EXPONENTIAL SINUSOIDAL MODELING OF TRANSITIONAL SPEECH SEGMENTS

Jesper Jensen, Søren Holdt Jensen, and Egon Hansen

Center for PersonKommunikation (CPK), Aalborg University, Denmark
E-mail: {jje,shj,eh}@cpk.auc.dk

ABSTRACT

A generalized sinusoidal model for speech signal processing is studied. The main feature of the model is that the amplitude of each sinusoidal component is allowed to vary exponentially with time. We propose to use the model in transitional speech segments such as speech onsets and voiced/unvoiced transitions. Computer simulations with natural speech signals indicate substantial better modeling performance in both transitional and voiced regions compared with the traditional constant-amplitude sinusoidal model.

1. INTRODUCTION

In recent years sinusoidal models for time domain speech signals have gained considerable interest in various applications including low bit-rate speech coding [6] and time-scale/pitch modifications of speech [7][3].

The basic sinusoidal model aims at representing a signal segment as a sum of constant-amplitude, constant-frequency sinusoids, i.e.,

$$\tilde{x}_n = \sum_{k=1}^{\tilde{K}} \tilde{a}_k \cos(\tilde{\omega}_k n + \tilde{\phi}_k), n = 0, \ldots, N-1, \quad (1)$$

where N denotes the segment length in samples, $\tilde{K}$ is the number of sinusoidal components, and $\tilde{a}_k$, $\tilde{\omega}_k$, and $\tilde{\phi}_k$ denote the amplitude, the angular frequency, and the initial phase, respectively, of the k'th sinusoidal component.

Due to the quasi-periodic nature of voiced speech sounds the sinusoidal model is particularly appropriate when modeling voiced speech segments. Moreover, it has been argued that the sinusoidal model is also valid for noise-like signals, such as some types of unvoiced speech sounds, provided that the frequencies of the sinusoidal components are spaced no more than approximately 100 Hz apart, see [6].

However, in speech signal segments where the stationarity assumption (constant-amplitude, constant-frequency) of the model is far from valid, the basic sinusoidal model in (1) may not be effective. Typically, such segments occur in transitional regions, e.g. at boundaries between unvoiced and voiced speech or at speech onsets. In this paper we study the modeling performance of the exponential sinusoidal model, a generalized version of the basic sinusoidal model. In particular, we aim at improving the modeling performance in transitional speech segments.

The outline of this paper is as follows. In Section 2 we introduce the exponential sinusoidal model. In Section 3 we present algorithms for robust estimation of the corresponding model parameters. In Section 4 we evaluate the proposed signal model by means of computer simulations with natural speech signals. Finally, conclusions and directions for future work are given in Section 5.

2. THE EXPONENTIAL SINUSOIDAL MODEL

Transitional speech segments are often characterized by relatively fast variations in amplitude, which cannot be modeled effectively by the basic sinusoidal model. To obtain better modeling performance in such segments, we generalize the basic sinusoidal model by allowing the amplitude of each sinusoidal component to vary exponentially with time within a signal segment. To be specific, the signal model we address is given by

$$\hat{x}_n = \sum_{k=1}^{K} a_k e^{-d_k n} \cos(\omega_k n + \phi_k), \quad (2)$$

for $n = 0, \ldots, N-1$. That is, the modeled signal segment $\hat{\mathbf{x}} = [\hat{x}_0\ \hat{x}_1\ \cdots\ \hat{x}_{N-1}]^T$ consists of a sum of K sinusoidal components, where a_k denotes the initial amplitude, d_k is the damping factor, w_k is the angular frequency, and ϕ_k is the initial phase of the k'th sinusoidal component. In (2) all sinusoidal parameters are scalar-valued. We refer to (2) as the exponential sinusoidal model.

Note that we do not restrict d_k to be positive, i.e., the amplitude of each component may be growing with time. This is, e.g. , suitable for modeling of speech onsets. Moreover, with the special case $d_k = 0$ for $k = 1, \ldots, K$, all sinusoidal components are undamped, and the exponential signal model reduces to the basic sinusoidal model (1).

3. PARAMETER ESTIMATION

Assume that a signal segment $\mathbf{x} = [x_0\ x_1\ \cdots\ x_{N-1}]^T$ is to be approximated by the model in (2). We use a two-step procedure in order to estimate the model parameters of (2). Initial estimates are obtained by applying a variant of Kung's state-space method [4]. Subsequently, the initial estimates are refined iteratively by solving a so-called structured total least norm (STLN) problem [8]. In the following we present the two steps of the parameter estimation procedure.

3.1. Initial estimates

In order to determine initial estimates of the exponential sinusoidal parameters a_k, d_k, ω_k, and ϕ_k we rewrite (2) in its complex form:

$$\hat{x}_n = \sum_{k=1}^{2K} a_k e^{j\phi_k} e^{(-d_k + j\omega_k)n} = \sum_{k=1}^{2K} c_k z_k^n, \qquad (3)$$

where $c_k = a_k e^{j\phi_k}$ is the complex-valued k'th amplitude and $z_k = e^{(-d_k+j\omega_k)}$ is the k'th signal pole.

Initially, we assume that the observed signal segment $\mathbf{x} = [x_0\ x_1\ \cdots\ x_{N-1}]^T$ can be modeled exactly by (3), i.e., $\mathbf{x}$ actually is a sum of $2K$ unknown damped complex-valued exponentials. Moreover, we assume that the number of complex exponentials $2K$ is known and that all signal poles are distinct.

We arrange the observed signal segment $\mathbf{x}$ in a Hankel-structured data matrix $\mathbf{X} \in \mathbb{R}^{L\times M}$:

$$\mathbf{X} = \begin{bmatrix} x_0 & x_1 & \cdots & x_{M-1} \\ x_1 & x_2 & & x_M \\ \vdots & & & \vdots \\ x_{L-1} & x_L & \cdots & x_{N-1} \end{bmatrix}$$

Using that $\mathbf{x}$, by assumption, can be written in the form of (3), it can easily be verified that the Vandermonde decomposition of $\mathbf{X}$ is given by:

$$\mathbf{X} \overset{\text{VD}}{=} \mathbf{S}\mathbf{C}\mathbf{T}^T,$$

where

$$\mathbf{S} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ z_1^1 & z_2^1 & \cdots & z_{2K}^1 \\ \vdots & \vdots & & \vdots \\ z_1^{L-1} & z_2^{L-1} & \cdots & z_{2K}^{L-1} \end{bmatrix}, \mathbf{T} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ z_1^1 & z_2^1 & \cdots & z_{2K}^1 \\ \vdots & \vdots & & \vdots \\ z_1^{M-1} & z_2^{M-1} & \cdots & z_{2K}^{M-1} \end{bmatrix},$$

$\mathbf{C} = diag(c_1, c_2, \ldots, c_{2K})$, and T denotes matrix transposition. Since all signal poles are assumed distinct, rank($\mathbf{S}$)= rank($\mathbf{T}$)=rank($\mathbf{X}$)=$2K$. The matrix $\mathbf{S}$ has the following shift-invariant property:

$$\mathbf{S}_{\downarrow}\mathbf{Z} = \mathbf{S}^{\uparrow}, \qquad (4)$$

where $\mathbf{Z} = diag(z_1, \ldots, z_{2K})$ contains the signal poles and $\mathbf{S}_{\downarrow}$ ($\mathbf{S}^{\uparrow}$) is the matrix $\mathbf{S}$ with the bottom (top) row deleted.

The matrix $\mathbf{S}$ cannot be estimated directly from the data matrix $\mathbf{X}$. Consider instead the truncated singular value decomposition (TSVD) of $\mathbf{X}$:

$$\mathbf{X} \overset{\text{SVD}}{=} \begin{bmatrix}\mathbf{U}_1 & \mathbf{U}_2\end{bmatrix} \begin{bmatrix}\mathbf{\Sigma}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{bmatrix} \begin{bmatrix}\mathbf{V}_1^H \\ \mathbf{V}_2^H\end{bmatrix} \overset{\text{TSVD}}{=} \mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^H,$$

where H denotes Hermitian transposition. The columns of $\mathbf{U}_1 \in \mathbb{C}^{L\times 2K}$ and $\mathbf{V}_1 \in \mathbb{C}^{M\times 2K}$ are the first $2K$ left and right singular vectors, respectively, while $\mathbf{\Sigma}_1 \in \mathbb{R}^{2K\times 2K}$ is a diagonal matrix with the $2K$ non-zero singular values of $\mathbf{X}$ in descending order.

The columns of $\mathbf{U}_1$ constitute an orthonormal basis for the column space of $\mathbf{X}$. Using that the $2K$ columns of $\mathbf{S}$ is another basis for the column-space of $\mathbf{X}$, it follows:

$$\mathbf{S} = \mathbf{U}_1\mathbf{F}, \qquad (5)$$

where $\mathbf{F} \in \mathbb{C}^{2K\times 2K}$ is an invertible change-of-basis matrix. By substituting (5) into (4) we get

$$\mathbf{U}_{1\downarrow}\mathbf{Z}^{(u)} = \mathbf{U}_1^{\uparrow}, \qquad (6)$$

where $\mathbf{Z}^{(u)} = \mathbf{F}\mathbf{Z}\mathbf{F}^{-1}$ is a similarity transform of the diagonal signal pole matrix $\mathbf{Z}$, i.e., the eigenvalues of $\mathbf{Z}^{(u)}$ are the signal poles $z_i, i = 1, \ldots, 2K$. Equation (6) is valid, when the observed signal segment actually is a sum of damped complex exponentials. For segments of real speech this is, generally, not the case, and (6) gives rise to an overdetermined system of equations, which we solve for $\mathbf{Z}^{(u)}$ in a total least squares sense. The initial estimates $d_k^{(1)}$ and $\omega_k^{(1)}$ of the damping factors and the frequencies of the exponential sinusoidal model in (3) are found from the magnitude and angle, respectively, of the eigenvalues of $\mathbf{Z}^{(u)}$.

The complex amplitudes c_k in (3) are determined by inserting the estimated values $d_k^{(1)}$ and $\omega_k^{(1)}$ in (3). This gives rise to an overdetermined system of N equations, which is linear in the $2K$ unknowns $c_k, k = 1, \ldots, 2K$. This system of equations is solved in least squares sense. Subsequently, initial estimates $a_k^{(1)}$ and $\phi_k^{(1)}$ of the amplitudes and phases are determined from the magnitude and angles of the estimated values of c_k.

3.2. STLN refinement

The approximation of $\mathbf{x}$ based on the initial parameter estimates is given by:

$$\hat{x}_n^{(1)} = \sum_{k=1}^{2K} a_k^{(1)} e^{j\phi_k^{(1)}} e^{(-d_k^{(1)} + j\omega_k^{(1)})n}$$

This signal segment may not be optimal, i.e., there may exist another sum of $2K$ damped complex exponentials, which is 'closer' to the original signal segment $\mathbf{x}$.

In general, a modeled segment $\hat{\mathbf{x}}$ may be written as

$$\hat{\mathbf{x}} = \mathbf{x} + \Delta\mathbf{x}, \tag{7}$$

where $\Delta\mathbf{x} = [\Delta x_0 \ \cdots \ \Delta x_{N-1}]^T$ is a perturbation vector. In order to have the best approximation of $\mathbf{x}$, we aim at minimizing $\|\Delta\mathbf{x}\|_2^2$, while keeping $\hat{\mathbf{x}}$ a sum of $2K$ damped complex exponentials. Here, $\|\cdot\|_2$ is the vector 2-norm

We exploit the fact, see e.g. [2, 1], that if a signal segment $\mathbf{s} = [s_0 \ s_1 \ \cdots \ s_{N-1}]^T$ is arranged in an augmented Hankel data matrix $[\mathbf{A}\ \mathbf{b}]$, where $\mathbf{A} \in \mathbb{R}^{J\times 2K}$, $\mathbf{b} \in \mathbb{R}^J$, $J > 2K, N = J + 2K$:

$$[\mathbf{A}\ \mathbf{b}] = \begin{bmatrix} s_0 & s_1 & \cdots & s_{2K-1} & s_{2K} \\ s_1 & & & s_{2K} & s_{2K+1} \\ \vdots & & & \vdots & \vdots \\ s_{J-1} & s_J & \cdots & s_{N-2} & s_{N-1} \end{bmatrix} \tag{8}$$

and $[\mathbf{A}\ \mathbf{b}]$ is rank-deficient, then the signal segment $\mathbf{s}$ must consist of a sum of $2K$ damped complex exponentials. On the other hand, if a signal segment consisting of $2K$ damped complex exponentials is arranged as in (8), then the augmented matrix $[\mathbf{A}\ \mathbf{b}]$ is rank-deficient.

Now assume that the samples of $\hat{\mathbf{x}}$ in (7) are arranged in a Hankel matrix $[\hat{\mathbf{A}}\ \hat{\mathbf{b}}]$ structured as in (8). Using (7), the matrix $[\hat{\mathbf{A}}\ \hat{\mathbf{b}}]$ containing the modeled segment may be written as:

$$[\hat{\mathbf{A}}\ \hat{\mathbf{b}}] = [\mathbf{A}\ \mathbf{b}] + [\Delta\mathbf{A}\ \Delta\mathbf{b}] = [\mathbf{A} + \Delta\mathbf{A}\ \mathbf{b} + \Delta\mathbf{b}],$$

where $[\mathbf{A}\ \mathbf{b}]$ and $[\Delta\mathbf{A}\ \Delta\mathbf{b}]$ are matrices of the type of (8) containing elements of the original segment $\mathbf{x}$ and the perturbation vector $\Delta\mathbf{x}$, respectively.

Now the problem of finding the sum of $2K$ damped complex exponentials, which is closest to $\mathbf{x}$, can be formulated as the following constrained minimization problem:

$$\min_{\Delta\mathbf{x}} \|\Delta\mathbf{x}\|_2^2 \text{ such that } \begin{cases} [\mathbf{A} + \Delta\mathbf{A}\ \mathbf{b} + \Delta\mathbf{b}] \\ \text{is rank-deficient} \\ [\mathbf{A}\ \ \mathbf{b}] \text{ and } [\Delta\mathbf{A}\ \Delta\mathbf{b}] \\ \text{have Hankel structure} \end{cases}$$

This problem may be interpreted and solved as a structured total least norm problem [8]. In order to solve the problem we use the iterative algorithm called STLNB in [8]. Initial values of $\Delta\mathbf{A}$ and $\Delta\mathbf{b}$ are found by arranging the elements of the initial perturbation vector $\Delta\mathbf{x}^{(1)} = \hat{\mathbf{x}}^{(1)} - \mathbf{x}$ in the Hankel matrix in (8). The iterations of STLNB are stopped, when the lowest singular value of $[\mathbf{A} + \Delta\mathbf{A}\ \mathbf{b} + \Delta\mathbf{b}]$ has reached its minimum and has stayed constant for more than 10 iterations. Then, the improved modeled segment is constructed from the elements of the first column and the last row of $[\mathbf{A} + \Delta\mathbf{A}\ \mathbf{b} + \Delta\mathbf{b}]$. The model parameters for this segment, which can be described exactly as a sum of $2K$ damped complex exponentials, are estimated by means of Kung's algorithm described in section 3.1.

4. SIMULATION RESULTS

We evaluate the potential of the exponential sinusoidal model (ESM) in (2) by comparing it with the basic sinusoidal model (BSM) in (1) and with an improved version of BSM (IBSM), which uses optimized BSM parameters as explained below. The models are evaluated with real speech signals sampled at a rate of 8 kHz and segmented into frames of 200 samples with an overlap of 40 samples between consecutive frames. The three signal models are used on each frame. Modeled signals are generated by overlap-adding the modeled frames.

The BSM parameters are derived from a peak-picking procedure applied to the magnitude of the short-time Fourier transform (STFT) of the Hamming weighted signal frame in question, see [6] for details. Peaks more than 60 dB below the largest peak are not considered, and the number of peaks is limited to 30. In this paper the STFT is evaluated by means of a 4096-point fast Fourier transform (FFT).

The IBSM parameters are obtained by minimizing:

$$V = \sum_{n=0}^{N-1} \left(x_n - \tilde{x}_n(\tilde{a}_k, \tilde{\omega}_k, \tilde{\phi}_k)\right)^2,$$

with respect to the sinusoidal parameters $\tilde{a}_k, \tilde{\omega}_k, \tilde{\phi}_k$. Here, $\tilde{x}_n$ is given by (1) and x_n denotes samples from the original signal frame. This minimization problem is a non-linear least squares problem. We solve it by using the BSM parameter estimates as initial values in the Levenberg-Marquardt algorithm, which is a hybrid Gauss-Newton/Steepest Descent iterative algorithm, see e.g. [5].

For each signal frame modeled with ESM, the same number of sinusoids was used as with BSM and IBSM.

In order to illustrate the performance of ESM, BSM and IBSM, the SNR defined below is calculated for each signal frame in a phrase:

$$\text{SNR} = 10\log_{10}\left(\frac{\sum_{n=0}^{N-1} x_n^2}{\sum_{n=0}^{N-1} e_n^2}\right),$$

where e_n denotes the modeling error at sample n. The result of this is shown in Figure 1. From this figure it is obvious that BSM and IBSM perform quite good in steady-voiced frames, but the performance decreases in transitional segments. Much better performance is obtained with ESM, especially in transitional regions, but also for voiced frames.

In Figure 2 is given an example of the modeling performance in a transitional signal frame. The performance with ESM is substantially better than with BSM and IBSM. In order to have a clearer indication of the modeling performance of ESM in transitional frames, we modeled 324 different transitional frames with BSM, IBSM and ESM. The average SNR for the three models is shown in Table 1. Clearly, ESM outperforms BSM and IBSM. Although

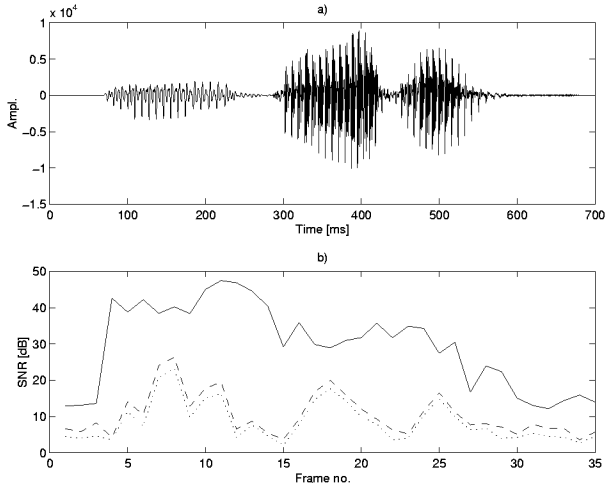

Figure 1: a) Original phrase, b) Modeling performance of BSM (dotted), IBSM (dashed) and ESM (solid).

	BSM	IBSM	ESM
avg. SNR [dB]	2.1	5.4	31.0

Table 1: Average SNR in transitional speech frames.

IBSM uses optimized parameter values, it performs poorly in transitional segments. This indicates that the basic sinusoidal model is not suitable in transitional segments.

In the experiments reported here, we used $K = \tilde{K}$, which means 4/3 more parameters pr. frame with ESM compared to BSM and IBSM. We emphasize, that if we used the same number of parameters pr. frame, the SNR would still be much higher for ESM compared with BSM and IBSM.

Informal listening tests confirm the results indicated by the SNR-values of Table 1. Signals modeled with BSM often have a "buzzy" sound in transitional regions, while ESM signals are almost indistinguishable from the originals.

5. CONCLUSIONS

In this paper we have presented a generalized sinusoidal speech signal model called the exponential sinusoidal model. The main feature of this model is that the amplitudes of the sinusoidal components are allowed to vary exponentially with time. This results in a considerable objective and subjective improvement, especially in transitional segments, compared to the basic sinusoidal model.

One drawback of the approach presented in this paper is the computationally expensive parameter estimation scheme. Alternatives to this scheme is a topic of current research.

6. REFERENCES

[1] R. De Beer, D. Van Ormondt, and F.T.A.W. Wajer. SVD-based modelling of medical NMR signals. In M. Moonen and B. De Moor, editors, *SVD and Signal Processing, III, Algorithms, Architectures and Applications*, pages 467–474. Elsevier Science B.V., 1995.

[2] I. Dologlou and C. Carayannis. LPC/SVD analysis of signals with zero modelling error. *Signal Processing*, 23(3):293–298, 1991.

[3] 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*, 5(5):389–406, 1997.

[4] S. Y. Kung, K. S. Arun, and D. V. Bhaskar Rao. State-space and singular-value decomposition-based approximation methods for the harmonic retrieval problem. *J. Amer. Opt. Soc.*, 73(12):1799–1811, 1983.

[5] S. A. Lill. A survey of methods for minimizing sums of squares of nonlinear functions. In L. C. W. Dixon, editor, *Optimization in action*, chapter 1. Academic Press, 1976.

[6] R. J. McAulay and T. F. Quatieri. Sinusoidal coding. In W. B. Kleijn and K. K. Paliwal, editors, *Speech Coding and Syntesis*, chapter 4. Elsevier Science B. V., 1995.

[7] T. F. Quatieri and R. J. McAulay. Shape invariant time-scale and pitch modification of speech. *IEEE Trans. Signal Processing*, 40(3):497–510, 1992.

[8] S. Van Huffel, H. Park, and J. B. Rosen. Formulation and solution of structured total least norm problems for parameter estimation. *IEEE Trans. Signal Processing*, 44(10):2464–2474, 1996.

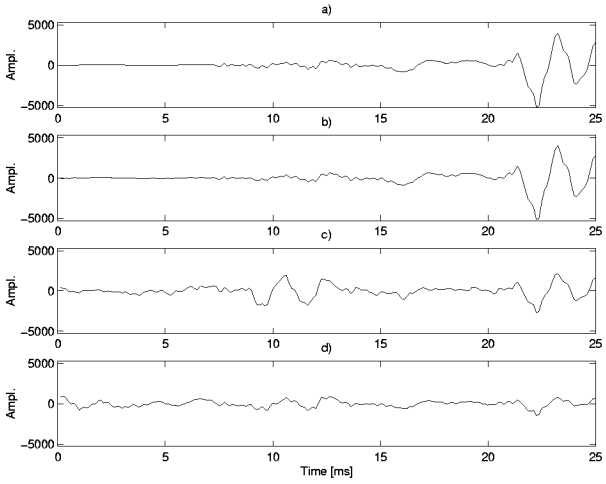

Figure 2: Modeling performance in speech onset (30 sinusoids). a) Original frame, b) ESM (SNR=29.2 dB), c) IBSM (SNR=3.7 dB), d) BSM (SNR=2.0 dB).