A PERCEPTUAL SUBSPACE METHOD FOR SINUSOIDAL SPEECH AND AUDIO MODELING

Jesper Jensen and Richard Heusdens

Dept. of Mediamatics
Delft University of Technology
Delft, The Netherlands
E-mail: {J.Jensen, R.Heusdens}@its.tudelft.nl

Søren Holdt Jensen

Dept. of Communication Technology
Aalborg University
Aalborg, Denmark
E-mail: SHJ@cpk.auc.dk

ABSTRACT

The problem of modeling a signal segment as a sum of exponentially damped sinusoidal components is of interest in a wide range of fields, including speech and audio processing. Often, model parameters are estimated using subspace based techniques that exploit the so-called shift-invariance property. A drawback of these estimation techniques in relation to speech and audio processing is that the *perceptual* relevance of the model components is not taken into account. In this paper we show how to combine well-known subspace based estimation techniques with a recently developed perceptual distortion measure, to obtain an algorithm for extracting perceptually relevant model components. In analysis-synthesis experiments with wideband audio signals, objective and subjective evaluations show that the proposed algorithm improves perceived signal quality considerable over traditional subspace based analysis methods.

1. INTRODUCTION

Sinusoidal models have proven to provide accurate and flexible representations of a large class of acoustic signals including audio and speech signals. For speech and audio processing, sinusoidal models have been applied in areas such as speech coding (e.g. [1]), speech enhancement (e.g. [2]), music synthesis (e.g. [3]), and more recently low bit-rate audio coding (e.g. [4]).

The applications above can be described in an analysis-modification-synthesis framework, where in the analysis stage model parameters are estimated for consecutive signal frames; in this stage it is typically assumed that each signal frame can be represented well as a linear combination of constant-amplitude, constant-frequency sinusoidal functions. In the modification phase, the estimated parameters may be quantized or otherwise modified. Finally, in the synthesis stage, the resulting parameters are used for reconstructing the possibly modified signal using interpolative or overlap/add synthesis (e.g. [1]).

Recently, several extended sinusoidal model variants have been proposed, which relax the constant-amplitude, constant-frequency assumption (e.g. [4, 5]). An extended model of particular interest is the so-called exponential sinusoidal model (ESM) which represents signal segments as sums of exponentially damped sinusoids. Observing that damped oscillations occur commonly in many natural signals including speech and audio, the ESM is often a physically reasonable model. The ESM has been applied to analysis-synthesis of audio (e.g. [6]) as well as speech signals (e.g. [7, 8]).

In many speech and audio applications it is of interest to represent only the perceptually relevant time/frequency regions of the signal in question by exploiting the masking properties of the human auditory system. By doing so, the signal can be represented by a reduced parameter set which e.g. may be exploited for efficient compression. The ESM parameter estimation schemes can roughly be divided into two main groups: analysis-by-synthesis schemes such as matching pursuit (MP) based algorithms (e.g. [9, 5]) and subspace-based schemes (e.g. [10, 11, 12]). While some work has been done for extracting perceptually relevant sinusoids using MP based schemes (e.g. [13]), less effort has been directed towards perceptual subspace based techniques.

In [14] an attempt is made to combine psycho-acoustical information with a subspace based ESM parameter estimation scheme. The signal to be modeled is divided into subbands and an independent (low-order) ESM is used for each subband. The ESM components are estimated in an iterative manner, one at a time, by assigning in each iteration an additional damped sinusoid to the subband having the largest residual error-noise to masking level, in much the same way as the bits are assigned to different subbands in MPEG-AUDIO [15]. The approach in [14] operates at a lower computational complexity than a corresponding full-band scheme, but is sub-optimal because subbands are treated independently. Furthermore, no perceptual knowledge is used for estimating the sinusoids within each subband.

In this paper we propose an algorithm which aims at minimizing a *perceptually* motivated distortion measure. In addition, it allows for joint estimation of perceptually relevant ESM parameters. The presented algorithm combines well-known subspace based estimation algorithms and a distortion measure derived from a recently developed psycho-acoustical masking model.

2. THE PERCEPTUAL DISTORTION MEASURE

To account for human auditory perception, we use the perceptual distortion measure described in [16]. The underlying psycho-acoustical model differs from existing spreading-function based models in the sense that it takes into account all auditory filters for computing the distortion, rather than considering the auditory filter receiving most of the distortion. The distortion measure D can be written as [13]:

$$D = \int_0^1 \hat{a}(f)|(\hat{w\varepsilon})(f)|^2 df, \quad (1)$$

The research was conducted within the ARDOR project, supported by the E.U. grant no IST-2001-34095.

where $\hat{}$ indicates the Fourier transform operation, $\hat{a}$ is a weighting function representing the frequency-dependent sensitivity of the human auditory system, w is the analysis window, and $\varepsilon = x - \tilde{x}$ is the modeling error, i.e., the difference between the original signal x and the modeled signal $\tilde{x}$. The weighting function $\hat{a}$ is usually chosen to be the reciprocal of the masking threshold. In this work we use the masking threshold derived from the psycho-acoustical model in [16], but other models can be used, e.g. the one from MPEG-Audio [15]. In order for Eq. (1) to define a norm, the weighting function $\hat{a}$ must be positive and real for all $f \in [0; 1)$ so that the distortion D in Eq. (1) can be rewritten as the convolution of two (infinite) discrete-time sequences:

$$D = \sum_n |(h * w\varepsilon)(n)|^2 = \|h * w\varepsilon\|_2^2, \tag{2}$$

where h is the inverse Fourier transform of $\sqrt{\hat{a}}$.

In the context of the ESM, the modeled signal frame $\tilde{x} = [\tilde{x}_0, \ldots, \tilde{x}_{N-1}]^T$ is given by

$$\tilde{x}_n = \sum_k a_k \exp(-d_k n) \cos(\omega_k n + \phi_k), \tag{3}$$

for $n = 0, \ldots, N-1$, where a_k, d_k, ω_k, and ϕ_k are amplitude, damping, (normalized) angular frequency, and phase parameters, respectively. The problem is for a given original signal frame x to find the set of ESM parameters which minimizes the perceptual distortion measure D in Eq. (2). Since a convolution operation can be formulated in terms of a matrix-vector multiplication, the minimization problem of interest can be stated as:

$$\min_{a_k, d_k, \omega_k, \phi_k} \|HW(x - \tilde{x}(a_k, d_k, \omega_k, \phi_k))\|_2^2, \tag{4}$$

where $W = diag(w)$ is a diagonal matrix containing the elements of the analysis window w, and H is a Toeplitz filtering matrix containing the elements of the, in this case symmetric, filter impulse response h. The effect of premultiplication with HW may be interpreted as a transformation from the linear domain where the l^2-norm does not necessarily correlate well with subjective quality to a perceptual domain where the l^2-norm is in better accordance with perceived quality. In principle, Eqs. (2) and (4) deal with signal sequences of infinite length. In practice, however, the ESM parameters must be estimated from finite sample sequences.

3. ESTIMATION OF PERCEPTUAL ESM PARAMETERS

The algorithms to be presented rely on the observation that the modeled segment $\tilde{x}$ in Eq. (3) can be expressed as a sum of complex exponentials:

$$\tilde{x}_n = \sum_{k=1}^{K} c_k z_k^n, \qquad n = 0, \ldots, N-1, \tag{5}$$

where $c_k = a_k \exp(j\phi_k)$ are complex amplitude parameters and $z_k = \exp(-d_k n + j\omega_k)$ are so-called signal poles. From Eq. (5) we see that the signal poles z_k contribute non-linearly to the objective function in Eq. (4). This non-linearity can be circumvented by using the so-called HTLS algorithm [11], which is a total least squares (TLS) based variant of Kung et al's original state space algorithm [10]. These algorithms belong to the class of single shift-invariant methods within the set of subspace-based signal analysis algorithms [12]. The HTLS algorithm is not immediately suited for solving the weighted problem in Eq. (4). Instead, to take the filtering effect of the matrix H in Eq. (4) into account, we consider the so-called prefiltered HTLS algorithm [17]. In the following we give a brief review of the HTLS algorithm and the prefiltered HTLS algorithm; for an in-depth treatment of the algorithms, we refer to [11] and [17], respectively.

3.1. Signal Poles with HTLS

Let us initially assume that the observed signal frame x can be represented by the ESM in Eq. (5) without error, and that the correct model order K is known. The HTLS algorithm [11] first arranges the observed signal frame x in a Hankel data matrix $X \in \mathbb{R}^{L \times M}, L, M > K$ whose first column is $[x_1 \cdots x_{L-1}]^T$ and whose last row is $[x_{L-1} \cdots x_N]$.

The singular value decomposition (SVD) of X is given by:

$$X = [U_1 U_2] \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix} = [U_1 \Sigma_1 V_1^T],$$

where Σ_1 is a diagonal matrix containing the K non-zero singular values, and the matrices U_1 and V_1 contain as columns the K corresponding left and right singular vectors, respectively. Finally, the shift-invariance property of U_1 (and V_1) ensures that the following matrix equations are satisfied:

$$U_1^{\downarrow} Z^{(u)} = U_1^{\uparrow}, \qquad V_1^{\downarrow} Z^{(v)} = V_1^{\uparrow}, \tag{6}$$

where the superscripts $\uparrow$ ($\downarrow$) denote deletion of the top (bottom) row of the matrix in question. The signal poles are found as the eigenvalues of the $K \times K$ matrix $Z^{(u)}$ (or $Z^{(v)}$). If the observed signal satisfies Eq. (5) and K is known, the underlying signal poles can be recovered without error.

In practice, however, x does *not* satisfy the ESM exactly, the matrix X will typically have a rank larger than K, and the shift invariance property in Eq. (6) will only be approximately valid. In this case, the matrix U_1 contains the left singular vectors corresponding to the K largest singular values of X, and the matrix $Z^{(u)}$ (or $Z^{(v)}$) is estimated as the TLS solution of the incompatible matrix equation in Eq. (6).

3.2. Signal Poles with Prefiltered HTLS

In the prefiltered HTLS algorithm [17], the Hankel data matrix X is postmultiplied with a full rank filter matrix $F \in \mathbb{R}^{M \times M}$ and the HTLS algorithm is applied to the matrix product $\bar{X} = XF$; a similar description can be derived when X is premultiplied with a filter matrix $G \in \mathbb{R}^{L \times L}$ [17]. It is straight-forward to show that $\bar{X}$, which generally is *not* Hankel structured, retains the rank-K and shift-invariant property. That is, when the observed signal frame satisfies Eq. (5) and K is known, we have:

$$\bar{U}_1^{\downarrow} \bar{Z}^{(u)} = \bar{U}_1^{\uparrow}, \qquad \bar{V}_1^{\downarrow} \bar{Z}^{(v)} = \bar{V}_1^{\uparrow}, \tag{7}$$

where $\bar{U}_1 \in \mathbb{R}^{L \times K}$ contains the left singular vectors corresponding to the K non-zero singular values of the filtered matrix $\bar{X}$, and the signal poles $\tilde{z}_k$ can be recovered without error as the eigenvalues of $\bar{Z}^{(u)}$; in this ideal scenario, the only requirement is that the filter matrix $F \in \mathbb{R}^{M \times M}$ has full rank.

The purpose of the filter matrix F is to implement the convolution in Eq. (2). Since, in practice, the sequences in Eq. (2)

have finite length, the convolution is circular; hence, we use a circular Toeplitz filter matrix F whose first column and first row is $[h_\tau \cdots h_1\ 0 \cdots 0\ h_q \cdots h_{\tau+1}]^T$ and $[h_\tau \cdots h_q\ 0 \cdots 0\ h_1 \cdots h_{\tau-1}]$, respectively, with $\tau = \lfloor \frac{q+1}{2} \rfloor$; experiments showed that this filter matrix structure leads to better performance than e.g. the Toeplitz structured filter matrix proposed in [17]. Forming the product $\bar{X} = XF$ corresponds to convolving circularly each row in X with the FIR filter impulse response $h = [h_1 \cdots h_q]^T$.

3.3. Estimation of Complex Amplitudes

Having estimated the signal poles with the prefiltered HTLS algorithm, the complex amplitudes $c_k = a_k \exp(j\phi_k)$ (and thus real amplitudes a_k and phases ϕ_k) are found as the solution to the weighted linear least squares problem

$$\tilde{c} = \arg\min_c \|HW(x - Vc)\|_2^2, \tag{8}$$

where $c = [c_1 \cdots c_K]^T$ is the complex amplitude vector, H is an $N \times N$ Toeplitz filtering matrix whose first column and row is $[h_\tau \cdots h_q\ 0 \cdots 0]^T$ and $[h_\tau \cdots h_1\ 0 \cdots 0]$, respectively, $W = diag(w)$ is an $N \times N$ diagonal matrix with the elements of the analysis window w on the main diagonal, and $V \in \mathbb{C}^{N \times K}$ is a Vandermonde matrix constructed from the signal pole estimates $\tilde{z}_k$. The ith column of V is of the form $[1\ \tilde{z}_i\ \tilde{z}_i^2 \cdots \tilde{z}_i^{N-1}]^T$.

3.4. Algorithm Outline

The proposed scheme for estimating perceptually relevant ESM parameters, denoted by 'P-ESM', can be outlined as follows.

Input: x, K.
Output: $\tilde{z}_k$, $\tilde{c}_k$

1. Compute perceptual weighting filter h from a psychoacoustical masking model (eg. [16]), and construct filter matrices F and H.
2. Construct Hankel structured data matrix X.
3. Compute prefiltered data matrix $\bar{X} = XF$.
4. Find perceptual signal pole estimates using the HTLS algorithm [11]: $\tilde{z}_k = HTLS(\bar{X})$.
5. Construct the Vandermonde matrix V from $\tilde{z}_k$, and estimate complex amplitude vector from weighted linear least squares problem: $\tilde{c} = \arg\min_c \|HW(x - Vc)\|_2^2$.

4. SIMULATION RESULTS

A number of simulation experiments was conducted to study and evaluate the performance of the proposed algorithm; objective as well as subjective tests were performed. Seven different audio signals, sampled at a frequency of 44.1 kHz, were used in the experiments, see Appendix A. A fixed frame length of $N = 1024$ samples (23.2 ms) was used, and in case of analysis-synthesis of entire signals, frames were extracted with an overlap of 50%. Typically, the filter impulse response h had a length of $q = 256$ samples (5.8 ms). A fixed model order of $K = 50$ was used for all frames, and the window w was a Hanning window.

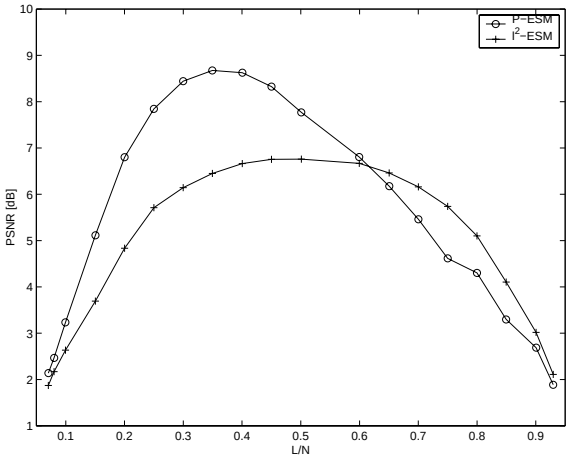

Fig. 1. Average modeling performance across 7 test signals as a function of dimension of data matrix $\bar{X} \in \mathbb{R}^{L \times M}$ for a fixed frame length of $N = 1024$ samples ($N = L + M - 1$).

To have an objective quality measure, we define the following 'perceptual' signal-to-noise ratio (PSNR) for the original signal frame x and its modeled counterpart $\tilde{x}$:

$$PSNR = 10 \log_{10} \left(\frac{\|HWx\|_2^2}{\|HW(x - \tilde{x})\|_2^2} \right) \text{ [dB]},$$

where the Toeplitz matrix H and the diagonal matrix W are identical to the ones used in Eq. (8). The PSNR measure aims at reflecting the quality of the modeled frame $\tilde{x}$ in the perceptual domain, and is valid to the extent that the perceptual model used for constructing H adequately represents the masking properties of the human auditory system. To measure the quality of an entire signal, we use the segmental PSNR defined as the average PSNR value, taken across the signal frames in question.

We compare the proposed parameter estimation scheme, P-ESM, to the standard non-perceptual algorithm ($F = I, H = I$). We denote this latter scheme l^2-ESM, where 'l^2' reflects that processing is done to minimize an unweighted l^2-norm (as opposed to a perceptually weighted l^2-norm).

4.1. Performance vs. Matrix Dimensions

In [18] is was argued that data matrices should be constructed 'as square as possible' for optimum performance with the standard HTLS algorithm. In this section we show by simulation that these settings are not optimal for the P-ESM scheme.

The signals in Appendix A were modeled with P-ESM and l^2-ESM using data matrices $\bar{X} \in \mathbb{R}^{L \times M}$ whose number of rows was varied in the range $L = 82, \cdots, 952$. Using a constant frame length resulted in taller and narrower data matrices for increasing values of L. Fig. 1 shows the modeling performance in terms of PSNR as a function of L/N. The performance curve for l^2-ESM has a broad optimum for L/N values in the range $1/3 - 2/3$. For P-ESM, the optimum is narrower and centered at $L/N \approx 1/3$.

The objective results in Fig. 1 are supported by informal listening tests. For l^2-ESM, signals generated with $1/4 < L/N < 3/4$ are perceptually identical to the reference signal generated with $L/N = 1/2$; for $L/N \leq 1/4$ and $L/N \geq 3/4$ a degradation is noticeable. Similarly, the P-ESM results in Fig. 1 are well in

line with perceived quality. While the quality at the reference setting $L/N \approx 1/2$ is good and certainly superior to the l^2-ESM signals, P-ESM quality is even better at $L/N \approx 1/3$. Furthermore, since the computational complexity of P-ESM is roughly $\mathcal{O}(\min(L, M)^3)$, the optimum at $L/N \approx 1/3$ implies a reduction in computations compared to the standard setting $L/N \approx 1/2$.

4.2. Listening Test with Audio Signals

To determine the subjective advantages of the proposed scheme, the seven signals listed in Appendix A were modeled with P-ESM and l^2-ESM and compared in a listening test. Ten listeners participated in the test. Signals were presented to the listeners as triplets OAB or OBA, where O was the original signal, A was the signal modeled with l^2-ESM and B was the signal modeled with P-ESM. The order (OAB or OBA) in which signals were presented was selected randomly for each presentation. The task of the listener was to decide, which signal (A or B) was perceptually closest to the original O. Each signal triplet was presented a total of 5 times during the test. The preference for P-ESM averaged across the ten listeners is shown in Table 1. Clearly, the P-ESM method performs better than the l^2-ESM method for all test signals.

Signal No.	1	2	3	4	5	6	7
Preference [%]	78	92	80	80	94	86	64

Table 1. Subjective preference for P-ESM over l^2-ESM.

5. CONCLUSION

We proposed a method for estimating perceptually relevant sinusoids for audio signal modeling. The method combines well-known subspace based estimation schemes with a recently developed perceptual distortion measure. In analysis-synthesis experiments with wideband audio signals, objective as well as subjective evaluations show that the proposed method leads to modeled signals of higher quality than standard subspace based estimation schemes.

Appendix A.

The audio signals used for evaluating the proposed parameter estimation scheme are outlined below.

Signal No.	Signal Name	Duration [s]
1	Abba	10.02
2	Trumpet	10.45
3	English female voice	6.84
4	Metallica	10.14
5	Contemporary pop music	10.05
6	Suzanne Vega	10.27
7	Carl Orff	10.90

6. REFERENCES

[1] R. J. McAulay and T. F. Quatieri, "Sinusoidal Coding," in *Speech Coding and Syntesis*, W. B. Kleijn and K. K. Paliwal, Eds., chapter 4. Elsevier Science B. V., 1995.

[2] J. Jensen and J. H. L Hansen, "Speech Enhancement Using a Constrained Iterative Sinusoidal Model," *IEEE Trans. Speech, Audio Processing*, vol. 9, no. 7, pp. 731–740, October 2001.

[3] E. B. George and M. J. T. Smith, "Analysis-by-synthesis overlap-add sinusoidal modeling applied to the analysis and synthesis of musical tones," *J. Audio Eng. Soc.*, June 1992.

[4] 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.

[5] M. Goodwin, "Matching pursuit with damped sinusoids," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1997, pp. 2037–2040.

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

[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*, 1998, pp. 353–356.

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

[9] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Processing*, vol. 41, no. 12, pp. 3397–3415, December 1993.

[10] 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.*, vol. 73, no. 12, pp. 1799–1811, 1983.

[11] 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*, vol. 110, pp. 228–237, 1994.

[12] A.-J. Van der Veen, E. F. Deprettere, and A. Lee Swindlehurst, "Subspace-based signal analysis using singular value decomposition," *Proc. IEEE*, vol. 81, no. 9, pp. 1277–1308, September 1993.

[13] R. Heusdens and S. van de Par, "Rate-distortion optimal sinusoidal modeling of audio and speech using psychoacoustical matching pursuits," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2002, pp. 1809–1812.

[14] K. Hermus, W. Verhelst, and P. Wambacq, "Psycho-acoustic modeling of audio with exponentially damped sinusoids," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2002, pp. 1821–1824.

[15] ISO/MPEG Committee, *Coding of moving pictures and associated audio for digital storage media at up to about 1.5Mbit/s – part 3: Audio*, ISO/IEC 11172-3, 1993.

[16] S. van de Par et al., "A new psychoacoustical masking model for audio coding applications," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2002, pp. 1805–1808.

[17] H. Chen, S. Van Huffel, and J. Vandewalle, "Bandpass prefiltering for exponential data fitting with known frequency regions of interest," *Signal Processing*, vol. 48, pp. 135–154, 1996.

[18] S. Van Huffel, "Enhanced resolution based on minimum variance estimation and exponential data modeling," *Signal Processing*, vol. 33, no. 3, pp. 333–355, 1993.