Frequency-domain Parameter Estimations for Binary Masked Signals

Zhang, Johan Xi; Christensen, Mads Græsbøll; Dahl, Joachim; Jensen, Søren Holdt; Moonen, Marc

Published in:
Proceedings of the International Conference on Spoken Language Processing

Publication date:
2008

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Zhang, J. X., Christensen, M. G., Dahl, J., Jensen, S. H., & Moonen, M. (2008). Frequency-domain Parameter Estimations for Binary Masked Signals. *Proceedings of the International Conference on Spoken Language Processing*.

Frequency-domain Parameter Estimations for Binary Masked Signals

J. X. Zhang[1] *, M. G. Christensen*[2] *, J. Dahl, S. H. Jensen and M. Moonen*$^{\perp}$

Department of Electronic Systems, Aalborg University, Aalborg, Denmark
jxz@es.aau.dk, mgc@es.aau.dk, joachim@es.aau.dk, shj@es.aau.dk
$^{\perp}$Department of Electrical Engineering, Katholieke Universiteit Leuven, Leuven, Belgium
marc.moonen@esat.kuleuven.be

Abstract

We present an approach for the extraction of parameters of a damped complex exponential model from a spectrogram modified by a binary mask. The parameters are estimated by a frequency domain based methods using subspace techniques, where the core algorithm is F-ESPRIT. The sub-band defined by the binary mask provides a reduced number of DFT-samples for the parameter extractions, which results in a computational efficient scheme with high parameter estimation accuracy. The proposed synthesis system has synthesis performance comparable to the so-called LSEE-MSTFT. The estimated parameters can be used in many applications such as audio/speech coding, pitch estimation and pitch scale modification.

Index Terms: STFT, Binary masks, Complex exponentials, Subspace techniques, Parameter estimation, F-ESPRIT

1. Introduction

The concepts of short-time Fourier analysis and synthesis are important for describing quasi-stationary (slowly time-varying) signals such as audio signals. Analysis/synthesis techniques have found many applications such as speech enhancement, speech or audio coding as well as source separation. In many applications an arbitrary modification on the spectrogram is used which leads to undefined STFT. This is because of the overlapping regions between adjacent short-time segments cannot have arbitrary variations. Historically, signal estimation from the modified STFT has been performed by applying the overlap-add (OLA) and filter-bank-summation (FBS) synthesis methods on the time-frequency functions [1, 2]. The standard OLA will normally give distortions at the boundaries if the spectrogram is modified arbitrarily. An improved synthesis method termed “least-squares error estimation from the modified STFT(LSEE-MSTFT)” was developed to minimize the synthesis distortions in [3], and this algorithm gives good estimates with almost the same computational complexity as standard OLA.

Processing audio and speech signals using binary masks have received considerable attention during the last decades in areas such as audio/speech enhancement and under-determined source separation. The binary mask is defined in the time-frequency (TF) domain where each TF points assigns a value either one or zero based on the signal in the region. In the modified spectrum, the binary mask is multiplied with the noisy signal spectrum to achieve the desired signal. For further analysis of the audio and speech signal a parametric modelling of the signal is required in applications such as coding, pitch estimation, musical synthesis ect. Sinusoidal models has been shown to be an accurate and flexible way to represent a large class of signals including audio and speech signals. The extended sinusoidal model which represents signal segments as sums of exponentially damped sinusoids is able to model real recorded acoustic signals more closely. A common problem with existing parametric extraction systems is the trade-off between parameter accuracy and computational complexity. The high-resolution estimation techniques based on subspace methods are normally related with high computational complexity.

In this paper we will present a parametric estimation scheme which estimates the parameters from the DFT samples of the masked sub-bands using a frequency-domain based subspace estimation algorithm. The presented scheme combines the binary mask with the extended version of Frequency-Selected ESPRIT (F-ESPRIT) methods to create a computationally efficient parametric estimation algorithm [4, 5, 6, 7]. An example where the proposed method can be used is in various coding applications; a noise corrupted signal is first enhanced by using a binary mask, and afterward the parameters of the signal are estimated using the samples from the DFT domain. One computationally demanding operation used in many subspace based estimation methods is the singular value decomposition (SVD) where the matrix is decomposed into subspaces. Our idea is to solve the same problem by first reducing the data amount by using binary mask and then perform SVD on smaller matrices, which reduces the overall computational burden.

In Section 2 we describe the general STFT framework and the binary mask, and the related equations and matrices will be defined. Section 3 describes the proposed method, and in Section 4 we apply the method to a recorded signal and demonstrate how the synthesized signal using estimated parameters can model the binary masked signal. Finally, discussions and conclusions are given in Section 5 and 6, respectively.

2. Framework

The binary masked signal $x(n)$ of length N is written in segmented vector form as $\mathbf{x}'_m$. The vector capture the signal from region $n = 1 + R(m-1), \cdots, N + R(m-1)$ and the window for the time-segment is sliding forward for every time-frame index m. Moreover, R is the STFT sampling period and defines the number of samples between two adjacent blocks. The time-frame index is denoted by m with values ranging from 1 to the total number of possible frames η. All the data vectors are Fourier transformed and collected together into a 2-dimensional

[1] The work of J. X. Zhang is supported by the Marie Curie EST-SIGNAL Fellowship (http://est-signal.i3s.unice.fr), contract no. MEST-CT-2005-021175.

[2] M. G. Christensen is supported by the Parametric Audio Processing project, Danish Research Council for Technology and Production Sciences grant no. 274060521.

STFT matrix

$$\mathbf{X} = \begin{bmatrix} \vdots & \vdots & & \vdots \\ \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_\eta \\ \vdots & \vdots & & \vdots \end{bmatrix} \qquad (1)$$

where the transformed vector is formulated as $\mathbf{x}_m = \boldsymbol{F}\{\mathbf{x}'_m\}$ and the symbol $\boldsymbol{F}$ denotes the DFT operation. Each elements of the STFT matrix is denoted with the notation $\mathbf{X}_{km}$ where k denotes row elements of different DFT frequencies and m denotes the column elements of the time-frame index.

The STFT can also be expressed as

$$\mathbf{X} = |\mathbf{X}| \odot \mathbf{\Psi}. \qquad (2)$$

Here $|\mathbf{X}|$ is the short-time magnitude spectrum, $\mathbf{\Psi}$ is the short-time phase spectrum and operator $\odot$ denotes element wise multiplication. The inverse-STFT using the OLA synthesis method is described in [1, 3].

From an implementation point of view the STFT is computed as a series of FFTs of windowed data frames, where the window slides forward through the time. The inverse-STFT synthesized using OLA can be implemented as the inverse of the spectrogram and added together with the decimation R.

2.1. Binary Time-Frequency Mask

A binary TF mask is a binary matrix where one indicates that the target signal is stronger than the interference within the corresponding TF region and zero indicates otherwise [8]. A modified spectrogram G is obtained as

$$\mathbf{G} = \mathbf{H} \odot \mathbf{X}, \qquad (3)$$

where $\mathbf{H}$ is the binary mask. For every fixed value of m in $\mathbf{H}$, there are regions with consecutive ones and zeros. Those region is defined as the frequency range for sub-bands. Consecutive non-zero frequency-domain samples from $\mathbf{G}$ are grouped into sub-bands, where every sub-band is feed into the proposed algorithm for parameter estimations; we elaborate on this in the following section. In this paper the binary mask is assumed to be known. Otherwise, the binary mask can easily be estimated from the modified spectrogram by grouping all non-zero elements into sub-bands.

3. Proposed Method

The model for the sum of all complex exponentials over all sub-bands for a modified time-frame segment is denoted $g_m(n)$ and modelled as (4). For notational simplicity $g_m(n)$ will be denoted as $g(n)$ in the rest of the paper, i.e.,

$$g(n) = \sum_{i=1}^{p} \left(A_i e^{\lambda_i n} + (A_i e^{\lambda_i n})^* \right) + v(t), \qquad (4)$$

where $\lambda_i = j\omega_i + \gamma_i$, ω_i and γ_i denotes respectively frequencies and damping factors, $*$ is the complex conjugate, p is the order and $v(t)$ denotes white complex zero-mean noise. Due to the symmetric properties of the estimated spectrogram on real acoustic signals only complex exponentials from region 0 to π in the normalized frequency scale are estimated.

A frame of STFT from the modified signal is segmented into different sub-bands defined by the binary mask. The DFT samples in a specific sub-band is used to estimate the parameters of complex exponential in that region and those DFT samples can be written in the frequency selective matrix form,

$$\mathbf{Y} = \begin{bmatrix} \mathbf{G}_{k_1 m} & & \mathbf{G}_{k_M m} \\ W_N^{k_1}\mathbf{G}_{k_1 m} & & W_N^{k_M}\mathbf{G}_{k_M m} \\ \vdots & \cdots & \vdots \\ W_N^{(S-1)k_1}\mathbf{G}_{k_1 m} & & W_N^{(S-1)k_M}\mathbf{G}_{k_M m} \end{bmatrix}, \qquad (5)$$

where $W_N^k = e^{-j\frac{2\pi k}{N}}$, $\mathbf{G}_{k_1 m}$,..., $\mathbf{G}_{k_M m}$ are the DFT samples on the transformed segment of g(n) in the frequency region defined by the binary mask. The matrix $\mathbf{Y}$ have dimensions $S \times M$, where M is the total number of DFT samples in the sub-band and S is normally referred to as the user parameter of the F-ESPRIT algorithm.

Following the definition from [4, 5], the measured frequency selective data in (6) is decomposed into their sub-spaces using the SVD,

$$\mathbf{Y}\mathbf{\Pi}_{\cup}^{\perp} = \begin{bmatrix} \mathbf{Z}_s & \mathbf{Z}_n \end{bmatrix} \begin{bmatrix} \mathbf{\Sigma}_s & 0 \\ 0 & \mathbf{\Sigma}_n \end{bmatrix} \begin{bmatrix} \mathbf{V}_s^H \\ \mathbf{V}_n^H \end{bmatrix}, \qquad (6)$$

where $\mathbf{\Pi}_{\cup}^{\perp}$ is referred to as a projection matrix defined by:

$$\mathbf{\Pi}_{\cup}^{\perp} = \mathbf{I} - \mathbf{U}^H(\mathbf{U}\mathbf{U}^H)^{-1}\mathbf{U}. \qquad (7)$$

The matrix $\mathbf{U}$ in (7) is formed from a subset of discrete Fourier transform bases.

$$\mathbf{U} = \begin{bmatrix} W_N^{k_1} & W_N^{k_2} & & W_N^{k_M} \\ W_N^{2k_1} & W_N^{2k_2} & & W_N^{2k_M} \\ \vdots & \vdots & \cdots & \vdots \\ W_N^{Sk_1} & W_N^{Sk_2} & & W_N^{Sk_M} \end{bmatrix}. \qquad (8)$$

$\mathbf{Z}_s$ contains signal subspace and $\mathbf{\Sigma}_s$ is the corresponding singular values. The dimension of signal subspaces p is estimated using the method proposed in [5]. The user parameter S can be selected in the range of $S \in (\lfloor \frac{M}{3} \rfloor, \lfloor \frac{M}{2} \rfloor)$.

In some applications the number of complex exponentials in the sub-band might be known, and a straightforward implementation without estimating the order can be applied [4]. The minimum required amount of DFT samples for the F-ESPRIT is $M = 2p + 1$ and the user parameter is set to be $S = \frac{M-1}{2}$ [4], where p is considered as known. In this paper, we only consider applications where the order in the sub-bands is unknown.

The frequency is estimated from the state transition matrix $\mathbf{T}$ using a technique known as shift-invariance estimation [9]. The shift-invariance relation is then

$$\mathbf{J}_1\mathbf{Z}_{s1}\mathbf{T} = \mathbf{J}_2\mathbf{Z}_{s2}, \qquad (9)$$

where $\mathbf{J}_1 = \begin{bmatrix} \mathbf{I}_{s-1} & \mathbf{0} \end{bmatrix}$, $\mathbf{J}_2 = \begin{bmatrix} \mathbf{0} & \mathbf{I}_{s-1} \end{bmatrix}$.

For a given $\mathbf{T}$, estimates of the λ_i parameters are found from the eigenvalues of μ_i = [eig($\mathbf{T}$)]$_i$, estimates of the ω_i =arg(μ_i) and γ_i =ln($|\mu_i|$).

In sub-bands without any harmonics, the order estimation algorithm will still give an order estimation which is $p \geq 1$. The frequency estimation of such sub-bands will use leakage signals in the sub-band to estimate their frequency parameters. So all frequency parameters which fall outside the defined sub-band should be discarded.

Once the frequencies and damping factors are known for all sub-bands in the time segment, the calculation of amplitudes and phases become a linear problem. The modified time-domain segment of $g(n)$ can be written in the vector form $\mathbf{g}$,

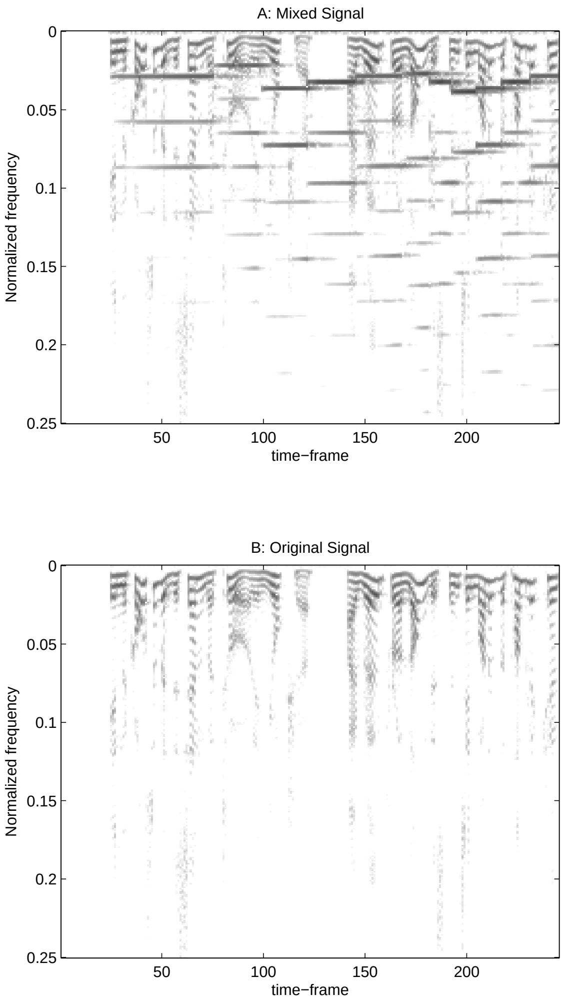

Figure 1: *Power spectrogram of: (A) Mixed signal (B) Original signal.*

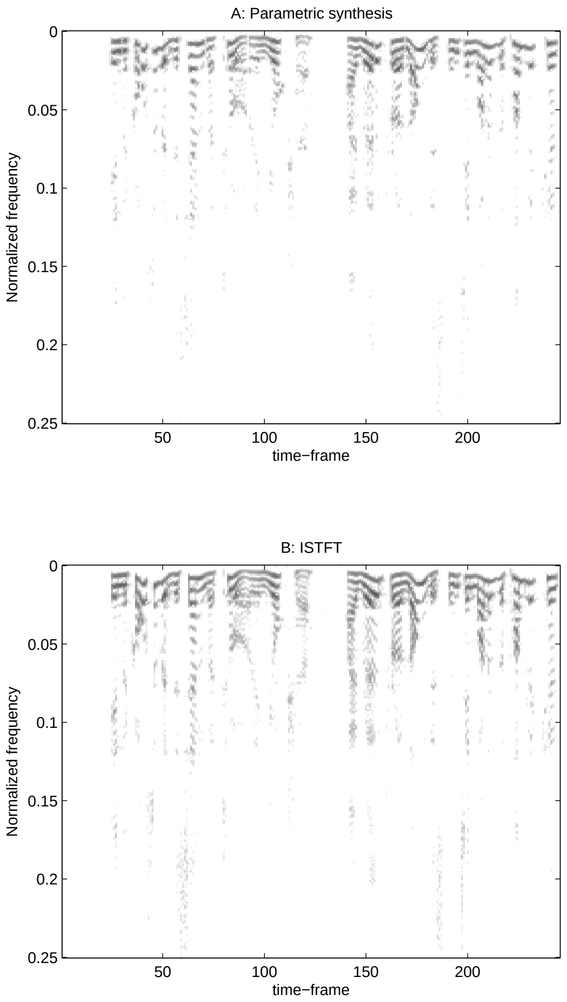

Figure 2: *Power spectrogram of the synthesized signal using the:(A) Estimated parameters (B) Inverse-STFT*

corresponding model for the segment is $\mathbf{Qa}$, where $\mathbf{a}$ is the amplitude vector and $\mathbf{Q}$ is the complex exponential matrix. Each column of $\mathbf{Q}$ is a realization of the complex exponential using the estimated parameters. The values of the complex amplitudes are calculated to minimize the quadratic error between the original signal $\mathbf{g}$ and the synthetic sinusoidal signal stated in (10).

$$\|\mathbf{W}(\mathbf{g}-\mathbf{Qa})\|^2 \,, \tag{10}$$

where $\mathbf{W}$ =diag$(\mathbf{w})$ and the diagonal vector $\mathbf{w}$ is the synthesis window.

The least-square minimization gives the following solution

$$\mathbf{a} = (\mathbf{Q}^H\mathbf{W}^T\mathbf{WQ})^{-1}\mathbf{Q}^H\mathbf{W}^T\mathbf{Wg} \tag{11}$$

For the weighting matrix estimated based on a standard window used in speech and audio processing such as Hamming or Hanning window, the estimated amplitudes will more accurately represent the amplitudes of harmonics in the center of the time-segment $\mathbf{g}$. The proposed method for estimation of model parameters from the spectrogram is outlined as follows:

1. Extract Sub-bands defined by $\mathbf{H}$ for a fixed time-frame.
2. Estimate order, frequency and damping factor for all subbands in the modified time-frame segment $\mathbf{g}$.
3. Estimate the complex amplitudes.
4. Repeat the procedure until all time-frames are processed.

For verification purposes the reconstruction of the estimated time-segment using extracted parameters is synthesized together into the time-domain signal using the extended OLA synthesis method described in [3], with the decimation used to create the STFT matrix. The recommended synthesis window $\mathbf{w}$ is either using a Hanning or Hamming window. From an implementation point of view, the synthesis signal is reconstructed from a series of time-domain estimation of the modified spectrogram, OLA, with the decimation R.

4. Numerical Examples

In the evaluations we investigate the estimated parameters from real recorded signals. A number of simulation experiments were conducted to study and evaluate the estimated parameters of the proposed method. Different audio and speech signals from [10] sampled at a frequency of 44.1 kHz were used in the experiments. A fixed frame length of N_w=2048 was used. In the case of analysis and synthesis, frames were extracted with an overlap of 50% (R=1024). A Hanning window was used in OLA. Mixed signals were created by adding two pre-recorded signals; see in Fig. 1(A). The signal is a combination of a sequence of male speech and a melodious phrase of clarinet, one of them is assumed to be the desired signal and the other is the noise. In this example we selected the speech signal to be the desired signal shown in Fig. 1(B).

The binary mask used for the signal enhancement is assumed to be known and estimated based on the ideal-binary mask criteria. Parameters extracted using the proposed method are synthesized using the OLA on estimated time-segments to create the time domain estimate of the desired signal. A power spectrogram of the estimated signal is shown in Fig. 2(A). The corresponding reference signal modified by the binary mask synthesized using the standard method [3] is shown in Fig. 2(B). Because of the high sample frequency of the signal, the energy of our desired signal is mainly concentrated in the lower frequency region of the plotted power spectrogram. Therefore only frequencies from 0 to 0.25π are plotted in Fig. 1 and 2. From informal listening test, we conclude that the proposed method gives a very high quality of reconstruction on the synthesized signal compared to synthesized signals based on traditional methods. The effect of good parametric reconstruction can be seen in the spectrogram shown in Fig. 2(B); we see that the spectrogram with parameters estimated using our proposed scheme is almost identical to the spectrogram synthesized using [3] except for regions with stochastic components.

5. Discussion

The proposed method has been tested under different merged sound files, which gives an accurate parameter estimation in most cases. The synthesized signal using estimated parameters preserves high perceptual quality compared with the binary mask modified signal. There are mainly two main factors which will affect the quality of the reconstructed signal using estimated parameters.

First, there are regions where the model assumed in the algorithm does not perfectly match the real recorded signals. The signal model is defined as complex exponentials in white noise which is a good model for voiced part of the speech and many instrumental sounds. From the recorded signals, we can no longer assume that the signal always contains harmonics embedded in white noise. In fact, some recorded signals have a colored background noise. To minimize disturbance due to the colored noise, a common solution used in many adaptive filtering applications is implemented where the total number of DFT samples fed to the proposed algorithm is limited to increase whiteness of the noise. Experiments have verified that the proposed method becomes more robust against colored noise. The trade off between increased robustness and limited amount of data samples is the "slightly" reduced parameter estimation accuracies. A balance should be selected depending on the estimated signal characteristics.

The algorithm requires a certain number of consecutive samples in the sub-band to estimate the parameters of the complex exponentials. Blindly masking the signal spectrogram using the ideal binary-mask will not guarantee that all masked sub-bands are satisfied by the algorithm. For increased robustness, resolution of the sub-bands must be decreased so that the minimum amount of masked data covered by the sub-band can satisfy the sample number requirement. Interference will be introduced according to the resolution of the sub-bands. The trade-off between resolution and suppression should be balanced based on the actual applications.

The proposed method gives high accuracy on the estimated parameters for both synthetic and recorded signals. However, in case of real recorded signals, the sub-bands which do not perfectly match our defined model are unavoidable so that some inaccuracies will still occur.

6. Conclusion

This paper presents a scheme for parameter extraction based on frequency domain data samples which is computationally cheaper than traditional time-domain ESPRIT. Informal listening test conclude that the quality of synthesized signals based on the estimated parameters can be comparable to the modified signal synthesized using [3]. The proposed method provides a computationally cheaper method for parameter estimations on the binary masked signals. The estimated parameters can be used in many existing applications where the sinusoidal model plays a vital part of the algorithms. Applications such as audio/speech coding and speech modification are especially interesting.

7. References

[1] J. Allen and L. Rabiner, "A unified approach to short-time fourier analysis and synthesis," *Proc. of the IEEE*, vol. 65, no. 11, pp. 1558–1564, 1977.

[2] T. F. Quatieri, *Discrete-Time Speech Signal Processing*. Prentice Hall PTR, 2002.

[3] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Trans. on Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[4] J. Gunnarsson and T. McKelvey, "High SNR performance analysis of F-ESPRIT," *Proc. Asilomar Conf. on Signals, Systems and Computers*, vol. 1, pp. 1003–1007, 2004.

[5] A. Jakobsson, M. G. Christensen, and S. H. Jensen, "Frequency selective sinusoidal order estimation," *Electronics Letters*, vol. 43, no. 21, 2007.

[6] R. Badeau, B. David, and G. Richard, "A new perturbation analysis for signal enumeration in rotational invariance techniques," *IEEE Trans. Signal Processing*, vol. 54, no. 2, pp. 450–458, 2006.

[7] M. G. Christensen, A. Jakobsson, and S. H. Jensen, "Sinusoidal order estimation using the subspace orthogonality and shift-invariance properties," *Proc. Asilomar Conf. Signals, Systems, and Computers*, 2007.

[8] N. Roman, D. Wang, and G. J. Brown, "Speech segregation based on sound localization," *Proc. IJCNN '01. Inter. Joint Conf. on Neural Networks*, vol. 4, pp. 2861–2866, 2001.

[9] S. Y. Kung, "A new identification and model reduction algorithm via singular value decomposition," *Proc. Asilomar Conf. on Circuits, Systems and Computers*, pp. 705–714, 1978.

[10] "Tech 3253-sound quality assessment material SQAM 1988." [Online]. Available: http://www.ebu.ch/en/technical/publications/tech3000_series