4

Modification of the Aperiodic Component of Speech Signals for Synthesis

Gaël Richard
Christophe R. d'Alessandro

ABSTRACT Modeling the excitation component of speech signals is a challenging problem for speech synthesis. Recently, several works have been devoted to periodic/aperiodic decomposition of the speech signal: a decomposition that permits a better characterization of the source. This chapter introduces a new analysis/synthesis algorithm for representing the aperiodic component of the excitation source in speech signals. This component is decomposed as a sum of random formant wave forms (FWF), which correspond to formant filters impulse responses. The time of arrivals of the FWF define the virtual excitation source. The signal is decomposed in subbands, and, according to the random modulation theory, each passband signal is represented as an envelope modulating an oscillating term. All parameters (formant filters and excitation sources) are estimated in the time domain. This new representation scheme gives a very good fusion of the aperiodic component with the quasi-periodic component of speech. The method proposed provides new relevant parameters for manipulating the voice quality features that are linked to noise. For example, it is possible to perform voice quality modifications such as time scaling, formant or modulation depth modifications of the aperiodic component, or modification of the periodic/aperiodic ratio.

4.1 Introduction

Modeling the excitation source of speech signal is a key problem for speech synthesis. The excitation source encompasses the glottal source (periodic flow and aspiration noise), frication noise (occurring at a constriction of the vocal tract), and burst releases (occurring after a sudden release of a closure in the vocal tract). Voice quality modification, voice conversion, and prosodic processing are highly dependent on our ability to analyze, model, and modify the excitation source. In the context of both rule-based and concatenation-based synthesis, it seems important to develop signal representation methods that are able to deal with natural excitation sources. Traditionally, the most important prosodic parameters are pitch, duration, and intensity. These parameters are intimately linked to voice quality.

Other important aspects of voice quality, which have not received as much attention in synthesis research, are the vocal effort and the phonatory quality. Ideally, a synthesizer should be able to simulate various types of phonatory styles, the extreme situations being whispered speech and shouting. To reach this goal, it appears necessary to deal with various types of excitation, including noise excitation, which can occur at the glottis or at various locations of the vocal tract.

The analysis and synthesis of the speech noise component has recently become a focus of interest for several reasons. On the one hand, it is well-known that this component is responsible for a part of the perceived voice quality (e.g., breathiness, creakiness, softness). There is a long history, especially in the fields of voice pathology and voice perception, of studies involving parameters such as jitter (random variation of the source periodicity), shimmer (random variation of the glottal flow amplitude), or diplophony. On the other hand, some new results [LSM93], [DL92] indicate that separate processing of the periodic and aperiodic components of speech signals may improve the quality of synthetic speech, in the framework of concatenative synthesis. Finally, for some methods, the two components obtained seem to be relevant from the acoustic point of view [Cha90, DAY95]: It is possible to associate the periodic and aperiodic components to physical components in the voice source. This is important in achieving realistic modifications of the source.

Different terminologies have been used by various authors (e.g., "harmonic + noise (H+N) model" in [LSM93], "multiband excitation (MBE) vocoder" in [GL88] and [DL92], "deterministic and stochastic components" in [SS90]). We prefer the terminology "periodic and aperiodic (PAP) components." The precise meaning attached to the terms "periodic" and "aperiodic" must be discussed in some detail. The acoustic model of speech production is a source/filter model, with an excitation source $e(t)$ and a filter $v(t)$. An exactly periodic vibration of vocal chords is not a reasonable assumption, even for sustained vowels, because of the complex nature of this phenomenon. More accurately, the excitation source can be decomposed into a quasi-periodic component $p(t)$ and an aperiodic component $a(t)$:

$$s(t) = e(t) * v(t) = [p(t) + a(t)] * v(t). \tag{4.1}$$

Along this line, the periodic component represents the regular vibratory pattern of the vocal chords, and the aperiodic component represents all the irregularities presented in both voiced and unvoiced sounds. It must be emphasized that the aperiodic component may represent signals of different natures. It is generally acknowledged that both modulation noise (i.e., noise due to the source aperiodicity, such as jitter, shimmer, and F_0 variations) and additive random noise are present in the aperiodic component. This additive random noise includes transients (e.g., bursts of plosives), steady noises (e.g., unvoiced fricatives), or time-modulated noises (e.g., noise in voiced fricatives or in breathy vowels). Ideally, a decomposition method should be able to separate these different sources of aperiodicity. However, having solely the speech signal $s(t)$, obtaining the two components $p(t)$ and $a(t)$ is not straightforward. For this study, a new PAP decomposition algo-

rithm is presented, which yields an aperiodic component with a realistic acoustic meaning [AYD95], [DAY95].

The first step is therefore to achieve an acoustically relevant decomposition. The second step is to define a model for the aperiodic component. However, poor modeling of the aperiodic component could introduce a lack of perceptual fusion between the quasi-periodic component and the aperiodic component. Recent studies show that this perceptual separation is a consequence of the weak coherence between these two components in the time domain [Cha90], [Der91]. Therefore, methods for representing the aperiodic component are needed that provide an accurate control in both time and frequency. In this chapter, a new analysis/synthesis model for the aperiodic component is introduced. The synthesis method is based on previous work on the elementary wave form representation of speech [Ale90], and on speech noise synthesis using random formant impulse responses [Ric92]. Furthermore, this new coding scheme provides relevant parameters for manipulating the voice quality features that are linked to noise. Breathiness, creakiness, or roughness of a voice represent such features.

The chapter is organized as follows. The next section gives a detailed description of the speech signal decomposition algorithm. Section 4.3 introduces the random formant wave form model and describes the various steps of the analysis/synthesis algorithm. Section 4.4 presents some evaluation results. Section 4.5 demonstrates some of the voice modification abilities of the method. Finally, the results are discussed and some conclusions are suggested in the last section.

4.2 Speech Signal Decomposition

Even though there is a long history of research on aperiodicities in speech, particularly in the field of voice analysis research, most studies do not explicitly perform a separation of the two components (a periodic component and an aperiodic component), but rather measure a harmonic-to-noise ratio (HNR) to describe different types of voices (see, e.g., [Hil87, Kro93]).

On the contrary, in the field of speech and music synthesis, explicit PAP decomposition of signals has become a focus of interest, without paying much attention to the underlying acoustic or perceptual aspects. Various algorithms based on sinusoidal or harmonic models have been proposed. The MBE vocoder [GL88] is based on linear predictive coding (LPC). The LPC residual signal is coded in the frequency domain in terms of different frequency bands, which are labeled either “harmonic” or “noise” depending on their resemblance to ideal harmonic structures. Although this is an efficient coding scheme, it is difficult, if not impossible, to interpret the different frequency bands in terms of speech production. In the H+N model ([SM93]), a low-pass harmonic signal is subtracted from the original signal. The synthetic noise signal is obtained by modulation by an energy envelope function of LPC-filtered noise. Nevertheless, in this technique, there is no noise for frequencies below 2-3 kHz and no harmonics above. Although it might improve

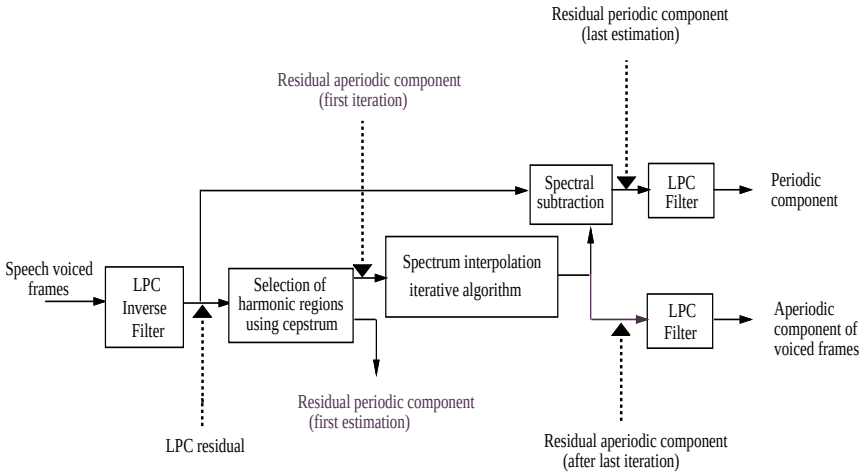

FIGURE 4.1. Schematized diagram of the PAP decomposition algorithm.

the quality of concatenative synthesis, this is not realistic from the acoustic point of view. Decomposition methods based on sinusoidal coding have also been proposed [SS90]. Some criteria for harmonicity are applied to the sinusoidal tracks to form the so-called deterministic component. The remaining frequency points are associated with the so-called stochastic component. Impressive decomposition is obtained for both speech and musical sounds. The problem with this approach is that a binary decision is taken for each frequency point: a region is either "deterministic" or "stochastic." This is not the case in actual speech, where noise is present even in harmonic regions.

A new algorithm for PAP decomposition has been proposed in [AYD95]. One of its aims is to obtain an aperiodic component that represents the real features of speech including voice quality features such as breathiness or roughness. This algorithm is applied only to voiced frames, the unvoiced frames being merged with the aperiodic component. Following the voiced-unvoiced decision, this algorithm may be decomposed into five main steps (see figure 4.1):

- An approximation of the excitation source signal is obtained by inverse filtering the speech signal (typically 8 kHz sampling rate, 10 LPC coefficients, 20 ms window size). The excitation source signal is then processed, on a frame-by-frame basis, using short-term Fourier analysis-synthesis (20 ms window size (200 pt), 5 ms overlap, 512 points fast Fourier transform (FFT)).

- For each frame, the ratio of periodic to aperiodic frequency points is measured in three steps, based on [Kro93]:

  1. The cepstrum of the original signal is computed.
  2. The region of the main peak (pitch) is isolated.

3. An inverse Fourier transform is then applied to this region. Ideally, the main peak is a Dirac distribution, and his inverse Fourier transform is a complex exponential. By considering only the real part of the spectrum, one obtains a sinusoid whose frequency is given by the location of the cepstrum main peak. The positive peaks of the sinusoid define the location of the harmonics, and all other positive values provide an estimation of the bandwidth of each harmonic. This last information is particularly important in practice as the cepstrum main peak is not an ideal Dirac distribution. The remaining part, the negative values of the sinusoid, provides an indication of frequency points where the valleys between harmonics are located. These frequency points will serve as a basis for a first approximation of the aperiodic component.

At this stage of processing, only a primary identification of the frequency points associated with the aperiodic component is formed and each frequency point is labeled as either periodic or aperiodic.

- The secondary estimation of the aperiodic component is then performed using an iterative algorithm based on Papoulis-Gershberg extrapolation algorithm ([Pap84]). Starting with the initial estimation, the signal is successively transformed from the frequency domain to the time domain and back to the frequency domain, imposing finite duration constraints in the time domain and the known noise samples in the frequency domain (frequency points in valleys between harmonics). After a few iterations (10 to 20), the obtained stochastic component possesses a continuous spectrum with extrapolated values in the harmonic regions (see figure 4.2). Thus, for each frequency point the complex values of both periodic and aperiodic components are available.

- The periodic component is then obtained by subtracting the complex spectrum of the aperiodic signal from the complex spectrum of the residual signal. The synthetic source components are obtained by inverse Fourier transform and overlap-add synthesis.

- Finally, the two components of the residual signal are filtered by the time-varying all-pole filter to obtain the final aperiodic and periodic components of the voiced frames. The complete aperiodic component is obtained by including the unvoiced frames of the original signal. The result of the PAP decomposition algorithm is depicted in figure 4.3.

This algorithm was tested using natural and synthetic signals [DAY95]. The results showed that the PAP decomposition algorithm is able to separate additive random noise and periodic voicing for a wide range of F_0 variation. Therefore, in normal natural speech, we feel justified in using the aperiodic component as an estimate of additive noise in the source, when jitter and shimmer are reasonably low. This is usually the case for speech synthesis databases. However, in the case

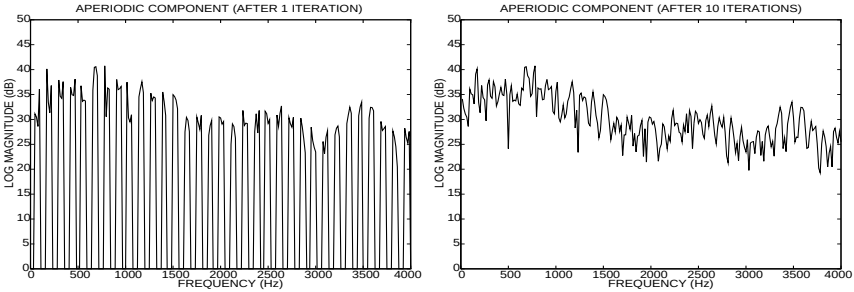

FIGURE 4.2. The effect of iterative noise reconstruction. The initial estimate of the aperiodic component has energy only between harmonic regions (a). After 10 iterations of the algorithm, a continuous spectrum is obtained (b).

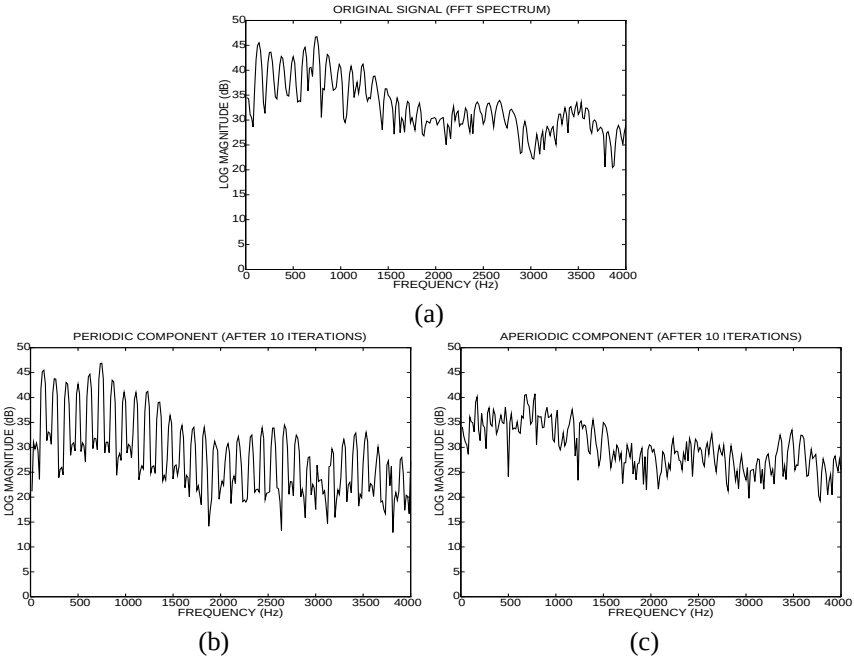

FIGURE 4.3. Result of the PAP decomposition algorithm: (a) displays the log magnitude spectrum of one frame of the original signal; (b) and (c) respectively represent the periodic and the aperiodic components obtained after decomposition.

of large jitter or shimmer values, additive random noise and modulation noise are merged in the aperiodic component. Although it is still possible to achieve separation of a periodic and an aperiodic component, it seems difficult in this case to separate the different physical components of the aperiodic component.

In the following discussion, we make the following assumptions:

1. The aperiodic and periodic components exist in speech signals. They are not artifacts due to a signal-processing method.
2. These components represent actual features of speech production that are linked to voice quality.
3. The components can be measured with accuracy using an appropriate PAP decomposition algorithm.

Sound example 1 of the audio demo (see Appendix) illustrates the results of the PAP decomposition algorithm on several sentences of natural speech.

4.3 Aperiodic Component Analysis and Synthesis

Several algorithms have been proposed for coding or synthesizing the aperiodic component of acoustic signals (see, for example, [Kla80, SS90, GL88, MQ92]). Most use a Gaussian excitation source and a slowly time-varying filter. It is clear that some time modulation of the aperiodic component is needed for speech because a white noise excitation source shaped by the spectral envelope of a slowly time-varying filter is not sufficiently precise in the time domain. As a matter of fact, it is acknowledged that it is important to take into account the temporal structure of noises if one wants to obtain a good perceptual fusion of the periodic and aperiodic components in the final reconstructed signal [CL91, Her91]. In formant synthesis, it is also common to modulate a noise source by the glottal flow [Kla80]. To trace this time modulation, [LSM93] proposed to time-modulate the high-pass noise excitation source by the time-domain envelope of the signal. However, this technique cannot be successfully applied to wideband noises and especially not for noises with significant energy in the lower part of the spectrum. This may be due to the fact that the rate of fluctuation of the envelope is of the same order of magnitude as the rate of fluctuation of the excitation noise signal. In other words, when some energy is present in the lower part of the spectrum, the maxima and minima of the time modulation (deduced from the envelope of the original noise signal) are almost never synchronized with the maxima and minima, respectively, of the white noise excitation signal. Thus, the precise time-domain control is lost and the resulting signal has a different modulation structure than the desired one.

Furthermore, for voice modification it may be important to control the spectral maxima (related to formants) and to get a description of the aperiodic component as a sum of well-localized spectro-temporal items.

For these reasons, we decided to develop an algorithm in the framework of source/filter decomposition, in which the filter is decomposed into several formant filters excited by separate random sources. Within a formant region, the passband noise signal is described as a random point process, which defines the random times of arrival of the formant filter impulse responses. The random point process is deduced from the maxima of the time-domain envelope.

The formant filters chosen are the Formant Wave Forms (FWF) introduced by [Rod80], which are close to second-order resonator impulse responses. A FWF is defined as a modulated sinusoid:

$$s(t) = \Lambda(t)\sin(2\pi f_c t + \phi) \tag{4.2}$$

where the FWF time domain envelope is given by:

$$\Lambda(t) = \begin{cases} 0 & \text{if } t \leq 0 \\ \frac{1}{2}A(1-\cos(\beta t))e^{-\alpha t} & \text{if } 0 < t \leq \pi/\beta \\ Ae^{-\alpha t} & \text{if } t > \pi/\beta \end{cases} \tag{4.3}$$

and π/β, $f_c, \alpha/\pi$, A, ϕ are the excitation duration, the formant center frequency, the -3 dB bandwidth, the formant amplitude, and the FWF initial phase, respectively.

An iterative algorithm was designed for automatic extraction of both source and filter parameters. The random excitation source is a set of points along the time axis, and the filter parameters are FWF parameters.

According to random modulation theory, any passband stochastic signal $x(t)$ can be represented using the real envelope $r(t)$ and the instantaneous phase $2\pi f_m t + \Psi(t)$, where f_m is arbitrary:

$$x(t) = r(t)\cos[2\pi f_m\, t + \Psi(t)]. \tag{4.4}$$

A more detailed theoretical background may be found in [Ric94]. The basic ideas of the algorithm are:

- to define the excitation point process according to the envelope maxima locations;
- to compute the FWF envelope $\Lambda(t)$ using the envelope $r(t)$ between two successive minima;
- to estimate the FWF center frequency from the center of gravity of the instantaneous frequency of $x(t)$.

More precisely, the analysis/synthesis algorithm is the following (see figure 4.4):

1. Band-pass filtering of the signal $x(t)$ (e.g., 6 bands, for a sampling rate of 8 kHz).

2. For each band-pass signal $x_b(t)$:

   a. Computation and low-pass filtering of the real envelope, using the Hilbert transform, $\hat{x}_b(t)$ of $x_b(t)$:

$$r(t) = \sqrt{x_b^2(t) + \hat{x}_b^2(t)} \tag{4.5}$$

   b. Definition of the excitation point process according to the real envelope maxima (see figure 4.5).

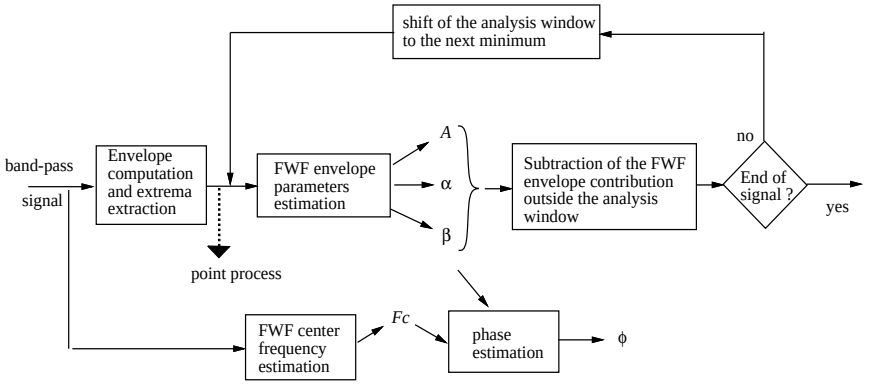

FIGURE 4.4. General diagram of the FWF estimation procedure for one band-pass signal. The same analysis must be applied to all band-pass signals (e.g., six bands for a sampling rate of 8 kHz) to obtain a complete description of the aperiodic component.

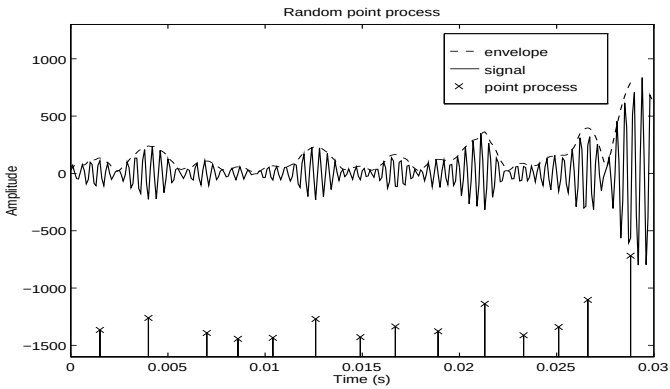

FIGURE 4.5. The point-process impulses are defined from the (time-domain) envelope maxima location.

c. Estimation of the FWF envelope parameters by fitting the FWF envelope and the real envelope between two successive minima of the envelope ($t \in [t_{m1}, t_{m2}]$). This gives α, β, and A.

d. Estimation of f_c as the optimal frequency f_m in equation (4.4). It is the weighted average of the instantaneous frequency $f_i(t)$:

$$f_c = \frac{\sum_{t\in[t_{m1},t_{m2}]} \Lambda^2(t - t_{m1}) f_i(t - t_{m1})}{\sum_{t\in[t_{m1},t_{m2}]} \Lambda^2(t - t_{m1})} \tag{4.6}$$

where the instantaneous frequency (time-derivative of the instantaneous phase) is given by:

$$f_i(t) = \frac{1}{2\pi} \times \frac{x_b(t)\hat{x_b}'(t) - x_b'(t)\hat{x_b}(t)}{r^2(t)} \tag{4.7}$$

e. The initial phase, ϕ, is set as a function of f_c and β in order to give a maximum at the exact place defined by the envelope maximum.

f. Subtraction of the FWF envelope contribution outside the analysis window (that is, for $t > t_{m2}$).

g. Iteration of steps c–f of the algorithm, until the end of the signal is reached.

3. FWF synthesis is performed using the estimated FWF parameters.

Sound example 2 of the audio demo illustrates the results of this algorithm on the aperiodic component of natural speech signals.

4.4 Evaluation

Perceptual tests (degrading category rating (DCR), see [CCI92]) were run to measure the quality obtained with the random FWF method compared to the LPC analysis/synthesis method. Ten subjects were asked to give an appreciation of the degradation of a synthetic signal (second of a pair) compared with a natural signal (first of the pair). Four conditions were tested:

Condition 1: Whispered speech (eight sentences of at least 1 s duration, 4 males/4 females).

Conditions 2,3 and 4: Normal speech (eight sentences of at least 1 s duration, 4 males/4 females). Periodic and aperiodic parts were separated. The aperiodic part was then modeled by either the LPC or random FWF model, scaled by a gain factor (1, 2, and 3 for tests 2, 3, and 4, respectively) before being added to the periodic component. The aim of this test was to measure the degree of fusion of the aperiodic and the periodic components and to test the robustness of this method when the aperiodic component is modified.

The results of the DCR test are given in figure 4.6. It is noticeable that both methods show similar results for condition 1 (whispered speech). This is not surprising, as the LPC model is excellent for this type of speech. The results for conditions 2 to 4 show a greater degradation for LPC than for random FWF. We think that these results are linked to the better time and frequency accuracy of our method: formants are well represented, and the time domain control gives a better perceptual fusion between the periodic and aperiodic components. In fact, the LPC analysis/synthesis method cannot trace the modulated structures that are present in the aperiodic component (see figure 4.7).

An informal listening test was also performed to compare the FWF representation to a simpler representation that takes into account the temporal structure of the noise. This simpler model (similar to the noise representation used in the

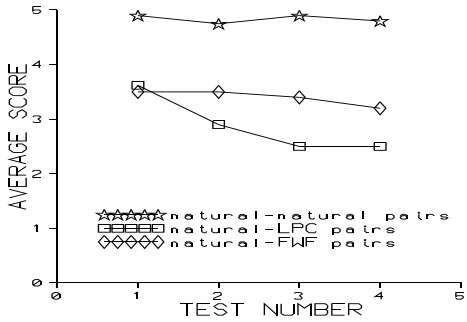

FIGURE 4.6. Perception test results. X-axis: condition number. Y-axis: average DCR score. A score of 5 corresponds to the answer "the signals within a pair are equals," and a score of 1 corresponds to the answer "the degradation in the second signal is very annoying." Stars denote pairs of identical signals. Diamonds denote pairs in which the second signal is reconstructed using the random FWF method. Squares denote pairs in which the second signal is reconstructed using LPC analysis/synthesis (from [Ric94]).

H+N model) consists of modulating (by an energy function) the signal obtained by filtering, with a normalized LPC filter, a white noise excitation source.

Our model seems to better represent the exact temporal structure of the noise and does not have the audible artifacts that can be seen in the other model (these artifacts are a consequence of a high amplitude of the excitation source (white noise) occurring at the same time as a high amplitude of the energy envelope function) (see figure 4.7). However, it seems that the two methods lead to results of comparable quality.

4.5 Speech Modifications

At the output of the analysis procedure, the aperiodic component is represented as a set of elementary wave forms well localized in the spectro-temporal domain. These wave forms are described by relevant acoustic parameters in the frequency domain (formant center frequencies, bandwidths, and amplitudes) as well as in the time domain (excitation times, instants of reference, initial phases) and thus provide various signal modification abilities.

In the context of realistic speech modifications, it is not sufficient to simply modify the speech signal. It is necessary to perform only those modifications that are possible in the speech production process. Although it is possible to separate a periodic and an aperiodic component, many voice-quality modifications affect both components. For example, the decay in intensity observed at the end of utterance results in changes in the glottal wave form, a higher spectral tilt, a lower periodic to aperiodic ratio, a lower aperiodic signal impulsiveness, etc. On the other hand, an increased vocal effort results in lower spectral tilt, a higher periodic to aperiodic

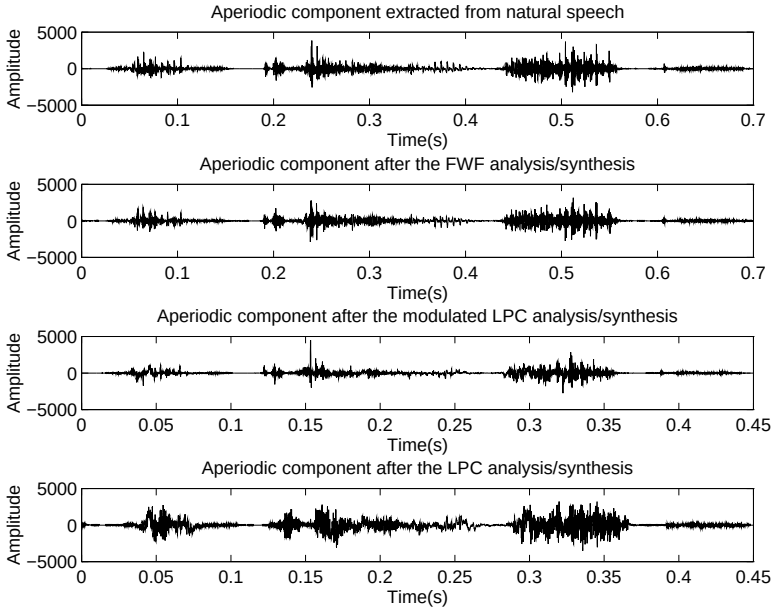

FIGURE 4.7. Time domain wave form of the (top curve) aperiodic component extracted from the speech signal with the PAP algorithm; (second curve) aperiodic component after the FWF analysis/synthesis method; (third curve) aperiodic component after the modulated LPC analysis/synthesis method; (bottom curve) aperiodic component after the LPC analysis/synthesis method.

ratio, a higher aperiodic signal impulsiveness, etc. Our current knowledge of these kinds of covariation of periodic and aperiodic parameters seems rather limited.

Due to the signal representation method proposed, several types of modification of the aperiodic components are straightforward. These modification capabilities are linked to the parameters that are available.

4.5.1 Time Scaling

Time scaling may be performed by simply modifying the reference instant (the time of generation) of each FWF. The results obtained are fairly good either for compression or dilation. However, for a large dilation coefficient, a more sophisticated procedure is needed, such as duplication in time of each wave form with lower amplitudes.

This type of time scaling results in a global dilation or compression of the signals without affecting the (possible) underlying periodicity of noise modulation.

4.5.2 *Spectral Modifications*

Format wave forms are defined by formant parameters. It is therefore easy to modify these parameters. Modifying formant center frequencies can be achieved simply by changing the corresponding parameter. It is also possible to change the formant spectral amplitudes. This allows us to change relevant parameters such as spectral tilt and noise amplitude in selected regions, and to shift the formants. These parameters are important for voice quality modification.

4.5.3 *Modification of the Aperiodic Component Impulsiveness*

For each FWF, it is also possible to control the individual time-domain envelope through the excitation time and bandwidth parameters. The time-domain envelope characterizes the modulation structure of the noise. Thus, it becomes possible to modify the overall depth of the time modulation of the stochastic component. This has an important consequence in the perceptual point of view as a deeper modulation gives a rougher voice with an impression of evident vocal effort, and a smoother modulation gives a softer and more whispery voice. Figure 4.8 illustrates the modification of the impulsiveness of a synthetic modulated signal produced by the FWF synthesizer.

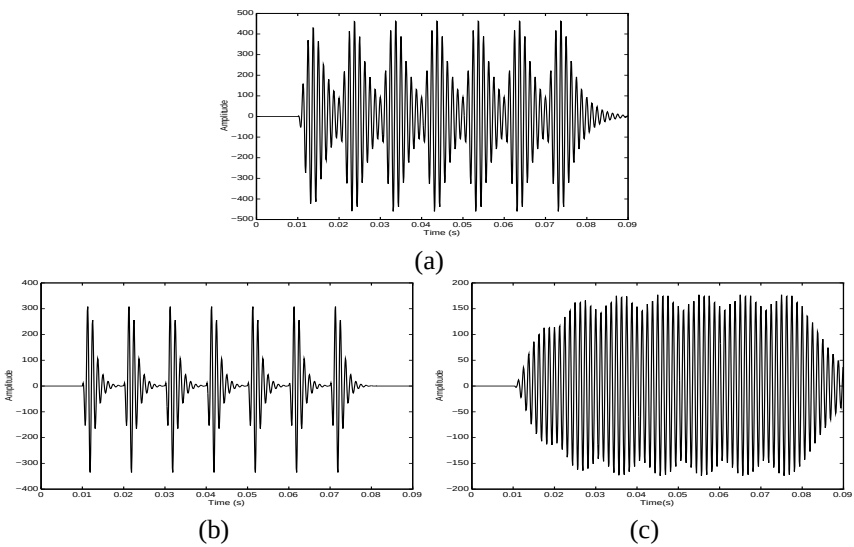

FIGURE 4.8. Modification of the aperiodic component impulsiveness: curve (a) displays the original synthetic signal; curve (b) gives an example with a deeper modulation structure obtained by simultaneously increasing the bandwidth and decreasing the excitation time (or onset time); in contrast, curve (c) gives an example with a smoother modulation structure obtained by simultaneously decreasing the bandwidth and increasing the excitation time.

4.5.4 *Modification of the Periodic/Aperiodic Ratio*

The periodic/aperiodic ratio can be easily modified. This can be done either globally (for all the formants), or locally (for particular frequency regions). A joint modification of impulsiveness and periodic/aperiodic ratios makes it possible to change continuously from voiced to whispered speech.

Sound examples 3 to 6 of the audio demo illustrate these various speech signal modifications using the PAP decomposition algorithm and the FWF model for the aperiodic component.

4.6 Discussion and Conclusion

The PAP decomposition of the speech signal seems relevant for studying voice quality features and, in particular, breathiness, roughness, or whisperiness of a voice. Compared to sinusoidal-coding-based methods, our decomposition method has the advantage of a better modeling of the aperiodic component. Both periodic and aperiodic components are defined at each frequency point in the complex frequency domain. Therefore, there is no binary decision between harmonic or noise regions, but a variable amount of noise at each frequency. This better reflects the acoustic reality.

As for aperiodic component coding, a time-frequency elementary wave form decomposition was preferred to the widely used LPC synthesis scheme.

The algorithm presented in this chapter for the analysis and representation of the aperiodic component proved to be efficient for modeling this component, including the strongly modulated segments of it.

Although the synthesized noise quality and naturalness is better with our method than with a conventional LPC model, the complexity (both in terms of computation and data rate) is much higher. The aim of this study was to design a technique that is able to represent noise with accuracy and with various voice quality modification capabilities, but not to perform a parameter rate reduction. However, the complexity in terms of data rate does not seem to be excessive for practical synthesis. Typically, the number of FWF per second is of the order of 1000, which leads to a data rate of 5000 parameters per second. Furthermore, this data rate can be easily lowered by suppressing the low-energy FWF. Actually, more than 50 percent of FWF are nearly inaudible.

The drawbacks of this new analysis-synthesis method are of two types. The method is more expensive than other methods, especially in terms of computation. In addition, it depends heavily on the success of the PAP decomposition method. This decomposition method is acoustically relevant in the case of little random modulation of the voice source. If this is not the case, it is more difficult to assign a meaning to the two components, and therefore the speech modification quality degrades. Unfortunately, it is likely that this last drawback will be shared by all decomposition methods.

We think that this first attempt to modify the aperiodic component of the voice source brings new capabilities for voice quality modification. Therefore it opens new ways for modeling different voice qualities and different voice styles. It also offers new challenges because so little is currently known about the production and perception of the aperiodic component of speech signals.

Acknowledgments: We wish to thank Daniel J. Sinder for reading and commenting on the manuscript.

REFERENCES

[Ale90] C. d'Alessandro. Time-frequency speech transformation based on an elementary wave form representation. *Speech Comm.* 9:419–431, 1990.

[AYD95] C. d'Alessandro, B. Yegnanarayana, and V. Darsinos. Decomposition of speech signals into deterministic and stochastic components. In *Proceedings of IEEE ICASSP'95*, Detroit, 760–764, 1995.

[DAY95] V. Darsinos C. d'Alessandro, and B. Yegnanarayana. Evaluation of a periodic/aperiodic speech decomposition algorithm. In *Proceedings of Eurospeech'95*, Madrid, Spain, 1995.

[Cha90] C. Chafe. Pulsed noise in self-sustained oscillations of musical instruments. In *Proceedings of IEEE ICASSP'90,* Albuquerque, 1157–1160, 1990.

[CCI92] CCITT. *Revised recommendation P.80 - "Methods for subjective determination of transmission quality."* SQEG, COM XII-118 E, International Telegraph and Telephone Consultative Committee (CCITT) (from Recommendation P.80, Blue Book, Volume V, 1989), 1992.

[CL91] D. G. Childers and C. K. Lee. Vocal quality factors: Analysis, synthesis and perception. *J. Acoust. Soc. Amer.* 9(5):2394–2410, 1991.

[DL92] T. Dutoit and H. Leich. Improving the TD-PSOLA text-to-speech synthesizer with a specially designed MBE re-synthesis of the segments database. In *Proceedings of EUSIPCO'92,* Brussels, Belgium, 343–346, 1992.

[GL88] D. Griffin D and J. S. Lim. Multiband excitation vocoder. *IEEE Trans. ASSP* ASSP-36(8):1223–1235, 1988.

[Her91] D. J. Hermes. Synthesis of breathy vowels : some research methods. *Speech Comm.* 10:497–502, 1991.

[Hil87] J. Hillenbrand. A methodological study of perturbation and additive noise in synthetically generated voice signals. *J. Speech and Hearing Res.* 30:448–461, 1987.

[Kla80] D. H. Klatt. Software for a cascade/parallel formant synthesizer. *J. Acoust. Soc. Amer.* 67(3):971–995, 1980.

[Kro93] G. de Krom. A cepstrum-based technique for determining a harmonics-to noise ratio in speech signals. *J. Speech and Hearing Res.* 36:254–266, 1993.

[LSM93] J. Laroche, Y. Stylianou, and E. Moulines. HNS: Speech modification based on a harmonic + noise model. In *Proceedings of IEEE ICASSP'93,* Minneapolis, 550–553, 1993.

[MQ92] R. J. McAulay and T. F. Quatieri. Low-rate speech coding based on the sinusoidal model. In *Advances in Speech Signal Processing,* S. Furui and M. M. Sondhi, eds., Marcel Dekker, New York, 165–208, 1992.

[Pap84] A. Papoulis. *Signal Analysis*. Mc Graw-Hill, New York, 1984.

[Ric92] G. Richard, C. d'Alessandro, and S. Grau. Unvoiced speech analysis and synthesis using Poissonian random formant-wave-functions. In *Proceedings of EUSIPCO'92,* Brussels, Belgium, 347–350, 1992.

[Ric94] G. Richard. *Modelisation de la composante stochastique de la parole.* PhD thesis, Universite de Paris-XI, Orsay, France, (in French), 1994.

[Rod80] X. Rodet. Time-domain formant-wave-function synthesis. In *Spoken Language Generation and Understanding*, J. C. Simon, ed., D. Reidel, Dordrecht, Netherlands, 1980. Also in *Comp. Music J.* 8(3):9–14, 1980.

[SS90] X. Serra and J. Smith. Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition. *Comp. Music J.* 14(4), 1990.

Appendix: Audio Demos

The audio demo contains sound examples, with explanations given.