Decomposition of Speech Signals into a Deterministic and a Stochastic part

Yannis Stylianou

Groupe ESIEE, Signal Department
Cite Descartes B.P. 99, 93162 Noisy Le Grand Cedex, FRANCE.
email:styliani@esiee.fr

ABSTRACT

In this contribution a novel method enabling en efficient decomposition of the speech signal into a deterministic and a stochastic part is presented. The deterministic part is modeled by harmonically related sinusoids; the proposed method makes use of a third order polynomial with real coefficients for the harmonic amplitudes and the phase is assumed to be linear. To obtain the stochastic part the deterministic part is simply subtracting from the original speech signal in the time domain. The results obtained on large speech database demonstrate effective decomposition of the speech signal between the two parts.

1. INTRODUCTION

The decomposition of the speech signal into a deterministic and a stochastic part has many applications. In speech synthesis where high-quality prosodic modifications (i.e. pitch and time scale) are required, the speech decomposition allows the possibility to apply different modification methods to each part, yielding more natural synthesis [9]. Because the signal is decomposed between two parts, different modification methods can be applied to each part, yielding more natural pitch and time-scale modifications. The naturalness of the prosodic modified speech signal is very important for high-quality text-to-speech synthesis based on acoustical units concatenation. For speech coding different coding schemes can be applied to each part in order to obtain low-rate speech coding systems [2]. Also, the harmonic (deterministic) to noise (stochastic) ratio has been used to detect pathological voices [4].
The speech signal decomposition has attracted a lot of research efforts; Griffin and Lim proposed the Multiband Excitation Model (MBE) [2] for speech coding. Smith and Serra [8] have developed a system for sound analysis/transformation/synthesis based on a deterministic and a residual part. Recently a novel iterative algorithm for decomposition of the speech signal into a periodic and an aperiodic part has been proposed in [1]. The method aims at decomposing the excitation signal (linear prediction residual) rather than the original speech signal.

In this contribution we describe a new method for speech decomposition. The deterministic (or periodic) part supposed to be harmonic; it makes use of third order polynomial with *real* coefficients for the harmonic amplitudes and the phase is supposed to be linear (first order polynomial). Given the harmonic part, the non-periodic part is obtained by subtracting the harmonic part from the original speech signal. The non-periodic part (or *residual signal*) thus accounts for everything in the signal that is not described by harmonic components. It includes the friction noise, the period-to-period fluctuations produced from the turbulences of the glottal airflow, etc [5].

The first part of the paper is devoted to the model description. Then, the estimation problem of model parameters is given. Next, some characteristics of the residual signal and the modeling error are discussed. An example of analysis of a speech signal is given to support our conclusions.

2. DESCRIPTION OF THE MODEL

The model that we present in this contribution can be viewed as a special version of the Harmonic + Noise Model, (HNM) [10] [9]. Like HNM, the current model assumes the speech signal to be composed of a periodic component $h[n]$ and a non-periodic component $r[n]$. The periodic component designated as sums of harmonically related sinusoids

$$h[n] = \sum_{k=0}^{L(n_a^i)} a_k(n) cos(\varphi_k(n)) \tag{1}$$

where

$$\begin{aligned} a_k(n) &= \alpha_k + \beta_k\,(n-n_a^i) + \gamma_k\,(n-n_a^i)^2 + \delta_k\,(n-n_a^i)^3 \\ \varphi_k(n) &= \epsilon_k + 2\pi\,k\zeta\,(n-n_a^i) \end{aligned} \tag{2}$$

The above parameters are updated at specific time-instants denoted by n_a^i. $f_0(n)$ and $L(n_a^i)$ represent the fundamental frequency and the number of pitch-harmonics included in the harmonic part respectively. These two parameters are held constants within each analysis frame and equal to their values at the center, n_a^i, of the analysis window. For convenience the index "n_a^i" will be omitted from the number of harmonics, L.

The non-periodic part is just the *residual* signal obtained by subtracting the periodic-part (harmonic part) from the original speech signal in the time-domain

$$r[n] = s[n] - h[n] \tag{3}$$

where $h[n]$ is the harmonic part and $s[n]$ represents the sampling version of the continuous speech signal $s(t)$ sampled at a rate of F_s samples per second.

3. ESTIMATION TECHNIQUE

As the speech signal is supposed to be harmonic, the first step of the analysis process consists of estimating the fundamental frequency and the number of harmonics included in the harmonic part. First, an initial pitch estimation is obtained by using a standard time-domain pitch detector based on a normalized cross-correlation function [3]. Next, an appropriate peak picking algorithm [9], using the initial pitch estimation, separates the "harmonic" peaks from the spurious peaks. Finally, a refined fundamental frequency is obtained by finding the fundamental frequency whose harmonics match better the frequencies of the peaks detected as "harmonics". Using the stream of the estimated pitch values, the position of the analysis time-instants, n_a^i, are set at a pitch-synchronous rate.

The second step of the analysis process consists, on voiced portions of speech, of estimating the values of the harmonic amplitudes and phases. This is done using a weighted least-squares method aiming at minimizing the following criterion with respect to unknown coefficients of the amplitude and phase polynomials

$$\epsilon = \sum_{n=n_a^i-N}^{n_a^i+N} w^2[n](s[n] - \hat{h}[n])^2 \tag{4}$$

$w[n]$ is a weighting window and N is the integer closest to the local pitch period $T(n_a^i)$. The above criterion has a quadratic form for the parameters and it is solved by inverting an over-determined system of linear equations [6]. At this point in the analysis, we find it convenient to switch to matrix notation. In particular we may write the harmonic part of the model as

$$\mathbf{Px} = \mathbf{h} \tag{5}$$

where the matrix $\mathbf{P}$ is defined by

$$\mathbf{P} = \left[\mathbf{B_1} \vdots\, diag(\mathbf{n})\mathbf{B_1} \vdots\, diag(\mathbf{n})^2\mathbf{B_1} \vdots\, diag(\mathbf{n})^3\mathbf{B_1}\right] \tag{6}$$

where $\mathbf{n} = [-N, -N+1, \cdots, N]^T$ denotes the $(2N+1)$-by-1 time vector, $diag(\mathbf{n})$ denotes a $(2N+1)$-by-$(2N+1)$ diagonal matrix with $\mathbf{n}$ the diagonal entries of this matrix and $\mathbf{B_1}$ is a $(2N+1)$- by-$(L+1)$ matrix given by

$$\mathbf{B_1} = \left[\mathbf{b_1} \vdots\, \mathbf{b_2} \vdots \cdots \vdots\, \mathbf{b_L} \vdots\, \mathbf{1}\right] \tag{7}$$

where $\mathbf{b_k}$ is a $(2N+1)$-by-1 vector defined by

$$\mathbf{b_k} = [cos(\varphi_k[-N])\, cos(\varphi_k[-N+1]) \cdots cos(\varphi_k[N])]^T \tag{8}$$

for $k = 1 \cdots L$ and $\mathbf{1}$ denotes $(N+1)$-by-1 unit vector : $\mathbf{1} = [1\,1\,1 \cdots 1]^T$. In (8), $\varphi_k[n]$ represents the first order polynomial of the phase.

Note that $\mathbf{x}$ is a $(4L+4)$-by-1 vector defined by

$$\begin{aligned} \mathbf{x} \;=\; & [\alpha_1\, \alpha_2 \cdots \alpha_L\, \alpha_0\, \beta_1\, \beta_2 \cdots \beta_L\, \beta_0 \\ & \gamma_1\, \gamma_2 \cdots \gamma_L\, \gamma_0\, \delta_1\, \delta_2 \cdots \delta_L\, \delta_0]^T \end{aligned} \tag{9}$$

Note also that as the coefficients ϵ_k and ζ are unknown and the amplitude polynomial has an order greater than zero, the above error criterion leads to non-linear equations and then, the solution must be calculated using relaxation methods. In order to avoid this iterative solution, the following analysis scheme (using only one iteration) was adapted. As the instantaneous frequency, $f_0(n)$, has been already calculated at the first step of the analysis and given that within the analysis frame $|n - n_a^i| \leq N$ we suppose that $f_0(n) = f_0(n_a^i)$, it is straightforward to show that the parameter ζ of the phase polynomial is equal to $f_0(n_a^i)$. Denoting by ϕ_k^i, the phase value of the k-th harmonic at the time-instant n_a^i, and evaluating the phase polynomial at n_a^i, using (2), then the coefficients ϵ_k are given by

$$\epsilon_k = \phi_k^i \text{ for } k = 1, \cdots, L \tag{10}$$

The ϕ_k^i can be efficiently estimated (see [9] for a fast method) solving a linear set of equations if we set to zero (as initial values) the coefficients β_k, γ_k and δ_k. Thus, having determined all the coefficients of the phase polynomial, the next step consists of estimating the coefficients of the amplitude polynomial. As $\varphi[n]$ is known for each $n \in [n_a^i - N \;\; n_a^i + N]$, it turns out that the solution to the least-squares error is given by

$$\mathbf{x} = \left(\mathbf{P}^T\mathbf{W}^T\mathbf{PW}\right)^{-1}\mathbf{P}^T\mathbf{W}^T\mathbf{Ws} \tag{11}$$

where $\mathbf{W}$ is a diagonal matrix with $w[n]$ as its entries. Note that the matrix to invert is symmetric and positive definite. Therefore, one method that could be used is the Cholesky decomposition[7] for solving the above linear set of equations. Note that this method is about a factor of two faster than alternative methods for solving linear equations[7].

Lastly, the harmonic part is readily obtained by

$$\hat{\mathbf{h}} = \mathbf{W}^{-1}\mathbf{Px} \tag{12}$$

and then the residual signal $r[n]$ is obtained by $r[n] = s[n] - \hat{h}[n]$.

3.1. Conditioning problem

At each analysis time instant, n_a^i, on voiced frames, $5\,L(n_a^i)$ real coefficients should be estimated. In order to avoid conditioning problems, one could choose the analysis frame length to be as short as $5\,L(n_a^i)$ samples. By way of illustration, in order to take into account sinusoids with frequencies up to $F_s/4$, the number of harmonics is equal to $L = T_0(n_a^i)/4$ where $T_0(n_a^i)$ is the pitch period in samples. Thus the analysis frame could be as short as $5\,T_0(n_a^i)/4$. In practice, we choose two pitch periods as frame length around an analysis time-instant, in order to avoid ill-conditioned problems as well as to allow sinusoids over $F_s/4$.

4. THE RESIDUAL SIGNAL

4.1. Variance of the residual signal

It can be shown [9] [5] that if the input signal, $s[n]$, is a white noise with unit variance then the variance of the residual signal, $r[n]$, is the diagonal of the follow matrix:

$$E(\mathbf{r}\mathbf{r}^h) = \mathbf{I} - \mathbf{W}\mathbf{P}(\mathbf{P}^h\mathbf{W}^h\mathbf{W}\mathbf{P})^{-1}\mathbf{P}^h\mathbf{W}^h \tag{13}$$

Using the same weighting window $w(t)$, a typical Hamming window, the variance of the residual signal from the above model and from the classic HNM [10] (zero order amplitude polynomial and linear phase) have been computed. The two variances are depicted in Fig.1. For convenience, HNM$_1$ will be referred to the classic HNM, while HNM$_2$ will be referred to the model presented here (third order polynomial for the amplitudes and linear phase). The variance of the HNM$_1$ residual signal is represented by a solid line and the variance of the HNM$_2$ residual signal by dashdotted line. As is made clear by Fig.1 the variance of the HNM$_1$ residual signal is not evenly distributed across the analysis frame (as it should ideally be) in contrast to the variance of the residual signal obtained from the HNM$_2$; it can be seen that the variance from HNM$_2$ is closest to the ideal least-squares variance for most of the analysis time.

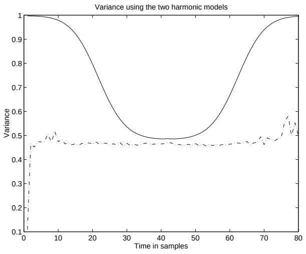

Figure 1: Variance of the least-squares residual signal form HNM$_1$ (solid line) and HNM$_2$ (dashdotted line). The weighting window was a typical hamming window.

4.2. Residual signal

In this section the residual signals from the HNM$_1$ and HNM$_2$ are compared. Note that the HNM$_1$ has no information about the parameters (harmonic amplitudes) variation within the analysis frame. This causes low frequencies to appear in the residual signal of the HNM$_1$. To show the above behaviour of the HNM$_1$ residual signal, a voiced fricative frame from an original speech signal sampled at $16kHz$ was selected. The frame used in the present study is plotted in Fig.2(a); the length of the analysis frame is two times the local pitch period. Harmonic peaks have been found up to $4000Hz$. Fig.2(b) shows the residual signal from the HNM$_1$. This figure, clearly shows the above indicated behaviour of the HNM$_1$ residual signal. For the same frame the deterministic part and the residual signal from the HNM$_2$ are presented in Fig.2(c) and (d) respectively.

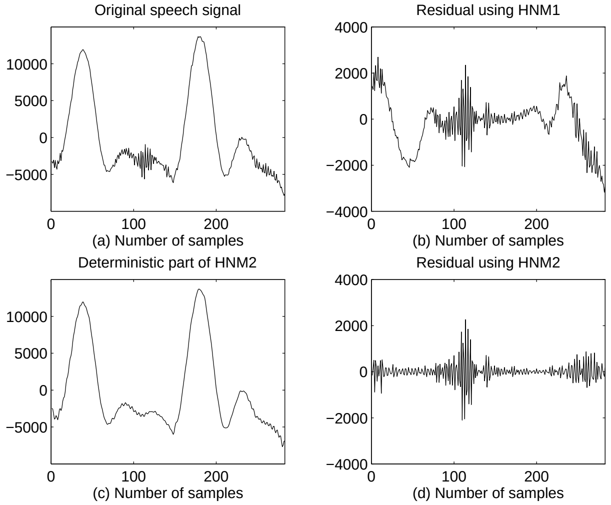

Figure 2: (a) A fricative voiced of an original speech signal and residual error signals from (b) HNM$_1$, and (d) HNM$_2$. The deterministic (periodic) part from HNM$_2$ is depicted in (c).

The measure of similarity between deterministic part and original signal used here, is given by

$$E = 10log_{10}\frac{\sigma^2_{r(t)}}{\sigma^2_{s(t)}} \tag{14}$$

where $\sigma^2_{r(t)}$ denotes the variance of the residual signal $r(t)$ and $\sigma^2_{s(t)}$ denotes the variance of the original speech signal $s(t)$. For example, the error produced from the two models for the original speech frame in Fig.2 was : $-15.8dB$ for HNM$_1$, and $-25.58dB$ for HNM$_2$.

In order to prove that the model HNM$_2$ is robust with respect to background noise the following test was done: a white noise was filtered by a low pass filter with cutoff frequency equal to the maximum voiced frequency ($4000Hz$) and it was added to the original speech signal; the HNM$_2$ parameters are estimated using the resulting noise corrupted speech signal. The frequency content of the two residual signal without and with the additive noise are shown in Fig.3. This figure clearly shows that the additive noise (about $25dB$) remains into the stochastic part (the noise level into the harmonic band, $0 - 4000Hz$, has been clearly increased). Note that the spectrum of the original signal has also been included in Fig.3 (dashed line).

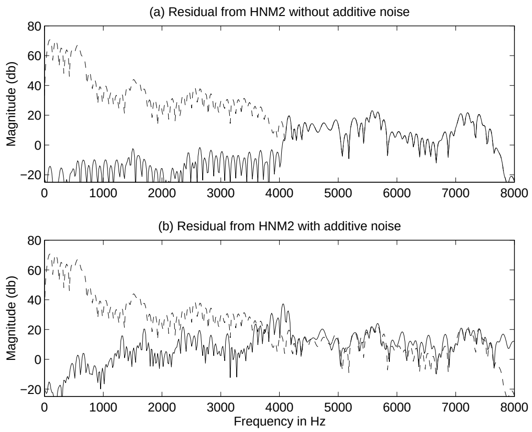

Figure 3: Magnitude of the Fourier transform of the residual signal (solid line) from HNM$_2$ (a)without and (b) with additive noise. The magnitude of the Fourier transform of the original speech signal has been also included (dashed line).

5. EXAMPLE OF APPLICATION

As an example of application, Fig. 4(a) shows a segment of a speech signal and in (b) the modeling error in dB (using the Eq.14) produced from the HNM$_1$ and HNM$_2$ is plotted. The modeling errors using a zero (HNM$_1$) and a third (HNM$_2$) order polynomial for the harmonic amplitudes are presented by a solid and a dashed line respectively in Fig. 4(b). As the full frequency band (from $0Hz$ up to half of the sampling frequency) of the residual signal is used in the definition of the modeling error, the error is large on voiced fricative regions (for example, between the 25^{th} and 35^{th} analysis frame). To give an acoustic notion to the modeling error, it is worth noting that if the modeling error is less than $-25dB$, the synthetic speech signal produces from the models is *indistinguishable* from the original speech signal.

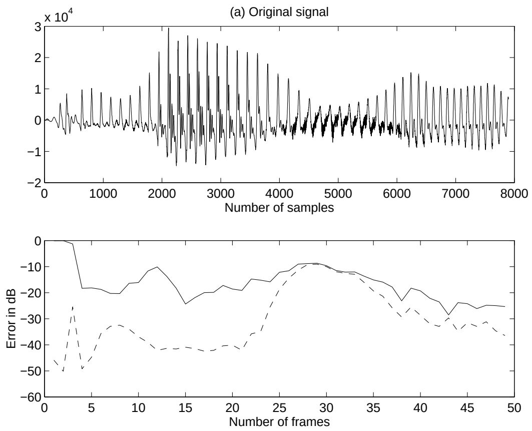

Figure 4: (a) Original speech signal *"wasi"* (b) modeling error in dB.

6. CONCLUSION

This paper has presented a version of the HNM which aims at decomposing the speech signal between two parts: (1) a deterministic (periodic) part using a sum of harmonically related sinusoids with time-varying harmonic amplitudes and linear phase and (2) a stochastic part which is defined as the difference (in the time domain) between the original speech signal and the deterministic part. The model has been extensively tested on a large number of speech signals, with male and female speakers given very satisfactory results. The analysis method has proved robust with respect to background noise. The proposed model is intended to be used for voice conversion, signal enhancement and monoral voice-separation.

Acknowledgment

The author would like to thank J. Laroche and E. Moulines for their encouragement and precious remarks during this work.

7. REFERENCES

1. C. d'Alessandro, B.Yegnanarayana, and V. Darsinos. Decomposition of speech signals into deterministic and stochastic components. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 760–763, 1995.
2. D.W. Griffin and J.S. Lim. Multiband-excitation vocoder. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-36(2):236–243, Fev 1988.
3. W. Hess. *Pitch determination of Speech Signals: Algorithmes and Devices.* Springer, Berlin, 1983.
4. F. Klingholz. The measurement of the Signal-to-Noise Ration (SNR) in Continuous Speech. *Speech Communication*, 6:15–26, 1987.
5. J. Laroche, Y. Stylianou, and E. Moulines. HNS: Speech modification based on a harmonic + noise model. *Proc. IEEE ICASSP-93, Minneapolis*, Apr 1993.
6. C. L. Lawson and R. J. Hanson. *Solving Least–Squares Problems.* Prentice Hall, Englewood Cliffs, New Jersey, 1974.
7. W.H. Press, S.A. Teukolsky, W.T Vettering, and B.P. Flannery. *Numerical Recipes in C, Second Edition.* Cambridge University Press, 1994.
8. X. Serra and J. Smith. Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition. *Computer Music J.*, 14(4):12–24, Winter 1990.
9. Y. Stylianou. *Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification.* PhD Dissertation. Ecole Nationale Supèrieure des Télécommunications, Paris, Jan 1996.
10. Y. Stylianou, J. Laroche, and E. Moulines. High-Quality Speech Modification based on a Harmonic + Noise Model. *Proc. EUROSPEECH*, 1995.