A COMPARISON OF SINUSOIDAL MODEL VARIANTS FOR SPEECH AND AUDIO REPRESENTATION

Jesper Jensen and Richard Heusdens
Dept. of Mediamatics, Technical University of Delft,
Delft–The Netherlands
E-mail: {J.Jensen, R.Heusdens}@ITS.TUDelft.NL

ABSTRACT

Two sinusoidal model variants for speech and audio representation are compared: the traditional constant-amplitude, constant-frequency sinusoidal model, and a generalized model where amplitudes can vary exponentially with time. Two classes of methods for estimation of model parameters are reviewed: matching pursuit (MP) and subspace based schemes. Furthermore, Newton optimized versions of these schemes are included in the study. The influence of model type and parameter estimation scheme on model performance was evaluated in simulation experiments with audio and speech signals. As expected, the exponential model outperforms the traditional sinusoidal model in segments with large signal level variations. For the non-optimized estimation schemes, the subspace method generally performs better than the MP method (an SNR gain of 2-7 dB was observed). Newton optimization improves the modeling performance significantly in all cases, and results in slightly better performance with MP (an SNR gain of 1-2 dB) compared to the subspace method.

1. INTRODUCTION

Sinusoidal models provide accurate and flexible parametric representations of many types of acoustical signals including speech and audio [9, 13]. In speech and audio processing, sinusoidal modeling has been used in a wide range of applications, e.g. signal synthesis [14, 13], coding [10, 1], and enhancement [12, 5].

Typically, sinusoidal modeling relies on the assumption that segments s_m of an original signal can be represented well as a sum of constant-amplitude, constant-frequency (CACF) sinusoids:

$$\hat{s}_m = \sum_{k=1}^{K_m} \hat{a}_{k,m} \cos(\hat{\omega}_{k,m} n + \hat{\phi}_{k,m}), \qquad (1)$$

for $n = 0, \ldots, N_m - 1$, where $\hat{s}_m$ is the model representation of the m'th signal segment, K_m is the model order, $\hat{a}_{k,m}$, $\hat{\omega}_{k,m}$, and $\hat{\phi}_{k,m}$ are the amplitude, frequency, and initial phase, respectively, of the k'th sinusoidal component, and N_m is the number of samples in the m'th segment. For later reference, we call Eq. (1) the Basic Sinusoidal Model (BSM). While the CACF assumption of the BSM may be satisfied well for many signal segments, it is far from valid in segments with rapid amplitude level variations, e.g. speech onsets or attacks in musical signals; in such segments the model does not perform well.

To have better modeling performance for a broader class of signals, a generalized version of the BSM has been proposed [11, 7, 6]. This model, which we denote the Exponential Sinusoidal Model (ESM), allows the amplitudes of each sinusoidal component to vary exponentially with time:

$$\tilde{s}_m = \sum_{k=1}^{K_m} \tilde{a}_{k,m} \exp(-\tilde{d}_{k,m} n) \cos(\tilde{\omega}_{k,m} n + \tilde{\phi}_{k,m}), \qquad (2)$$

for $n = 0, \ldots, N_m - 1$, where the additional parameters $\tilde{d}_{k,m}$ are so-called damping factors. The damping factors are not restricted to be positive, i.e., some sinusoidal amplitudes may be growing with time. Moreover, in the special case where $\tilde{d}_{k,m} = 0$ for $k = 1, \ldots, K_m$, the ESM in Eq. (2) reduces to the BSM in Eq. (1).

This paper provides a comparative performance study of the BSM and ESM for speech and audio representation. The study has two main objectives. First, since model performance is highly dependent on the quality of parameter estimates, two schemes for extracting the model parameters are reviewed and evaluated. Secondly, to gain further insights into the performance characteristics of the models, the influence of the model order K_m and the signal segment length N_m is analyzed for different signal types.

2. ESTIMATION OF MODEL PARAMETERS

The problem of fitting a sum of (possibly exponentially damped) sinusoids to a signal segment arises in many engineering areas, and consequently, a wide range of algorithms exists for estimating the parameters of Eqs. (1) and (2). For speech and audio modeling, estimation algorithms can roughly be divided into three categories: 1) spectral peak-picking (e.g. [9]), 2) iterative analysis-by-synthesis (AbS) (e.g. [2]), and 3) subspace processing (e.g. [11]).

We consider analysis algorithms from categories 2) and 3), since these perform better than category 1) algorithms (although usually at a higher computational cost). For BSM parameter estimation we use a particular AbS method called *matching pursuit* (MP) [8]. Further, for ESM parameter estimation we review an MP-based method as well as a subspace-based method. In addition, optimized algorithm variants are included in the study, where parameter estimates are refined using Newton optimization.

2.1. BSM Parameters (Matching Pursuit)

Matching pursuit is an algorithm for approximating a signal by a finite expansion into elements (functions) chosen from a redundant dictionary [8]. Let $D = (g_\gamma)_{\gamma \in \Gamma}$ be a dictionary, that is, a set of functions indexed by $\gamma \in \Gamma$, where Γ is an index set. In the case of BSM parameter estimation, the dictionary is populated by CACF sinusoidal functions (windowed by a sequence w). The MP algorithm is a greedy iterative algorithm which searches the dictionary

The research is supported by Philips Research and the Technology Foundation STW, applied science division of NWO and the technology programme of the ministry of Economics Affairs

for the function that best matches the signal, and subtracts this function (properly scaled) to form a residual signal to be approximated in the next iteration. The iterations are continued until a number of K_m dictionary entries have been selected.

The BSM parameter estimation algorithm can be outlined as:

```
Algorithm 1:BSM-MP
Input: $s_m, K_m, w$
Output: $\hat{\omega}_{k,m}, \hat{a}_{k,m}, \hat{\phi}_{k,m}, \quad k = 1, \dots, K_m$
Initialize: $x := diag(s_m)w$;
for $k := 1 : K_m$
    1: Find the (scaled) dictionary element $g_{\gamma'} \in D$ closest
       to $x$, i.e., find $\gamma' = \arg\max_{\gamma \in \Gamma} |\langle x, g_\gamma \rangle|$, and save the
       corresponding parameters $(\hat{a}_{k,m}, \hat{\omega}_{k,m}, \hat{\phi}_{k,m})$.
    2: Update residual: $x := x - \alpha g_{\gamma'}$ with $\alpha = \langle x, g_{\gamma'} \rangle$.
end
```

Here, $diag(s_m)$ denotes a square matrix with the elements of the segment s_m on the main diagonal. In Step 1 we have used that finding the (scaled) dictionary element that minimizes the two-norm distance to x corresponds to choosing the element with largest magnitude correlation with the signal x (assuming that the dictionary elements have unit norm) [3]. In step 2, $\alpha = \langle x, g_{\gamma'} \rangle$ is a scaling factor obtained from step 1. If we further assume that the sinusoidal frequencies belong to a discrete set of L equispaced frequencies, the dictionary search is implemented efficiently using an L points FFT [3].

Alg. 1 provides an efficient way of obtaining the BSM parameters. However, since the algorithm is greedy, the parameters are generally sub-optimal, i.e., another sum of K_m undamped sinusoids may exist, that approximates the segment s_m better. In an attempt to find the optimal set of BSM sinusoids, we include Alg. 2 (outlined below), which is a refined version of Alg. 1. The key feature of Alg. 2 is that, at each iteration, the BSM parameters obtained so far are refined simultaneously using Newton optimization (step 2).

2.2. ESM Parameters (Matching Pursuit)

We consider two MP based algorithms for ESM parameter estimation: *Algorithm 3:ESM-MP* and *Algorithm 4:ESM-MP-OPT*. The structures of these algorithms are identical to those outlined for Algs. 1 and 2, respectively. Thus, in Alg. 3, the MP dictionary is populated by (windowed) exponentially damped sinusoids, and the sinusoidal components are extracted one at a time using the same iterative and greedy procedure as outlined for Alg. 1. In Alg. 4, the ESM parameters obtained so far at a given iteration are refined using Newton optimization. Then, the refined ESM parameters are used to synthesize an optimal modeled segment (windowed), which is subtracted from the windowed original segment x to form the residual used as input to the next iteration.

Similarly to Algs. 1 and 2, the dictionary search in Algs. 3 and 4 can be implemented efficiently using FFT techniques [3].

2.3. ESM Parameters (Shift Invariance)

Finally, we consider the subspace-based approach described in [15] for estimating the ESM parameters. This approach is based on the so-called shift invariance (SI) property which characterizes certain vector subspaces of Hankel (or Toeplitz) matrices constructed from linear combinations of complex exponentials. From the SI property, damping $\tilde{d}_{k,m}$ and frequency $\tilde{\omega}_{k,m}$ parameters are determined using a total least squares scheme. Given the damping and frequency estimates, the amplitude $\tilde{a}_{k,m}$ and phase $\tilde{\phi}_{k,m}$ parameters can be determined from the solution of linear least squares problem. An outline of the algorithm (Alg. 5) is given below. For an in-depth treatment of the algorithm, we refer to [15].

```
Algorithm 2:BSM-MP-OPT
Input: $s_m, K_m, w$
Output: $\hat{\omega}_{k,m}, \hat{a}_{k,m}, \hat{\phi}_{k,m}, \quad k = 1, \dots, K_m$
Initialize: $x := diag(s_m)w$; $r := x$;
for $k := 1 : K_m$
    1: Find the (scaled) dictionary element $g_{\gamma'} \in D$ closest to
       $r$, i.e., find $\gamma' = \arg\max_{\gamma \in \Gamma} |\langle r, g_\gamma \rangle|$, and collect the
       corresponding sinusoid parameters.
    2: Optimize the $k$ sinusoids found so far:
       $\{\hat{a}_i, \hat{\omega}_i, \hat{\phi}_i\}_{i=1,\dots,k} = \arg\min_{\{a_i, \omega_i, \phi_i\}_{i=1,\dots,k}} ||x - x^{(k)}||^2$,
       where $x^{(k)}(n) = w(n) \sum_{i=1}^{k} a_i \cos(\omega_i n + \phi_i)$.
    3: Synthesize optimized $k$-order model:
       $\hat{x}^{(k)}(n) = w(n) \sum_{i=1}^{k} \hat{a}_i \cos(\hat{\omega}_i n + \hat{\phi}_i)$.
    4: Update residual signal: $r := x - \hat{x}^{(k)}$;
end
$\hat{a}_{k,m} := \hat{a}_k$; $\hat{\omega}_{k,m} := \hat{\omega}_k$; $\hat{\phi}_{k,m} := \hat{\phi}_k$ for $k = 1, \dots, K_m$.
```

```
Algorithm 5:ESM-SI
Input: $s_m, K_m, w$
Output: $\tilde{a}_{k,m}, \tilde{d}_{k,m}, \tilde{\omega}_{k,m}, \tilde{\phi}_{k,m}, \quad k = 1, \dots, K_m$
Initialize: $x := diag(s_m)w$
1: Form Hankel matrix from $s_m$ and use SI property to
   estimate damping factors $\tilde{d}_{k,m}$ and frequencies $\tilde{\omega}_{k,m}$.
2: Given estimated $\tilde{d}_{k,m}$ and $\tilde{\omega}_{k,m}$, find amplitudes
   $\tilde{a}_{k,m}$ and phases $\tilde{\phi}_{k,m}$ to minimize $||x - W\tilde{s}_m(\cdot)||^2$
   with $\tilde{s}_m(\cdot)$ given by Eq. (2), and $W = diag(w)$.
```

As with the previous algorithms, we include a Newton optimized version (Alg. 6) in our study. This algorithm refines the parameter estimates of Alg. 5 by solving the non-linear problem:

$$\min_{\tilde{a}_{k,m}, \tilde{d}_{k,m}, \tilde{\omega}_{k,m}, \tilde{\phi}_{k,m}} ||x - W\tilde{s}_m(\cdot)||^2,$$

with $\tilde{s}_m(\cdot)$ given by Eq. (2), and $W = diag(w)$.

3. EXPERIMENTAL RESULTS

3.1. Performance vs. Model Order

The performance of the BSM and ESM in combination with Algs. 1–6 was studied as a function of the model order K_m. A number of speech segments s_m of length N_m=160 samples (20 ms at a sampling frequency of 8 kHz) were represented with the BSM and the ESM for model orders $K_m = 1, \dots, 30$. For the MP based parameter estimation algorithms, a Hanning window w was used to extract the signal segments, and the dictionary search was implemented using a 4096-points FFT.

To evaluate model performance, the SNR quality measure defined as

$$\text{SNR}(\hat{s}_m, s_m) = 10 \log_{10} \left(\frac{||W s_m||^2}{||W(s_m - \hat{s}_m)||^2} \right) \text{[dB]} \quad (3)$$

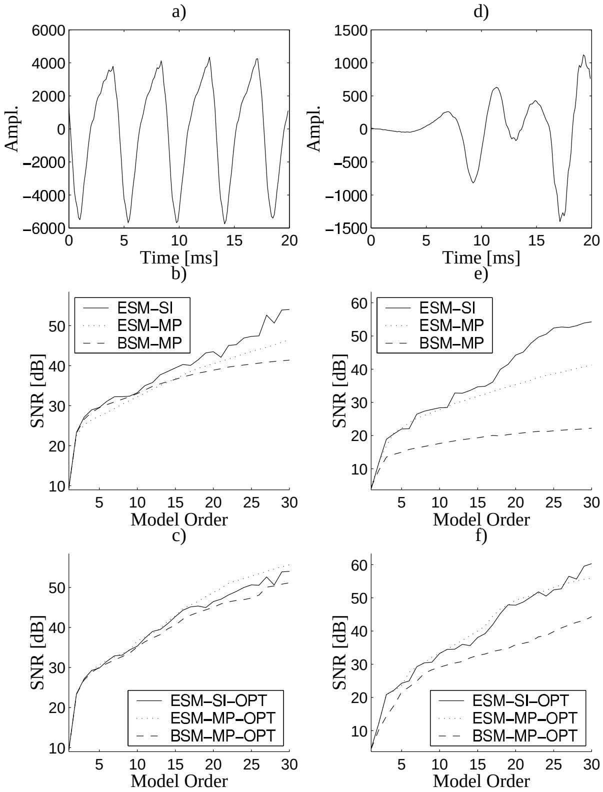

Figure 1: SNR vs. Model Order K_m. a) Voiced segment (female), b)-c) SNR vs. model order for various models and parameter estimation schemes, d) Voiced onset (male), e)-f) SNR vs. model order for various models and parameter estimation schemes.

was used for the BSM, where $W = diag(w)$. A similar measure with $\hat{s}_m$ substituted by $\tilde{s}_m$ was used for the ESM.

A large number of speech segments of different types (voiced, unvoiced, onsets, etc.) from different speakers were processed. Fig. 1 illustrates typical results of this study for a female fully voiced segment (Figs. 1a–c) and for a male voiced onset (Figs. 1d–f). For fully voiced segments, we only see a slight improvement with the ESM over the BSM, while in segments where the amplitude level varies more rapidly, the ESM performs much better. Further, for the ESM, the subspace-based estimation scheme *ESM-SI* performs better than the MP based scheme *ESM-MP*, particularly for large model orders, while for their Newton optimized counterparts, performance is nearly identical. As expected for stationary unvoiced segments (not shown in the figure), performance here is much lower than for voiced segments.

3.2. Performance vs. Frame Length

In order to study the modeling performance as a function of the segment length, four different speech signals with a duration of approximately 3 seconds each were represented with the BSM and the ESM using Algs. 1–6 for parameter estimation. Analysis segments were extracted using a Hanning window with an overlap of 50 % between consecutive segments. After parameter estimation, modeled segments were synthesized using Eqs. (1) and (2), and concatenated using a Hanning window based overlap-add (OLA) procedure. In order to quantify the modeling performance for several test sentences, we use the segmental SNR (SEG-SNR) defined as the average SNR of concecutive segments. The SEG-SNR was calculated using segment lengths of 240 samples (30 ms) taken with an overlap of 75 % after OLA.

Fig. 2 shows SEG-SNR as a function of the segment length for the various parameter estimation schemes. From Fig. 2a we see that for the non-optimized parameter estimation schemes, the subspace-based algorithm (*ESM-SI*) is better than the MP-based algorithm (*ESM-MP*), while the *BSM-MP* algorithm gives much lower performance. When the parameters are refined using Newton optimization, the performance gap between BSM and ESM reduces to approximately 3-5 dB, and *ESM-MP-OPT* performs slightly better (≈ 1-2 dB) than *ESM-SI-OPT* for all segment lengths.

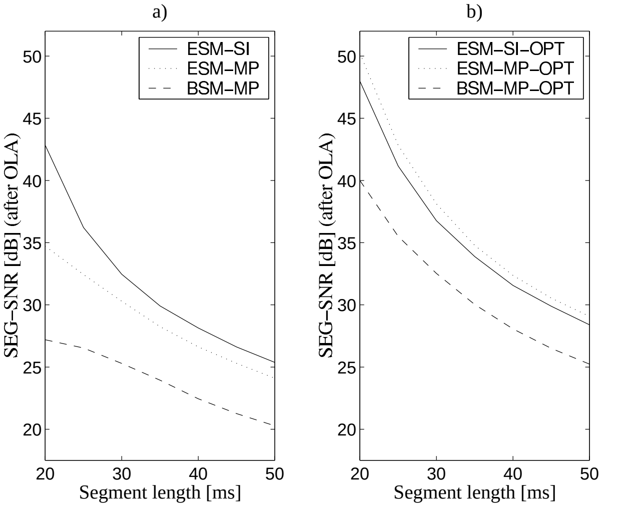

Figure 2: SEG-SNR vs. Segment Length N_m for speech signals (with $K_m = 30$). a) Non-optimized parameter estimation schemes, b) Newton optimized estimation schemes.

3.3. Performance vs. Time

Finally, to compare the performance of the BSM and ESM for audio signal representation, an audio signal was processed using Algs. 1–6 for parameter estimation. Modeled signals were generated using the same Hanning window based overlap-add procedure as described above, but with a segment length of N_m = 882 samples (20 ms at a sampling rate of 44.1 kHz). The performance of Algs. 1–6 is shown in Fig. 3. The figure suggests that the same performance ordering of the algorithms is valid for audio signals as for speech signals, although the performance difference between BSM and ESM is smaller here.

4. CONCLUDING REMARKS

The comparison of (non-optimized) algorithms for sinusoidal parameter estimation showed the subspace based approach to perform better than the MP based method. However, the MP method appears to be more flexible in a number of ways. First, the MP method is capable of extracting windowed (tapered) sinusoidal components, which is of importance when OLA based synthesis is used. This seems difficult with the subspace based approach, since tapered sinusoids do not satisfy the SI property on which the subspace method relies. Secondly, the MP method can be modified to use more perceptually relevant distortion measures [4] than the least square criterion used in this paper; this may not be easy with the subspace method. Finally, while the MP method can be implemented using FFT techniques, the subspace method requires estimation of a set of dominant singular vectors. The complexity of the latter may be

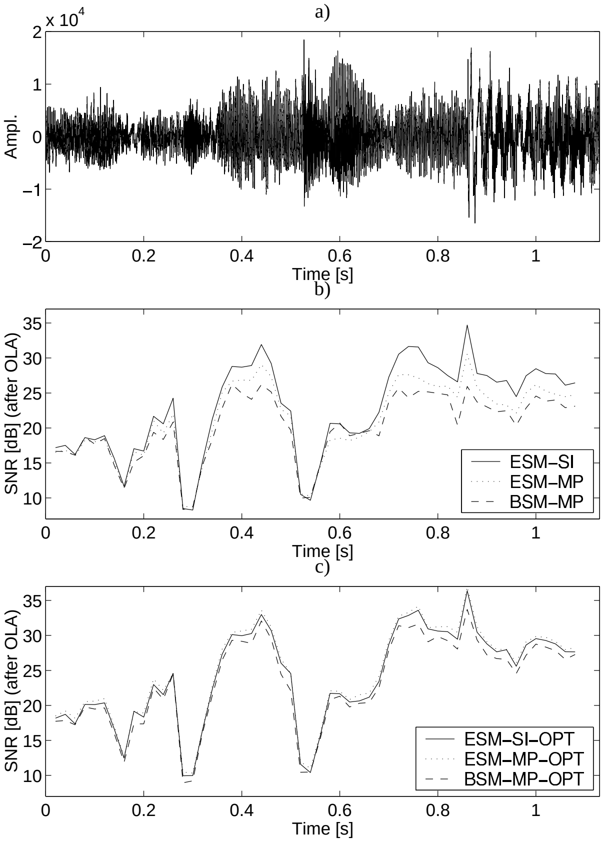

Figure 3: SNR vs. Time for audio signal (with $K_m = 30$). a) Time domain signal, b) Non-optimized parameter estimation schemes, c) Newton optimized estimation schemes.

prohibitive with large segment lengths encountered at a sampling frequency of e.g. 48 kHz.

The comparison of the BSM with the ESM showed that for signals with relatively many onsets (e.g., some speech signals), the ESM performs better than the BSM, while in signals with few amplitude variations (e.g, the audio signal in Fig. 3) the BSM achieves performance similar to ESM. Thus, a combined BSM-ESM scheme may be advantageous, where the ESM is only applied in segments where the modeling performance is significantly better than that of the BSM.

5. REFERENCES

[1] B. Edler, H. Purnhagen, and C. Ferekidis. ASAC – Analysis/Synthesis Codec for very low Bit Rates. In *Preprint 4179 (F-6) 100th AES Convention*, 1996.

[2] 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), 1997.

[3] M. Goodwin. Matching Pursuit With Damped Sinusoids. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 3:2037–2040, 1997.

[4] R. Heusdens, R. Vafin, and W.B. Kleijn. Sinusoidal modeling using psychoacoustic-adaptive matching pursuits. Submitted to *IEEE Signal Processing Lett.*

[5] J. Jensen and J. H. L Hansen. Speech Enhancement Using a Constrained Iterative Sinusoidal Model. *Accepted for Publication in IEEE Trans. Speech, Audio Processing*, 2001.

[6] J. Jensen, S. H. Jensen, and E. Hansen. Exponential Sinusoidal Modeling of Transitional Speech Segments. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 473–476, 1999.

[7] P. Lemmerling, I. Dologlou, and S. Van Huffel. Speech Compression Based on Exact Modeling and Structured Total Least Norm. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 353–356, 1998.

[8] S. Mallat and Z. Zhang. Matching Pursuits With Time-Frequency Dictionaries. *IEEE Trans. Signal Processing*, 41(12):3397–3415, 1993.

[9] R. J. McAulay and T. F. Quatieri. Speech Analysis/Synthesis Based on a Sinusoidal Representation. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-34(4):744 – 754, 1986.

[10] 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.

[11] J. Nieuwenhuijse, R. Heusdens, and E. F. Deprettere. Robust Exponential Modeling of Audio Signals. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 3581–3584, 1998.

[12] T. F. Quatieri and R. J. McAulay. Noise Reduction using a Soft-Decision Sine-Wave Vector Quantizer. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 821 – 824, 1990.

[13] X. Serra and J. Smith III. Spectral Modeling Synthesis: A Sound Analysis/Synthesis System Based on a Deterministic plus Stochastic Decomposition. *Computer Music Journal*, 14(4):12 – 24, 1990.

[14] Y. Stylianou. Applying the Harmonic Plus Noise Model in Concatenative Speech Synthesis. *IEEE Trans. Speech, Audio Processing*, 2001.

[15] S. Van Huffel, H. Chen, C. Decanniere, and P. Van Hecke. Algorithm for Time-Domain NMR Data Fitting Based on Total Least Squares. *J. Magn. Reson. A*, 110:228–237, 1994.