A Method for Filter Shaping in Convolutive Blind Source Separation

Radoslaw Mazur and Alfred Mertins

Institute for Signal Processing, University of Lübeck, 23538 Lübeck, Germany
{mazur,mertins}@isip.uni-luebeck.de

Abstract. An often used approach for separating convolutive mixtures is the transformation to the time-frequency domain where an instantaneous ICA algorithm can be applied for each frequency separately. This approach leads to the so called permutation and scaling ambiguity. While different methods for the permutation problem have been widely studied, the solution for the scaling problem is usually based on the minimal distortion principle. We propose an alternative approach that shapes the unmixing filters to have an exponential decay which mimics the form of room impulse responses. These new filters still add some reverberation to the restored signals, but the audible distortions are clearly reduced. Additionally the length of the unmixing filters is reduced, so these filters will suffer less from circular-convolution effects that are inherent to unmixing approaches based on bin-wise ICA followed by permutation and scaling correction. The results for the new algorithm will be shown on a real-world example.

1 Introduction

The blind source separation (BSS) problem has been widely studied for the instantaneous mixing case and several efficient algorithms exist [1,2,3]. However, in a real-world scenario in an echoic environment, the situation becomes more difficult, because the signals arrive several times with different time lags, and the mixing process becomes convolutive. Although some time-domain methods for solving the convolutive mixing problem exist [4,5], the usual approach is to transform the signals to the time-frequency domain, where the convolution becomes a multiplication [6] and each frequency bin can be separated using an instantaneous method. This simplification has a major disadvantage though. As every separated bin can be arbitrarily permuted and scaled, a correction is needed. When the permutation is not correctly solved the separation of the entire signals fails. A variety of different approaches has been proposed to solve this problem utilizing either the time structure of the signals [7,8,9] or the properties of the unmixing matrices [10,11,12]. When the scaling is not corrected, a filtered version of the signals is recovered. In [13,14] the authors proposed a postfilter method that aims to recover the signals as they have been recorded at the microphones, accepting the distortions of the mixing system while not adding new ones. This concept appears to be quite reasonable, but the desired goal is

T. Adali et al. (Eds.): ICA 2009, LNCS 5441, pp. 282–289, 2009.

often not exactly achieved in practice due to circular convolution artifacts that stem from the bin-wise independent design of the unmixing filters which does not obey the filter-length constraints known from fast-convolution algorithms. This problem has been addressed in [15], where the authors applied a smoothing to the filters in the time domain in order to reduce the circular-convolution effects.

In this paper, we propose a new method for solving the scaling ambiguity with the aim of shaping the unmixing filters to have an exponential decay. This mimics the behavior of room impulse responses and reduces the reverberation. In order to achieve this, we calculate the dependency between the scaling factors and the impulse responses of the unmixing filterbank and calculate the scaling factors that shape the desired form.

2 The Framework for Mixing and Blind Unmixing

The instantaneous mixing process of N sources into N observations can be modeled by an $N \times N$ matrix $\mathbf{A}$. With the source vector $\boldsymbol{s}(n) = [s_1(n), \ldots, s_N(n)]^T$ and negligible measurement noise, the observation signals are given by

$$\boldsymbol{x}(n) = [x_1(n), \ldots, x_N(n)]^T = \mathbf{A} \cdot \boldsymbol{s}(n). \tag{1}$$

The separation is again a multiplication with a matrix $\mathbf{B}$:

$$\boldsymbol{y}(n) = \mathbf{B} \cdot \boldsymbol{x}(n) \tag{2}$$

with $\boldsymbol{y}(n) = [y_1(n), \ldots, y_N(n)]^T$. The only source of information for the estimation of $\mathbf{B}$ is the observed process $\boldsymbol{x}(n)$. The separation is successful when $\mathbf{B}$ can be estimated so that $\mathbf{BA} = \mathbf{D\Pi}$ with $\mathbf{\Pi}$ being a permutation matrix and $\mathbf{D}$ being an arbitrary diagonal matrix. These two matrices stand for the two ambiguities of BSS. The signals may appear in any order and can be arbitrarily scaled. For the separation we use the well known gradient-based update rule according to [1].

When dealing with real-world acoustic scenarios it is necessary to consider the reverberation. The mixing system can be modeled by FIR filters of length L. Depending on the reverberation time and sampling rate, L can reach several thousand. The convolutive mixing model reads

$$\boldsymbol{x}(n) = \mathbf{H}(n) * \mathbf{s}(n) = \sum_{l=0}^{L-1} \mathbf{H}(l)\boldsymbol{s}(n-l) \tag{3}$$

where $\mathbf{H}(n)$ is a sequence of $N \times N$ matrices containing the impulse responses of the mixing channels. For the separation we use FIR filters of length $M \geq L - 1$ and obtain

$$\boldsymbol{y}(n) = \mathbf{W}(n) * \boldsymbol{x}(n) = \sum_{l=0}^{M-1} \mathbf{W}(l)\boldsymbol{x}(n-l) \tag{4}$$

with $\mathbf{W}(n)$ containing the unmixing coefficients.

Using the short-time Fourier transform (STFT), the signals can be transformed to the time-frequency domain, where the convolution approximately becomes a multiplication [6]:

$$\boldsymbol{Y}(\omega_k, \tau) = \boldsymbol{W}(\omega_k)\boldsymbol{X}(\omega_k, \tau), \qquad k = 0, 1, \ldots, K-1, \tag{5}$$

with K being the FFT length. The major benefit of this approach is the possibility to estimate the unmixing matrices for each frequency independently, however, at the price of possible permutation and scaling in each frequency bin:

$$\boldsymbol{Y}(\omega_k, \tau) = \boldsymbol{W}(\omega_k)\boldsymbol{X}(\omega_k, \tau) = \boldsymbol{D}(\omega_k)\boldsymbol{\Pi}(\omega_k)\boldsymbol{S}(\omega_k, \tau) \tag{6}$$

where $\boldsymbol{\Pi}(\omega)$ is a frequency-dependent permutation matrix and $\boldsymbol{D}(\omega)$ an arbitrary diagonal scaling matrix.

The correction of the permutation is essential, because the entire unmixing process fails if different permutations occur at different frequencies. A number of approaches has been proposed to solve this problem. [7,8,9,10,11,12].

When the scaling ambiguity is not solved, filtered versions of the sources are recovered. A widely used approach has been proposed in [13]. The authors aimed to recover the signals as they were recorded at the microphones accepting all filtering done by the mixing system. A similar technique has been proposed in [14] under the paradigm of the minimal distortion principle, which uses the unmixing matrix

$$\boldsymbol{W}'(\omega) = \mathrm{dg}(\boldsymbol{W}^{-1}(\omega)) \cdot \boldsymbol{W}(\omega) \tag{7}$$

with $\mathrm{dg}(\cdot)$ returning the argument with all off-diagonal elements set to zero. However, as mentioned in the introduction, the independent filter design for each frequency bin may result in severe circular convolution artifacts in the final unmixed time-domain signals. In this paper, we therefore propose a method to re-scale the frequency components in such a way that the resulting unmixing filters obey a desired decay behavior. This new approach will be described in the next section.

3 Filter Shaping

The proposed method is to introduce a set of scaling factors $c(\omega)$ for the unmixed frequency components that ensure that the unmixing filters obey a desired decay behavior. The motivation for this comes from the fact that the impulse responses achieved by the minimal distortion principle have a quite arbitrary form. In particular, they often show many large coefficients after the main peak, which results in a significant amount of added reverberation and can even lead to problems of circular-convolution artifacts. For addressing both above-mentioned problems we propose to shape the unmixing filters to have an exponential decay. This reduces the perceived echoes as well as the problems of circular convolution.

In Fig. 1 the overall BSS system is shown. It consists of $N \times N$ single channels as depicted in Fig. 2. In this representation the permutation has already been corrected. The dependency of time-domain filter coefficients of a filter vector

Fig. 1. Overview of frequency-domain BSS

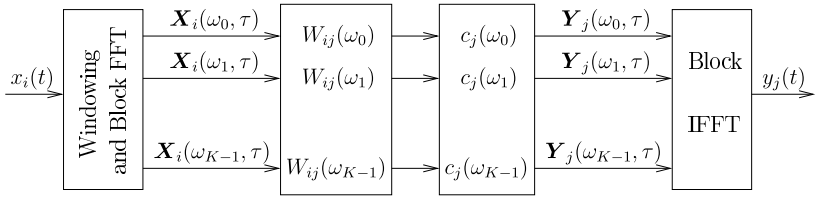

Fig. 2. Data flow from input i to output j

$\boldsymbol{w}_{ij}$ and scaling factors $\boldsymbol{c}_j = [c_j(\omega_0), c_j(\omega_1), \ldots, c_j(\omega_{K-1})]^T$ for output j can be calculated as follows:

$$\begin{aligned}\boldsymbol{w}_{ij} &= \sum_l \boldsymbol{E}_l \cdot \mathcal{F}^{-1} \cdot \boldsymbol{C}_j \cdot \boldsymbol{W}_{ij} \cdot \mathcal{F} \cdot \boldsymbol{D}_l \cdot \delta \\ &= \sum_l \boldsymbol{E}_l \cdot \mathcal{F}^{-1} \cdot \mathrm{diag}(\mathcal{F} \cdot \boldsymbol{D}_l \cdot \delta) \cdot \boldsymbol{W}_{ij} \cdot \boldsymbol{c}_j \\ &= \boldsymbol{V}_{ij} \cdot \boldsymbol{c}_j\end{aligned} \tag{8}$$

where $\mathrm{diag}(\cdot)$ converts a vector to a diagonal matrix. The term δ is a unit vector containing a single one and zeros otherwise. $\boldsymbol{D}_l$ is a diagonal matrix containing the coefficients of the STFT analysis window shifted to the lth position according to the STFT window shift. $\mathcal{F}$ is the DFT matrix. $\boldsymbol{W}_{ij}$ is a diagonal matrix containing the frequency-domain unmixing coefficients. $\boldsymbol{c}_j$ is a vector of the sought scaling factors, and $\boldsymbol{C}_j$ is a diagonal matrix made up as $\boldsymbol{C}_j = \mathrm{diag}(\boldsymbol{c}_j)$. $\boldsymbol{E}_l$ is a shifting matrix corresponding to $\boldsymbol{D}_l$, defined in such a way that the overlapping STFT blocks are correctly merged. Note that for real-valued signals and filters, the above equation can be modified to exploit the conjugate symmetry in the frequency domain.

Using the formulation of [16,17] a desired impulse response $\boldsymbol{d}_{d_{ij}}$ can now be expressed as

$$\boldsymbol{d}_{d_{ij}} = \mathrm{diag}(\boldsymbol{\gamma}_{d_{ij}}) \cdot \boldsymbol{V}_{ij} \cdot \boldsymbol{c}_j \tag{9}$$

with $\boldsymbol{\gamma}_{d_{ij}} = [\gamma_{d_{ij}}(0), \gamma_{d_{ij}}(1), \ldots, \gamma_{d_{ij}}(M-1)]$ a vector with the desired shape of the unmixing filter. Here we use a two-sided exponentially decaying window

$$\gamma_{d_{ij}}(n) = \begin{cases} 10^{q_1(n_o - n)} & \text{for } 0 \le n \le n_0 \\ 10^{q_2(n - n_o)} & \text{for } n_0 \le n \end{cases} \tag{10}$$

with n_0 being the position of the maximum of $|\boldsymbol{w}_{ij}|$. The factors q_1 and q_2 have been chosen heuristically as $q_1 = -0.1$ and $q_2 = -0.05$.

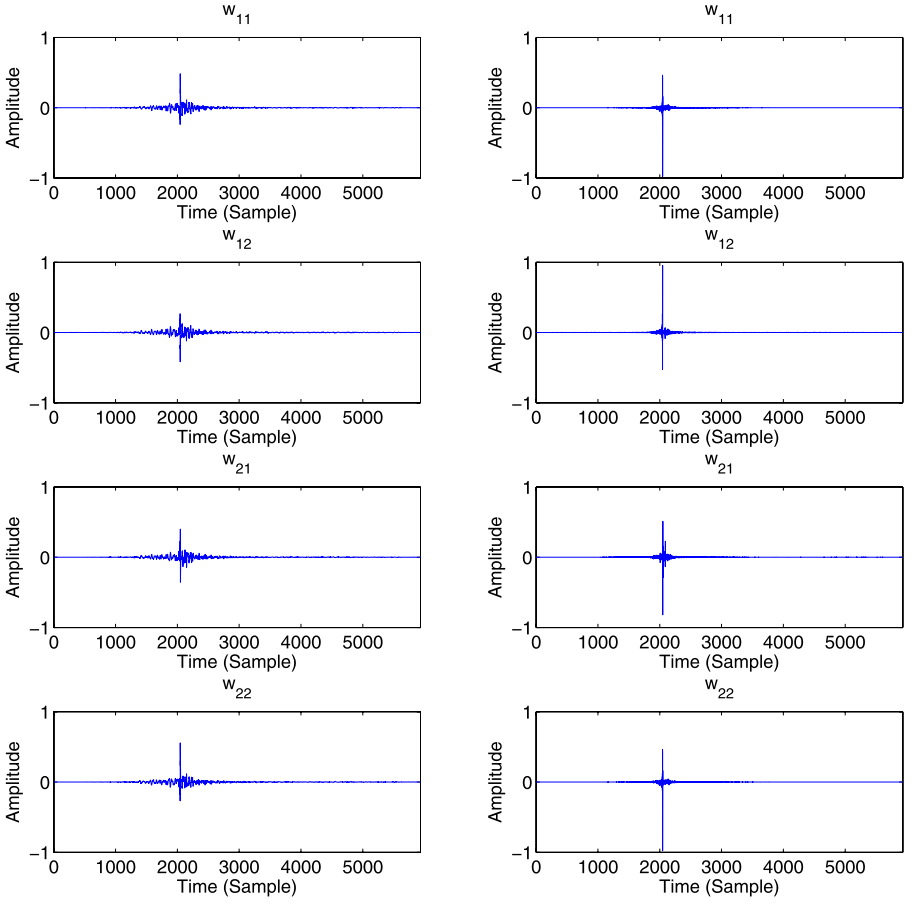

Fig. 3. Comparison of filter sets using the minimal distortion principle (left) and the new method (right)

An undesired part is formulated analogously as

$$\boldsymbol{d}_{u_{ij}} = \mathrm{diag}(\boldsymbol{\gamma}_{u_{ij}}) \cdot \boldsymbol{V}_{ij} \cdot \boldsymbol{c}_j \tag{11}$$

with

$$\gamma_{u_{ij}}(n) = \begin{cases} 10^{q_3(n_o-n)} & \text{for } 0 \le n \le n_0 \\ 10^{q_4(n-n_o)} & \text{for } n_0 \le n \end{cases} \tag{12}$$

Here the factors have been chosen heuristically to be $q_3 = 0.001$ and $q_4 = 0.0005$.

As the filters corresponding to the same output channel have the same scaling factors, $\boldsymbol{c}_j(\omega)$ has to be optimized simultaneously for these filters. Therefore $\boldsymbol{V}_{ij}$ corresponding to the same output j are stacked to $\bar{\boldsymbol{V}}_j$. The same applies for $\boldsymbol{\gamma}_{d_{ij}}$ and $\boldsymbol{d}_{d_{ij}}$ which are stacked into $\bar{\boldsymbol{\gamma}}_{d_j}(n)$ and $\bar{\boldsymbol{d}}_{d_j}$ respectively.

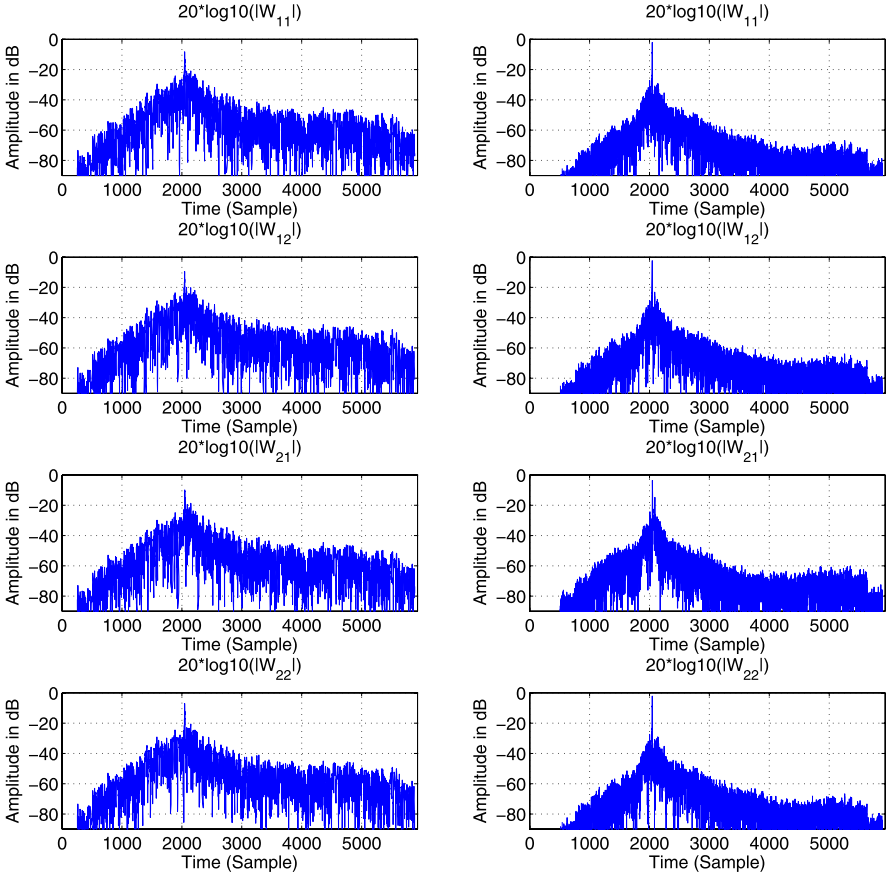

Fig. 4. Magnitudes of filters designed via the minimal distortion principle (left) and the new method (right)

Now the matrices $\bar{\boldsymbol{A}}$ and $\bar{\boldsymbol{B}}$ can be calculated as in [18]:

$$\boldsymbol{d}_u^H \boldsymbol{d}_u = \boldsymbol{d}_{u_j}^H \cdot \bar{\boldsymbol{V}}_j^H \cdot \mathrm{diag}(\boldsymbol{\gamma}_{u_j}^H) \cdot \mathrm{diag}(\boldsymbol{\gamma}_{u_j}) \cdot \bar{\boldsymbol{V}}_j \cdot \boldsymbol{d}_{u_j} = \boldsymbol{d}_{u_j}^H \cdot \bar{\boldsymbol{A}} \cdot \boldsymbol{d}_{u_j} \quad (13)$$

$$\boldsymbol{d}_d^H \boldsymbol{d}_d = \boldsymbol{d}_{d_j}^H \cdot \bar{\boldsymbol{V}}_j^H \cdot \mathrm{diag}(\boldsymbol{\gamma}_{d_j}^H) \cdot \mathrm{diag}(\boldsymbol{\gamma}_{d_j}) \cdot \bar{\boldsymbol{V}}_j \cdot \boldsymbol{d}_{d_j} = \boldsymbol{d}_{d_j}^H \cdot \bar{\boldsymbol{B}} \cdot \boldsymbol{d}_{d_j} \quad (14)$$

Finally, the optimal scaling factors $\boldsymbol{c}_{opt}$ are the solution of the generalized eigenvalue problem [18]

$$\bar{\boldsymbol{B}} \cdot \boldsymbol{c}_{opt} = \bar{\boldsymbol{A}} \cdot \boldsymbol{c}_{opt} \cdot \lambda_{max} \quad (15)$$

with λ_{max} being the largest eigenvalue and $\boldsymbol{c}_{opt}$ being the corresponding eigenvector.

4 Simulations

Simulations have been done on real-world data available at [19]. This data set consists of eight-seconds long speech recordings sampled at 8 kHz. The chosen parameters were a Hann window of length 2048, a window shift of 256, and an FFT-length of $K = 4096$. Every frequency bin has been separated using 200 iterations of the gradient-based rule from [1]. As the original sources are available for the considered data set, the permutation problem has been ideally solved, so that permutation ambiguities do not influence the results, and the scaling problem can be studied exclusively.

Table 1. Comparison of the signal-to-interference ratios in dB between the minimal distortion principle and the new algorithm

	Left	Right	Overall
MDP	16.07	16.72	16.41
New Alg.	24.88	28.81	26.75

In Figs. 3 and 4 the filters designed with the traditional method (7) and the proposed method are shown, respectively. The main difference is the clearly visible and significantly bigger main peak and the faster decay of the impulse responses designed with our method. As one can observe by comparing Fig. 4, the energy difference between the main peak and the tail of the impulse response could be increased by about 25 dB.

The new filters are also able to significantly enhance the separation performance as shown in Table 1.

5 Conclusions

In this paper, we have proposed the use of the scaling ambiguity of convolutive blind source separation for shaping the unmixing filters. We calculate a set of scaling factors that shape exponentially decaying impulse responses with less reverberation. On a real-world example, the energy decay could be improved by 25dB, which also translated into better signal-to-interference ratios.

References

1. Amari, S., Cichocki, A., Yang, H.H.: A new learning algorithm for blind signal separation. In: Touretzky, D.S., Mozer, M.C., Hasselmo, M.E. (eds.) Advances in Neural Information Processing Systems, vol. 8, pp. 757–763. MIT Press, Cambridge (1996)
2. Hyvärinen, A., Oja, E.: A fast fixed-point algorithm for independent component analysis. Neural Computation 9, 1483–1492 (1997)

3. Cardoso, J.F., Soulomiac, A.: Blind beamforming for non-Gaussian signals. Proc. Inst. Elec. Eng., pt. F. 140(6), 362–370 (1993)
4. Douglas, S.C., Sawada, H., Makino, S.: Natural gradient multichannel blind deconvolution and speech separation using causal FIR filters. IEEE Trans. Speech and Audio Processing 13(1), 92–104 (2005)
5. Aichner, R., Buchner, H., Araki, S., Makino, S.: On-line time-domain blind source separation of nonstationary convolved signals. In: Proc. 4th Int. Symp. on Independent Component Analysis and Blind Signal Separation (ICA 2003), Nara, Japan, pp. 987–992 (April 2003)
6. Smaragdis, P.: Blind separation of convolved mixtures in the frequency domain. Neurocomputing 22(1-3), 21–34 (1998)
7. Rahbar, K., Reilly, J.: A frequency domain method for blind source separation of convolutive audio mixtures. IEEE Trans. Speech and Audio Processing 13(5), 832–844 (2005)
8. Anemüller, J., Kollmeier, B.: Amplitude modulation decorrelation for convolutive blind source separation. In: Proceddings of the second international workshop on independent component analysis and blind signal separation, pp. 215–220 (2000)
9. Mazur, R., Mertins, A.: Solving the permutation problem in convolutive blind source separation. In: Davies, M.E., James, C.J., Abdallah, S.A., Plumbley, M.D. (eds.) ICA 2007. LNCS, vol. 4666, pp. 512–519. Springer, Heidelberg (2007)
10. Sawada, H., Mukai, R., Araki, S., Makino, S.: A robust and precise method for solving the permutation problem of frequency-domain blind source separation. IEEE Trans. Speech and Audio Processing 12(5), 530–538 (2004)
11. Wang, W., Chambers, J.A., Sanei, S.: A novel hybrid approach to the permutation problem of frequency domain blind source separation. In: Puntonet, C.G., Prieto, A.G. (eds.) ICA 2004. LNCS, vol. 3195, pp. 532–539. Springer, Heidelberg (2004)
12. Mukai, R., Sawada, H., Araki, S., Makino, S.: Blind source separation of 3-d located many speech signals. In: 2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 9–12 (October 2005)
13. Ikeda, S., Murata, N.: A method of blind separation based on temporal structure of signals. In: Proc. Int. Conf. on Neural Information Processing, pp. 737–742 (1998)
14. Matsuoka, K.: Minimal distortion principle for blind source separation. In: Proceedings of the 41st SICE Annual Conference, vol. 4, pp. 2138–2143, August 5-7 (2002)
15. Sawada, H., Mukai, R., de la Kethulle, S., Araki, S., Makino, S.: Spectral smoothing for frequency-domain blind source separation. In: International Workshop on Acoustic Echo and Noise Control (IWAENC 2003), pp. 311–314 (September 2003)
16. Arslan, G., Evans, B.L., Kiaei, S.: Equalization for discrete multitone transceivers to maximize bit rate. IEEE Trans. on Signal Processing 49(12), 3123–3135 (December)
17. Kallinger, M., Mertins, A.: Multi-channel room impulse response shaping - a study. In: Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing, Toulouse, France, vol. V, pp. 101–104 (May 2006)
18. Melsa, P., Younce, R., Rohrs, C.: Impulse response shortening for discrete multitone transceivers. IEEE Trans. on Communications 44(12), 1662–1672 (1996)
19. `http://www.kecl.ntt.co.jp/icl/signal/sawada/demo/bss2to4/index.html`