TRANSIENT MODELING FOR OVERLAP-ADD SINUSOIDAL MODEL OF SPEECH

Slava Shechtman

Speech Technologies, IBM Haifa Research Lab, Haifa, Israel

ABSTRACT

Speech sinusoidal modeling has been successfully applied to a broad range of speech analysis, synthesis and modification tasks. At most, it reproduces a high quality speech, however for speech transients (e.g. plosives, glottal stops) it suffers from reduced fidelity due to lack of intra-frame modeling of irregularities. Various extensions had been proposed for the stationary sinusoidal model to cope with this problem. One of simple and well-known in the art approaches is incorporating of an intra-frame *magnitude envelope* into the sinusoidal model. It used to be done by iterative analysis-by-synthesis procedure. In this paper we derive an optimal analytic solution for this problem. We will show that this solution yields significantly better model fit than the known-in-the-art analysis-by-synthesis approach.

Index Terms— Speech analysis, Sinusoidal modeling, Speech transient modeling, Magnitude envelope

1. INTRODUCTION

Speech sinusoidal modeling has been long in the core of many speech production models, used in a wide range if applications, such as speech synthesis, speech coding and speech transformation. Stationary sinusoidal modeling (SSM), representing a signal as a finite sum of sine waves, is widely used to describe a harmonic part of voiced speech, due to its simplicity and accuracy [1][2][3]. Stationary unvoiced signals can also be reliably represented by this model, assuming dense enough sampling of the spectrum (i.e. arbitrarily setting of "unvoiced pitch" $f_{0,uv} \le 100Hz$) [1]. With various noise modeling extensions (e.g. frequency jittering [2], phase randomization [2][4], noise addition [3]) it is capable of high quality synthesis of any quasi-stationary speech portions (voiced/unvoiced/mixed).

The SSM approach is rather efficient and practical. Speech can be synthesized with constant frame update rate using simple overlap-add operation (i.e. no need for costly sample-wise parameter interpolation), provided one uses a reasonable frame rate (e.g. 100-200 *Hz*) and precise frame alignment procedure at the synthesis [2][3][4].

However, the stationary models feature reduced fidelity at speech transients, such as voiced/unvoiced energy bursts, that are common in plosives and glottal stops. The so-called pre-echo effect (i.e. energy smearing at burst instants) is a typical artifact observed in SSM-generated speech [5].

Various methods were explored to improve handling of transients by the sinusoidal model. Some of them exploit special transient handling (i.e. transform coding [6][7] or increasing frame rate locally [6]), thus requiring some transient detection algorithms [6]. In [5] a very accurate adaptive model is proposed, but it is rather complicated both for analysis and for synthesis. On the contrary, the model proposed in [8] is less precise but more simple and practical for synthesis.

In the latter method that was proposed by George & Smith [8] and utilized for speech coding in [4], an intra-frame *magnitude envelope* is incorporated into the sinusoidal model to track speech energy variations across the frame. The speech synthesis for this sinusoidal model extension (named hereby *Magnitude Envelope Sinusoidal Modeling* or MESM) is kept simple (i.e. it can be performed in overlap-add constant-frame-rate manner). However, the proposed sinusoidal parameter estimation (sinusoidal amplitudes and phases) is iterative and sub-optimal [8].

In this paper we derive an optimal analytic solution for simultaneous extraction of MESM sinusoidal parameters (amplitudes and phases). We will show that this solution yields significantly better model fit (SNR) than the original iterative solution in [8].

The paper is structured as follows. First we review the SSM [2][3] and the MESM with iterative solution [8]. Then the optimal solution is derived. Further, the performance of the optimal solution is evaluated and compared to the SSM and the iterative MESM.

2. STATIONARY SINUSOIDAL MODELING (SSM)

Within the SSM formulation [2], the windowed portion of speech $s_w(n)$ is approximated by a finite sum of sine waves:

$$s_w(n) \cong \hat{s}_w(n) = w(n)\sum_{k=0}^{L} A_k \cos(\theta_k n + \varphi_k), \quad -N \le n \le N, \tag{1}$$

where $w(n)$ is a symmetric window , e.g. Hamming or Hanning, of $2N+1$ length, $\{A_k\}$ and $\{\varphi_k\}$ are harmonic amplitudes and phases correspondingly and θ_k is the position of the highest local maximum found on the short

time amplitude spectrum $\|S_w(\theta)\|$ in a close vicinity of $\theta_0 k$, i.e. the k-th multiple of the angular pitch frequency θ_0. Consequently, the determination of θ_k is based on a harmonic peak picking operation, requiring a preceding high-resolution pitch frequency estimation stage [1][2][3].

The dual representation of SSM approximation in frequency domain is given by

$$S_w(\theta) \cong \hat{S}_w(\theta) = \sum_{k=-L}^{L} \frac{A_k}{2} e^{j\varphi_k} W(\theta-\theta_k) \triangleq \sum_{k=-L}^{L} c_k \tfrac{1}{2} W(\theta-\theta_k), \quad (2)$$

where $S_w(\theta)$ is a short time spectrum, $W(\theta)$ is the DFT of $w(n)$ and the vector $\mathbf{c} \triangleq \{c_k\}_{k=0}^{L} \triangleq \{c_{\mathrm{Re},k} + jc_{\mathrm{Im},k}\}_{k=0}^{L}$, referred to as *line spectrum*, is to be estimated by an error criterion minimization. The error can be expressed either in time or in frequency domain:

$$\sum_{m=-N}^{N} \left\| s_w(m) - \hat{s}_w(m) \right\|^2 = \tfrac{1}{N_{FFT}} \sum_{m=0}^{N_{FFT}-1} \left\| S_w\left(2\pi m/N_{FFT}\right) - \hat{S}_w\left(2\pi m/N_{FFT}\right) \right\|^2, \quad (3)$$

The frequency domain estimation (2) is further developed to a matrix form as follows [2]:

$$\begin{aligned} &\hat{S}_w(\theta) = \sum_{k=0}^{L} c_k \tfrac{1}{2} W(\theta-\theta_k) + \overline{c}_k \tfrac{1}{2} W(\theta+\theta_k) = \\ &= \sum_{k=0}^{L} c_{\mathrm{Re},k} \tfrac{1}{2}\left(W(\theta-\theta_k) + W(\theta+\theta_k)\right) + \\ &+ jc_{\mathrm{Im},k} \tfrac{1}{2}\left(W(\theta-\theta_k) - W(\theta+\theta_k)\right). \end{aligned} \quad (4)$$

$$\hat{\mathbf{S}}_w = \mathbf{W}_1 \mathbf{c}_{\mathrm{Re}} + j\mathbf{W}_2 \mathbf{c}_{\mathrm{Im}},$$

where $\mathbf{c}_{\mathrm{Re}}$ and $\mathbf{c}_{\mathrm{Im}}$ are respectively the real and the imaginary parts of the line spectrum, and $\mathbf{W}_1$ and $\mathbf{W}_2$ are matrices containing shifted replicas of $W(\theta)$ as their columns:

$$\begin{cases} w_1(m,k) = \frac{1}{2} W(\frac{2\pi m}{N_{FFT}} - \theta_k) + \frac{1}{2} W(\frac{2\pi m}{N_{FFT}} + \theta_k) & 0 \le m \le {}^{N_{FFT}}\!/_{2} \\ w_2(m,k) = \frac{1}{2} W^i(\frac{2\pi m}{N_{FFT}} - \theta_k) - \frac{1}{2} W(\frac{2\pi m}{N_{FFT}} + \theta_k) & 0 \le k \le L \end{cases}. \quad (5)$$

In the case of symmetric windows (real-valued $\mathbf{W}$ matrices), the substitution of (4) in (3) and its minimization with respect to vectors $\mathbf{c}_{\mathrm{Re}}, \mathbf{c}_{\mathrm{Im}}$ results in the following equation set:

$$\begin{bmatrix} \mathbf{W}_1^T \mathbf{W}_1 & 0 \\ 0 & \mathbf{W}_2^T \mathbf{W}_2 \end{bmatrix} \begin{bmatrix} \mathbf{c}_{\mathrm{Re}} \\ \mathbf{c}_{\mathrm{Im}} \end{bmatrix} = \begin{bmatrix} \mathbf{W}_2^T \mathbf{S}_{\mathrm{Re}} \\ \mathbf{W}_2^T \mathbf{S}_{\mathrm{Im}} \end{bmatrix}, \quad (6)$$

In [3] it was shown that the time domain minimization can be expressed by a Toeplitz set of equations for efficient solution, however, here we present a different time domain formulation that resembles a frequency domain solution [2] and is further generalized in Section 4 of this paper. The time domain signal can be expressed in a matrix form as follows:

$$\begin{aligned} &\hat{s}_w(n) = \\ &= \sum_{k=0}^{L} A_k \cos(\varphi_k) w(n) \cos(\theta_k n) - A_k \sin(\varphi_k) w(n) \sin(\theta_k n), \\ &\qquad -N \le n \le N, \end{aligned} \quad (7)$$

$$\hat{\mathbf{s}}_w = \mathbf{T}_1 \mathbf{c}_{\mathrm{Re}} + \mathbf{T}_2 \mathbf{c}_{\mathrm{Im}}, \qquad \mathbf{c}_{\mathrm{Re}} = \left[A_0 \cos(\varphi_0) \cdots A_L \cos(\varphi_L) \right]^T \quad \mathbf{c}_{\mathrm{Im}} = \left[A_0 \sin(\varphi_0) \cdots A_L \sin(\varphi_L) \right]^T, \quad (8)$$

where $\mathbf{T}_1$ and $\mathbf{T}_2$ are matrices containing "windowed quadrature" $\mathbf{t}_1(k), \mathbf{t}_2(k)$, $0 \le k \le L$ columns vectors:

$$\begin{cases} \mathbf{t}_1(k) = \left[w(-N)\cos(-\theta_k N) \ldots w(N)\cos(\theta_k N) \right]^T \\ \mathbf{t}_2(k) = \left[-w(-N)\sin(-\theta_k N) \ldots -w(N)\sin(\theta_k N) \right]^T \end{cases} \quad (9)$$

The substitution (8) in (3) and its minimization with respect to vectors $\mathbf{c}_{\mathrm{Re}}, \mathbf{c}_{\mathrm{Im}}$ results in the following equation set

$$\begin{bmatrix} \mathbf{T}_1^T \mathbf{T}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{T}_2^T \mathbf{T}_2 \end{bmatrix} \begin{bmatrix} \mathbf{c}_{\mathrm{Re}} \\ \mathbf{c}_{\mathrm{Im}} \end{bmatrix} = \begin{bmatrix} \mathbf{T}_1^T \mathbf{s}_w \\ \mathbf{T}_2^T \mathbf{s}_w \end{bmatrix}. \quad (10)$$

Please note that we used here the fact that $\mathbf{T}_1^T \mathbf{T}_2 = \mathbf{T}_2^T \mathbf{T}_1 = \mathbf{0}$ for any symmetric $w(n)$.

3. ITERATIVE MAGNITUDE ENVELOPE SINUSOIDAL MODELING (ITERATIVE MESM)

Let $\sigma^i(n)$ be a magnitude envelope, i.e. an intra-frame magnitude modulation curve of $s[n]$, centered over i-th frame mid-point. Its purpose is to provide a concise representation of energy variations within an analysis frame and to reduce the effects of these variations on parameter estimation [8]. The magnitude envelope may be estimated by low-pass filtering of the input signal [8] or by calculating moving weighted average of its magnitude followed by down-sampling [4]. Typical magnitude envelopes and their spectral transforms are displayed on Figure 1.

The windowed portion of speech centered over i-th frame center, is well approximated by:

$$s_w(n) \approx \sigma^i(n) x(n) w(n) \approx \sum_{k=1}^{L} w(n) \sigma^i(n) A_k \cos(\theta_k n + \varphi_k), \quad (11)$$

where *x(n)* is a stationary signal, that can be represented by SSM. Then the minimization task is defined by:

$$\arg\min_{\{A_k, \theta_k, \varphi_k\}_k} \left\| w(n)\left(s(n) - \sum_{k=1}^{L} \sigma^i(n) A_k \cos(\theta_k n + \varphi_k)\right) \right\|^2, \quad (12)$$

The solutions of (12), proposed in prior art works, are iterative [4][8]. In each iteration, a new sinusoidal term is estimated, and then the model residual is formed. The parameters for each sinusoid are optimized to minimize a measure of the residual error energy. Thus, the recursion for the error residual after m iterations, is given by:

$$r_0(n) = s_w(n),$$
$$r_m(n) = r_{m-1}(n) - w(n)\sigma^i(n)A_m\cos(\theta_m n + \varphi_m)) = \\ = s_w(n) - \sum_{k=1}^{m}\sigma^i(n)w(n)A_k\cos(\theta_k n + \varphi_k)) \quad (13)$$

And the *m*-th iteration minimization task is defined as:

$$\min_{\theta_m}\left(\min_{A_m,\varphi_m}\sum_n\left(r_{m-1}(n) - w(n)\sigma^i(n)A_m\cos(\theta_m n + \varphi_m)\right)^2\right). \quad (14)$$

Assuming the frequency, θ_m is given, the *m*-th spectral line is derived analytically [8]:

$$\begin{bmatrix}\mathbf{t}_1^T(m)\mathbf{t}_1(m) & \mathbf{t}_1^T(m)\mathbf{t}_2(m)\\ \mathbf{t}_1^T(m)\mathbf{t}_2(m) & \mathbf{t}_2^T(m)\mathbf{t}_2(m)\end{bmatrix}\begin{bmatrix}c_{\mathrm{Re},m}\\ c_{\mathrm{Im},m}\end{bmatrix} = \begin{bmatrix}\mathbf{t}_1^T(m)\mathbf{r}_{m-1}\\ \mathbf{t}_2^T(m)\mathbf{r}_{m-1}\end{bmatrix}, \quad (15)$$

where $\mathbf{t}_1(m)$ and $\mathbf{t}_2(m)$ are defined in (9) and $\mathbf{r}_{m-1} \triangleq \{r_{m-1}(n)\}_{n=-N}^{N}$. The frequency minimization step in (14) is performed by explicit search over a redundant codebook of M frequencies ($M \gg L$), so the equation set (15) is solved about *M* times in each iteration. However, it is desirable both in terms of performance and optimality to reduce the frequency search codebook in each iteration to prevent selection of frequencies in close vicinity to each other [8][4]. In matching pursuits system [4] it is proposed to select just a single frequency per frequency bin, similar to what comes out of the spectral peak picking algorithm [1].

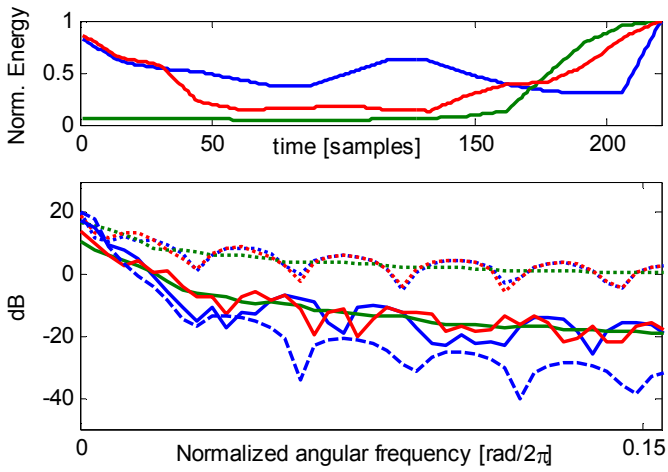

Figure 1: Typical magnitude envelopes, and their spectral transforms. Hanning window spectral transform denoted by dashed line, magnitude envelope transforms denoted by dotted lines, and corresponding *envelope window* (i.e. a product of a magnitude envelope and Hanning window) transforms denoted by solid lines.

4. JOINT MAGNITUDE ENVELOPE SINUSOIDAL MODELING (JOINT MESM)

Given a magnitude envelope $\sigma^i(n)$, a windowed speech approximation can be defined using a non-symmetric analysis window, varying from frame to frame. Indeed, defining $w_{env}^i(n) \triangleq w(n)\sigma^i(n)$, referred to as *envelope window*, we can rewrite (11) as:

$$s_w(n) \approx \sum_{k=1}^{L} w_{env}^i A_k\cos(\theta_k n + \varphi_k), \quad (16)$$

Typical spectral transforms of the *envelope windows* are displayed at Figure 1.

We can use the SSM formulations (7), (8) and (9), substituting the constant symmetric window $w(n)$ by the varying *envelope window* $w_{env}^i(n)$. In that case, the substitution of (8) to (3) for time-domain minimization, brings us to (we omitted the frame index *i* for simplicity):

$$\begin{bmatrix}\mathbf{T}_1^T\mathbf{T}_1 & \mathbf{T}_2^T\mathbf{T}_1\\ \mathbf{T}_1^T\mathbf{T}_2 & \mathbf{T}_2^T\mathbf{T}_2\end{bmatrix}\begin{bmatrix}\mathbf{c}_{\mathrm{Re}}\\ \mathbf{c}_{\mathrm{Im}}\end{bmatrix} = \begin{bmatrix}\mathbf{T}_1^T\mathbf{s}_w\\ \mathbf{T}_2^T\mathbf{s}_w\end{bmatrix}. \quad (17)$$

The line spectrum frequency set is defined by the peak picking technique similar to the SSM [2], prior to solution of (17).

Figure 2 displays a typical reconstructed transient signals for SSM, iterative MESM, and joint MESM, compared to the corresponding PCM waveform. One can notice that a strong pre-echo artifact, which is present in the SSM reconstruction, is reduced by I-MESM and further reduced by J-MESM.

For practical considerations, there is no need to apply the magnitude envelope for quasi-stationary speech frames. So, in this work we chose to apply MESM only to those frames having their *magnitude fluctuation R* (18) above a predetermined threshold.

$$R \triangleq \max(\sigma^i(n)) \Big/ \min(\sigma^i(n)) > R_{th} \quad (18)$$

For the rest of the speech the magnitude envelope is discarded, and the stationary solutions ((6) or (10)) are applied.

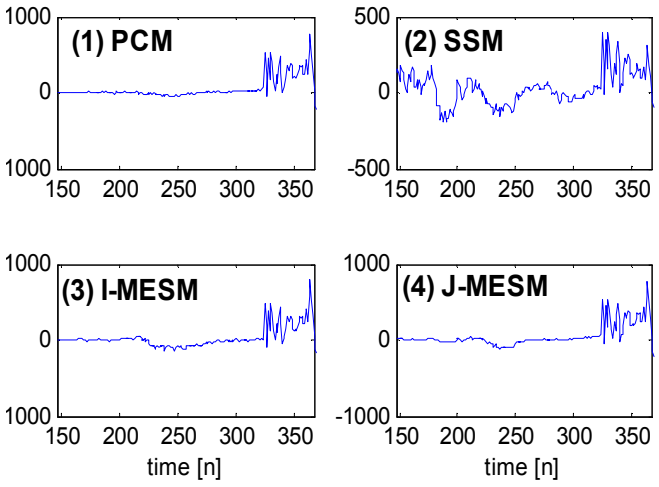

Figure 2: Typical reconstructed waveforms for SSM, I-MESM, and J-MESM.

5. EXPERIMENTAL RESULTS

Various sinusoidal models described here were applied upon a set of 100 US English male wideband sentences (sampling rate of 22050 *Hz*). The frame update rate was selected to be 200 *Hz* with 10ms synthesis window (Hanning). Both voiced and unvoiced were analyzed with the SSM [2], the iterative MESM (I-MESM) [4] with frequency codebook length M = 1024, and the joint MESM (J-MESM). For each one of the systems, the SNR_{seg}, as defined in (19), was estimated for transient frames solely. The transient frames were chosen to be those with the *magnitude fluctuation* $R > 4$. This resulted in 13% of analyzed data (above 6000 frames).

$$SNR_{seg} \triangleq \frac{1}{D}\sum_{i=1}^{D} 10\log_{10}\frac{\left(\mathbf{s}_w^i\right)^T \mathbf{s}_w^i}{\left(\mathbf{s}_w^i - \hat{\mathbf{s}}_w^i\right)^T \left(\mathbf{s}_w^i - \hat{\mathbf{s}}_w^i\right)} \qquad (19)$$

The cross-system model fit comparison results are displayed in Table 1. One can notice that the joint MESM yields above 10dB of improvement for unvoiced frames and about 7 dB of improvement for voiced frames over the stationary system (SSM), compared to just about 2-3 dB improvement over the SSM for the iterative MESM. As a whole, the model fit of joint MESM is better than the SSM by about 10dB, and better than the iterative MESM by about 8 dB.

One may notice, however, that the model fit for voiced is lower than unvoiced. This is due to the fact that there are a lot of non-harmonic components in voiced transients, which are not appropriately modeled. One might want to treat all the transients as unvoiced; this is a feasible approach for pure reconstruction (no voice transformation is expected). Alternatively, the model can be extended, similar to proposed in [4] or [3].

	SSM	I-MESM	J-MESM
Unvoiced	13.78	15.46	24.51
Voiced	5.78	8.63	12.43
All	12.17	14.08	22.07

Table 1: The segmental SNR values for various sinusoidal model systems. Measured for voiced "transient" frames, unvoiced "transient" frames, and for all the "transients".

6. SUMMARY

In this paper we presented an optimal analytic solution for estimation of sinusoidal parameters in presence of magnitude envelope, applied for speech transient modeling. The proposed solution significantly outperforms the known in the art iterative solution in terms of model fit, yielding 8dB improvement in segmental SNR for non-stationary speech.

7. AKNOWLEDGEMENTS

The author would like to thank Nuance Company for supporting this research, and acknowledge Mr. Alex Sorin for fruitful discussions related to this work.

12. REFERENCES

[1] McAulay, R.J., Quatieri, T.F., Sinusoidal coding. In: Kleijn, W.B., Paliwal, K.K. (Eds.), Speech Coding and Synthesis. Elsevier, Amsterdam, pp. 121–173., 1995.

[2] D. Chazan, R. Hoory, A. Sagi, S. Shechtman, A. Sorin, Z. Shuang and R. Bakis, "High quality sinusoidal modeling of wideband speech for the purpose of speech synthesis and modification", ICASSP 2006, Toulouse, May 2006

[3] Stylianou, Y., "Applying the harmonic plus noise model in concatenative speech synthesis", IEEE Trans. Speech and Audio Processing, vol. 9, no. 1, pp. 21-29, Jan. 2001.

[4] C. Ö. Etemoglu and V. Cuperman, "Matching pursuits sinusoidal speech coding," IEEE Trans. Speech Audio Process., vol. 11, no. 5, pp. 413–424, Sep. 2003.

[5] G. Kafentzis, O. Rosec, Y. Stylianou, "On the Modeling of Voiceless Stop Sounds of Speech using Adaptive Quasi-Harmonic Models", Interspeech 2012, Portland, Sept. 2012

[6] S. Levine, "Audio representations for data compression and compressed domain processing," Ph.D. dissertation, Stanford University, 1999.

[7] A. Spanias, "Speech Coding: A tutorial review," Proceeding of he IEEE, vol. 82, pp. 1541–1582, Oct 1994.

[8] E. B. George and M. J. T. Smith, "Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model," IEEE Trans. Speech Audio Processing, vol. 5, pp. 389–406, Sept. 1997.