Asymptotically Exact AM-FM Decomposition
Based on Iterated Hilbert Transform
Francesco Gianfelici, Giorgio Biagetti, Paolo Crippa, and Claudio Turchetti
Dipartimento di Elettronica, Intelligenza Artificiale e Telecomunicazioni,
Università Politecnica delle Marche,
Via Brecce Bianche 12, I-60131 Ancona, Italy.

Abstract
This paper presents a multicomponent sinusoidal model of speech signals, obtained through a rigorous mathematical formulation that ensures an asymptotically exact reconstruction of these nonstationary signals, despite the presence of transients, voiced segments, or unvoiced segments. This result has been obtained by means of the iterated use of the Hilbert transform, and the convergence properties of the proposed method have been both analytically investigated and empirically tested. Finally, an adaptive segmentation algorithm used to accurately compute instantaneous frequencies from unwrapped phases, suited to complete the proposed AM-FM model, is presented.

1. Introduction
Sinusoidal models, as defined by Quatieri and McAulay in [1], are highly parametric representations of speech signals, based on physiologic properties of speech production and perception. These representations can be assimilated to the joint effect of both Amplitude Modulation (AM) and Frequency Modulation (FM), where carriers are to be determined together with amplitudes and frequency envelopes.

Several extensions to the basic sinusoidal model have been proposed, in order (i) to improve the modeling accuracy during transients, (ii) to avoid the problem of using too long time frames, for which the time resolution is degraded, or too short time frames, for which the frequency resolution is degraded and sinusoidal component estimation becomes difficult, and (iii) to prevent distortions and pre-echoes from appearing in the signal reconstruction due to parameter nonstationarity, both inside the frame and between frames. Some of the most known extensions are based on exponential damping and delays, such as Exponential Sinusoidal Model (ESM), Exponentially Damped Sinusoids (EDS), Damped Delayed Sinusoids (DDS), and Partial Damped and Delayed Sinusoids (PDDS).

Different techniques to estimate the model parameters have been developed too, based either on the Teager-Kaiser operator [2] or on the Hilbert transform [3]. Nevertheless, the approximation inherent to these techniques generates problems, especially in the reconstruction of transients.

More recently, Huang et al. [4] proposed a different analysis technique suited for nonlinear and non-stationary data, the Empirical Mode Decomposition (EMD), which is based on the iterative extraction of intrinsic modes, followed by the application of the Hilbert transform to compute a spectrum from them. This iterative technique looks very promising and has been applied to many different application fields, such as seismology, oceanography, and the processing of biological data, but its applicability in the speech processing area has not yet received a very thorough investigation in the literature.

In this paper we present a different iterative approach to compute an asymptotically exact multicomponent sinusoidal model of speech signals, based on the iterated application of the Hilbert transform to a filtered version of the amplitude envelope, and on the exact computation of these amplitudes and associated phases. The algorithm can be applied to signals without limitations on their duration, on the number of components to be extracted, and on the desired modeling accuracy during both stationary signal portions and transients. For the purpose of completing the FM decomposition, an a posteriori adaptive segmentation algorithm is finally used to extract arbitrarily accurate instantaneous frequencies from the phases previously obtained.

2. Sinusoidal Modeling
The representation of a given signal $x(t)$ in terms of a sinusoidal multicomponent model (or AM-FM) is given by:

$$x(t) = \sum_{i=1}^{M} A_i(t) \cos[\omega_i t + \phi_i(t)], \quad 0 < t < T \quad (1)$$

where $A_i(t)$ are the signal envelopes, $\omega_i$ are the instantaneous frequencies (or frequency centroids), and $\phi_i(t)$ the instantaneous phases. In case of $M = 1$, (1) reduces to a monocomponent model and the summation argument is named resonance [5]. Generally $A_i(t)$ should be slowly varying and $\omega_i$ should be a constant or should vary slowly.

Two different techniques are widely used in order to extract the aforementioned parameters ($A_i(t)$, $\omega_i$, $\phi_i(t)$). The first one is based on a differential nonlinear approach that uses the Teager energy operator. The second one is based on the Gabor analytical signal obtained from the Hilbert transform. This work uses a technique similar to the latter. Let $z(t)$ be a generic signal and $\hat{z}(t)$ its associate complex Gabor signal, defined as:

$$z(t) = x(t) + iH[x(t)] = x(t) + i\hat{x}(t) \quad (2)$$

where $\hat{x}(t) = H[x(t)]$ is the Hilbert transform\(^1\). The envelope and the instantaneous frequency are obtained as the amplitude and the derivative of the instantaneous phase of the complex signal, respectively.

Technically, in the implementation phase of the above extraction techniques, band-pass or low-pass filtering processes are needed to regularize the large variations of the envelope

\(^1\)Defined, for instance, by using the Fourier transforms $X(\omega)$ of $x(t)$ and $\hat{X}(\omega)$ of $\hat{x}(t)$ as $\hat{X}(\omega) = -i \text{sign}(\omega) X(\omega)$. 

and instantaneous frequency estimations. These filtering processes are effective in the modeling of stationary signals, but their effectiveness is reduced for highly nonstationary signals like speech signals. Therefore these techniques are used only after segmentation of the time span. But, at the best of authors’ knowledge, the methodologies so far adopted for parameter extraction are approximated: as a result a non-exact reconstruction of the original signal is obtained due to distortions and pre-echoes.

3. Iterative AM-FM Decomposition

Let \( x(t) \) be a generic speech signal. By virtue of (2) it is possible to rewrite it as:

\[
x(t) = \Re[z(t)] = a_0(t) \cos[\alpha_0(t)]
\]

(3)

where \( a_0(t) = |z(t)| \) and \( \alpha_0(t) = \arg[z(t)] \). Our aim here is to obtain a multicomponent decomposition of \( x(t) \) by means of iterated applications of representations like (3) to the amplitude component. Of course, since \( \alpha_j(t) \) is always non-negative, it is necessary to separate its “almost constant” component \( \pi_j(t) \) from its “alternating” component \( \tilde{\alpha}_j(t) \) beforehand, with a suitable adaptive filtering algorithm acting upon \( \alpha_j(t) \) itself, so that \( \alpha_j(t) = \tilde{\alpha}_j(t) + \pi_j(t) \). The first step, with \( j = 0 \), will thus be:

\[
x(t) = a_0(t) \cos[\phi_0(t)] = [\pi_0(t) + \tilde{\alpha}_0(t)] \cos[\phi_0(t)]
\]

(4)

where \( \phi_0(t) = \alpha_0(t) \). By denoting with:

\[
z_{j+1}(t) = \tilde{\alpha}_j(t) + iH[\tilde{\alpha}_j(t)]
\]

(5)

the (complex) Gabor signal associated with the alternating component, it is possible to proceed with the decomposition using the relations:

\[
a_j(t) = |z_j(t)| \quad \alpha_j(t) = \arg[z_j(t)] \quad j = 1, \ldots, N
\]

(6)

so that it results:

\[
\tilde{a}_0(t) = a_1(t) \cos[\alpha_1(t)].
\]

(7)

By substituting (7) into (4) we obtain, after simple manipulations:

\[
x(t) = \pi_0(t) \cos[\phi_0(t)] + \frac{a_1(t)}{2} \sum_{i=1}^{2^1} \cos[\phi_1(t)]
\]

(8)

where:

\[
\phi_1(t) = \alpha_1(t) - \phi_0(t) \quad \phi_1(t) = \alpha_1(t) + \phi_0(t).
\]

(9)

From (9) it is apparent that the number of components increases geometrically with the number of iterations. Letting this latter number be \( N + 1 \), and using the generalization of (7):

\[
\tilde{\alpha}_j(t) = a_{j+1}(t) \cos[\alpha_{j+1}(t)] \quad j = 0, \ldots, N - 1
\]

(10)

it is possible to write:

\[
x(t) = \sum_{j=0}^{N} \pi_j(t) \sum_{i=1}^{2^j} \cos[\phi_j(t)] + \frac{\tilde{a}_N(t)}{2N} \sum_{i=1}^{2^N} \cos[\phi_N(t)]
\]

(11)

\[\text{The filter, for instance, could be defined so as to keep in the alternating component only a fraction } \kappa < 1/2 \text{ of the total signal energy.}\]

that is a generalized multicomponent sinusoidal model, in which the phases \( \phi_N(t) \) can be iteratively computed as:

\[
\phi_N^{2L-1}(t) = \alpha_N(t) - \phi_{N-1}(t)
\]

(12)

\[
\phi_N^{2L}(t) = \alpha_N(t) + \phi_{N-1}(t)
\]

(13)

for \( 1 \leq L \leq 2^{N-1} \). It can be shown that, as \( N \) increases, the last term in (11):

\[
r_N(t) = \frac{\tilde{a}_N(t)}{2N} \sum_{i=1}^{2^N} \cos[\phi_i(t)]
\]

(14)

rapidly vanishes:

\[
\|r_N\|^2 \leq \|\tilde{a}_N\|^2 \leq \kappa \|a_N\|^2 = 2\kappa \|\tilde{a}_{N-1}\|^2
\]

(15)

where \( \kappa \) (less than one half) is the filter energy loss, and the factor 2 stems from (6) and the fact that the Hilbert transform, which is used in (5), preserves energy. This topic will be discussed with more detail in Sect. 5, where an example of this truncation error will be shown as a function of \( N \). From that it will be possible to state that good approximations are possible even with low values of \( N \). Having shown that:

\[
\lim_{N \to \infty} \frac{\tilde{a}_N(t)}{2^N} \sum_{i=1}^{2^N} \cos[\phi_i(t)] = 0,
\]

(16)

equation (11) can be reformulated as:

\[
x(t) \approx \sum_{k=1}^{2^{N+1}-1} \overline{\pi}_k(t) \cos[\Phi_k(t)]
\]

(17)

where:

\[
\Phi_k(t) = \phi_k(t), \quad k = 2^j - 1 + i
\]

(18)

and \( \overline{\pi}_k(t) = \pi_j(t)/2^j \) for \( 2^j - 1 < k < 2^{j+1} - 1 \).

Eq. (11) is an exact decomposition of the signal \( x(t) \) in terms of amplitude and phase envelopes, and by virtue of (16) it can be almost always reduced to the simplified form (17). The next section will show how to obtain a decomposition in terms of instantaneous frequencies from the phase envelopes derived here.

4. Segmentation Algorithm for the Instantaneous Frequency Calculus

Given the model stated in (17), let:

\[
\Phi_k(t) = \overline{\Phi}_k(t) + \bar{\Phi}_k(t)
\]

(19)

where:

\[
\overline{\Phi}_k(t) = \int_{t-1}^{t} \overline{\pi}_k(t) \, dt
\]

(20)

and \( \bar{\Phi}_k(t) \) is the modeling error, while \( \overline{\Phi}_k(t) \) is the instantaneous frequency we look for.

In the literature several methods have been proposed to estimate \( \overline{\pi}_k(t) \), such as Short Time Fourier Transform, Multiband Demodulation Analysis (Time-Varying Gabor Filterbank), Matching Pursuit Technique, Instantaneous Frequency Attractors, which are almost all based on a maybe adaptive a priori segmentation of the time span \( [0, T] \) [6].

Our approach is based on the assumption that \( \overline{\pi}_k(t) \) can be assumed to be piecewise constant if the intervals over which it
is constant are adaptively estimated a posteriori. This makes it easy to impose an upper bound to the error like:

\[ |\Phi_k(t)| < \varepsilon(t) \quad (21) \]

where \( \varepsilon(t) \) is the desired accuracy.

From (21) the adaptive segmentation problem can be stated as the problem of finding a set of disjoint time spans \( I_k = [T_{an}, T_{bk}] \) that covers the whole \([0, T]\) interval, such that \( \Phi_k(t) \) is constant in each \( I_k \) and (21) holds.

The proposed algorithm is sketched in Fig. 1, and it is based on the following two points:

(i) instantaneous frequencies may be obtained from unwrapped phases \( \Phi_k(t) \) by means of discrete approximation of the relation:

\[ \omega_k(t) = (d/dt)\Phi_k(t) \quad (22) \]

(ii) \( \Phi_k \) can be estimated over finite intervals by means of linear regression using data provided by \( \Phi_k(t) \). Linear regression is also known to be a robust technique to compute instantaneous frequencies from noisy phase signals.

By making use of the aforementioned ideas, the proposed algorithm starts with \( T_a = 0 \), and computes the derivative of the unwrapped phase as in (i) by means of the method in (ii), increasing \( T_b \) until the condition (21) is no longer met \( \forall t \in [T_a, T_b] \). Then it advances \( T_a \) to the last \( T_b \) that satisfied the condition and iterates until \( T_a = T \).

Figure 2: The synthetic signal (23) used for testing (a), its first two amplitude envelopes \( \pi_0(t) \) (solid line) and \( \pi_1(t) \) (dashed line) (b), and its unwrapped phases \( \alpha_0(t) \) (solid) and \( \alpha_1(t) \) (dashed) (c).

The result so obtained is an estimation of the instantaneous frequency, and the modeling error \( \Phi_k(t) \) can be made arbitrarily small. The accuracy is of course directly related to the number of intervals produced, and the algorithm easily allows the introduction of signal-dependant bounds \( \varepsilon(t) \) to take into account phenomena like pre-echoes, distortions, and so on.

### 5. Application Examples

In this section the application of the proposed algorithms to both synthetic and speech signals is presented.

#### 5.1. Synthetic Signal

In order to better understand the decomposition behavior of the proposed approach, we applied it to a simple two-component AM synthetic signal of the form:

\[ x(t) = A_1(t) \sin(2\pi f_1 t) + A_2(t) \sin(2\pi f_2 t) \quad (23) \]

with:

\[ A_1(t) = 1 + 0.5 \cos(2\pi f_0 t) \quad , \quad A_2(t) = 0.2 \quad (24) \]

where \( f_0 = 10 \text{ Hz}, f_1 = 500 \text{ Hz}, \) and \( f_2 = 1650 \text{ Hz} \). This signal is shown in Fig. 2, together with the first two amplitude envelopes and the corresponding phases. It is easy to note that the amplitude envelopes accurately match the original modulating signals \( A_1(t) \) and \( A_2(t) \). The phase curves reported in Fig. 2(c) correspond to \( \alpha_0(t) \) (solid line) and to \( \alpha_1(t) \) (dashed line), and have a mean slope of 500.0 Hz and 1149.7 Hz, respectively. The first slope corresponds exactly to the carrier frequency \( f_1 \), whilst the latter needs to be combined with the former (added, in this case) to obtain the carrier frequency \( f_2 \), as it should be according to the formulation given in Sect. 3.
Figure 3: The Italian word “settima” (a), its elementary amplitude envelopes $\pi_j(t)$ (b), and phases $\alpha_j(t)$ (c).

5.2. Speech Signal

The signals used are part of the Italian portion of the Multitext Prosodic Database [7], which is composed of about 150 sentences spoken by 10 different speakers and segmented at the word level.

Figure 3 shows elementary amplitude envelopes $\pi_j(t)$ and phases $\alpha_j(t)$ obtained applying $N = 5$ iterations of the sinusoidal model to the Italian word “settima”, while Fig. 4 shows the result of the adaptive segmentation algorithm applied to the first extracted phase with an error bound $\varepsilon(t) = 2\pi$. The results show how the required accuracy has been obtained through a highly irregular segmentation of the time axis, which would have been quite hard to find had the segmentation been made a posteriori.

The asymptotically exactness of the proposed method arises from (16), but it can be empirically stated that the convergence is achieved even with low values of $N$, as it is shown in Fig. 5. The relative energy of the residual $\|r_N(t)\| / \|x(t)\|$ is plotted as a function of the iteration number $N$, and from that it can be seen that $20 < N < 30$ suffice to give an error that can be assimilated to round-off noise. Moreover, on the grounds of a subjective listening test performed with 40 people, it is possible for us to state that, already with $N \geq 6$, the model can be deemed equivalent to the original signal, with a residual error on the order of $10^{-3}$.

6. Conclusion

This paper presents an asymptotically exact multicomponent sinusoidal model of speech signals based on the iterated application of the Hilbert transform. Instantaneous frequencies have been obtained by means of linear regression over time intervals adaptively detected a posteriori so as to permit arbitrary accuracy to be achieved.

The proposed algorithm has been applied to a synthetic signal to highlight its capabilities of separating different components and to a speech signal to show its modeling effectiveness.

The result obtained overcomes most of the limitations that are proper of traditional Quatieri-McAulay sinusoidal models for speech signals, when the model parameters are exactly estimated and the adaptive segmentation is not performed a priori.

7. References