Audio Engineering Society

Convention Paper 8189

Presented at the 129th Convention
2010 November 4–7 San Francisco, CA, USA

Determined Source Separation for Microphone Recordings Using IIR Filters

Christian Uhle[1], Josh Reiss[2]

[1] *Fraunhofer Institute for Integrated Circuits, Erlangen, Germany*

[2] *Center for Digital Music, Queen Mary University of London, London, UK*

Correspondence should be addressed to Christian Uhle (christian.uhle@iis.fraunhofer.de)

ABSTRACT

A method for determined blind source separation for microphone recordings is presented which attenuates the direct path cross-talk using IIR filters. The unmixing filters are derived by approximating the transmission paths between the sources and the microphones by a delay and a gain factor. For the evaluation, the proposed method is compared to three other approaches. Degradation of the separation performances are caused by fractional delays and the directivity of microphones and sources, which are discussed here. Advantages are low latency, low computational complexity and high sound quality.

1. INTRODUCTION

Blind source separation (BSS) is the task of recovering the latent source signals given observations of audio, sonar, radio, biological (such as EEG and MEG) or other signals. For convolutive mixtures of speech signals this problem is often referred to as the *cocktail party problem*, which derived its name from a classical example of an auditory scene with multiple sources. Applications of BSS of microphone signals are automated speech recognition (ASR), communication (e.g. tele-conferencing), hearing aids, forensics, audio coding and music production. Different approaches to this problem exist. The following overview of prior work focuses on the separation of *convolutive* mixtures, i.e. the observations are the sum of multiple differently delayed versions of the source signals, and *determined* mixtures, i.e. the number of microphones equals the number of sources.

A widely used approach to BSS is Independent Component Analysis (ICA), as introduced by Herault, Jutten and Ans, and formulated by Comon [1] for instantaneous linear mixtures. For an overview on ICA one is referred to [2, 3, 4]. ICA estimates source signals given the observations of a mixing process under the assumptions that the source signals are statistically independent and have non-gaussian distributions. It has been shown that an unmixing process which optimizes independence criteria of the output signal (and minimizing the mutual information) can restore the source signals since the mixing results in signals with more mutual dependency than the underlying source signals. The results are obtained with ambiguities of scaling and permutation. Given the above assumptions, the sources can also be estimated by maximizing criteria of non-gaussianity (e.g. quantified using the kurtosis, approximations of negentropy), since according to the Central Limit Theorem any mixture will be closer to gaussian than the sources.

From this it can be concluded that ICA methods have high computational load (due to the numerical optimization of the criteria, typically making heavy use of non-linear functions), and make certain assumptions about the source signals (e.g. statistical independence, non-gaussianity). The statistical criteria are computed from a representative portion of the input signals.

ICA has been originally proposed for instantaneous mixtures [1]. The separation of multiple speakers recorded by multiple microphones poses the problem of *convolutive* mixtures, since the source signals arrive at the microphones with a time delay, and, besides the direct path cross-talk, the microphones will also capture the room reflections, a multitude of time-delayed versions of the same source (and, of course, background noise).

Weinstein et.al. investigated a determined BSS method for two sources using a recursive structure based on the assumption of decorrelation of the source signals [5] and concluded that this criterion alone is not sufficient. Recursive unmixing systems were also investigated in [6, 7]. Thi and Jutten [6] presented a method based on criteria of statistical independence and yielded an attenuation of the interfering source of 20 dB for synthetic mixtures of two sources but less separation for real recordings. Lee et.al. also derive learning rules from independence criteria (information maximization, maximum likelihood and negentropy) and showed an improvement in the recognition performance of an ASR system when using the BSS as a pre-processing step [8].

The generic framework TRINICON for adaptive multiple-input/multiple-output processing for applications to BSS, dereverberation and parameter estimation has been presented in [9]. This unified approach exploits the three fundamental statistical source properties, non-gaussianity, non-whiteness and non-stationarity. An extensive review of ICA and related methods is out of the scope of this paper, the following two methods are mentioned because they were used for comparison with the presented method. Both methods process the signals in the frequency domain using the Short-Term Fourier Transform. This allows more efficient implementations than its time domain counterparts and requires a solution to deal with permutation indeterminacies that appear from different frequency bins. Parra and Spence [10] presented a frequency-domain method for microphone recordings (determined mixing) exploiting the non-stationarity of the source signals. The method is based on decorrelation at multiple times for the instantaneous case and extended to convolutive mixtures by solving the separation for each frequency. They reported a crosstalk attenuation of up to 14 dB for real-world signals. A MATLAB implementation of the algorithm has been implemented and made publicly available by Harmeling [11]. Mitianoudis and Davies have developed another frequency domain method for determinend BSS based on ICA [12], where permutation problem is addressed by means of frequency coupling in the source model.

Another well-known approach to separate sources from two observations is based on the estimation of inter-channel level differences (ICLD) and inter-channel time differences (ICTD) or inter-channel phase differences (ICPD) in each time-frequeny bin. These cues play an important role for binaural human hearing as formulated in the Duplex Theory [13, 14] and spatial audio processing applications [15].

The source signals are separated using a spectral weighting (as in speech enhancement using spectral subtraction [16] or Wiener filtering, or with binary weights, also referred to as binary masking). The spectral weights used to retrieve a source signal are computed such that time-frequency bins with similar ICLD and ICLD are weighted with similar weights. Typically, the spectral weights are real-valued numbers, i.e. the magnitude spectrogram is modified and the original phase of the input signal is applied for the synthesis of the output time signal. These methods rely to varying extent on W-disjoint orthogonality of the sources. The condition of W-disjoint orthogonality states that the source signals do not overlap in the time-frequency representation [17, 18].

A very prominent method following this approach is the Degenerate Unmixing Estimation Technique (DUET) [17, 19]. The separation is achieved by clustering the time-frequency bins into sets with similar ICLD and ICTD and binary masking. Although this method is based on the assumption of anechoic recordings of W-disjoint orthogonal sources it can also cope with a small degree of reverberation and approximately W-disjoint orthogonality. A restriction of the original method is that the maximum frequency which can be processed due to phase ambiguities equals half the speed of sound over maximum microphone spacing, which has been addressed in [20].

Other methods based on ICLD and ICTD are

- the Modified ADRess algorithm [21], which extends the Azimuth Discrimination and Resynthesis (ADRess) [22] algorithm for the processing of microphone signals
- the method based on time-frequency correlation for time-delayed mixtures (AD-TIFCORR) [23]
- Direction Estimation of Mixing Matrix (DEMIX) for anechoic mixtures [24], which includes a confidence measure that only one source is active at a particular time-frequency bin
- Model-based Expectation-Maximization Source Separation and Localization (MESSL) [25]
- methods mimicking the binaural human hearing mechanism as in e.g. [26, 27]

Furthermore, methods have been proposed which assume specific microphone settings. A method for modifying a stereo microphone recording using the acoustic information obtained with a spot microphone (i.e. a microphone close to the source of interest) has been presented by Faller and Erne [28]. An estimate of the impulse response between the spot microphone and the stereo microphone is derived and used to identify the signal components of the source of interest in the stereo recording for further modification.

Kokkinis and Mourjopoulos [29] compared Wiener Filtering to the method by Parra and Spence [10, 11] for recordings with two microphones and two sources. The estimate of the noise power spectral density is derived directly from the microphone which is close to the interfering source. They conclude that the method based on Wiener Filtering yielded better results with respect to separation and sound quality with less computational load.

Finally, it should be mentioned that the source separation problem is closely related to the attenuation or amplification of a source within a mixture, as for example in speech enhancement (see [30] for a comprehensive recent review), especially if one considers the fact that the state-of-the-art methods are in general not able to achieve perfect separation, at least for real-world signals.

This paper is organized as follows: Section 2 describes the proposed method. Closely related to the source separation task is the problem of comb-filtering when mixing two or more microphone signals containing differently delayed source signals. An efficient solution is derived from the presented source separation method as explained in Section 3. Section 4 presents experiments, results and a discussion of the sources of error. Finally, Section 5 gives the conclusions.

2. PROPOSED METHOD FOR DETERMINED BLIND SOURCE SEPARATION

Consider the situation where Q audio source signals s_q are recorded by P microphones, resulting in observations x_p. Each microphone signal is the sum of filtered versions of the source signals and background noise, which will be neglected in the following. The transmission paths between the source q and the microphone p can be modeled by impulse responses h_{qp}.

This can be expressed in the z-domain in matrix notation as

$$\mathbf{X}(z) = \mathbf{H}(z)\mathbf{S}(z) \tag{1}$$

with

$$\mathbf{X}(z) = \begin{pmatrix} X_1(z), & \dots, & X_P(z) \end{pmatrix}^T \tag{2}$$

$$\mathbf{H}(z) = \begin{pmatrix} H_{11}(z), & \dots, & H_{1Q}(z) \\ \vdots & \ddots & \vdots \\ H_{P1}(z), & \dots, & H_{PQ}(z) \end{pmatrix} \tag{3}$$

$$\mathbf{S}(z) = \begin{pmatrix} S_1(z), & \dots, & S_Q(z) \end{pmatrix}^T \tag{4}$$

Here and in the following the determined case is assumed, where $Q = P$. For recovering the source signals the unmixing system $\mathbf{W}$ in Equation (5) is used.

$$\mathbf{Y}(z) = \mathbf{W}(z)\mathbf{X}(z) = \mathbf{W}(z)\mathbf{H}(z)\mathbf{S}(z) := \widehat{\mathbf{S}}(z) \tag{5}$$

The mixing system and the unmixing system are shown in Figure 1 for $Q = 2$.

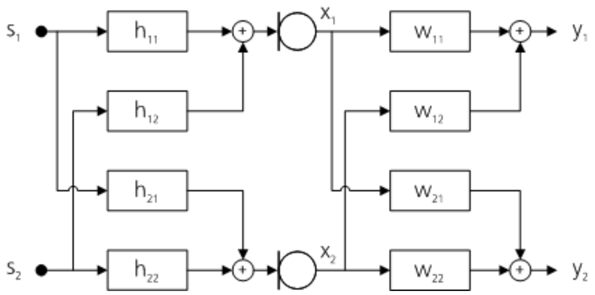

Fig. 1: Mixing system and unmixing system for 2 sources and 2 microphones.

The presented source separation method is derived from the general solution of the unmixing system and by assuming the simplified scenario of two microphones and two sources in anechoic conditions. Source separation in anechoic conditions deals with delayed mixtures and aims at the attenuation of the direct path cross-talk of the interfering sources. The free-field assumption enables the proposed solution to the BSS problem, as will be shown in the following section. In reverberant conditions, the presented method attenuates the interfering sources to varying degree depending on the amount of reverberation or direct to reverberation ratio, i.e. it will yield more improvements of the signal to interference ratio the shorter the reverberation time is and the smaller the distances between sources and microphones are.

The unmixing system in Equation (5) is derived by inverting the mixing system $\mathbf{H}$.

$$\mathbf{W}(z) = \mathbf{H}^{-1}(z) \tag{6}$$

The solution according to Cramer's rule leads to the unmixing system as shown in Equation (7).

$$\mathbf{W}(z) = \frac{1}{\det \mathbf{H(z)}} \begin{pmatrix} H_{22}(z) & -H_{12}(z) \\ -H_{21}(z) & H_{11}(z) \end{pmatrix} \tag{7}$$

with the determinant of the mixing system

$$\det \mathbf{H}(z) = H_{11}(z)H_{22}(z) - H_{12}(z)H_{21}(z) \tag{8}$$

Assuming freefield conditions, ideal omnidirectional radiation patterns of the sources and ideal omnidirectional directivity patterns of the microphones, each of the mixing filters can be approximated as shown in Equation (9) as a delay τ_{pq} and a scaling factor α_{pq}.

$$H_{pq}(z) = \alpha_{pq} z^{-\tau_{pq}} \tag{9}$$

This solution requires knowledge of all α_{pq} and τ_{pq}. The same solution is obtained by using the ICTD and ICLD instead of the absolute values, such that only two parameters need to be estimated for each source.

It can be seen that since $\mathbf{H}(z)$ is an FIR system, the elements of $\mathbf{W}(z)$ are IIR filters (more precisely, they are feedback comb-filters) whose feedback coefficients are derived from the determinant of $\mathbf{H}(z)$. This solution can be interpreted as an extension of phase cancellation to multiple sources. Phase cancellation can be applied to remove an interfering signal from a mixture if a clean observation of the interferer is available. In other words, if only one source is active, it can ideally be removed from one microphone signal by subtracting a delayed and scaled version of the other microphone signal. In the case considered here (where all observation are mixtures of all source signals) the basic principle of phase cancellation can be used by incorporating the feedback path of the unmixing filters.

An example of the unmixing filters to recover $s_1(n)$ is shown in Figures 2 and 3, for an examplary mixing system described by a matrix of delays $\mathbf{T} = \left(\begin{smallmatrix} 19 & 49 \\ 51 & 17 \end{smallmatrix}\right)$ and a matrix of gains $\mathbf{A} = \left(\begin{smallmatrix} 0.89 & 0.74 \\ 0.72 & 0.96 \end{smallmatrix}\right)$.

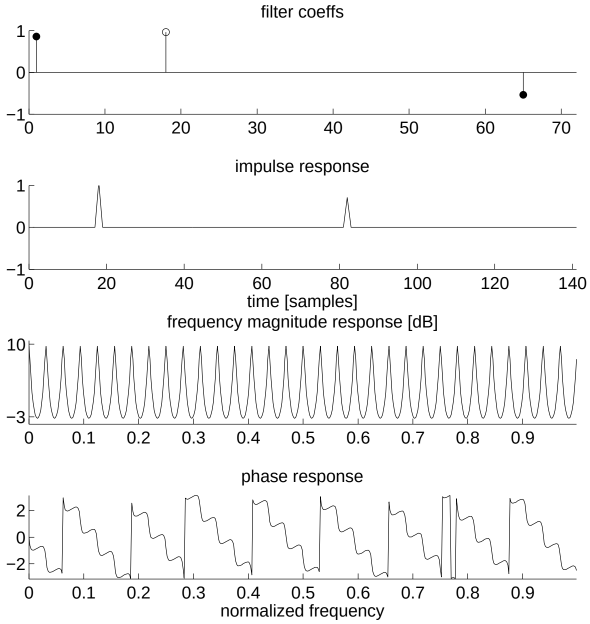

Fig. 2: Example for unmixing filter $W_{11}(z)$. Filter feedforward (circle) and feedback (filled) coefficients, impulse, frequency and phase response (from top to bottom).

It is shown that

- the feedback coefficients are the same for both filters
- only 3 coefficients of each filter are non-zero, which enables a computational efficient implementation

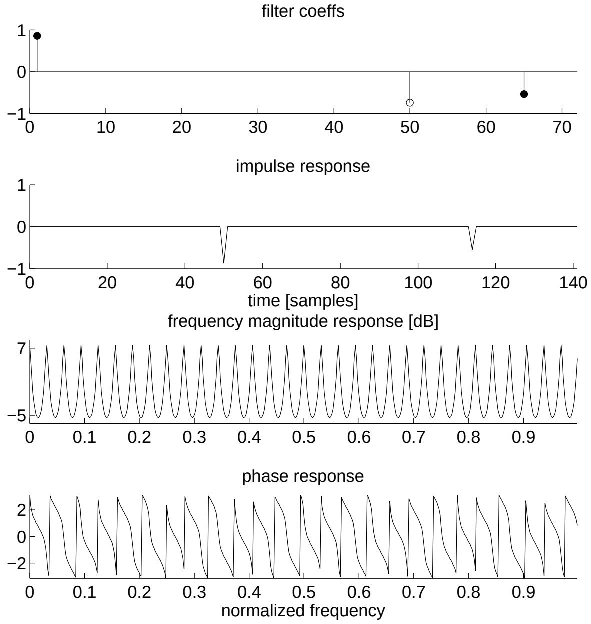

Fig. 3: Example for unmixing filter $W_{12}(z)$. Filter feedforward (circle) and feedback (filled) coefficients, impulse, frequency and phase response (from top to bottom).

- the frequency magnitude response has a comb-filter characteristic

For the stability of the unmixing filters it is required that $\alpha_{11}\alpha_{22} > \alpha_{12}\alpha_{21}$, otherwise the poles are not inside the unit circle. This condition is satisfied in the free field for a typical microphone set-up where e.g. $\tau_{11} < \tau_{12}$ and $\tau_{22} < \tau_{21}$, i.e. source 1 is closer to microphone 1 than microphone 2 and source 2 is closer to microphone 2 than microphone 1, due to the attenuation of sound when traveling in air. The SPL decreases by ΔL when increasing the distance from τ_1 to τ_2, e.g. for point sources according to Equation (10).

$$\Delta L = 20 \log_{10} \frac{\tau_1}{\tau_2} \qquad (10)$$

It is worth noticing the following features of this approach:

- The adaption method does not depend on signal characteristics like non-gaussianity, non-whiteness or non-stationarity. It can therefore robustly deal with any type of signals for whose the parameters can be estimated robustly.

- The unmixing system is an LTI system if the position of the sources and microphones do not change. It can be analyzed with well-known tools and is easy to implement.
- The unmixing method relies on a small number of parameters which makes it computationally efficient.
- Unlike many other methods it does not use spectral weighting which often leads to artifacts like musical noise.

2.1. Parameter Estimation

The performance of the presented method relies on a robust estimation of the parameters α_{pq} and τ_{pq}. For a wide range of signals, the Generalized Cross-Correlation (GCC) using the phase transform (PHAT) introduced by Knapp and Carter [31] gives robust estimates of the ICTD for a wide range of input signals. The ICTD is found as the time-lag for which the weighted cross-correlation function $R_{12}(\tau)$ between the microphone signals is at a maximum, where the weighted cross-correlation function is computed as the inverse Fourier transform of the phase of the cross-spectrum.

$$\tau_{12} = \arg\max_{\tau} R_{12}(\tau) \tag{11}$$

with

$$R_{12}(\tau) = \sum_{k=0}^{N-1} \frac{X_1(\omega) X_2(\omega)^*}{|X_1(\omega)|\,|X_2(\omega)|} e^{\frac{2\pi k n j}{N}} \tag{12}$$

The parameter estimation of the DUET method can potentially yield both, estimates for the ICTD and the ICLD. The processing is done in the frequency domain using the Short-term Fourier Transform. The parameters $\alpha_{12}(\omega)$ and $\tau_{12}(\omega)$ are estimated from the ratios of the short-term spectra according to Equations (13) and (14) for each frequency bin.

$$\alpha_{12}(\omega) = \left| \frac{X_1(\omega)}{X_2(\omega)} \right| \tag{13}$$

$$\tau_{12}(\omega) = -\frac{1}{\omega} \angle \left(\frac{X_1(\omega)}{X_2(\omega)} \right) \tag{14}$$

In realistic cases where the W-disjoint orthogonality of the sources is only approximately fulfilled and the signals are recorded in reverberant environments, only a small number of time-frequency bins yield correct parameter values.

2.2. Details of the Implementation

In this early stage of the project the method assumes that the number of sources equals the number of microphones. Although the parameter estimation can in principle be performed on-line, its current implementation estimates the parameter beforehand, i.e. a dual path process is performed where in a first stage the parameters (ICLD and ICTD) are estimated and in a second stage the separation is performed in real-time. The current implementation of the method assumes time-invariant parameters. It should be noted that this is not a restriction of the source separation method, but of its current implementation. At the current stage of this project, a reliable parameter estimation working on-line and smooth transitions between filter coefficients are not implemented yet.

The described dual-path processing allows for an assessment of the separation method independently of the robustness of the parameter estimation. In general, the parameter estimation is an integral part of many BSS methods and as such it is difficult to decouple it from the separation in all cases. Here it is feasible and useful due to the facts that the number of parameters is very small and there are various applications for and approaches to the estimation of the parameters needed here. For example, a real-time implementation of DUET is described in [32] which performs the parameter estimation per time-frequency bin. ICLD estimation for reducing the comb-filtering in down-mixing of multi-track signals is described in [33].

Furthermore, to cope with moving sources a smooth transition between the unmixing filters is required, which is not addressed here.

3. MULTISOURCE COMB-FILTER COMPENSATION

In the following section the problem of comb-filtering during downmixing is addressed. Considering the mixing system in Figure 1, the microphone signals can be written as

$$X_1(z) = \alpha_{11} z^{-\tau_{11}} S_1(z) + \alpha_{12} z^{-\tau_{12}} S_2(z) \qquad (15)$$

$$X_2(z) = \alpha_{21} z^{-\tau_{21}} S_1(z) + \alpha_{22} z^{-\tau_{22}} S_2(z) \qquad (16)$$

Adding both microphone signals yields a combined signal which (if $\tau_{11} \neq \tau_{21}$ and $\tau_{12} \neq \tau_{22}$) is the sum of the comb-filtered source signals. If only one source is active the comb-filtering can be eliminated by delaying the microphone signal which picks up the respective source signal first by the ICTD. If both sources are active the source separation together with appropriate delay compensation prior to downmixing is capable of preventing the sum signal from having the (typically undesired) comb-filtering.

Another solution without separation of the source signals with the benefit of less computational load is presented in the following. The sum signal $X_s(z)$ without comb-filter artifacts is derived by applying compensation filters $G_n(z)$ to each microphone signal $X_n(z)$ prior to downmixing. This processing is in the following called Comb-Filter Compensation (CFC).

$$X_s(z) = G_1(z) X_1(z) + G_2(z) X_2(z) \qquad (17)$$

The compensation filters $G_n(z)$ are derived similarily to the unmixing filters in the z-domain. The sum signal from Equation (17) can be written as

$$\begin{aligned} X_s(z) &= G_1(z)\left(H_{11}(z) S_1(z) + H_{12}(z) S_2(z)\right) \\ &\quad + G_2(z)\left(H_{22}(z) S_2(z) + H_{21}(z) S_1(z)\right) \\ &\equiv S_1(z) + S_2(z) \end{aligned}$$

This leads to the system of equations

$$\mathbf{H}(z)^T \mathbf{G}(z) = \begin{pmatrix} 1 & 1 \end{pmatrix}^T \qquad (18)$$

with $\mathbf{G}(z) = \begin{pmatrix} G_1(z) & G_2(z) \end{pmatrix}^T$.

The solution for the compensation filters is derived from the solution of the system of equations in (18) according to Cramer's rule as

$$G_1(z) = \frac{H_{22}(z) - H_{21}(z)}{\det(H)} \qquad (19)$$

$$G_2(z) = \frac{H_{11}(z) - H_{12}(z)}{\det(H)} \qquad (20)$$

Similar to the unmixing filters, this solution can be generalized to any determined mixing system with more microphones and sources.

Examples for compensation filters for the mixing system used in the previous section are illustrated in Figures 4 and 5.

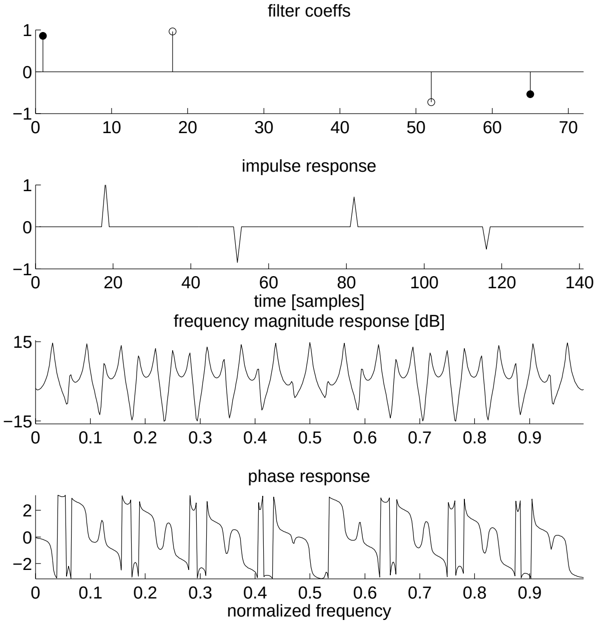

Fig. 4: Example for compensation filter $G_1(z)$. Filter feedforward (circle) and feedback (filled) coefficients, impulse, frequency and phase response (from top to bottom).

4. EVALUATION

This section gives an evaluation of the BSS and CFC methods using synthetic and real-world recordings. Subsequently, sources of error are identified and their impact is analyzed. The synthetic mixtures are created using a software package for simulating microphone recordings in rooms using the image-source method [34]. This method also gives

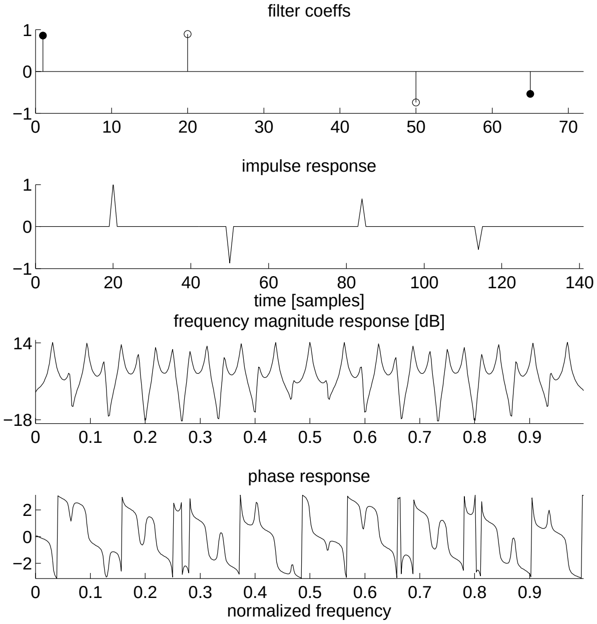

Fig. 5: Example for compensation filter $G_2(z)$. Filter feedforward (circle) and feedback (filled) coefficients, impulse, frequency and phase response (from top to bottom).

access to the impulse responses of each mixing filter, which can be used to measure the parameters gain and delay with high precision. These parameters can then be used to investigate the method independently from the parameter estimation. The recordings with real microphones in real rooms were done in the listening room of the Centre for Digital Music at the Queen Mary University of London, England, and in the listening room at the Fraunhofer IIS in Erlangen, Germany. The acoustical properties of the latter room are compliant with the recommendation ITU-R BS.1116-1 [35].

4.1. Synthetic Signals

Figure 6 illustrates two sawtooth time signals which were mixed without artificial reverberation and separated using the described method, where the parameters were estimated from the impulse responses between the sources and the microphones. It is shown that very high separation is achieved for synthetic mixtures with no reverberation if the parameters are estimated correctly.

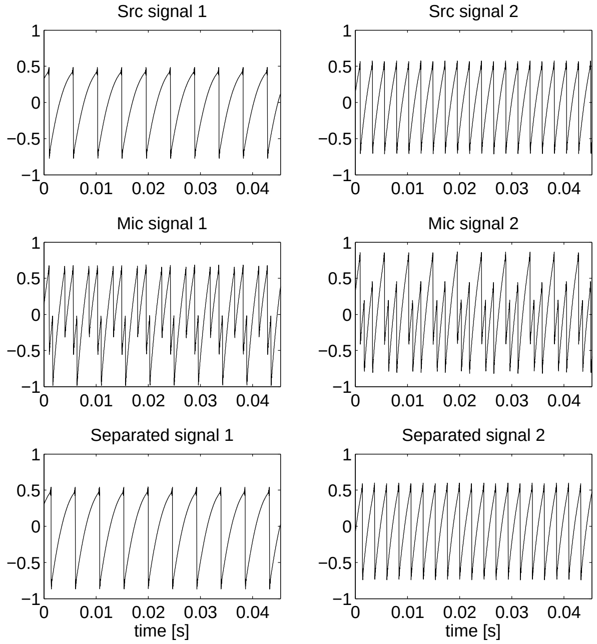

Fig. 6: Source signals, microphone signals and separated signals (from top to bottom).

It comes as no surprise that if the separation works successfully, the mixdown using comb-filter compensation yields identical results compared to a mixdown of the source signals, as shown in Figure 7. Listening to the unprocessed mixdown of microphone recordings of different signals (e.g. speech, musical instruments) revealed strong comb-filter artifacts. This effect is eliminated completely by using the proposed CFC method.

Since the presented method considers the direct path crosstalk only, its advantages are reduced the higher the amount of reverberation in the recorded signals. However, listening to processed recordings of speech signals simulated with different reverberation times shows that even for very reverberant recordings the intelligibility of the speech benefits from the source separation.

4.2. Microphone Recordings

The evaluation with real-world microphone recordings typically needs to be performed without having access to the reference source signal. But if the positions of sources and microphones do not change, the separation method is an LTI system and can be

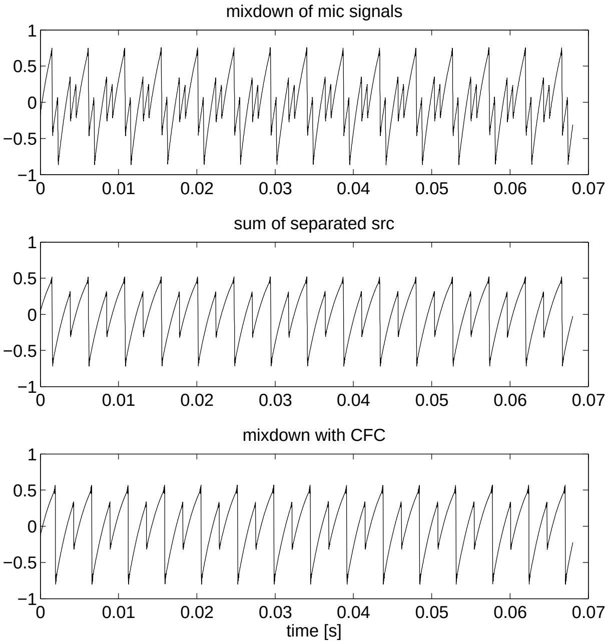

Fig. 7: Mixdown of microphone signals, mixdown of source signals with delay compensation (for reference), and mixdown using the proposed comb-filter compensation (from top to bottom).

analyzed as such. Tranfer functions characterizing the transmission of the interfering and the desired sources can be derived by activating only one source at a time. The attenuation of the active source in the microphone which is intended to capture the other source (for this microphone the active source will then be referred to as interfering source) will be the same as if the other source would be active at the same time. The transfer function of the desired source is derived in the same way.

Microphone signals were recorded with two human speakers using two omnidirectional small diaphragm condenser microphones DPA 4006. Results for a processed example are shown in Figures 8 and 9. The speakers were asked not to move. Small deviations from an average delay can be observed during parameter estimation as shown in Figure 10 indicating small movements of the speakers.

The separation performance of the presented method can be degraded due to

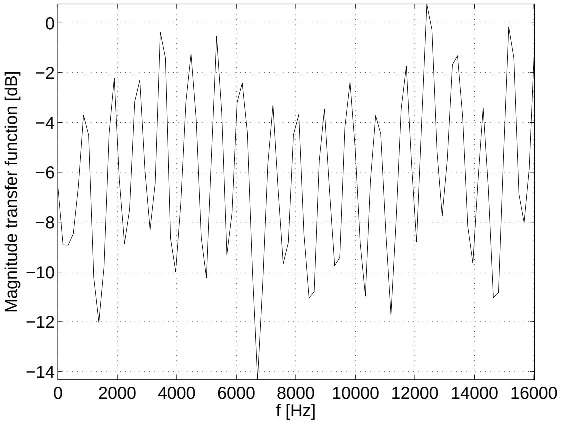

Fig. 8: Frequency magnitude response of the attenuation of the interfering source.

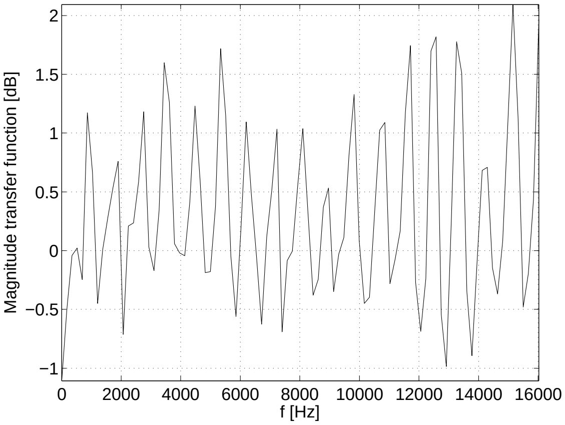

Fig. 9: Frequency magnitude response of the transmission of the desired source.

- Reverberation
- Errors in the estimation of the parameter (i.e. delays and gains)
- Fractional delays
- Frequency dependence of the directivity patterns of the microphones and the radiation patterns of the sources

The influence of parameter estimation errors and fractional delays are analyzed in the following.

4.3. Listening Test

A listening test was performed for comparison of the proposed method to three previous methods, namely

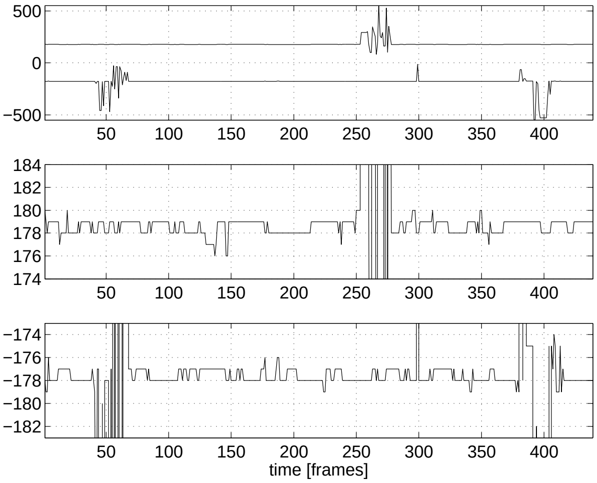

Fig. 10: Exemplary results of the delay estimation for both sources (top) and for each of the sources separately.

DUET, the method by Parra and Spence (referred to as FD-ICA1) and the method by Mitianoudis and Davies (referred to as FD-ICA2). Additionally, the unprocessed microphone signal (from the microphone closer to the desired source) was presented. The reference signal contained only the desired source. In the case of a real microphone recording, the recording was repeated with having only the desired source active.

Four items were presented:

- one recording as described in the previous section with human speakers
- a synthetic mixture of delayed speech signals
- a mixture of two string instruments using the room simulation
- a mixture of two human speakers using the room simulation

The levels of all signals were adjusted to have equal loudness according to ITU-R BS1770. The test was very similar to a MUSHRA test, with the difference that no lower anchor has been used. The listeners were asked to rate the "perfomance of the separation" by taking the degree of attenuation of the interfering source and the processing artifacts into account. The signals were presented using headphones. Figure 11 shows the combined ratings of all 13 listeners using the median and 95% confidence interval, for each item separately and combined for all items. The combined result for the proposed method is better than for any of the other methods. For the one microphone recording, FD-ICA1 shows the highest median of all ratings, although no statistical significant difference can be observed between the best methods.

4.4. Parameter Estimation Errors

For this purpose, the unmixing filters $W_{pq}(z)$ are derived from mixing filters $H_{pq}(z)$ which differ from the real mixing filters $\widetilde{H}_{pq}(z)$ due to one or more of the error sources mentioned above [1].

Then, the unmixed signals will be the sum of both filtered source signals

$$\mathbf{Y}(z) = \mathbf{V}(z)\mathbf{S}(z) \tag{21}$$

with

$$\mathbf{V}(z) = \widetilde{\mathbf{W}}(z)\mathbf{H}(z) \tag{22}$$

For example if $Q = 2$ then $y_1(n)$ is the sum of the desired signal $s_1(n)$ filtered with $V_{11}(z)$ and of the interfering signal $s_2(n)$ filtered with $V_{12}(z)$.

$$V_{11}(z) = \frac{H_{22}(z)\widetilde{H}_{11}(z) - H_{12}(z)\widetilde{H}_{21}(z)}{\det \mathbf{H}(z)} \tag{23}$$

$$V_{12}(z) = \frac{H_{22}(z)\widetilde{H}_{12}(z) - H_{12}(z)\widetilde{H}_{22}(z)}{\det \mathbf{H}(z)} \tag{24}$$

This shows that perfect separation is achieved if and only if $\widetilde{\mathbf{H}}(z) = \mathbf{H}(z)$. We will call filter $V_{12}(z)$ the interference filter since it is related to the signal-to-interference ratio (SIR) and filter $V_{11}(z)$ the artifacts filter since it is related to the signal-to-artifacts ratio (SAR). The assessment of the separation performance using this LTI analysis is advantageous compared to SNR measurements like SIR and SAR. From the transfer functions, the respective SNR values can easily be obtained if the source signals are known.

[1] The tilde labels the true mixing system instead of the wrongly identified one for ease of displaying.

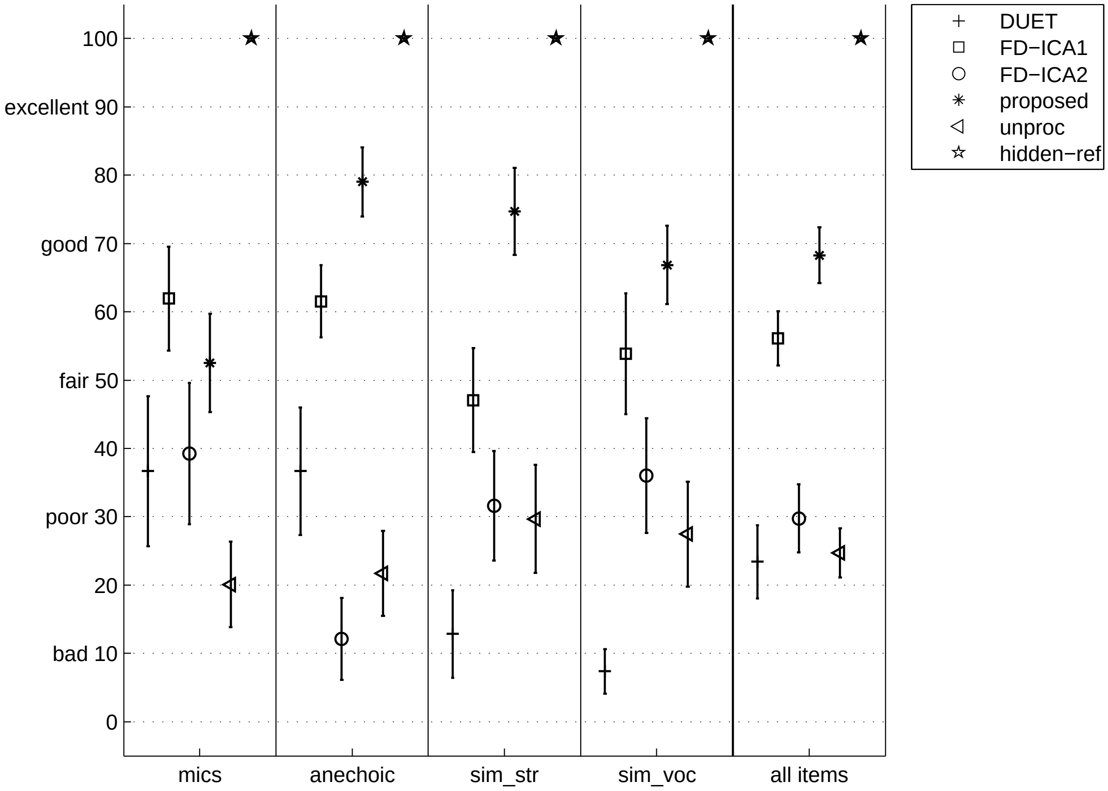

Fig. 11: Results of the listening test. The items are a microphone recording (mics), an anechoic mixture (anechoic), and two recordings with room simulation, with string instruments (sim str) and speech (sim voc).

This is illustrated with examples in Figures 12 to 15. If the true parameters are $\mathbf{T} = \left(\begin{smallmatrix} 0 & 12 \\ 16 & 0 \end{smallmatrix}\right)$ and $\mathbf{A} = \left(\begin{smallmatrix} 1 & 0.6 \\ 0.6 & 1 \end{smallmatrix}\right)$,

It is shown that

- underestimation of gains leads to more crosstalk than overestimation.
- delay estimation error leads to ripple and decreasing separation in high frequency

In particular, Figures 8 and 9, as compared to the magnitude as a function of frequency in Figures 12 and 13, indicate that incorrect estimation of the gains is a probable source of error. This is likely to be the cause of the relatively poor performance of the proposed method in the listening test for the microphone recording of Section 4.3.

Typically, the delay parameters for microphone recordings are not integer multiples of the sampling period. The error which is caused by a fractional delay of $\frac{n}{m}$ samples equals the error for delay estimation errors of m samples. Consequently, the presented method benefits from processing at high sampling rates.

5. CONCLUSIONS AND FUTURE WORK

A source separation method has been presented for microphone recordings where the number of sources equals the number of microphones. The presented method uses IIR filters for attenuating the interfering sources in the microphone signals similar to phase cancellation. A small number of parameters needs to be estimated, namely ICLD and ICTD for each transmission paths. Its performance is limited by fractional delay, frequency dependence of the directivity of the microphones and radiation patterns of the sources. Based on the newly presented BSS method, a computationally efficient method for comb-filter compensation has been presented.

Several steps may be taken to improve on the presented method. Alternative methods to DUET and GCC-PHAT may provide better estimation of the parameters. Fractional delay filters could be implemented to address the delay error. Use of reference microphones, with known positions, would permit very accurate estimation of the delay and gain parameters. A windowed approach to time delay and gain estimation may allow an implementation to work in real-time, and on moving sources.

6. ACKNOWLEDGEMENT

The first author would like to thank the Fraunhofer IIS and the Queen Mary University of London for supporting his research sabbatical during which the presented work has been done, and Alice Clifford for her help with recording the test signals and fruitful discussions on time delay estimation. Furthermore we thank Markus Kallinger for valuable comments on a draft of this paper.

7. REFERENCES

[1] Pierre Comon, "Independent component analysis, a new concept?," *Signal Processing, Elsevier*, vol. 36, no. 3, pp. 287–314, 1994.

[2] J.-F. Cardoso, "Blind source separation: Statisticle principles," *Proc. of the IEEE*, vol. 86, pp. 2009–2025, 1998.

[3] C. Jutten and A. Taleb, "Source separation: From dusktill dawn," in *Proc. Int. Workshop on Independent Component Analysis and Blind Source Separation*, 2000.

[4] A. Hyvaerinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, Wiley, 2001.

[5] E. Weinstein, M. Feder, and A.V. Oppenheim, "Multi-channel signal separation by decorrelation," *IEEE Trans. on Speech and Audio Processing*, vol. 1, pp. 405–413, 1993.

[6] H.-L.N. Thi and C. Jutten, "Blind source separation for convolutive mixtures," *Signal Proc.*, vol. 45, pp. 209–229, 1995.

[7] T.-W. Lee, A.J. Bell, and R.H. Lambert, "Blind separation of delayed and convolved sources," in *Proc. of NIPS*, 1996.

[8] T.-W. Lee, A.J. Bell, and R. Orglmeister, "Blind separation of real-world signals," in *Proc. of Int. Conf. on Neural Networks*, 1997.

[9] H. Buchner, R. Aichner, and W. Kellermann, "TRINICON: A versatile framework for multichannel blind signal processing," in *Proc. of ICASSP*, 2004.

[10] L. Parra and C. Spence, "Convolutive blind separation of non-stationary signals," *IEEE Trans. on Speech and Audio Proc.*, vol. 8, pp. 320–327, 2000.

[11] S. Harmeling, "Parra/Spence's blind source separation algorithm," http://people.kyb.tuebingen.mpg.de/harmeling, 2001, accessed June 21, 2010.

[12] N. Mitianoudis and M.E. Davies, "Audio source separation of convolutive mixtures," *IEEE Trans. on Speech and Audio Processing*, vol. 11, pp. 489–497, 2003.

[13] Lord Rayleigh, "On our perception of sound direction," *Philosophical Magazine*, vol. 6, pp. 214–232, 1907.

[14] J. Blauert, *Spatial Hearing*, MIT Press, 1996.

[15] Jeroen Breebart and Christof Faller, *Spatial Audio Processing*, Wiley, 2007.

[16] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[17] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures," in *Proc. of ICASSP*, 2000.

[18] S. Rickard and O. Yilmaz, "On the approximate w-disjoint orthogonality of speech," in *Proc. of ICASSP*, 2002.

[19] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. on Signal Proc.*, vol. 52, pp. 1830–1847, 2004.

[20] S. Rickard, "The duet blind source separation algorithm," in *Blind Speech Separation*, S: Makino, T.-W. Lee, and H. Sawada, Eds. Springer, 2007.

[21] N. Cahill, R. Cooney, K. Humphreys, and R. Lawlor, "Speech source enhancement using a modified adress alorithm for applications in mobile communications," in *Proc. of the AES 121st Conv.*, 2006.

[22] D. Barry, B. Lawlor, and E. Coyle, "Sound source separation: Azimuth discrimination and resynthesis," in *Proc. of DAFx*, 2004.

[23] M. Puigt and Y. Deville, "A time-frequency correlation-based blind source separation method for time-delay mixtures," in *Proc. of ICASSP*, 2006.

[24] Simon Arberet, Remi Gribonval, and Frederic Bimbot, "A robust method to count and locate audio sources in a stereophonic linear anechoic micxture," in *Proc. of ICASSP*, 2007.

[25] M.I. Mandel, R.J. Weiss, and D.P.W. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Trans. on Audio, Speech and Language Proc.*, vol. 18, pp. 382–394, 2010.

[26] H. Viste and G. Evangelista, "On the use of spatial cues to improve binaural source separation," in *Proc. of DAFx*, 2003.

[27] A. Favrot, M. Erne, and C. Faller, "Improved cocktail-party processing," in *Proc. of DAFx*, 2006.

[28] C. Faller and M. Erne, "Modifying stereo recordings using acoustic information obtained with spot recordings," in *Proc. of the AES 118th Conv.*, 2005.

[29] E. Kokkinis and J. Mourjopoulos, "Unmixing acoustic sources in real reverberant environments for close-microphone applications," *J. Audio Eng. Soc., to appear*, 2010.

[30] P.C. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, 2007.

[31] C.H. Knapp and G.C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. on Acoust., Speech, and Sig. Proc.*, vol. 24, pp. 320–327, 1976.

[32] M. Baeck and U. Zoelzer, "Real-time implementation of a source separation algorithm," in *Proc. of the 6th Int. Conf. on Digital Audio Effects (DAFx)*, 2003.

[33] E. Perez Gonzalez and J. Reiss, "Determination and correction of individual channel time offsets for signals involved in an audio mixture," in *Proc. of the 125th AES Conv.*, 2008.

[34] J.B. Allen and D.A. Berkeley, "Image method for efficiently simulating small–room acoustics," *J. Acoust. Soc. Am.*, vol. 65, pp. 943–950, 1979.

[35] International Telecommunication Union, Radiocomunication Assembly, "Methods for subjective assessment of small impairments in audio systems including multichannel sound systems," Recommendation ITU-R BS.1116, 1997, Geneva, Switzerland.

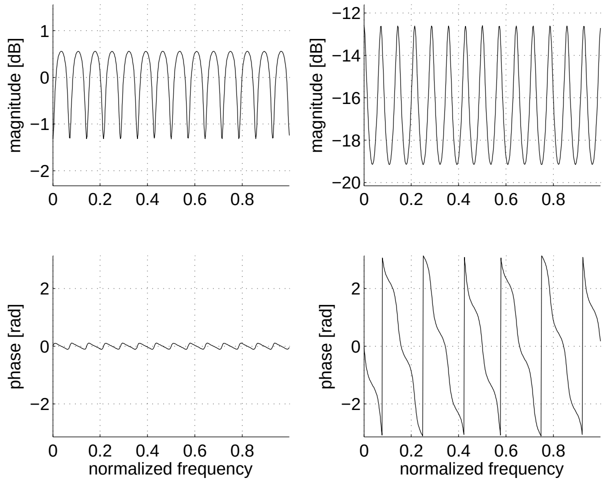

Fig. 12: Effect of underestimation of off-diagonal gains by a factor of 0.8.

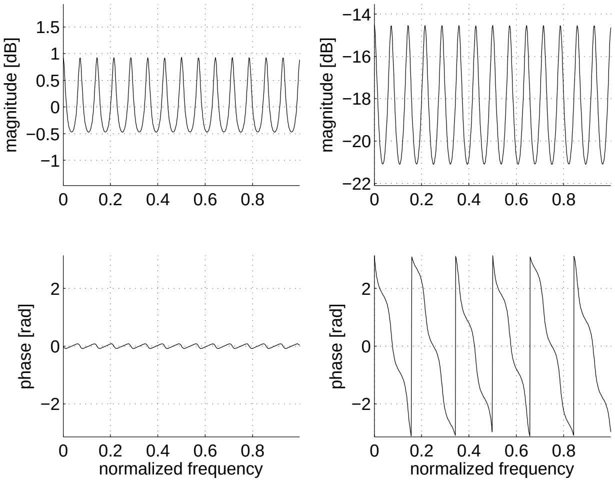

Fig. 13: Effect of overestimation of off-diagonal gains by a factor of 1.25.

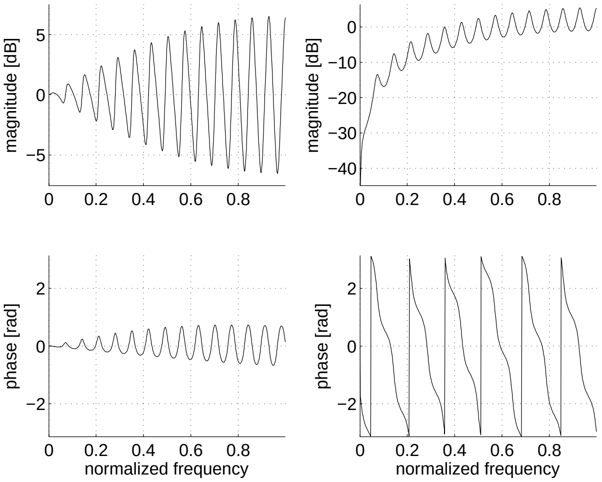

Fig. 14: Effect of underestimation of off-diagonal delays by 1 sample.

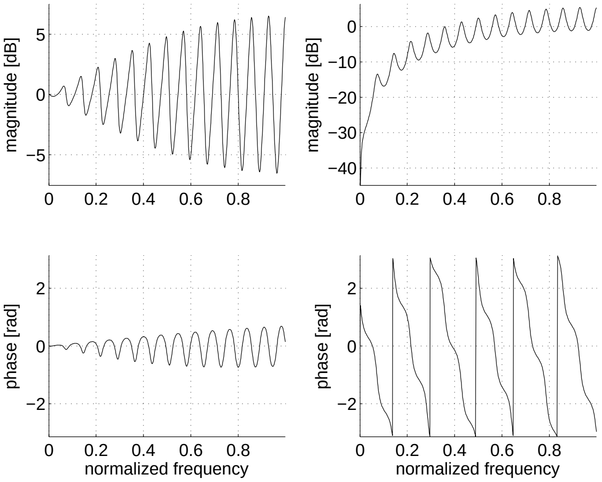

Fig. 15: Effect of overestimation of off-diagonal delays by 1 sample.