DECOMPOSITION OF SPEECH SIGNALS INTO DETERMINISTIC AND STOCHASTIC COMPONENTS

C. d'Alessandro †, B. Yegnanarayana ‡& V. Darsinos †§

†LIMSI-CNRS- BP 133 - F-91403 Orsay, France.
‡Dept of CSE, Indian Institute of Technology, Madras-600036, India
§Wire Communications Laboratory, GR-26500 University of Patras, Greece

ABSTRACT

This paper presents a new method for decomposition of the speech signal into a deterministic and a stochastic component. The method is based on iterative signal reconstruction. The method involves: (1) Separation of speech into an approximate excitation and filter components using Linear Predictive (LP) analysis; (2) Identification of frequency regions of noise and deterministic components of excitation using cepstrum; (3) Reconstruction of the two excitation components of the residual using an iterative algorithm; (4) Finally, the deterministic and stochastic components of the excitation are then obtained by combining the reconstructed frames of data using an overlap-add procedure. The deterministic and stochastic components are then passed through the time varying all-pole filter to obtain the components of the speech signal. The algorithm is able to decompose varying mixtures of stochastic and deterministic signals, like the noise bursts produced at the glottal closure and the deterministic glottal pulses. This new algorithm is a powerful tool for analysis of relevant features of the source component of speech signals.

1. Introduction

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 for voiced, as quasiperiodic sequence of glottal pulses. But in real speech even the voiced part consists of some random component, especially at the glottal closure due to turbulence, and in weak voicing and voiced fricatives. For synthesis, this random component also must be included in the excitation in order to produce a natural sounding synthetic speech. Moreover, this random component may also help characterizing the nature of the speech segment such as breathiness, roughness, etc. Detailed characterization of source may also help in generating synthetic speech with desired voice characteristics [1].

In this paper we propose a new algorithm for decomposition of speech signals into deterministic and stochastic components. Several algorithms have been proposed in the literature which are based on: sinusoidal representation [2], linear prediction (LP) and voiced/unvoiced decision for each frequency band [3], harmonic + noise model [4]. The main differences in our approach compared to previous algorithms are: (1) Decomposition is performed on an approximate source signal (LP residual); (2) With this algorithm, the impulse-like behavior around the significant instants is captured, therefore there is no need for imposing the harmonicity or continuity of sinusoids across the frames. One does not need any frame synchronization or continuity, as the new deterministic component from each overlapping region of the frame is simply concatenated; (3) An iterative algorithm is developed for estimating the noise part of excitation. The noise part is subtracted from the signal to obtain the deterministic component.

An important question with regard to the significance of the deterministic and stochastic components is whether these components represent some features of speech production or they are merely a convenient representation of signals. We have tested our algorithm on various natural and synthetic speech signals containing a mixture of quasiperiodic excitation and noise excitation. In all the cases the algorithm demonstrated its ability in separating the two components fairly accurately.

2. Description of the algorithm

We assume the following model for speech production:

$$s(t) = e(t) * v(t) = (p(t) + r(t)) * v(t) \qquad (1)$$

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,
$r(t)$ is the random part of the excitation.

In spectral domain we can write:

$$S(\omega) = \mid S(\omega) \mid e^{j\theta(\omega)} \qquad (2)$$

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

Notice that: (1) The complex addition shows the relative importance of the magnitude and phase of each of the components in the signal; (2) Depending on the Signal to Noise Ratio (SNR) at each frequency, the different components of the excitation source get prominence. The decomposition algorithm contains 5 main steps:

2.1 Extraction of linear prediction residual

Since the objective is to separate the components of the excitation, an approximation to the source part of the signal is derived by using the linear prediction residual (10 kHz sampling rate, 12th order LP analysis on overlapping segments of 25.6 msec at 200 Hz analysis frame rate). The

0-7803-2431-5/95 $4.00 © 1995 IEEE

residual is obtained by passing the speech signal through an inverse filter. Each frame (25.6 msec) of the residual is expressed in terms of the DFT coefficients. A 512-point DFT was used in these studies.

2.2 Identification of the frequency regions of the stochastic component using cepstrum

A voiced/unvoiced decision was done separately for each frame. The decision is biased in favour of voiced frames over the unvoiced frames. That is, it is designed to make errors only in one direction (i.e., to label unvoiced frames as "voiced", rather than labelling voiced frames as "unvoiced"). The errors in voiced/unvoiced labelling will be corrected by further processing the voiced frames after decomposition. One will obtain a deterministic component with almost no energy in frames incorrectly labelled as "voiced". All such "unvoiced" frames are moved to the stochastic component. For voiced frames, each DFT coefficient is contributed by both the periodic part and the random noise part. It is important to identify which subset of the DFT coefficients should be combined to form the deterministic component and which ones to form the stochastic component. To reach this goal, we may consider those frequency samples for which the deterministic component is higher than the random component, and use those frequency components in the summation. This is accomplished using cepstrum and a pitch detection algorithm for identifying the regions corresponding to the system (vocal tract) and the regions corresponding to periodic and random parts of the excitation. Because of their distinct locations in the quefrency domain, the distributions of these energies in the frequency domain can be obtained.

2.3 Identification of the stochastic component using extrapolation

The ratio of the energies of the periodic and random parts can be obtained at each frequency point. From the ratio it should be possible to retain only those frequency points which have a higher periodic component over the noise component. Note that just knowing the SNR (ratio of periodic to random parts) at each frequency point does not enable us to separate the two components by subtraction, since at each frequency point there is contribution due to both the periodic and random parts. According to equation (3), it is necessary to use knowledge of both the amplitude and phase of the two components for separating the components.

We propose an iterative procedure to overcome the above difficulty. From the frequency distribution of the periodic part in the log magnitude spectrum, we can hypothesize, as a starting point, that the valleys between two harmonics are mostly due to noise. Figures 1-3 illustrate the analysis process, for a particular frame extracted of a vowel ($F_0 \simeq 120$ Hz). Figures 1 and 2 demonstrate the effect of the noise extrapolation algorithm. These figures correspond to the magnitude spectrum of the noise components of the LP residual signal after cepstral separation. Noise is set to zero in the harmonic regions, and to the measured spectral values in the regions between harmonics. It is clear that such a noise component cannot be representative of the speech production mechanism. Using these noise regions, the noise samples in the harmonic regions are estimated using an iterative algorithm. The iterative algorithm is based on a spectral extrapolation algorithm similar to the Papoulis-Gerchberg algorithm [5, pp-244-248]. Starting with zero values in the harmonic regions and the actual DFT coefficients in the noise regions, an estimate of the noise samples in the harmonic regions is obtained by iteratively moving from frequency domain to time domain and vice versa, imposing finite duration constraint in the time domain, and the known noise samples constraint in the frequency domain.

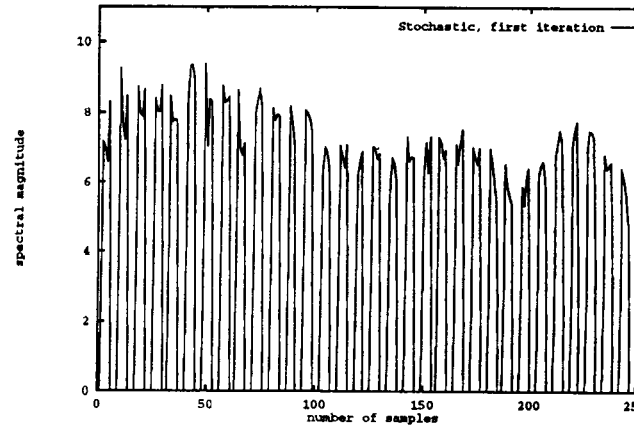

Figure 1: Spectral magnitude of a frame of the noise component for a vowel, before extrapolation

The following are the steps in the extrapolation algorithm:

1. Take a 256 sample frame of LP residual signal.
2. Perform a 512 pt DFT.
3. Identify the harmonic regions in the frequency domain using cepstrum and a pitch determination algorithm.
4. Set the DFT values in the harmonic regions to zero and perform an IDFT to get an estimate of the noise samples.
5. Set the noise samples beyond the 256 points to zero and recompute the DFT.
6. Replace the DFT values in the noise regions to the original values, and perform an IDFT to obtain the next estimate of the noise samples.
7. Repeat steps 5 and 6 for a few(10-20) iterations. It is interesting to see that during each iteration the noise samples build up in the harmonic regions.

The extrapolation algorithm is continued until the difference (in terms of magnitude of the noise spectrum) between two successive steps becomes less than a given threshold value. Figure 2 shows the spectral magnitude of the noise component obtained after spectral extrapolation (using 10 iterations, compare with Figure 1).

2.4 Identification of deterministic signal

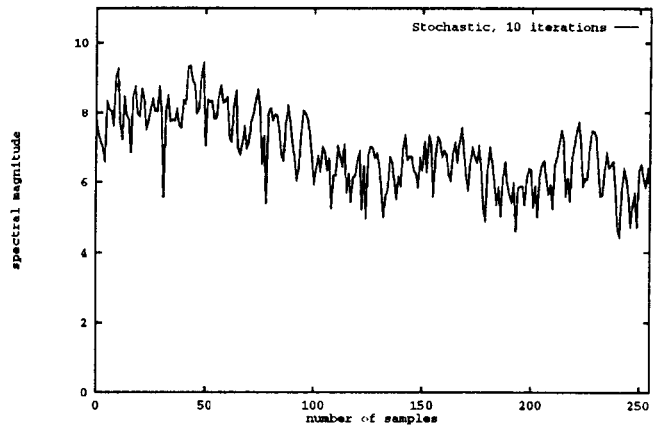

Figure 2: Spectral magnitude of a frame of the noise component for a vowel, after extrapolation (10 iterations)

Noise is reconstructed (in amplitude and phase) for every frequency sample. Figure 4 represents the noise component and the original DFT of the signal (dashed line: noise component, continuous line: original DFT).

The deterministic component is obtained by subtracting the noise samples from the residual signal. Nearly the same result will be obtained starting with the DFT coefficients in the harmonic regions, and iteratively building up the deterministic component, after subtracting an estimate of the noise from the DFT coefficients in the harmonic regions at each iteration. Figure 4 shows the same reconstruction as in Figure 3, for the harmonic component. In this case, the harmonic component energy is reduced between the harmonic peaks, because some energy that was present at the starting point was transfered to the noise component during the iterative extrapolation phase.

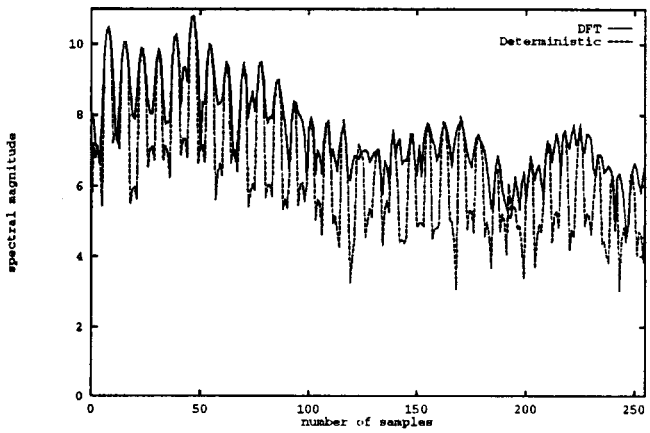

Figure 3: DFT for a voiced frame (continuous line), and deterministic component (dashed line)

2.5 Synthesis.

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

3. Results

The algorithm presented above was applied to several natural and synthetic speech signals. One of the aims of this work was to develop a speech signal decomposition method which is relevant from the speech production point of view. Figure 5 illustrates the capabilities of the method for non-stationary noise decomposition.

The first picture (top) represents a realistic excitation signal for a vowel or a voiced consonant. This signal is made up of synthetic glottal pulses (Liljencrants-Fant Model), with synthetic pulses of noise synchronized with the glottal closure epochs. The breathiness of the resulting voice depends on the amplitude and duration of the aspiration noise, which is introduced by synchronized noise pulses. Therefore, it is important to decompose the signal in such a way that both the frequency domain and the the time domain characteristics of the glottal pulse and noise excitation signals are well represented. The second picture (middle) shows these synthetic pulses of noise in isolation. They were generated as the product of white noise signal and a rectangular wave. This synthetic voiced excitation signal was passed through an all-pole filter corresponding to the vowel /a/. The synthetic speech signal was analyzed using our decomposition algorithm. The third picture (bottom) represents the stochastic component extracted from the residual signal of the synthetic vowel. Middle and bottom pictures are very similar. The algorithm successfully separated the deterministic and stochastic components in the time domain.

Systematic assessment of the algorithm have been conducted. Noise is a rather vague concept that encompass various physical situations in voiced excitation. Among the sources of noise that can be considered, we paid particular attention to additive pulsed noise in the residual (aspiration noise), additive continuous noise in the residual (frication noise), effect of jitter, effect of shimmer, effect of F_0 variations. The algorithm demonstrated excellent capabilities in decomposing the components in aspiration and frication noise analysis. Large jitter and shimmer values (jitter $> F_0$, shimmer > 1.5 dB), or large F_0 variations within a frame increase the energy in the stochastic part. This effect might be partly considered as an artifact, because F_0 changes are not random excitation noises. The status of shimmer is more intricate. The effect of large shimmer values is a degradation of the stochastic signal, and a smoothing of the deterministic signal. Ideally, one would prefer to analyze perturbations of the glottal pulse (i.e., F_0 changes, jitter and shimmer) on the one hand, and random noise excitation (i.e. aspiration and frication noise, transients), on the other hand. With our method, both are merged in the stochastic component.

Finally, the results demonstrated that the new algorithm is a powerful tool for analysis of the noise component in the excitation part of speech signals.

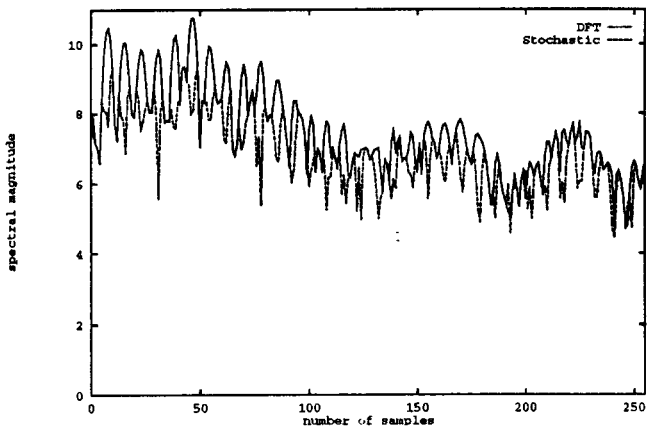

Figure 4: DFT for a voiced frame (continuous line), and stochastic component component (dashed line)

Acknowledgment

This work was conducted while Prof. B. Yegnanarayana and V. Darsinos were visiting LIMSI. It was supported in part by a grant of the University Paris XI, Orsay, and by the CEC ERASMUS program in phonetics and speech communication.

5. References

[1] Richard, G., & d'Alessandro C., 1994. "Time-domain analysis-synthesis of the aperiodic component of speech signals." *Proc. 2nd ESCA/IEEE Workshop on Speech Synthesis.*

[2] Serra X., Smith J., 1990. "Spectral modeling synthesis : a sound analysis/synthesis system based on a deterministic plus stochastic decomposition." *Computer Music Journal,* Vol. 14, No. 4.

[3] Griffin D., Lim J.S., 1988. "Multiband excitation vocoder." *IEEE Trans. Acoust., Speech, Signal Processing,* Vol. ASSP-36, No. 8.

[4] Laroche J., Stylianou Y., Moulines E., 1993. "HNS: Speech modification based on a harmonic + noise model." *Proc. IEEE ICASSP 94.*

[5] Papoulis A., 1984. *Signal Analysis.* Mc Graw-Hill.

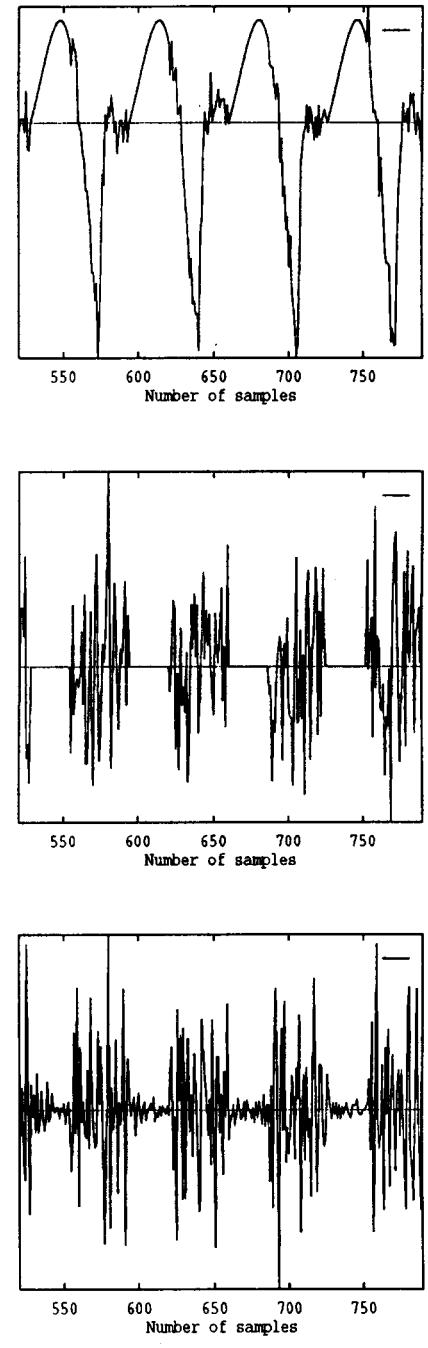

Figure 5: Synthetic glottal pulse with aspiration noise (top). Synthetic aspiration noise of top signal alone (middle, amplitude increased). Stochastic component extracted from top signal (bottom, same amplitude as middle).