AM-FM DECOMPOSITION OF SPEECH SIGNALS: AN ASYMPTOTICALLY EXACT APPROACH BASED ON THE 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.

Keywords: AM-FM Model, Sinusoidal Model, Instantaneous Frequency Estimation, Envelope Estimation, Iterated Hilbert Transform, Gabor Signal.

1. INTRODUCTION

In their pioneering work [1], Quatieri and McAulay define sinusoidal models as highly parametric representations of speech signals, based on physiologic properties of speech production and perception. These representations are effectively similar to joint Amplitude Modulation (AM) and Frequency Modulation (FM), where carriers are to be determined together with amplitude and frequency envelopes.

A number of 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 frequency resolution is degraded and sinusoidal component estimation become 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 are based on exponential damping and delays, such as Exponential Sinusoidal Model (ESM) [2], Exponentially Damped Sinusoids (EDS) [3], Damped Delayed Sinusoids and Partial Damped Delayed Sinusoids (DDS and PDDS) [4, 5].

Different techniques to estimate the model parameters have been developed too, based on the Teager-Kaiser operator [6] or the Hilbert transform [7]. Nevertheless the approximation inherent to these techniques causes problems especially in the reconstruction of transients.

This paper present an asymptotically exact multicomponent sinusoidal model, based on the iterated application of the Hilbert transform and on the exact computation of amplitude envelopes and arbitrarily accurate instantaneous frequencies, without regard to the length, number of components, and desired modeling accuracy during both stationary signal portions and transients. An *a posteriori* adaptive segmentation algorithm is used to limit the phase error in the FM decomposition.

2. SINUSOIDAL MODELING: LIMITATIONS AND EXTENTIONS

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

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

where $A_i(t)$ are the signal envelopes, ω_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 called *resonance* [8]. Generally $A_i(t)$ should be slowly varying and ω_i should be a constant or should vary slowly.

To extract the aforementioned parameters ($A_i(t)$, ω_i, $\phi_i(t)$), two different techniques are widely used. The first one is based on a differential non-linear 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 $x(t)$ a generic signal and $z(t)$ its associate complex Gabor signal, defined as:

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

where $H[x(t)]$ 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 and instantaneous frequency estimations. These filtering processes are effective in the modeling of stationary signals, but their effectiveness is reduced for highly non-stationary signals like speech signals. Therefore these techniques are used only after segmentation of the time span. But, at the best 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. AM-FM DECOMPOSITION: THE MATHEMATICAL FORMULATION

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

$$x(t) = \Re[z(t)] = a_0(t)\cos[\alpha_0(t)] \tag{3}$$

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 $a_j(t)$ is always non-negative, it is necessary to separate its "almost constant" component $\overline{a}_j(t)$ from its "alternating" component $\widetilde{a}_j(t)$ beforehand, with a suitable adaptive filtering algorithm[2] acting upon $a_j(t)$ itself, so that:

$$a_j(t) = \widetilde{a}_j(t) + \overline{a}_j(t)\,. \tag{4}$$

The first step, with $j = 0$, will thus be:

$$x(t) = a_0(t)\cos[\phi_0^1(t)] = [\overline{a}_0(t) + \widetilde{a}_0(t)]\cos[\phi_0^1(t)] \tag{5}$$

where:

$$\phi_0^1(t) = \alpha_0(t)\,. \tag{6}$$

By denoting with:

$$z_{j+1}(t) = \widetilde{a}_j(t) + iH[\widetilde{a}_j(t)] \tag{7}$$

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

$$a_j(t) = |z_j(t)| \quad \alpha_j(t) = \arg[z_j(t)] \qquad j = 1, \ldots, N \tag{8}$$

so that it results:

$$\widetilde{a}_0(t) = a_1(t)\cos[\alpha_1(t)] \tag{9}$$

which, once placed inside of (5) and having made use of basic trigonometry, yields:

$$x(t) = \overline{a}_0(t)\cos[\phi_0^1(t)] + \frac{a_1(t)}{2}\sum_{i=1}^{2^1}\cos[\phi_1^i(t)] \tag{10}$$

where:

$$\phi_1^1(t) = \alpha_1(t) - \phi_0^1(t) \qquad \phi_1^2(t) = \alpha_1(t) + \phi_0^1(t)\,. \tag{11}$$

From (11) 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 (9):

$$\widetilde{a}_j(t) = a_{j+1}(t)\cos[\alpha_{j+1}(t)] \qquad j = 0, \ldots, N-1 \tag{12}$$

it is possible to write:

$$x(t) = \sum_{j=0}^{N}\frac{\overline{a}_j(t)}{2^j}\sum_{i=1}^{2^j}\cos[\phi_j^i(t)] + \frac{\widetilde{a}_N(t)}{2^N}\sum_{i=1}^{2^N}\cos[\phi_N^i(t)] \tag{13}$$

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

$$\phi_N^{2L-1}(t) = \alpha_N(t) - \phi_{N-1}^L(t) \tag{14}$$

$$\phi_N^{2L}(t) = \alpha_N(t) + \phi_{N-1}^L(t) \tag{15}$$

for $1 \le L \le 2^{N-1}$. It can be shown that, as N increases, the last term in (13):

$$r_N(t) = \frac{\widetilde{a}_N(t)}{2^N}\sum_{i=1}^{2^N}\cos[\phi_N^i(t)] \tag{16}$$

rapidly vanishes. In fact:

$$\|r_N\|^2 \le \|\widetilde{a}_N\|^2 \le \kappa\,\|a_N\|^2 = 2\kappa\,\|\widetilde{a}_{N-1}\|^2 \tag{17}$$

where κ (which, as mentioned before, must be less than one half) is the filter energy loss, and the factor 2 stems from (7)–(8) and the fact that the Hilbert transform preserves energy. Section 5 will discuss this topic in more depth, and examples of this truncation error viewed as a function of N will be shown. From that it will be possible to state that good approximations are possible even with low values of N.

Having shown that:

$$\lim_{N\to\infty}\frac{\widetilde{a}_N(t)}{2^N}\sum_{i=1}^{2^N}\cos[\phi_N^i(t)] = 0 \tag{18}$$

[1] defined by using the Fourier transforms $X(\omega)$ of $x(t)$ and $\hat{X}(\omega)$ of $\hat{x}(t)$ as $\hat{X}(\omega) = -i\ \mathrm{sign}(\omega)\,X(\omega)$ [9].

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

equation (13) can be rewritten as:

$$x(t) \approx \sum_{k=1}^{2^{N+1}-1} \overline{A}_k(t) \cos[\Phi_k(t)]. \quad (19)$$

where:

$$\Phi_k(t) = \phi_j^i(t) \text{ , with } k = 2^j - 1 + i \quad (20)$$

and:

$$\overline{A}_k(t) = \frac{\overline{a}_j(t)}{2^j} \text{ , for } 2^j - 1 < k < 2^{j+1} - 1 \, . \quad (21)$$

Eq. (13) is an exact decomposition of the signal $x(t)$ in terms of amplitude and phase envelopes, and by virtue of (18) it can be almost always reduced to the simplified form (19). In the next section it will be shown how to obtain a decomposition in terms of instantaneous frequencies from the phase envelopes derived here.

4. INSTANTANEOUS FREQUENCY CALCULUS: THE ADAPTIVE SEGMENTATION ALGORITHM

With reference to (19), let:

$$\Phi_k(t) = \overline{\Phi}_k(t) + \widetilde{\Phi}_k(t) \quad (22)$$

where:

$$\overline{\Phi}_k(t) = \int_0^t \overline{\omega}_k(\tau) \, d\tau \quad (23)$$

and $\widetilde{\Phi}_k(t)$ is the modeling error, while $\overline{\omega}_k(t)$ is the instantaneous frequency we look for. A number of methods have been proposed in the literature to estimate $\overline{\omega}_k(t)$, 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 an *a priori*, maybe adaptive [10], segmentation of the time span $[0, T]$. Our approach is based on the assumption that $\overline{\omega}_k(t)$ 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:

$$\left| \widetilde{\Phi}_k(t) \right| < \varepsilon(t) \quad (24)$$

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

From (24) the adaptive segmentation problem can be straightforwardly stated as the problem of finding the minimum set of compact disjoint time spans that covers the whole $[0, T]$ interval, where in every one of which both $\overline{\omega}_k(t)$ is constant and (24) holds.

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

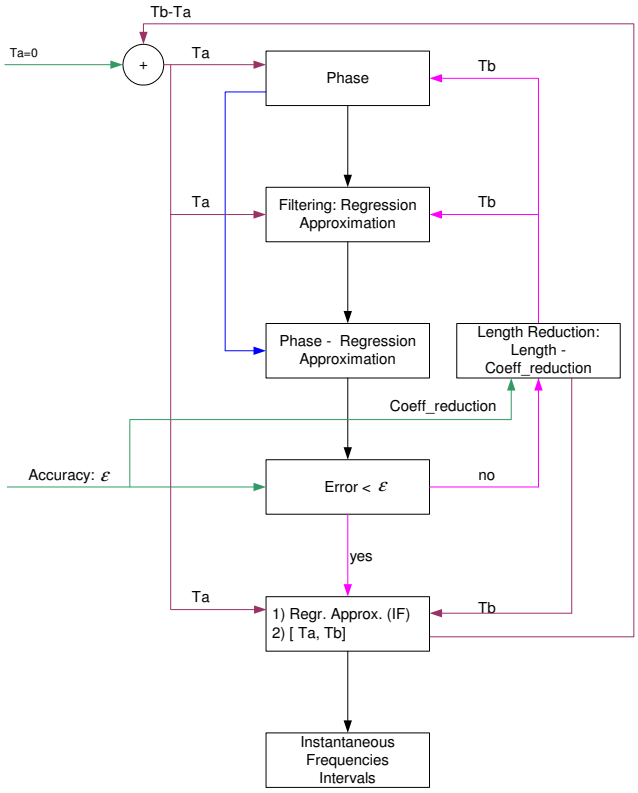

Fig. 1. Sketch of the adaptive segmentation and Instantaneous Frequency extraction algorithm.

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

$$\omega_k(t) = (d/dt)\Phi_k^S(t) \quad (25)$$

(ii) $\overline{\omega}_k$ can be estimated over finite intervals by means of linear regression using data provided by $\Phi_k(t)$. This also provides a direct relationship between interval length and error since it is obvious that:

$$\max\{|\widetilde{\Phi}_k([T_a, T_{b1}])|\} \leq \max\{|\widetilde{\Phi}_k([T_a, T_{b2}])|\} \quad (26)$$

for $T_{b1} \leq T_{b2}$. 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)*, finding the maximum T_b that satisfies (24) $\forall t \in [T_a, T_b]$. Then it advances T_a to T_b and iterates until $T_a = T$.

The result so obtained is an estimate of the instantaneous frequency, and the modeling error $\widetilde{\Phi}_k(t)$ can be made

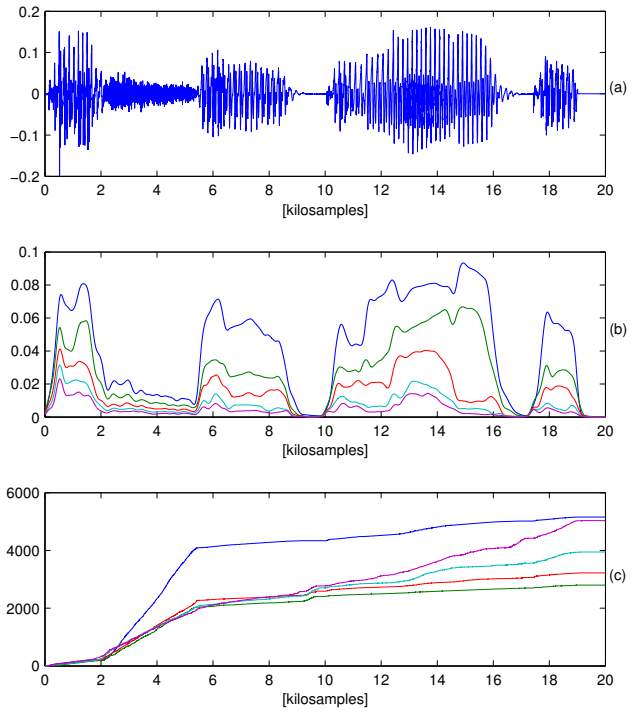

Fig. 2. The Italian word "assolutamente" (a), its elementary amplitude envelopes $\overline{a}_j(t)$ (b), and phases $\alpha_j(t)$ (c), for $j = 0, \ldots, 4$.

arbitrarily small. The accuracy of course is 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. ALGORITHM APPLICATION: THE ESTIMATION RESULTS

In this section the application of the proposed algorithms to speech signals is presented. The signals used are part of the Italian portion of the *Multext Prosodic Database* [11], which is composed of about 150 sentences spoken by 10 different speakers, sampled at 20 kHz and segmented at the word level.

Figure 2 shows elementary amplitude envelopes $\overline{a}_j(t)$ and phases $\alpha_j(t)$ obtained applying $N = 5$ iterations of the sinusoidal model to the Italian word "assolutamente", while Fig. 3 shows the result of the adaptive segmentation algorithm applied to the first extracted phase with an error bound $\varepsilon(t) = 10$. The results show how the required accuracy has been obtained through a highly irregular segmentation of the time axis, which would be quite hard to find if the segmentation was made *a priori*.

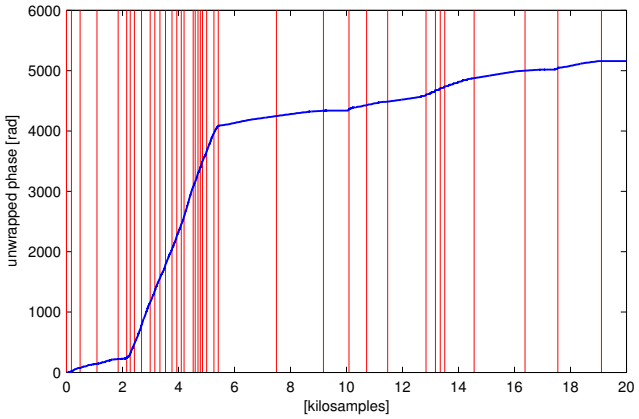

Fig. 3. Results of the adaptive segmentation algorithm applied to $\alpha_0(t)$. Vertical bars are interval boundaries.

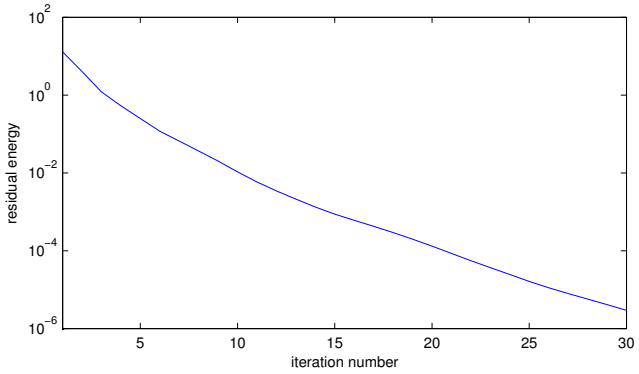

Fig. 4. Energy of the residual amplitude envelope $\|\widetilde{a}_N(t)\|^2$ as a function of the iteration number N.

The asymptotically exactness of the proposed method arises from (18), but it can be empirically stated that the convergence is achieved even with low values of N, as it is shown in Fig. 4. The energy of the residual amplitude envelope $\|\widetilde{a}_N(t)\|^2$ 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 the quantization noise of the audio signal. Moreover, on the grounds of a subjective listening test performed with 40 people, it is possible to state that, already with $N \geq 6$, the model can be deemed equivalent to the original signal.

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 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

[1] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34, no. 4, pp. 744 – 754, 1986.

[2] J. Jensen, S.H. Jensen, and E. Hansen, "Exponential sinusoidal modeling of transitional speech segments," in *In Proc. of IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP '99)*, 1999, vol. 1, pp. 473–476.

[3] K. Hermus, W. Verhelst, P. Lemmerling, P. Wambacq, and S.V. Huffel, "Perceptual audio modeling with exponentially damped sinusoids," *Signal Processing*, vol. 85, no. 1, pp. 163–176, 2005.

[4] R. Boyer and K. Abed-Meraim, "Estimation of damped and delayed sinusoids: algorithm and Cramer-Rao bound," in *Proc. of IEEE Int. Conf. Acoust., Speech, and Signal Processing (ICASSP '03)*, 2003, vol. 6, pp. 137–140.

[5] R. Boyer and K. Abed-Meraim, "Audio modeling based on delayed sinusoids," *IEEE Trans. Speech Audio Processing*, vol. 12, no. 2, pp. 110–120, 2004.

[6] P. Maragos, J.F. Kaiser, and T.F. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Trans. Speech Audio Processing*, vol. 41, no. 10, pp. 3024 – 3051, 1993.

[7] S.L. Hahn, *Hilbert Transforms in Signal Processing*, Artech House, Boston, 1996.

[8] A. Potamianos and P. Maragos, "A comparison of energy operators and the hilbert transform approach to signal and speech demodulation," *Signal Processing*, vol. 37, no. 1, 1994.

[9] L. Marple Jr., "Computing the discrete-time "analytic" signal via FFT," *IEEE Trans. Signal Processing*, vol. 47, no. 9, pp. 2600–2603, Sept. 1999.

[10] M. Goodwin and J. Laroche, "Audio segmentation by feature-space clustering using linear discriminant analysis and dynamic programming," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2003, vol. 1, pp. 131 – 134.

[11] Estelle Campione and Jean Véronis, "A multilingual prosodic database," in *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP'98)*, Sydney, Australia, 1998, vol. 7, pp. 3163–3166.