omega)|^2]|V(\omega)|^2. \tag{3}$$

Depending on the proportion of the periodic and the aperiodic components at each frequency the different components of the excitation source get prominence. The method for deriving the components of the source signal contains the following steps of processing.

- *Linear prediction (LP) residual*: The speech signal is separated into an approximate excitation and filter components using LP analysis. The LP residual signal is then decomposed into short (10–20 ms) segments of overlapping signals.
- *Periodic and aperiodic regions*: For each of the overlapping segments of the excitation signal, the frequency bands of the periodic and aperiodic regions are determined using the DFT and the cepstrum of the signal in the analysis segment.
- *Decomposition*: For each segment, the aperiodic component is reconstructed using the iterative algorithm described in Section III. The aperiodic component of the plain excitation signal is obtained by adding in the time domain the components for each of the overlapping segments. Subtracting this aperiodic component from the LP residual signal gives the periodic component of the excitation.
- *Synthesis*: The aperiodic and periodic components are used separately to excite the time varying all-pole filter to obtain the corresponding components of the speech signal.

Fig. 1 gives a schematic diagram of the proposed algorithm. In the following we will discuss details of the above steps.

B. Linear Prediction Residual

The objective is to separate the speech signal into components corresponding to the periodic and aperiodic components of the excitation source. Most of the available methods for decomposition attempt to process the speech signal directly. But windowing the speech signal produces undesirable features in the analysis due to truncation effects, because we are truncating a signal consisting of highly correlated samples. The effects of truncation can be reduced significantly, if the decomposition is attempted on the residual excitation obtained by removing the correlated part from the speech signal [18]. This is because the excitation signal samples are nearly uncorrelated. Therefore, in this paper we propose to

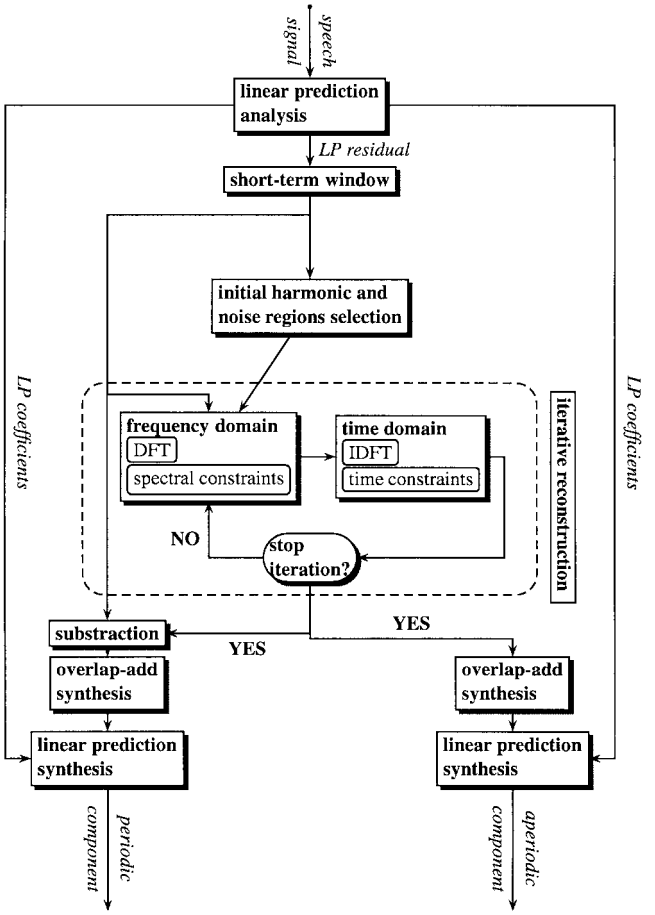

Fig. 1. Schematic diagram of the periodic–aperiodic decomposition algorithm.

decompose the signal corresponding to the excitation signal first. Then the components of the excitation signal are used to generate the corresponding components of the speech signal.

Note that even an approximation to the residual signal will significantly improve the effectiveness of the resulting decomposition. LP analysis is performed on each of the overlapping segments of speech data. The output of this analysis is a set of LP coefficients for each segment. The residual signal is obtained by passing the speech signal through the inverse filter defined by these linear predictive coefficients (LPC's).

These overlapping frames of the LP residual signal are considered for further analysis. For each of these frames, the data is multiplied with a data window (the Hamming window was preferred [19]), and a DFT is computed. The size of the DFT was chosen to be twice the size of the data in the analysis frame. Each DFT coefficient is in general contributed to by both the periodic part and the aperiodic part. It is necessary to determine the relative strengths of the periodic and aperiodic parts of the residual signal. If a DFT coefficient corresponds mostly to the aperiodic part, there is no point in considering that coefficient in the summation to obtain an initial estimate of the periodic component. Therefore, to obtain an approximate periodic component, $e_p(n)$, of the residual, sum only those

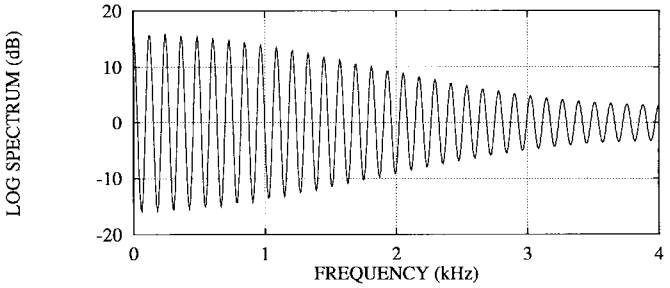

Fig. 2. Log magnitude spectrum of a frame of the harmonic part of the excitation signal for a vowel, derived from liftered spectrum. Frequency dots corresponding to positive values are associated to the harmonic regions, and frequency dots corresponding to negative values are associated to the noise regions.

DFT coefficients ($k \in F_p$) for which $|P(k)| > |R(k)|$. That is

$$e_p(n) \approx \sum_{k \in F_p} E(k) \exp\left(\frac{j2\pi}{N}\, nk\right) \tag{4}$$

$$\approx = \sum_{k \in F_p} [P(k) + R(k)] \exp\left(\frac{j2\pi}{N}\, nk\right) \tag{5}$$

where N is equal to the size of the DFT and $0 < k < N-1$. Here, F_p is the set of frequency points at which $|P(k)| > |R(k)|$.

If we can derive the values of either the periodic part $P(k)$ or the aperiodic part $R(k)$, then the other part can be obtained by merely subtracting the known part from the overall residual signal. But in general it is not easy to determine either of these individual components because they are combined by a complex addition in the residual signal. However, it seems possible to determine the ratio of $|P(k)|$ and $|R(k)|$. In such a case, we may consider those values of k for which the periodic part is higher than the random part, and use those frequency samples in the summation to estimate the periodic part.

C. Identification of Noise and Harmonic Regions Using Cepstrum

In our approach, we propose to decompose the frequency domain initially into two regions, one belonging predominantly to the periodic part and the other to the random part of the spectrum. We use a straightforward harmonic-based selection for this purpose, in which the knowledge of the pitch period is used to mark the harmonic and noise regions in the spectrum as shown in Fig. 2.

In the figure, the sinusoidal log spectrum is due to the peak in the cepstrum at the pitch period. The frequency intervals corresponding to the positive values of the log spectrum are identified as the harmonic regions and the frequency intervals with negative values of the log spectrum are identified as the noise regions.

The idea of using cepstrum to derive a comb filter in the spectral domain for computing the harmonic to noise signal energy ratio was proposed in [6]. Here, we use the cepstrum analysis to discuss the manner in which the periodic $(P(k))$ and aperiodic $(R(k))$ components are combined, and to derive the ratio of the two spectral components as a function of frequency. Note that we can only obtain the relative values

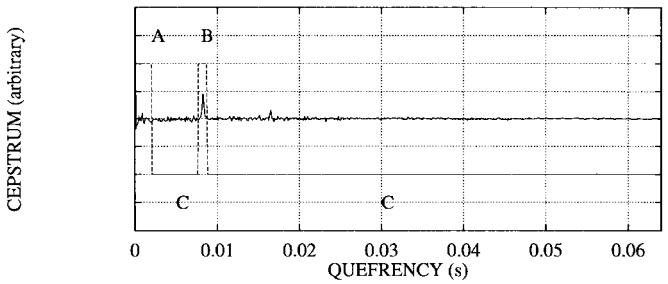

Fig. 3. Real cepstrum of a segment of the LP residual signal for voiced speech. Three marked regions are corresponding to: (a) vocal-tract system, (b) harmonic part of excitation, and (c) noise part of excitation.

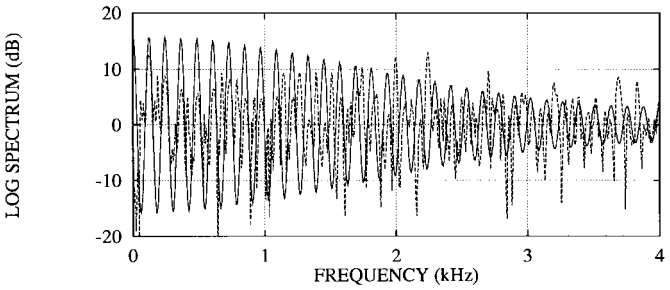

Fig. 4. Log magnitude spectra of a frame of the harmonic and noise parts of the LP residual signal for a vowel, derived from liftered cepstrum.

of these components due to the log operation involved in the cepstral analysis.

Let us consider the speech production model given in (3) again. If we consider the log magnitude spectrum, we get

$$log|S(\omega)| = log|P(\omega) + R(\omega)| + log|V(\omega)|. \quad (6)$$

The IDFT of this log magnitude spectrum gives the real cepstrum $c(t)$ in the quefrency domain. That is

$$c(t) = IDFT[\log|S(w)|]. \quad (7)$$

Fig. 3 shows the real cepstrum for a segment of voiced speech. The frequency domain of the cepstrum can be split approximately into three distinct regions. The low quefrency region (marked A in Fig. 3) in the range 0–0.002 s is mainly due to the vocal tract system characteristics. The rest of the quefrency region can be attributed to the excitation part. In the region corresponding to the excitation, the portion marked B in Fig. 3, i.e., 0.01 s around the cepstral peak at the pitch period, can be attributed to the harmonic or periodic part. The remaining excitation portion of the quefrency region can be attributed to the noise or random part. Note that the region identified as random may contain cepstral peaks around twice and thrice the pitch periods. But the energy contribution due to these peaks can be assumed to be insignificant compared to the energy contribution due to the region around the cepstral peak at the pitch period. The choices for the widths for the three parts in the quefrency domain are only approximate. The reason for the choice of a region around the pitch region is due to broadening of the cepstral peak when two successive periods of the residual signal are not exactly equal. The broadening also takes place due to windowing of the analysis frame of the residual signal.

The main objective of this cepstral analysis is to show that one can derive an approximation to the harmonic-to-noise ratio (HNR) at each frequency sample. Assuming that the energy due to the region around the pitch peak in the cepstrum corresponds to the harmonic part and the rest of the energy corresponds to the noise part, the ratio of the energies of the respective components is derived as HNR. The ratio of the harmonic and random parts of the spectrum can be obtained at each frequency point by computing the IDFT of the corresponding liftered components (see Fig. 4). From the ratio, it should be possible to retain only those frequency samples which have a higher periodic component over the aperiodic component in order to derive an approximation to the periodic component.

As mentioned earlier, we have adopted a simple harmonic-based selection in our studies. Since it may be difficult to identify the location of the cepstral peak at the pitch period exactly due to the broadening effect or due to noise, this location can be determined by a separate analysis of the speech signal for pitch extraction. The analysis automatically identifies whether the frame is voiced or unvoiced (see Section IV for details on the pitch detection and voiced/unvoiced decision algorithms).

In the voiced/unvoiced decision, obtained separately for each frame, the decision is biased in favor of voiced frames over the unvoiced frames. That is, the algorithm is designed to make errors only in one direction (i.e., to label unvoiced frames as "voiced," rather than labeling voiced frames as "unvoiced"). The errors in voiced/unvoiced labeling can be corrected by further processing the voiced frames after decomposition. For example, one will obtain a periodic component with almost no energy for frames incorrectly labeled as "voiced."

Let us denote the sets of frequency samples in the harmonic and noise regions as F_p and F_r, respectively. A straightforward method is to use the DFT coefficients in the harmonic regions and perform an IDFT to obtain the periodic component. Likewise, one can use the DFT coefficients in the noise regions to derive the aperiodic component. But a better approach is to use this information as an initial estimation and to reconstruct the periodic and aperiodic components iteratively as discussed in the next section.

III. Iterative Reconstruction of the Random Component

A first approximation of the aperiodic components can be obtained in the frequency domain by selecting suitable frequency samples F_r. But this first approximation is not satisfactory: it is not physically significant, because the spectrum of the aperiodic component is a spectral comb. The problem addressed in this section is the reconstruction of the aperiodic component for all frequencies in the complex spectral domain. This problem can be viewed as an extrapolation problem, and we will make use of an extrapolation algorithm to solve it.

We propose an iterative procedure to reconstruct the aperiodic component first, and then use it to determine the periodic component. Note that the harmonic regions F_p correspond to

the positive portion of the log spectrum in Fig. 2. But the DFT coefficients in some parts of these harmonic regions may be contributed to mainly by noise. This is true especially in the low signal-to-noise ratio (SNR) regions of the spectrum of the speech signal. On the other hand, in the valley regions of the harmonics, namely, in the noise regions F_r, the DFT coefficients are mainly due to noise. This is evident from Fig. 4, where the valley regions are mostly dominated by the noise part (dotted line), whereas the harmonic or peak regions are sometimes affected by the noise part as well.

Thus, we hypothesize that the DFT coefficients in the valleys between harmonic regions (i.e., noise regions) are mostly due to aperiodic component only. In practice, the sidelobe effects of the windowing may produce significant values in the noise regions. We discuss the effects of windowing in Section IV, in the light of experimental results. For the present, we assume that the subset F_r of the frequency samples provide an initial estimation of the aperiodic part. The problem is to estimate the complex aperiodic component in the harmonic regions in the frequency domain.

An iterative method for bandlimited signal extrapolation problems was proposed by Gerchberg and Papoulis [20, pp. 244–248]. We extend the method for the case of comblike filtered noise spectrum to reconstruct the aperiodic component. Starting with zero values in the harmonic regions and the actual DFT coefficients in the noise regions, an estimate of the aperiodic component in the harmonic regions is obtained by iteratively moving from the frequency domain to the time domain and vice versa, imposing finite duration constraint in the time domain, and the known noise samples constraint in the frequency domain.

Let N be the number of points in the DFT computation. Then the number of data samples should be less than or equal to $N/2$. In our case, we assume $N/2-1$ data samples (with $N/2$ even). Let $r(n)$ and $R(k)$ represent the true aperiodic component and its DFT, respectively, that we are trying to reconstruct by the iterative algorithm. Note that $r(n)=0$, for $n \geq N/2$ and $R(k)=E(k)$, for $k \in F_r$ (noise regions in the frequency domain), where $E(k)$ are the DFT coefficients of the analysis segment of the LP residual.

The iterative algorithm for reconstruction of the aperiodic component is as follows:

First Iteration: We form the initial estimate of the DFT samples of the aperiodic component as

$$R_0(k)=\begin{cases} E(k), & \text{for } k \in F_r \text{ (noise regions)} \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

and compute its IDFT $r_o(n)$. Since we have started with a segment of $N/2-1$ data samples, and we are using a N-point DFT, the time samples beyond $N/2-1$ are set to zero. That is, form a signal

$$\hat{r}_0(n)=\begin{cases} r_0(n), & \text{for } n < N/2 \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

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

$$R_m(k)=\begin{cases} E(k), & \text{for } k \in F_r \\ \hat{R}_{m-1}(k), & \text{otherwise} \end{cases} \tag{10}$$

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

$$\hat{r}_m(n)=\begin{cases} r_m(n), & \text{for } n < N/2 \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

We shall show that the functions $R_m(k)$ tends to $R(k)$ as m $\rightarrow \infty$, in the mean square (MS) sense.

Since $\hat{r}_m(n)=r_m(n)$, for $n < N/2$, and $r(n)=\hat{r}_m(n)=0$, for $n \geq N/2$, we have

$$\sum_{n=1}^{N} |r(n)-r_m(n)|^2 > \sum_{n=1}^{N} |r(n)-\hat{r}_m(n)|^2. \tag{12}$$

Likewise, in the frequency domain, since $R_{m+1}(k)=R(k)=E(k)$, for $k \in F_r$ (noise regions), and $R_{m+1}(k)=\hat{R}_m(k)$, for $k \notin F_r$, we have

$$\frac{1}{N}\sum_{k=1}^{N} |R(k)-\hat{R}_m(k)|^2 > \frac{1}{N}\sum_{k=1}^{N} |R(k)-R_{m+1}(k)|^2. \tag{13}$$

Using Parseval's formula for the discrete case, we have the following result:

$$\begin{aligned} &\frac{1}{N}\sum_{k=1}^{N} |R(k)-R_m(k)|^2 \\ &\quad = \sum_{n=1}^{N} |r(n)-r_m(n)|^2 > \sum_{n=1}^{N} |r(n)-\hat{r}_m(n)|^2 \\ &\quad = \frac{1}{N}\sum_{k=1}^{N} |R(k)-\hat{R}_m(k)|^2 \\ &\quad > \frac{1}{N}\sum_{k=1}^{N} |R(k)-R_{m+1}(k)|^2. \end{aligned} \tag{14}$$

This result shows that successive iterations will reduce the mean-square value of the error

$$\frac{1}{N}\sum_{k=1}^{N} |R(k)-R_m(k)|^2. \tag{15}$$

Furthermore, the limit of this error is zero (i.e. the functions $R_m(k)$ tends to $R(k)$ as $m \rightarrow \infty$). The limit exists because the error is nonnegative and decreasing.

Suppose that the limit is strictly positive (and not zero). In this case there exists a function $L(k) \neq R(k)$, and the functions $R_m(k)$ tend to $L(k)$ as $n \rightarrow \infty$. We have

$$L(k)=\begin{cases} E(k)=R_m(k)=R(k), & \text{for } k \in F_r \\ \neq R(k), & \text{otherwise.} \end{cases} \tag{16}$$

In the time domain, we have also $l(n)=0$ for $n \geq N/2$. This is because the functions $\hat{r}_m(n)$ tend to $l(n)$ as $n \rightarrow \infty$. Form the difference h between l and g. This function must satisfy

$$H(k)=\begin{cases} 0, & \text{for } k \in F_r \\ \neq 0, & \text{otherwise} \end{cases} \tag{17}$$

and $h(n)=0$ for $n \geq N/2$. In other words, h must be a comb-filtered signal in the frequency domain, because it is zero for $k \in F_r$, and it must also be bounded in time to the

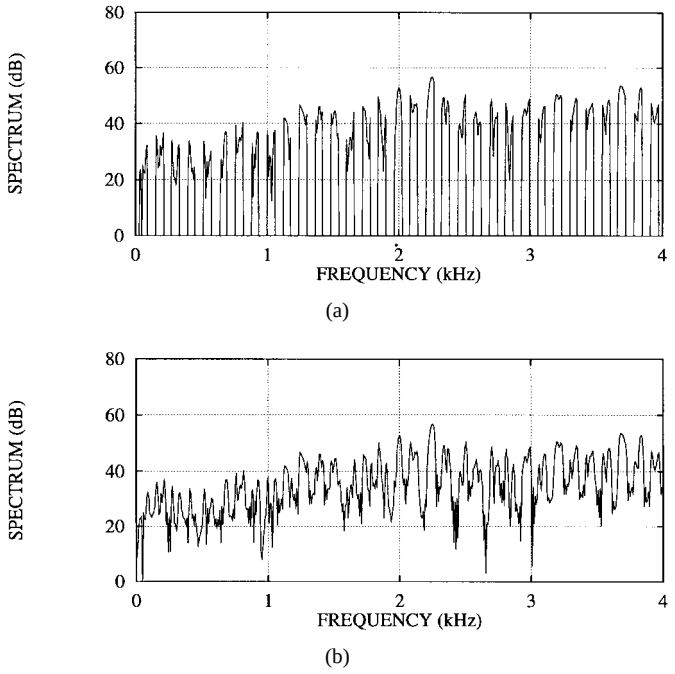

Fig. 5. Log magnitude spectra of a frame of the noise component of excitation for a vowel: (a) before the start of iterations; (b) after ten iterations.

interval $[1, N/2-1]$. Clearly, these two constraints can not be satisfied simultaneously. Therefore, $L = R$, and the functions R_m converge to R.

Therefore, the iterations can be repeated until the difference between the energies of the aperiodic component $\hat{r}_m(n)$ for two successive iterations is below a prefixed threshold.

Fig. 1 summarizes the algorithm for the proposed method of decomposition. Fig. 5 shows the reconstructed aperiodic component spectra for a vowel segment before the start of the iteration and after ten iterations. In this example, the selection of harmonic regions was done using the regions of positive values in the function plotted in Fig. 2. It is interesting to note that the samples of the aperiodic component build up in the harmonic regions after every iteration.

Similar iterative algorithms can be developed for reconstruction of the periodic component in the noise regions or reconstruction of both aperiodic component in the harmonic regions and periodic component in the noise regions, simultaneously. We have found that the results are not significantly different in all these cases. Hence, we have used only the aperiodic component signal reconstruction algorithm.

The periodic component is obtained by subtracting the aperiodic component $r_m(n)$ obtained at the last iteration from the residual signal. Fig. 6 shows the spectra of a segment of the LP residual, the periodic and aperiodic components using the proposed iterative algorithm. It is interesting to see that the proposed method has indeed modified the spectral shape, especially in the low amplitude periodic regions.

Note that, in this method, the decomposition is practically imposed by the choice of the harmonic and noise regions in the frequency domain. The iterative algorithm helps in obtaining realizable component signals satisfying the imposed constraints in the frequency domain and the finite support constraint in the time domain. In this process, the unknown

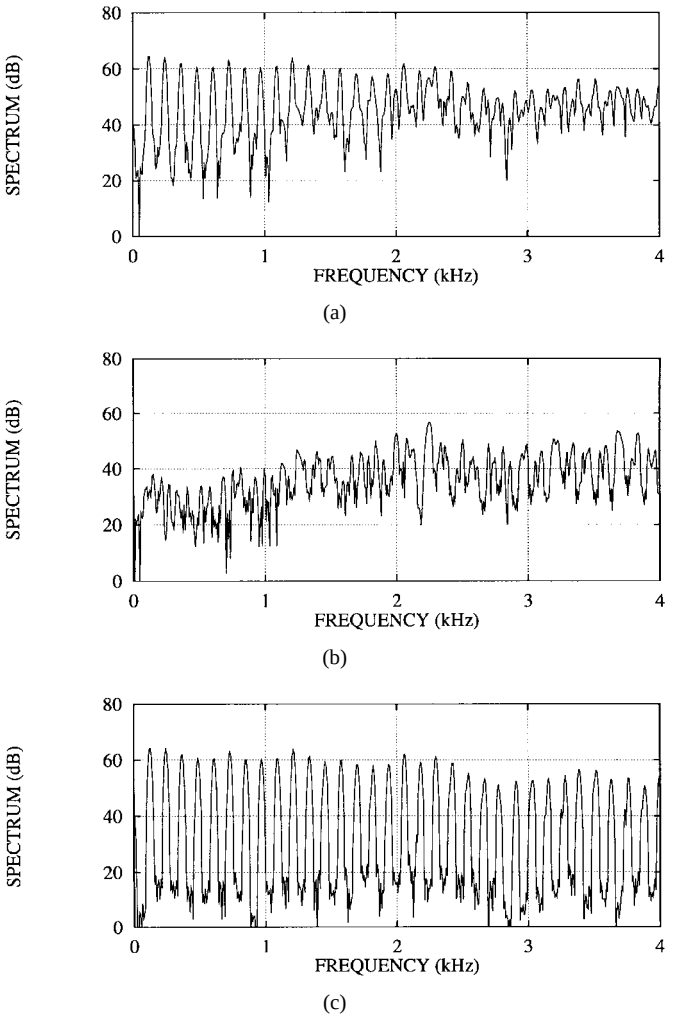

Fig. 6. Log magnitude spectra derived from a voiced segment. (a) Spectrum of LP residual. (b) Spectrum of the aperiodic component of the LP residual signal. (c) Spectrum of the periodic component of the LP residual signal.

aperiodic component values are built up in the harmonic regions. The iterative algorithm thus can be viewed as deriving a physically realizable signal starting with an ideal comb filter in the frequency domain.

The periodic and aperiodic components of the residual are obtained for each overlapping analysis frame, and the component for the entire utterance is derived by simply adding the sample values in the overlapping regions for successive frames. The speech signals corresponding to these components can be generated by passing these component residual signals separately through the time varying all-pole filter.

IV. Experiments

In this section, we discuss some experimental results obtained with the proposed algorithm. Synthetic speech is considered first. Then application of the method to natural speech is discussed.

A. Decomposition of Synthetic Speech: Methodology

First, we have considered the case of a synthetic voiced segment to study the behavior of the algorithm, and partic-

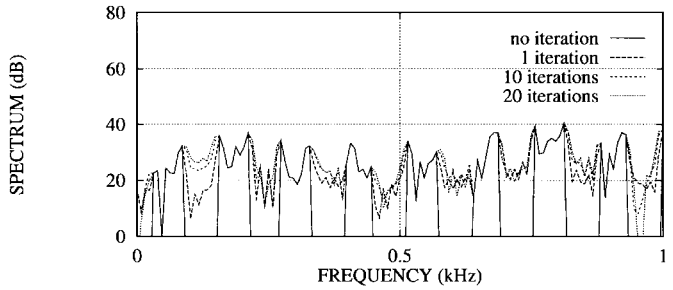

Fig. 7. Effect of the iterations on the reconstruction of the aperiodic component in spectral domain. Log magnitude spectrum of the aperiodic component of the LP residual signal (in the 0–1 kHz band) computed with zero, one, ten, and 20 iterations.

ularly the sidelobe effects due to windowing the residual. An all-pole model was excited with a glottal pulse with different levels of noise around the glottal closure. In this first series of experiments, synthetic voice was preferred for evaluation, because it allowed for an accurate control on all the signal parameters. Moreover, the case of breathy vowels was considered because it is a difficult and interesting situation of periodic and aperiodic mixture in actual speech.

All the synthetic stimuli were generated by means of a formant synthesizer [21]. The excitation part of the synthetic signals was the sum of two components: glottal pulse (using the classical LF-model [22]) and noise burst. Noise bursts were made of Gaussian white noise modulated by a rectangular time window, centered around the instant of glottal closure. Formant filters were set according to the acoustic values of the French vowel /*a*/. Pitch was either 120 or 200 Hz (the average values for male and female speakers). Duration of the noise burst was varied is three steps: 0, 60, and 100% of the fundamental period (i.e., 0, 5, and 8.3 ms for the 120 Hz pitch condition and 0, 3, and 5 ms for the 200 Hz pitch condition). The periodic to aperiodic ratio (or HNR) was varied in four steps: ∞ dB (no aperiodic synthetic signal), 20, 10, and 5 dB. This HNR corresponds to the power ratio between the glottal pulses signal and the noise bursts signal.

Three synthetic signals were computed for each condition. The synthetic aperiodic component was obtained by passing the aperiodic part of the source (a train of pitch-synchronous noise bursts) through the vocal tract formant filters. The synthetic periodic component was obtained by passing the periodic part of the source (a train of glottal pulses) through the vocal tract formant filters. The plain synthetic signal was the sum of the synthetic aperiodic component and the synthetic periodic component. Three signals where also obtained after decomposition of the total synthetic signal: the measured aperiodic component, the measured periodic component, and the plain measured component, which is formed by summing the measured aperiodic component and the measured periodic component. Figs. 8 and 9 give an example of five versions of a signal used in the experiments. The signal at the top of the graph is the plain synthetic signal. The second and third signals are the synthetic aperiodic and measured aperiodic components. The two signals at the bottom of the figure are the synthetic periodic and measured periodic components.

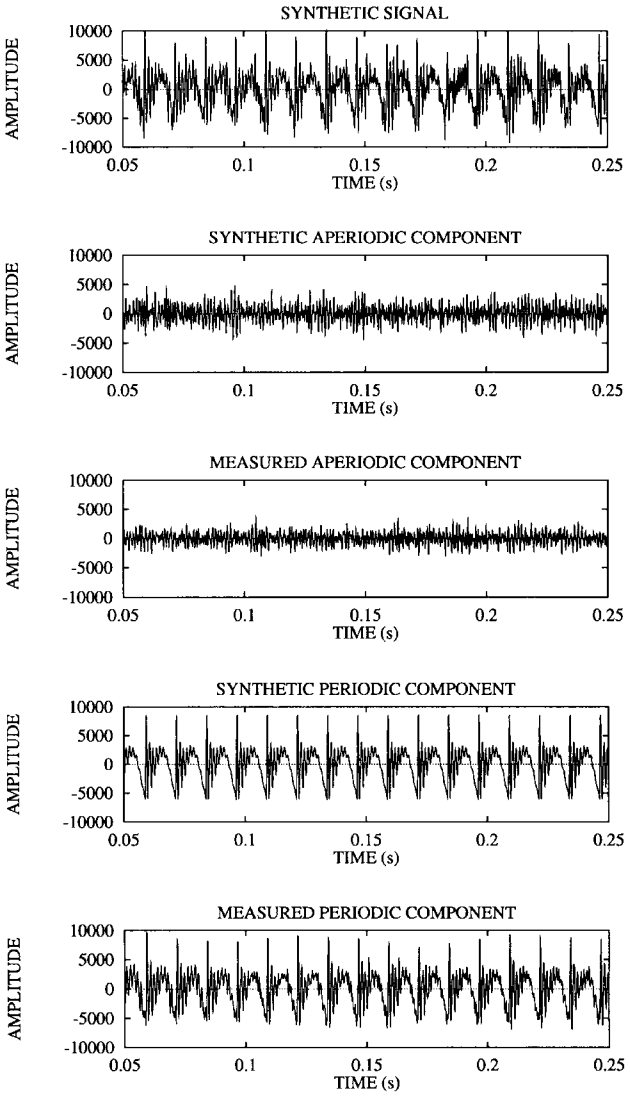

Fig. 8. Oscillograms (from top to bottom) of: synthetic signal; the corresponding synthetic aperiodic component; measured aperiodic component; synthetic periodic component; measured periodic component. $F0 = 80$ Hz, burst duration = 100% of $1/F0$, HNR $= 5$ dB.

B. Decomposition of Synthetic Speech: Results

Quantitative assessment of the algorithm was obtained by comparing the HNR of the synthetic signals (power ratio of the synthetic periodic signal and synthetic aperiodic signal) with the corresponding HNR measured for decomposed signals (power ratio of the measured aperiodic signal and the measured periodic signal). If the decomposition method was perfect, we should expect the measured HNR to be equal to the input HNR, for all the conditions. The results, reported in Table I, show that the measured HNR is actually close to the input HNR for all the conditions. The difference is on average less than 1.7 dB, for the 20, 10, and 5 dB conditions. When there is no aperiodic component in the synthetic signal, the measured aperiodic component is about 40 dB lower than the measured periodic component. This gives an idea of the accuracy of the method: the artificial aperiodic component resulting from the decomposition of a synthetic periodic signal is actually very

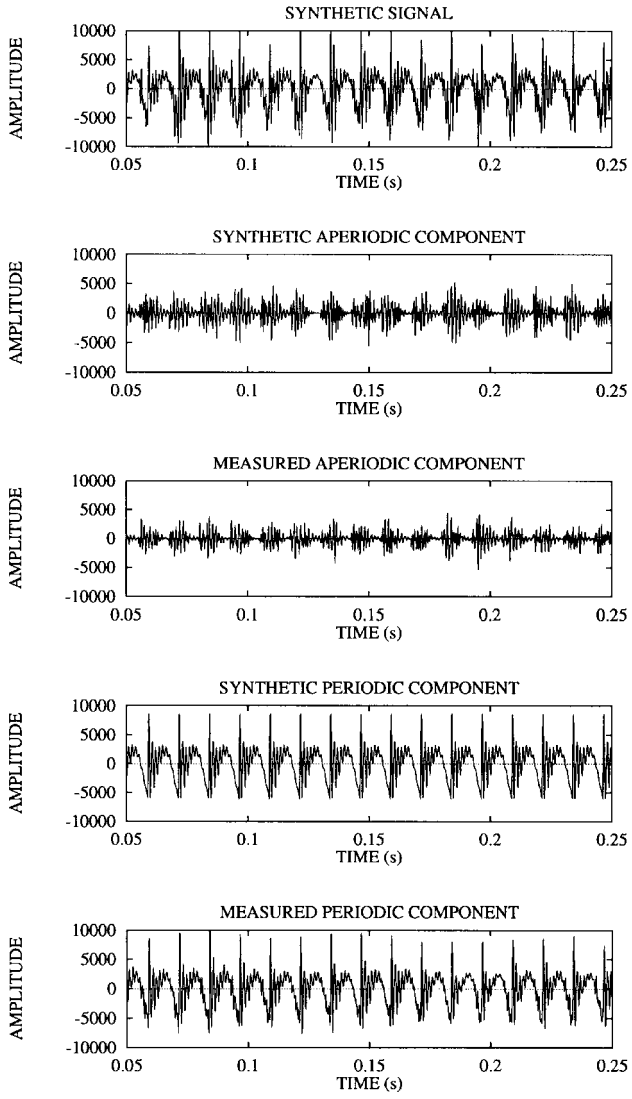

Fig. 9. Oscillograms of (from top to bottom): synthetic signal; the corresponding synthetic aperiodic component; measured aperiodic component; synthetic periodic component; measured periodic component. $F0 = 80$ Hz, burst duration = 60% of $1/F0$, HNR = 5 dB.

weak. Therefore, the method seems acceptable for practical HNR measurements.

The results show that accuracy of the decomposition does not seem dependent on the pitch of the synthetic signal, or on the duration of the noise burst. Figs. 8 and 9 compare the synthetic signals with the measured signals in time domain. In Fig. 8, there is no time modulation of the aperiodic component. In Fig. 9, the noise bursts last 60% of the fundamental period. This indicates that time-modulation of the aperiodic component is preserved by the decomposition.

Fig. 7 shows the effect of the iterative algorithm for reconstruction of the aperiodic component in the spectral domain. The magnitude spectrum of a frame of the synthetic signal is displayed in the 0–1 kHz frequency band. The straight line is the magnitude spectrum of the synthetic aperiodic component. Dotted lines correspond to the measured aperiodic component, which is obtained after zero, one, ten, and 20 iterations. Twenty iterations were used in the experiments because some trials showed that almost no improvement resulted from more

TABLE I
HNR MEASUREMENTS FOR PERIODIC/APERIODIC DECOMPOSITION OF SYNTHETIC SIGNALS. $F0$: FUNDAMENTAL FREQUENCY. BDR: BURST DURATION RATIO (RATIO OF BURST DURATION TO FUNDAMENTAL PERIOD). HNR: HNR FOR SYNTHETIC SIGNALS. MEASURED HNR: HNR OBTAINED AFTER DECOMPOSITION. DIFFERENCE: DIFFERENCE BETWEEN HNR FOR SYNTHETIC SIGNAL AND HNR OBTAINED AFTER DECOMPOSITION

F0 (Hz)	BDR (%)	HNR (dB)	measured HNR (dB)	difference (dB)
120	-	∞	42.5	-
120	60	20	22.1	2.1
120	60	10	11.8	1.8
120	60	5	6.7	1.7
120	100	20	20.2	0.2
120	100	10	12.0	2.0
120	100	5	7.2	2.2
200	-	∞	41.4	-
200	60	20	21.5	1.5
200	60	10	10.9	0.9
200	60	5	5.4	0.4
200	100	20	24.1	4.1
200	100	10	11.6	1.6
200	100	5	6.7	1.7

iterations (the average difference in HNR between ten and 20 iterations was only 0.1 dB).

Direct decomposition using the signal rather than the LPC residual was also tried. The result showed that decomposition is still good, but that the difference between input HNR and measured HNR degrades slightly. This loss in accuracy is about 1 dB on average. This indicates that LPC decomposition is useful, but that rather good results can also be obtained without it.

In practice, the decomposition algorithm is dependent on several parameters (fast Fourier transform size, zero-padding size, window type, and window size) as usual with processing methods based on the short term Fourier transform (STFT). The analysis-synthesis parameters used for this experiments were: sampling rate, 8 kHz; FFT size, 1024 points; window type, Hamming; window length, 511 points $\simeq$ 64 ms; LPC, 10 coefficients, autocorrelation method; overlap between frames, 8 ms; 20 iterations. The number of iterations was chosen when the average difference between iteration was less than 0.1 dB. All the synthetic signals were voiced, therefore no voiced/unvoiced decision was necessary. Although pitch was known for the synthetic signals, a pitch detection algorithm was used [23] for the initial estimation of harmonic and noise regions. Therefore, experiments conducted with synthetic and natural speech are comparable.

The experiments conducted with synthetic speech demonstrated that the algorithm is able to decompose with a good accuracy a synthetic mixture of periodic and aperiodic components. Therefore, it can be used for analysis of natural speech. Some examples are given in the next section.

C. Decomposition of Natural Speech

We have decomposed the speech from sentences uttered by various speakers into periodic and aperiodic components. Throughout, we have considered speech signals sampled at 8 kHz. The speech utterance was first analyzed to determine voiced and unvoiced segments and the pitch period for voiced segments. The spectral comb pitch detection analysis

method was used in the experiments [23]. The voiced/unvoiced decision was taken on the basis of thresholds for signal energy, lowpass energy and zero-crossing. Each frame was classified either "voiced" or "unvoiced." Unvoiced frames were not further processed. Periodic/aperiodic decomposition was applied to voiced frames only.

The LP residual segments belonging to voiced segments were decomposed into periodic and aperiodic components (tenth-order LP analysis, frame size 32 ms, shift of 4 ms). For each frame of these segments, the data was multiplied by a Hamming window of size 255 samples. The algorithm was based on computation with 512-point DFT's and IDFT's. After reconstruction of the aperiodic component, it was appended with the residual of the unvoiced segments. The resulting periodic and aperiodic components of the LP residual were used to excite the time-varying all-pole LP filter. The periodic and aperiodic components of the speech signal were then obtained.

Fig. 10 shows the wideband spectrograms of the original signal and of the periodic and aperiodic component signals, for an utterance of a sentence by a male speaker. It can be seen that most of the noisy part of the spectrogram in the top of Fig. 10 has been removed in the periodic component, displayed in the middle of Fig. 10. Also, one can hardly see any periodic component in the spectrogram for the aperiodic component given in the bottom of Fig. 10. Thus, we could effectively separate these two components using the proposed decomposition algorithm. Subjective listening to these component signals also confirm this observation: the aperiodic component sounds like whispered speech. Note that the significant low-frequency energy in the aperiodic component can be attributed to turbulence noise at the glottal closure in the voiced segments.

Fig. 11 shows narrowband spectrograms for an utterance of the same sentence by a female speaker. Here also we have been able to accomplish the decomposition effectively, using the proposed algorithm. It must be noted that the aperiodic component may appear more important in the spectrograms than it actually is, because automatic gain control is used in the spectrographic display. However, the signal amplitudes are consistent among spectrograms.

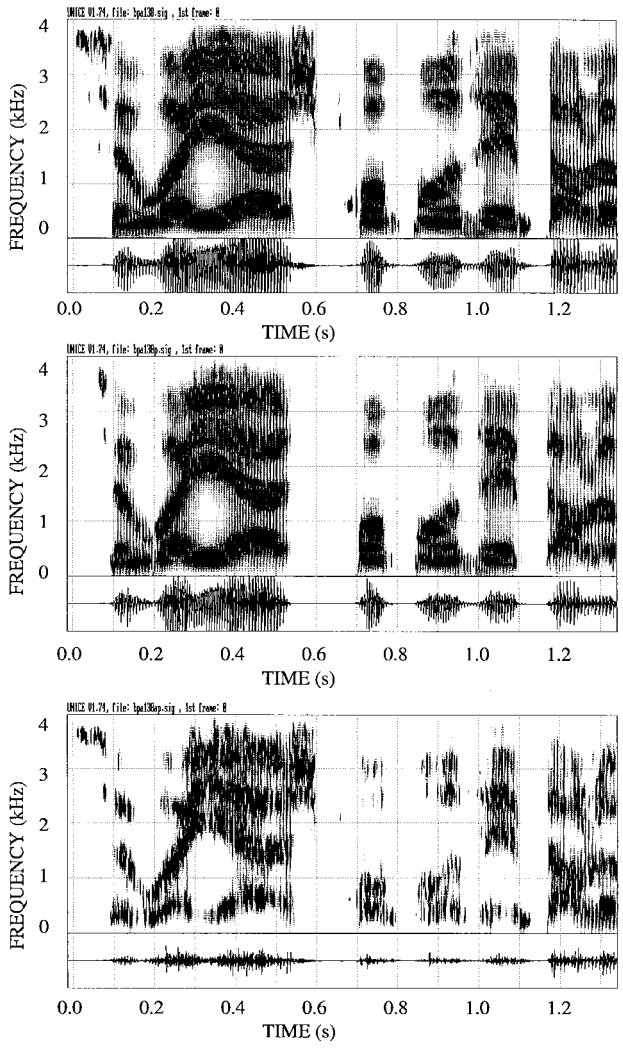

Fig. 10. Wideband spectrogram (male voice), *ce voyage, proposè* . . .. Top: original speech. Middle: periodic component. Bottom: aperiodic component.

V. Conclusions

In this paper, we have proposed a new method for decomposing the excitation part of voiced speech into deterministic and stochastic components. These components correspond to the quasiperiodic and aperiodic parts of the excitation. Compared to the sinusoidal coding based methods, our decomposition method has the advantage of better modeling of the aperiodic component, because both periodic and aperiodic components are assumed to be present at each sample in the frequency domain. The method uses the known noise regions in the spectrum, and reconstructs the aperiodic component in the harmonic regions. A signal that is primarily due to quasiperiodicity is then obtained by complex substraction of the aperiodic component in the spectral domain. The method performs the decomposition on an estimate of the excitation source signal rather than on the speech signal itself directly. This reduces the sidelobe effects of windowing on the decomposition. The linear prediction residual is used as an estimate of the excitation, although any method can be used to obtain an estimate of the excitation signal from speech.

Systematic assessment of the proposed decomposition algorithm has been carried out on simulated data. This was necessary to determine the usefulness of the proposed decomposition method to study natural speech. Synthetic data was chosen as it enabled us to control the voice characteristics by varying parameters. The result of these studies showed that the decomposition algorithm is able to separate aspiration noises and the periodic noise in the voice source. Further experiments devoted to the effect of other sources of voice aperiodicity (like jitter, shimmer, or large changes in pitch) are reported in [24] and in a companion paper [25].

The proposed method is conceptually simple, and is easy to implement. However, the amount of computation is large: a pair of DFT/IDFT is needed for each iteration, due to the extrapolation algorithm used. Therefore, real-time implementation of the method might require some special attention if a large number of iterations is to be used. The experiments

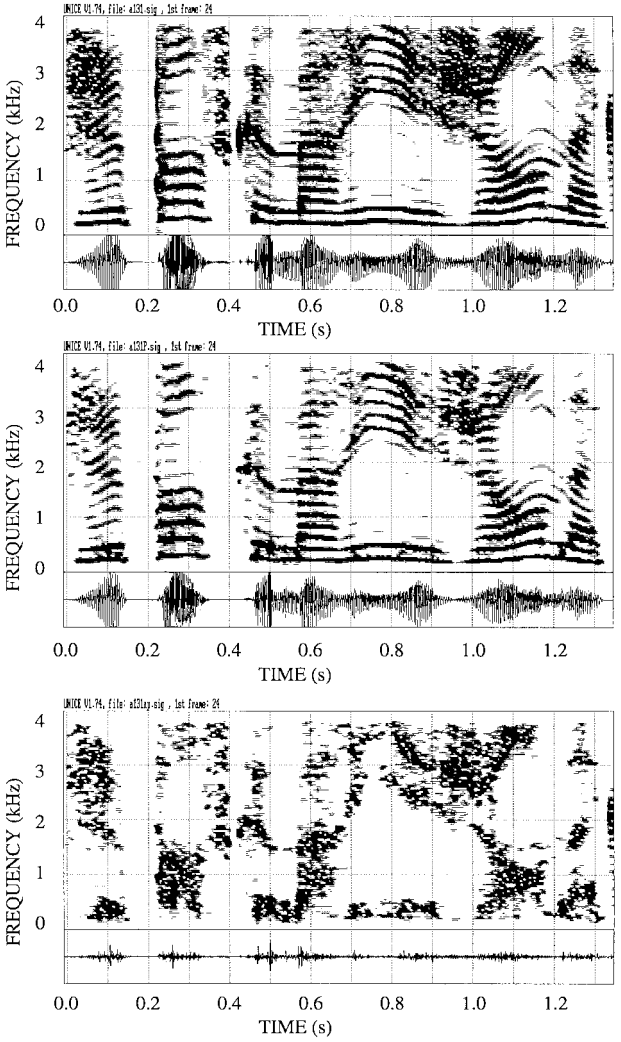

Fig. 11. Narrowband spectrogram (female voice), *Je pense que Marie et Jean* . . .. Top: original speech. Middle: periodic component. Bottom: aperiodic component.

indicated that less than ten iterations are sufficient in practice: then, close to real-time versions of the algorithm can easily be written on modern general purpose computers.

Periodic/aperiodic decomposition of the speech signal seems relevant for studying voice quality features, and in particular, breathiness or roughness of voices. It can also be used for modification of voice quality in the context of speech synthesis [8], or to produce voices with desired source characteristics. The decomposition algorithm was studied in the context of speech signals. However, it is a general technique that may be valid in many other situations as well. As a matter of fact, signals made of a mixture of periodic and aperiodic components are rather common in musical acoustics [26], [27], industrial sound and vibration, and biomedical signal processing, for instance.

ACKNOWLEDGMENT

Part of this work was conducted while Prof. B. Yegnanarayana and V. Darsinos were visiting LIMSI. The authors express their sincere gratitude to the five anonymous reviewers that helped to improve the form and content of the paper.

REFERENCES

[1] G. Fant, *Acoustic Theory of Speech Production*. The Hague, The Netherlands: Mouton, 1960.
[2] D. H. Klatt, and L. C. Klatt, "Analysis, synthesis a, and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Amer.*, vol. 87, pp. 820–857, 1990.
[3] E. B. Holmberg, J. S. Perkell, R. E. Hillman, and C. Gress, "Individual variation in measures of voice," *Phonetica*, vol. 51, pp. 30–37, 1994.
[4] 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.
[5] D. G. Childers and C. K. Lee, "Vocal quality factors: Analysis, synthesis and perception," *J. Acoust. Soc. Amer.*, vol. 90, pp. 2394–2410, 1991.
[6] 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.
[7] D. J. Hermes, "Synthesis of breathy vowels: some research methods," *Speech Commun.*, vol. 10, pp. 497–502, 1991.
[8] G. Richard and C. d'Alessandro, "Modification of the aperiodic component of speech signals for synthesis," in *Progress in Text-to-Speech Synthesis*, J. P. H. Van Santen, R. W. Sproat, J. P. Olive, and J. Hirschberg, Eds. New York: Springer-Verlag, 1996.
[9] 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, pp. 12–24, 1990.
[10] R. J. McAulay and T. F. Quatieri, "Speech analysis-synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 744–754, 1986.
[11] T. F. Quatieri and R. J. McAulay, "Speech transformations based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 1449–1464, 1986.
[12] E. B. George and M. J. T. Smith, "Analysis-by-synthesis/overlap-add sinusoidal modeling applied to the analysis and synthesis of musical tones," *J. Audio Eng. Soc.*, vol. 40, pp. 497–516, 1992.
[13] J. Laroche, Y. Stylianou, and E. Moulines, "HNS: Speech modification based on a harmonic + noise model," *Proc. IEEE-ICASSP'93*, pp. 550–553.
[14] D. Griffin and J. S. Lim, "Multiband excitation vocoder," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 1223–1235, 1988.
[15] 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.
[16] X. Rodet, P. Depalle, and G. Poirot, "Speech analysis and synthesis methods based on spectral envelopes and voiced/unvoiced functions," in *Proc. Europ. Conf. Speech Technology*, 1987, pp. 155–158.
[17] C. d'Alessandro, B. Yegnanarayana, and V. Darsinos, "Decomposition of the speech signal into deterministic and stochastic components," *Proc. IEEE-ICASSP'95*, pp. 760–763.
[18] J. D. Markel and A. H. Gray, *Linear Prediction of Speech.* New York: Springer-Verlag, 1976.
[19] F. Harris, "On the use of windows for harmonic analysis with the discrete Fourier transform," *Proc. IEEE*, vol. 66, pp. 51–83, 1978.
[20] A. Papoulis, *Signal Analysis*, Int. ed. New York: McGraw-Hill, 1984.
[21] D. H. Klatt, "Software for a cascade/parallel formant synthesizer" *J. Acoust. Soc. Amer.*, vol. 67, pp. 971–995, 1980.
[22] G. Fant, J. Liljencrants, and G.-G. Lin, "A four parameter model of glottal flow," *Speech Trans. Lab. Q. Prog. Stat. Rep.*, vol. 4, R. Inst. Technol., Stockholm, Sweden, pp. 1–17, 1985.
[23] P. Martin, "Comparison of pitch detection by cepstrum and spectral comb analysis," in *Proc. IEEE-ICASSP'82*, pp. 180–183.
[24] V. Darsinos, C. d'Alessandro, and B. Yegnanarayana, "Evaluation of a periodic/aperiodic speech decomposition algorithm," in *Proc. Eurospeech'95*, pp. 393–396.
[25] C. d'Alessandro, V. Darsinos, and B. Yegnanarayana, "Effectiveness of a periodic and aperiodic decomposition method for analysis of voice sources," this issue, pp. 12–23.
[26] C. Chafe, "Pulsed noise in self-sustained oscillations of musical instruments," in *Proc. IEEE-ICASSP'90*, pp. 1157–1160.
[27] P. Cook, "Noise and aperiodicity in the glottal source a study of singer voices," in *Proc. 12th Int. Cong. Phonetic Sciences*, 1991.

B. Yegnanarayana (M'78–SM'84) was born in India on January 9, 1944. He received the B.E., M.E., and Ph.D. degrees in electrical communication engineering from the Indian Institute of Science, Bangalore, India, in 1964, 1966, and 1974, respectively.

He was a Lecturer from 1966 to 1974 and an Assistant Professor from 1974 to 1978 in the Department of Electrical Communication Engineering, Indian Institute of Science. From 1977 to 1980, he was a Visiting Associate Professor of Computer Science at Carnegie Mellon University, Pittsburgh, PA. He was a visiting scientist at ISRO Satellite Center, Bangalore, from July to December 1980. Since 1980, he has been a Professor in the Department of Computer Science and Engineering, Indian Institute of Technology, Madras. He was a visiting Professor at the Institute for Perception Research, Eindhoven Technical University, Eindhoven, The Netherlands, from July 1994 to January 1995. From 1966 to 1971, he was engaged in the development of environmental test facilities for the Acoustics Laboratory, Indian Institute of Science. Since 1972, he has been working on problems in the area of speech signal processing. He is presently engaged in research activities in digital signal processing, speech recognition, and neural networks.

Dr. Yegnanarayana is a fellow of he Indian National Academy of Engineering.

Christophe d'Alessandro (M'95) was born in Marseille, France, on December 16, 1961. He received the B.S. degree in mathematics and the M.S and the Ph.D degrees in computer science from Paris VI University, in 1983, 1984, and 1989, respectively.

He has been a permanent Researcher at LIMSI, a laboratory of the French National Agency for Scientific Research (CNRS), since October 1989. Prior to joining CNRS, Dr. d'Alessandro was a Lecturer in computer science at Paris XI University from October 1987 to October 1989. His research interests include text-to-speech synthesis, signal processing for speech analysis and synthesis, perception and synthesis of intonation in speech and singing, voice source analysis and synthesis, speech synthesis assessment, and musical acoustics.

Dr. d'Alessandro is a member of the ASA and the French Acoustical Society, where he is presently Chairman of the Musical Acoustics Committee.

Vassilis Darsinos was born in 1969 in Athens, Greece. In 1993, he received the Dipl. degree in electrical engineering from the Electrical Engineering and Computer Science Department, University of Patras, Patras, Greece, in 1993. He received the post-graduate degree in 1995. The same year, in the framework of the Erasmus Project, he visited LIMSI in Paris as a post-graduate student for a period of six months. He is currently a Ph.D student and researcher at the Wire Communications Laboratory, University of Patras, and his current research interests include speech analysis and synthesis, modeling of speaker characteristics, and voice quality control.