EURASIP Journal on Applied Signal Processing 2003:11, 1091–1109

Exploiting Acoustic Similarity of Propagating Paths for Audio Signal Separation

Bin Yin

Faculty of Electrical Engineering, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

Storage Signal Processing Group, Philips Research Laboratories, P.O. Box WY-31, 5656 AA Eindhoven, The Netherlands
Email: bin.yin@philips.com

Piet C. W. Sommen

Faculty of Electrical Engineering, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands
Email: p.c.w.sommen@tue.nl

Peiyu He

Faculty of Electrical Engineering, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

University of Sichuan, Chengdu 610064, China
Email: hepeiyu@hotmail.com

Received 20 September 2002 and in revised form 26 May 2003

Blind signal separation can easily find its position in audio applications where mutually independent sources need to be separated from their microphone mixtures while both room acoustics and sources are unknown. However, the conventional separation algorithms can hardly be implemented in real time due to the high computational complexity. The computational load is mainly caused by either direct or indirect estimation of thousands of acoustic parameters. Aiming at the complexity reduction, in this paper, the acoustic paths are investigated through an acoustic similarity index (ASI). Then a new mixing model is proposed. With closely spaced microphones (5–10 cm apart), the model relieves the computational load of the separation algorithm by reducing the number and length of the filters to be adjusted. To cope with real situations, a blind audio signal separation algorithm (BLASS) is developed on the proposed model. BLASS only uses the second-order statistics (SOS) and performs efficiently in frequency domain.

Keywords and phrases: blind signal separation, acoustic similarity, noncausality.

1. INTRODUCTION

In recent years, blind signal separation (BSS) has grasped the attention of lots of researchers because of its numerous attractive applications in speech processing, digital communications, medical science, and so on. BSS, within the framework of independent component analysis (ICA) [1, 2], deals with the problem of separating statistically independent sources only from their observed mixtures while both the mixing process and source signals are unknown.

For acoustical applications, it can be used to extract individual audio sources from multiple microphone signals when several sources are simultaneously active [3]. In other words, it becomes possible, for instance, in a teleconferencing system, to pick up one desired speech signal under a relatively low signal-to-noise ratio (SNR) (so called "cocktail party effect").

For a certain combination of source-sensor positions, instead of solving three-dimensional wave equations, the acoustic transmission from the source to the sensor can be simply described using an impulse response, which is obtained by measuring the signal received by the sensor after a sound pulse has been emitted from the source. An example is shown in Figure 1.

Thus, an acoustic mixing process in a reverberant environment can be modelled as

$$\underline{x}[k] = (h * \underline{s})[k], \tag{1}$$

where $\underline{s}[k] = (s_1[k], \ldots, s_n[k])^T$ and $\underline{x}[k] = (x_1[k], \ldots, x_n[k])^T$ denote the vectors of audio sources and microphone signals, respectively, and

$$h[k] = \begin{pmatrix} h_{11}[k] & h_{12}[k] & \cdots & h_{1n}[k] \\ \vdots & \vdots & \ddots & \vdots \\ h_{n1}[k] & h_{n2}[k] & \cdots & h_{nn}[k] \end{pmatrix} \tag{2}$$

is a transfer function matrix whose element h_{ij} expresses the room impulse response (RIR) from the jth source to the ith microphone, k is the discrete-time index which corresponds to the sampling moment, $*$ denotes linear convolution and $(\,)^T$ denotes matrix transpose. Here we assume the numbers of sources and microphones are the same and the environment is noise free.

The sources can be separated either by inverting the transfer function matrix $h[k]$ after having obtained the estimate of the individual h_{ij} (known as forward model methods), or by directly finding a demixing matrix $G[k]$ which satisfies $(G * h)[k] = I[k]P$, where $I[k]$ is a diagonal transfer function matrix and P denotes a matrix of permutation (known as backward model methods).

In principle, a sound pulse emitted from the source will be reflected infinite times by the walls and other obstacles, so an IIR filter seems to be suitable to describe the characteristics of an RIR. However, as shown in Figure 1, an RIR reveals a decaying waveform so that after a certain number of taps the residual signal becomes too weak to be detected by the sensor (e.g., human ears). Therefore, in practice, an FIR filter can be a quite acceptable approximation. In audio separation and many other applications, an FIR filter, for instance, having 1000–2000 taps with 8 kHz sampling frequency in a usual office, gives a good performance. An FIR filter is preferred because it provides much convenience when applied in digital signal processing.

For RIRs of such a considerably long length, in both forward and backward model methods, audio separation becomes a huge task due to the estimation of thousands of coefficients. It gets even more challenging in real-time implementations which are often needed in audio signal processing.

In this paper, aimed at the feasibility of real-time applications, a simplified mixing model is proposed which takes advantage of acoustic propagation similarities. In literature, a model which is close to the proposed one has also been used in signal separation analysis, especially in 2×2 case, for example, in [4, 5, 6]. By only considering the antidiagonal terms in the mixing matrix, the theoretical analysis of BSS became much more simplified. However, its feasibility has never been explicitly studied from an application point of view. In [4], two possibilities were given. One was that, two sources were standing near their own sensors so that only the transfer functions in two coupling paths (antidiagonal terms in the mixing matrix) should be taken into account; the other was that, in an anechoic, isotropic, and homogeneous environment, a coupling path could spatially equal the direct path in cascade with an auxiliary path. The latter resembles the form of the proposed model in this paper. In general, these two hypotheses will not hold in a natural reverberant environment. Besides, the noncausality introduced during the simplification has not been considered, which is inevitable according to the analysis in Section 4. Another paper that should be mentioned is [7], where the authors used a compact (1 cm) microphone array in order to describe the difference between the acoustical paths with a pure time delay. The signal separation was done in two stages: first separation by estimating and inverting the delay matrix and further separation with a feedback network. The idea is somewhat related, but we replace the pure delay with a filter of which the characteristics will be carefully studied. The proposed model provides the possibility of achieving signal separation in one step and relieving the excessive constraints on the dimension of the microphone array. On the basis of the new model, a BSS algorithm is described which only uses second-order statistics (SOS) and is efficiently realized in frequency domain. By applying the simplified mixing model, it is shown that the number of filters to be estimated is reduced to some extent. Besides, the taps of filters are significantly decreased in the case where microphones are intentionally closely placed. Several other advantages are mentioned as well. As a whole, they effectively give a possibility to a real-time implementation of audio source separation.

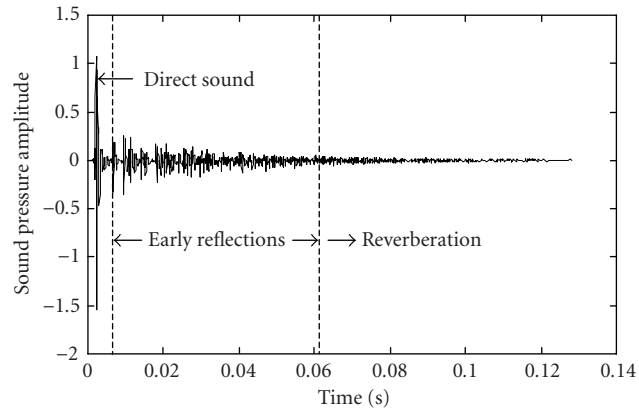


Figure 1: An example of an impulse response between two points in a room.

The remainder of this paper is organized as follows. In Section 2, concentrating on a 1-speaker-2-microphones system, we study the similarity between acoustic paths by defining an acoustic similarity index (ASI). Section 3 gives a simplified mixing model for blind audio source separation in both time and frequency domain. In Section 4, a nonblind speech signal separation scheme is designed in order to demonstrate the feasibility of the proposed model. To be able to cope with a real audio signal separation problem, we develop an adaptive BSS algorithm for the new mixing model in Section 5. Finally, Section 6 gives the conclusions.

2. SIMILARITY OF ACOUSTIC PATHS

2.1. Room impulse responses

An RIR describes the characteristics of a propagating path when a sound propagates from its source to a sensor (say a microphone). The bandwidth needed for an RIR differs according to a specific application. For voiced speech, the upper bound drops below 4 kHz. Therefore, in speech signal separation, a sampling frequency f_s of 8 kHz is adequate. Without extra indication, $f_s = 8$ kHz will be used in this paper.

Throughout the whole paper except for the experiments on the recorded real-world data, we employ a Matlab program "Room" [8], which applies the image theory [9], to generate RIRs. This choice is made for several reasons.

(1) The experimental environment is clean and we are able to have accurate control over the experimental conditions, such as the room dimensions, wall reflection coefficients, and, especially, the subtle changes of the source and microphone positions.
(2) The image solution of a rectangular enclosure, like the office environment, rapidly approaches an exact solution of the wave equation as the walls of the room become rigid. Under typical conditions, the frequency range of 100 Hz–4 kHz, wall reflection coefficients greater than 0.7, and both source and microphone not close to the wall, it does not introduce serious problems into the final result [9].

Now let us have a close look at the characteristics of an RIR. The room dimensions are uniformly adopted as 4 m × 5 m×3 m (width×depth×height), like that of a normal office. Placing one of the ground corners at the origin of a three-dimensional coordinate system, we can express any location in the room with a triplet (x, y, z), with x, y, and z the width, depth, and height, respectively. By default, the sources and microphones are located at the plane of $z = 1.5$ m.

Listed in Figure 2 are two RIRs with different wall reflections, corresponding to an almost anechoic environment with a reverberation time $T_{60} = 0.116$ second and a strongly reverberant environment with $T_{60} = 0.562$ second. The reverberation time T_{60} is defined as the time needed for the sound pressure level to decay by 60 dB when a steady-state sound source in the room is suddenly switched off [10]. The first observation is that the RIR with $T_{60} = 0.562$ second has much denser reflections and much longer decaying time than that with $T_{60} = 0.116$ second. The second is that the RIR with $T_{60} = 0.562$ second becomes nonminimum phase. Usually a nonminimum phase RIR occurs when a microphone picks up an echo stronger than the direct signal. This can happen in a strongly reverberant environment, as shown in this case, and also when a microphone is placed more closely to a wall or other obstacles than to the source.

2.2. Susceptibility of a room impulse response

Due to the wide range of wavelengths (from about 17 mm to 17 m) and the low propagating speed of a sound, a slight change of the source or sensor position may influence the fine structure of the impulse response significantly.

To study this susceptibility, consider a 1-speaker-2-microphone setup in the aforementioned room. The RIRs from the speaker at (x_s, y_s, z_s) to the microphone 1 at (x_1, y_1, z_1) and 2 at (x_2, y_2, z_2) are described by $h_{11}[k]$ and $h_{21}[k]$, respectively. Both are of length L_0. A difference room impulse response (DRIR) $\Delta h_{21}[k]$ can be defined as

$$h_{21}[k] = (\Delta h_{21} * h_{11})[k]. \tag{3}$$

The DRIR is used to describe the variation of the RIR when the microphone position is shifted from (x_1, y_1, z_1) to (x_2, y_2, z_2). It is exists in the form of an IIR filter as

$$\Delta h_{21}[k] = (h_{21} * h_{11}^{-1})[k]. \tag{4}$$

For convenience of later analysis and processing, we like to express it in the form of an FIR filter. We understand from Section 2.1 that an RIR could lose the minimum phase characteristic in certain acoustical conditions. So, without any prior knowledge, we have to assume $h_{11}[k]$ a nonminimum phase FIR filter. The impulse response of its stable inversion will be a noncausal infinite double-sided converging sequence. After the convolution in (4), $\Delta h_{21}[k]$ also becomes a noncausal double-sided IIR filter. The exception arises only when the zeros of $h_{11}[k]$ outside the unit circle are cancelled by those of $h_{21}[k]$, which is unlikely to happen in reality. To make it suitable for practical use, we execute two operations: first shift it by a delay of τ samples, and then truncate it such that

$$\Delta\tilde{h}_{21}[k] = \mathrm{Trc}\{(h_{21} * h_{11}^{-1})[k-\tau]\}, \tag{5}$$

where $\mathrm{Trc}\{\cdot\}$ denotes the truncation that cuts off all the taps before $k = 0$ and after $k \geq L$. The relationship between $\Delta\tilde{h}_{21}[k]$ and $\Delta h_{21}[k]$ can be written as

$$\Delta h_{21}[k-\tau] = (\Delta\tilde{h}_{21} + \epsilon_{21})[k], \tag{6}$$

$\epsilon_{21}[k]$ denotes an error filter varying with different choices of τ and L, obviously $\epsilon_{21} \to 0$ with $\tau, L \to \infty$. Convolving both sides of (6) with $h_{11}[k]$ and using (3), we have

$$h_{21}[k-\tau] = (\Delta\tilde{h}_{21} * h_{11})[k] + \epsilon_{21}^{R}[k], \tag{7}$$

where $\epsilon_{21}^{R}[k] = (\epsilon_{21} * h_{11})[k]$. When the parameters τ and L are chosen large enough so that the term $\epsilon_{21}[k]$ becomes negligible for certain applications, (7) can be simplified as

$$h_{21}[k-\tau] \approx (\Delta h_{21} * h_{11})[k], \tag{8}$$

where the tilde in $\Delta\tilde{h}_{21}[k]$ has been omitted for simplicity of expression. Therefore, from now on, $\Delta h_{21}[k]$ denotes a causal FIR filter of length L. To distinguish, we denote the IIR $\Delta h_{21}[k-\tau]$ in (6) as $\Delta h_{21}^{o}[k]$.

In the following simulation, we study the feasibility of the above expression in various acoustical situations. Depicted in Figure 3a is the simulation setup, where the source and one microphone stand still to fix h_{11} and the other microphone moves along the hollow arrow to give various h_{21}'s. We use the efficient block frequency domain adaptive filtering (BFDAF) algorithm for the DRIR estimation. The block diagram of the estimation scheme is plotted in Figure 3b. The input signal $s[k]$ is white and the delay adopted is always half the filter length, that is, $\tau = L/2$. The simulation is done with $T_{60} = 0.116$ second, 0.270 second, and 0.562 second, corresponding to a weakly, a mildly, and a strongly reverberant environment, respectively. The results are recorded in

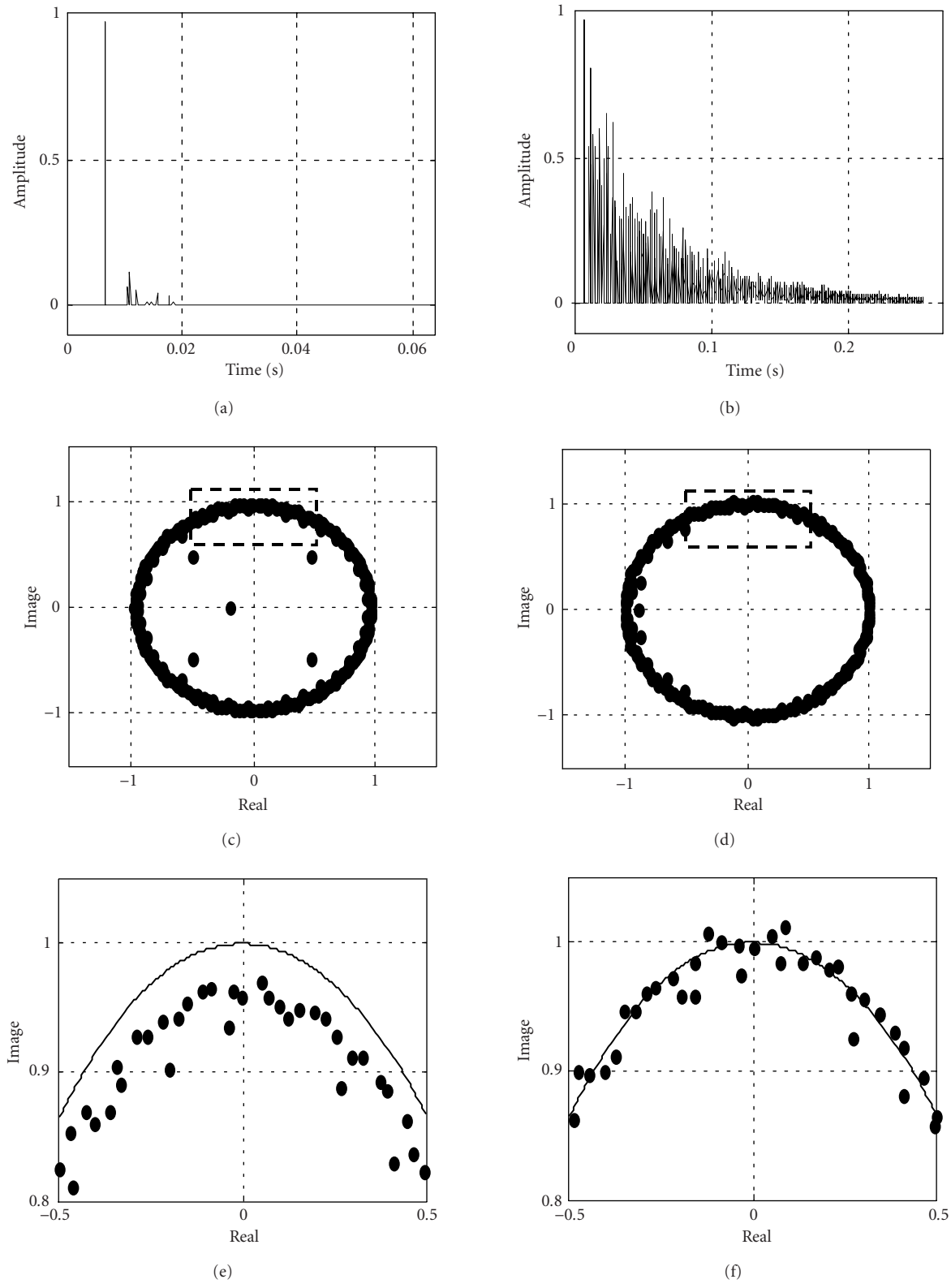


Figure 2: Comparison of zero distributions. Left column ($T_{60} = 0.116$ second): (a) RIR, (c) zero distribution at the z plane, (e) zooming in the area with broken lines in (c). Right column ($T_{60} = 0.562$ second): (b) RIR, (d) zero distribution at the z plane, (f) zooming in the area with broken lines in (d).

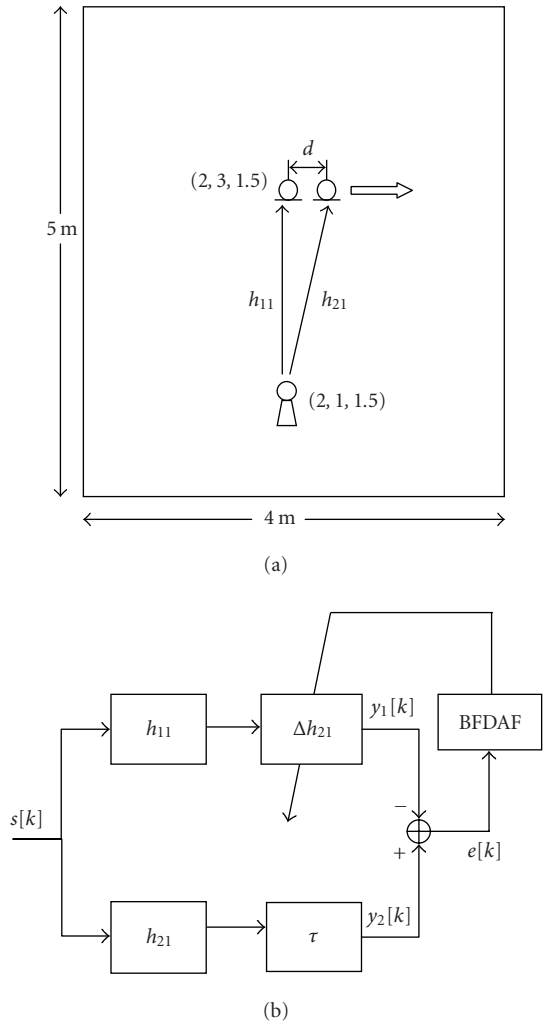


Figure 3: The acquisition of the DRIR Δh_{21}: (a) the simulation setup in the room, (b) the block diagram of the acquisition scheme.

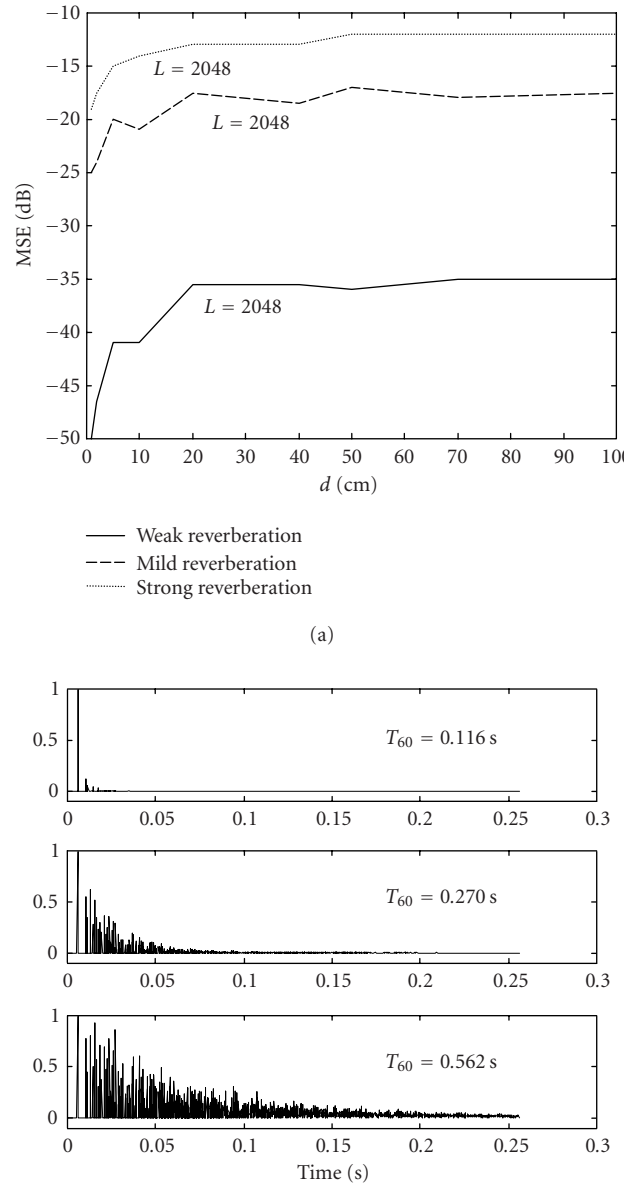


Figure 4: The feasibility study of the expression in (8). (a) Magnitude of the residual signals. (b) RIRs in the three acoustical conditions.

Figure 4a, where the mean square error (MSE) at the vertical axis is defined as

$$\text{MSE} = \lim_{k\to\infty} 10\log\frac{\sum_{l=-T}^{T}|e[k+l]|^2}{\sum_{l=-T}^{T}|y_2[k+l]|^2}, \qquad (9)$$
$$e[k+l] = y_2[k+l] - y_1[k+l],$$

where T denotes a certain number of samples chosen for averaging. The corresponding h_{11}'s are also plotted in Figure 4b. In fact, the residual signal $e[k]$ in (9) satisfies

$$e[k] = (\epsilon_{21}^{R} * s)[k], \qquad (10)$$

which reflects the normalized modelling error when we express h_{21} as h_{11} convolved with an FIR Δh_{21}.

For a given filter length L, two facts can be observed. First, the more reverberant the environment is, the more modelling error exists. This phenomenon may be intuitively explained as follows. With the increase of reverberation, the RIR gets longer; besides, its inverse becomes double sided and both tails take quite some time to converge due to the fact that its zeros tend to distribute more closely to the unit circle, and even exceed it in the case of a large reverberation. This can be seen in Figure 2d. Thus, truncating in (5) introduces more errors. Secondly, the further the two microphones are placed away from each other, the more modelling error we have. This trend happens rapidly, especially when the microphone spacing increases from 1 cm to 20 cm. We will analyze this in more detail in Section 2.3.

For practical applications, one cares mostly not about the existence of this error, but how small it should be so as to provide a satisfactory result to a specific application. In audio signal separation, despite a certain modelling error introduced when the expression (8) takes place, the separation can still be achieved in the sense that the cross talk left is inaudible or not disturbing any more to human ears. Therefore, by defining a DRIR Δh_{21}, we have effectively related the RIR h_{21} to h_{11}.

2.3. Acoustic similarity index

The second fact above suggests that as two microphones get closer, a similarity may start to play a role between the two acoustic paths, despite the susceptibility of an RIR. It simplifies the description of the DRIR and in turn decreases the modelling error in (7). These two aspects are not conflicting because whether the similarity or difference prevails depends on the accuracy of our interest. Let us have a closer look at several DRIRs with $T_{60} = 0.270$ second obtained in the last simulation, which is considered to be a normal situation.

With $d = 1$ cm, the DRIR appears like a pure time delay (Figure 5a), and accordingly, its amplitude frequency response looks quite flat in the most part of the spectrum and really fluctuate only at a limited number of high frequency components (Figure 5b). When the spacing increases, besides the central tap, more taps start to grow in magnitude so that more frequency components get influenced (Figures 5c, 5d, 5e, and 5f). This is understandable because as d increases the low frequency components of a sound signal see propagating path difference later and less than the high frequency components due to their longer wavelengths. The simulation implies that in general the characteristics of RIRs are not very much influenced by a small shift (within 5 cm in this case) of the objects because the wavelengths of the audio signals (greater than 9 cm for voiced speech) are well above this scale. The two acoustical paths before and after the shift can be regarded to be alike up to a time delay.

Now we are in the position of defining an ASI that reflects the degree of this similarity. We put the coefficients of $\Delta h_{21}[k]$ in a vector $\underline{c} = [c_1, \ldots, c_L]^T$. The ASI can be defined as

$$\mathrm{ASI}_{[h_{11},h_{21}]} = \exp\left\{ -\frac{\|\underline{c} - E_m\underline{c}\|_2^2}{\|E_m\underline{c}\|_2^2} \right\}, \tag{11}$$

where E_m represents a matrix with one at the m-mth position and zeros elsewhere, and $E_m\underline{c} = [0, \ldots, c_m, \ldots, 0]^T$ where $m = \arg\max_i\{|c_i|\}$. The exponential part expresses the ratio between the power of the central tap and the sum of the powers of the rest, which, in frequency domain, can be interpreted as the flatness of the spectrum. The nonlinear function $\exp\{\cdot\}$ is adopted in order to reflect the rapid drop of the ASI as soon as the DRIR starts to differ from a pure time delay. We calculate the corresponding ASI values for the DRIRs obtained in the last subsection and record them in Figure 6a.

For all situations, the general trend is the same: the ASI decreases as the microphone spacing d increases. When d approaches zero, the ASI approaches one, the highest value of the similarity. It can be obtained from (11) with Δh_{21} a single pulse in that case. In an almost anechoic environment, the ASI keeps very close to one even with d large up to 20 cm (solid line). It is because any RIR resembles a single-pulse-like form due to very few reflections, so that any two RIRs can be similar regardless of the object positions. While in more reverberant cases the ASI declines drastically at the first several centimeters (dashed and dotted lines), and after that it stays almost zero, meaning that the similarity between the two acoustic paths has gone.

Through the following analysis, we can see a bit more how the ASI varies according to the shape of an RIR. Suppose we have two RIRs, each of two taps, and they are written in z domain as

$$\begin{aligned} h_{11}(z^{-1}) &= z^{-p_1} + r_1 z^{-(p_1+g_1)}, \\ h_{21}(z^{-1}) &= z^{-p_2} + r_2 z^{-(p_2+g_2)}, \end{aligned} \tag{12}$$

where we assume $1 > r_1, r_2 > 0$, the time delays p_i and g_i $(i = 1, 2)$ are positive integers, and $p_1 < p_2$. By means of long division, we get

$$\begin{aligned} \Delta h_{21}(z^{-1}) &= \frac{h_{21}(z^{-1})}{h_{11}(z^{-1})} \\ &= z^{-(p_2-p_1)} + r_2 z^{-(p_2-p_1+g_2)} - r_1 z^{-(p_2-p_1+g_1)} \\ &\quad - r_1 r_2 z^{-(p_2-p_1+g_2+g_1)} + r_1^2 z^{-(p_2-p_1+2g_1)} + \cdots. \end{aligned} \tag{13}$$

The first term is the so-called central tap. Since the rest of the taps converge in magnitude, the next couple of terms become very critical for determining the ASI. If the microphone spacing d is so small that $g_2 = g_1$ holds, then we have

$$\begin{aligned} \Delta h_{21}(z^{-1}) &= z^{-(p_2-p_1)} + (r_2 - r_1) z^{-(p_2-p_1+g_1)} \\ &\quad - r_1(r_2 - r_1) z^{-(p_2-p_1+2g_1)} + \cdots. \end{aligned} \tag{14}$$

The values of the side taps are reduced because of the subtraction $(r_2 - r_1)$. When r_1 and r_2 are comparable, meaning that the two RIRs are quite similar, the reduction could be very significant. This will lead to a high ASI value. Otherwise, it will be very much likely to get a low ASI except that the values of r_i $(i = 1, 2)$ themselves are much smaller than one, for instance, in the case of a very weak reverberation. This rough analysis is basically also applicable to the practical situations although the expression gets more complicated because of the longer RIRs.

Hence, in order to get a higher ASI value, we have to either let the environment less reverberant or make the microphones more closely spaced. These two effects can be observed in Figure 6a.

Naturally, the fact that with a small microphone spacing a DRIR looks single-pulse like provides a possibility to use less filter taps. We repeat the simulation in Figure 4 with $T_{60} = 0.270$ second and various L's. The results are plotted in Figure 6b. Three microphone spacings are chosen. For $d = 0.5$ cm, the modelling error stays below -18 dB even with $L < 150$. The reason is that the DRIR resembles a single pulse

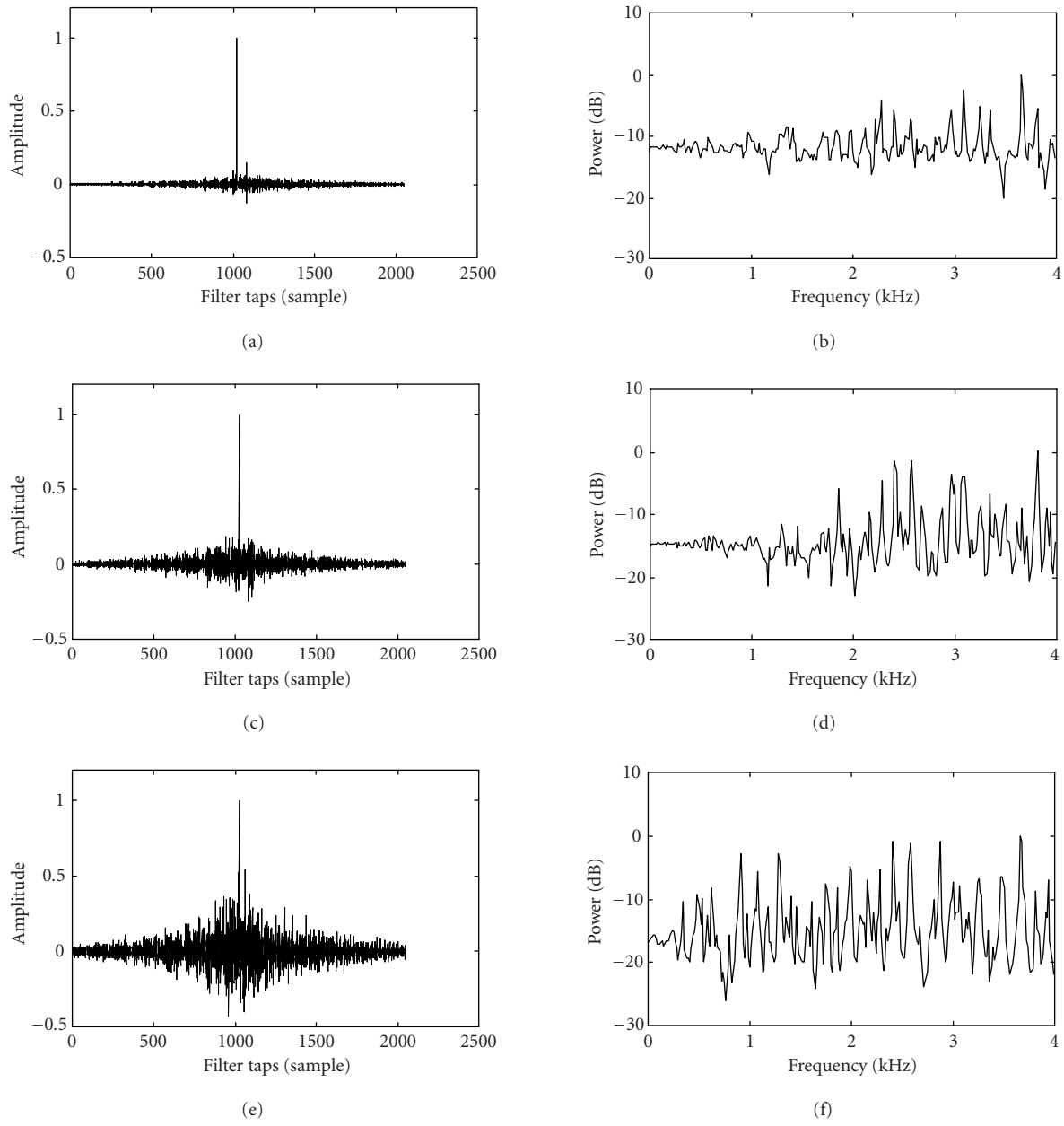
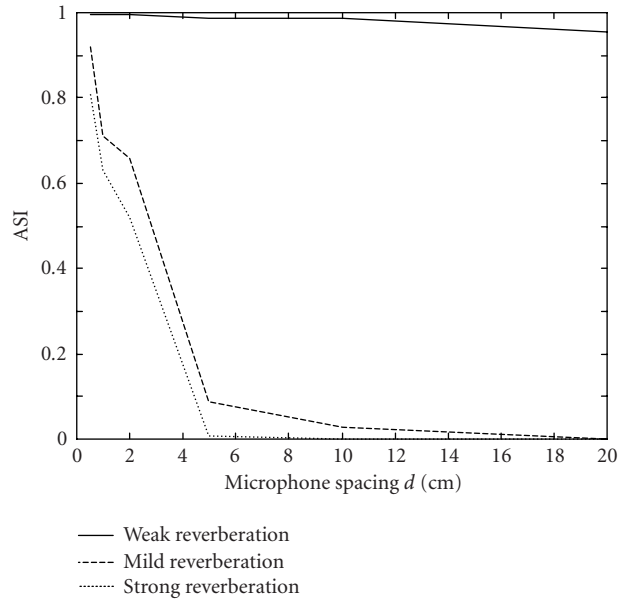


FIGURE 5: The DRIRs for different microphone spacing d with $T_{60} = 0.270$ second. Left column: impulse responses in time domain, (a) $d = 1$ cm, (c) $d = 5$ cm, (e) $d = 20$ cm. Right column: corresponding amplitude responses in frequency domain, (b) $d = 1$ cm, (d) $d = 5$ cm, (f) $d = 20$ cm.

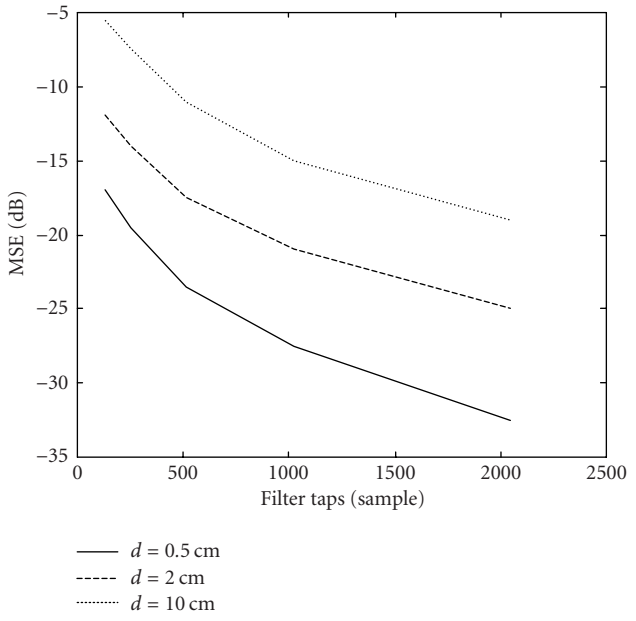
so much that most of side taps can be practically neglected. The ASI equals 0.89, reflecting the high similarity of the two RIRs. For $d = 2$ cm, the MSE needs $L > 750$ to remain below -20 dB since the tail of the DRIR includes stronger taps and, when truncated, significant errors will occur. Correspondingly, the ASI decreases to 0.58. For $d = 10$ cm, the ASI is 0.03, meaning actually that no similarity exists. The conclusion is that for a certain MSE requirement, fewer taps are needed with a smaller microphone spacing.

We must point out that the simulation results indicate some general rules, but these concrete numbers can fluctuate in different acoustical environments. For instance, for the same microphone spacing d, the ASI value can be different with the variation of the distance w from source to microphone. In the former simulations, the w was set around 2 m. Here we change the w to see how the ASI changes accordingly. The results are obtained with $T_{60} = 0.270$ second and plotted in Figure 7.

When the w is smaller than 1 m, the ASI becomes above 0.5 even with 10 cm microphone spacing (compared to



(a)



(b)

Figure 6: (a) The ASI versus the microphone spacing d in different acoustic situations ($L = 2048$). (b) The modelling error versus the length of the DRIR with $T_{60} = 0.270$ second (2048 taps used for the RIRs). The source-to-microphone distance w equals 2 m in both figures.

ASI = 0.03 with $w = 2$ m), meaning that the similarity starts to play a role. The reason is that when the microphones move to the source, the RIRs tend to be minimum phase because the distance w is small compared to that from the microphones to the walls so that the direct sound is more likely to

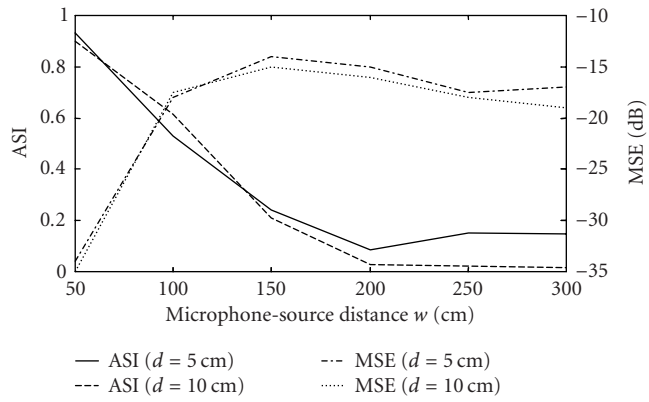


Figure 7: The ASI and MSE versus the source-to-microphone distance w in different microphone spacings.

be dominant over the reflections. This leads to shorter RIRs as well as their inverses. Consequently, the ASI becomes high. This provides us with another possibility to acquire a high ASI value. An interesting phenomenon can be observed that, for $d = 5$ cm, the ASI begins to increase at a low level when w is larger than 2 m. It can be explained by the effect that a certain microphone spacing becomes relatively small if the microphones move away from the source, resulting in an increase of the ASI. This effect is always there, but until the w gets large enough, it does not prevail over the other opposite effect that ASI decreases due to the nonminimum phase tendency of RIRs.

In general, the simulation results indicate that one should expect a more single-pulse-like DRIR within 5–10 cm microphone spacing under the normal room acoustics, corresponding to an ASI value above 0.5.

3. A SIMPLIFIED MIXING MODEL

In the case of a high ASI, introducing a simplified mixing model becomes very attractive to audio signal separation which usually suffers from high complexity. To derive the new model, we first generalize the relationship in (8) to the case of n sources and n microphones

$$h_{ml}[k-\tau] = (\Delta h_{ml} * h_{ll})[k], \quad m, l = 1, \ldots, n,\ m \neq l, \tag{15}$$

where the modelling error is omitted. Then we can rewrite the model in (1) as

$$\underline{x}[k-\tau] = (\Delta h * \underline{s}')[k], \tag{16}$$

where $\underline{x}[k-\tau] = (x_1[k-\tau], \ldots, x_n[k-\tau])^T$, $\underline{s}'[k] = ((h_{11} * s_1)[k], \ldots, (h_{nn} * s_n)[k])^T$, and

$$\Delta h[k] = \begin{pmatrix} \delta_\tau[k] & \Delta h_{12}[k] & \cdots & \Delta h_{1n}[k] \\ \vdots & \vdots & \ddots & \vdots \\ \Delta h_{n1}[k] & \Delta h_{n2}[k] & \cdots & \delta_\tau[k] \end{pmatrix}. \tag{17}$$

For convenience of the latter expression, $\delta[k-\tau]$ is written as $\delta_\tau[k]$ representing a time delay of τ samples. Since the microphones should be closely spaced relative to each source, a microphone array will be a reasonable solution. The components in the vector $\underline{s'}$ are mutually independent due to the assumed independence between the sources, so the signal separation can be achieved after obtaining the estimation of the mixing matrix $\Delta h[k]$.

Using this simplified model in audio signal separation has several specific advantages.

(1) What we attempt to recover are the signals propagating and arriving just in front of the microphones before mixing, that is, the sources convolved by the RIRs from their emitting points to the respective microphones, which often sound more natural than the clean sources themselves when there is not too much reverberation present.
(2) The number of filters to be estimated is reduced from n^2 to $n(n-1)$.
(3) Furthermore, with the existence of the similarity between acoustic paths, fewer coefficients are required to describe Δh_{ml}'s because they appear very much like a single-pulse function. As a result, the computational load for the mixing model estimation can be significantly reduced.

As seen in (16), in a reverberant environment, microphone signals are convolutive mixtures of original sources. For much more efficient implementation, we will transform the problem into the frequency domain so as to realize signal separation simultaneously for every frequency component as in the case of an instantaneous mixing [11, 12].

The discrete Fourier transform (DFT) allows us to express circular convolutions as products in frequency domain, while in (16) linear convolutions are assumed. A linear convolution can be approximated by a circular convolution if $L \ll N$, where N denotes the number of points within one data frame in the DFT. Also a linear time shifting in Δh can be approximated by a circular time shifting if $\tau \ll N$. Therefore, we can write approximately

$$\begin{aligned} \underline{X}(\omega_i, p-\tau) &\approx \Delta\mathcal{H}(\omega_i)\underline{S}'(\omega_i, p), \\ \omega_i &= \frac{(i-1)}{N}2\pi, \quad i = 1, \ldots, N, \end{aligned} \tag{18}$$

where ω_i denotes the ith frequency component;

$$\underline{X}(\omega_i, p-\tau) = (X_1(\omega_i, p-\tau), \ldots, X_n(\omega_i, p-\tau))^T \tag{19}$$

represents the DFT of the microphone signals where $X_m(\omega_i, p-\tau)$ comes from the DFT of the vector of signals from the mth microphone, that is,

$$\underline{x}_m[p-\tau] = (x_m[p-\tau], \ldots, x_m[p-\tau+N-1])^T, \tag{20}$$

starting at $p-\tau$ and of length N, which is given by

$$X_m(\omega_i, p-\tau) = \sum_{\kappa=0}^{N-1} e^{-j\omega_i\kappa} x_m[p-\tau+\kappa]; \tag{21}$$

$\underline{S}'(\omega_i, p-\tau)$ is obtained from the vector of the filtered source signals $\underline{s'}[k]$ in the same way as $\underline{X}(\omega_i, p-\tau)$; $\Delta\mathcal{H}(\omega_i)$ denotes the frequency domain counterpart of the filter matrix $\Delta h[k]$ and can be expressed as

$$\Delta\mathcal{H}(\omega_i) = \begin{pmatrix} e^{-j\omega_i\tau} & \cdots & \Delta H_{1n}(\omega_i) \\ \vdots & \ddots & \vdots \\ \Delta H_{n1}(\omega_i) & \cdots & e^{-j\omega_i\tau} \end{pmatrix}, \tag{22}$$

where $(\Delta H_{ml}(\omega_1), \Delta H_{ml}(\omega_2), \ldots, \Delta H_{ml}(\omega_N))^T$ represents the Fourier transform of the mlth $(m, l = 1, \ldots, n,\ m \neq l)$ DRIR $\Delta h_{ml}[k]$ of length L.

4. SIGNAL SEPARATION IN THE 2×2 CASE

In this section, we take a 2-speaker-2-microphone system as an example to demonstrate the feasibility of the proposed mixing model in speech signal separation.

A simulation scheme of the separation is shown in Figure 8. The left diagram expresses the parameter measuring part where two BFDAF algorithms are used in parallel, and the right describes the separation part. The $\Lambda^{-1}[k]$ acts as a postprocessing filter that is the inversion of

$$\Lambda[k] = (\delta_{2\tau} - \Delta h_{21} * \Delta h_{12})[k], \tag{23}$$

where Δh_{ij} must be measured when only s_j is active, so the measurement may be first done with two sources made alternatively active, and after convergence of the filter parameters, the separation is then switched on. If we rewrite the mixing process in (16) without the modelling error as

$$\begin{pmatrix} x_1[k-\tau] \\ x_2[k-\tau] \end{pmatrix} = \begin{pmatrix} \delta_\tau & \Delta h^o_{12} \\ \Delta h^o_{21} & \delta_\tau \end{pmatrix} * \begin{pmatrix} (h_{11} * s_1)[k] \\ (h_{22} * s_2)[k] \end{pmatrix}, \tag{24}$$

where $\Delta h^o_{ml}[k]$ $(m \neq l)$ denotes the IIR DRIR that makes $h_{ml}[k] = (\Delta h^o_{ml} * h_{ll})[k]$ accurately hold, after the demixing in Figure 8, we have

$$\begin{aligned} \begin{pmatrix} \tilde{s'}_1[k] \\ \tilde{s'}_2[k] \end{pmatrix} = \Lambda^{-1} * \begin{pmatrix} \delta_{2\tau} - \Delta h_{12} * \Delta h^o_{21} & \delta_\tau * \epsilon_{12} \\ \delta_\tau * \epsilon_{21} & \delta_{2\tau} - \Delta h_{21} * \Delta h^o_{21} \end{pmatrix} \\ * \begin{pmatrix} (h_{11} * s_1)[k] \\ (h_{22} * s_2)[k] \end{pmatrix}, \end{aligned} \tag{25}$$

where $\epsilon_{ml}[k] = (\Delta h^o_{ml} - \Delta h_{ml})[k]$ $(m \neq l)$ denotes the modelling error as defined in (6). When the modelling errors are zero, the separation part in Figure 8 functions exactly as an inversion of the mixing process and a perfect signal separation will be achieved. The cross talk left depends on the magnitude of nonzero modelling errors.

The separation may be implemented efficiently in frequency domain as well. Its corresponding frequency domain structure is given in Figure 9, where the input $\underline{x}_i$ to the FFT block is the vector of the ith microphone signals obtained

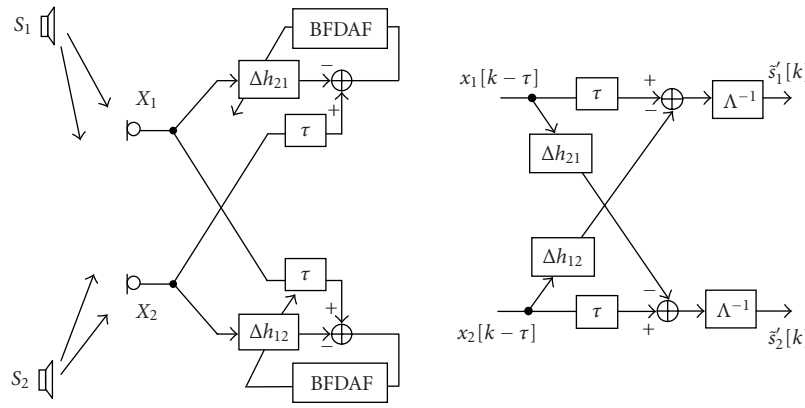


Figure 8: The signal separation scheme (2×2 case).

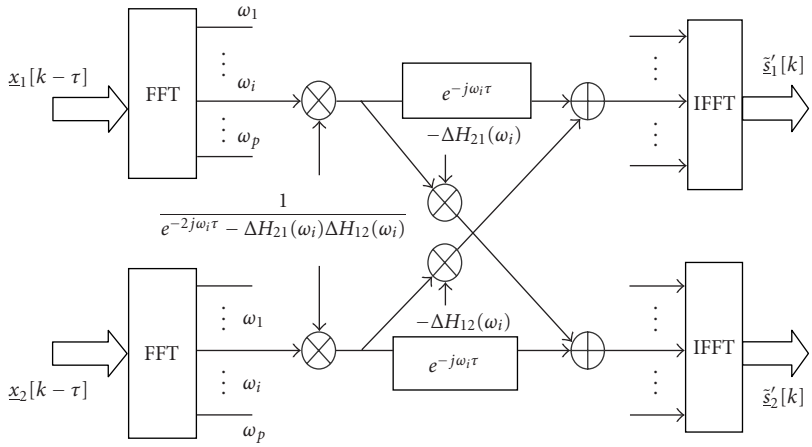


Figure 9: The separation implemented in frequency domain (2×2 case).

by buffering N consecutive data samples. The operations between FFT and IFFT blocks equal the inversion of the mixing matrix in frequency domain given in (18). One can see that the separation is independently operated for each frequency component ω_i, which converts the convolutively mixing problem into an instantaneous one.

As mentioned in Section 2.1, the time delay τ is introduced for the causal stable inversion of a nonminimum phase RIR. In general one can simply let $\tau = L/2$, but it is not necessarily like that. The proper choice of τ relies on several factors, for instance, the wall reflection and the distance between sources and microphones. In particular, if the reverberation is quite weak and the audio source is located close to its microphone (say within several 10 centimeters), τ may be chosen as zero since in this case RIRs are normally minimum phase. The advantage is that less taps are needed or with the same taps one can provide the right tail with more freedom, which probably gives more significant information. The detailed experimental results can be found in [13].

As for the postprocessing filter Λ^{-1}, it again concerns the inversion of a nonminimum phase filter. To solve the problem, one possibility is simply moving it away (correspondingly omitting the term $1/(e^{-2j\omega_i\tau} - \Delta H_{12}(\omega_i)\Delta H_{21}(\omega_i))$ in Figure 9) because it has nothing to do with the effectiveness of the separation; the other possibility is keeping it there to improve the sound quality at the cost of introducing another extra time delay.

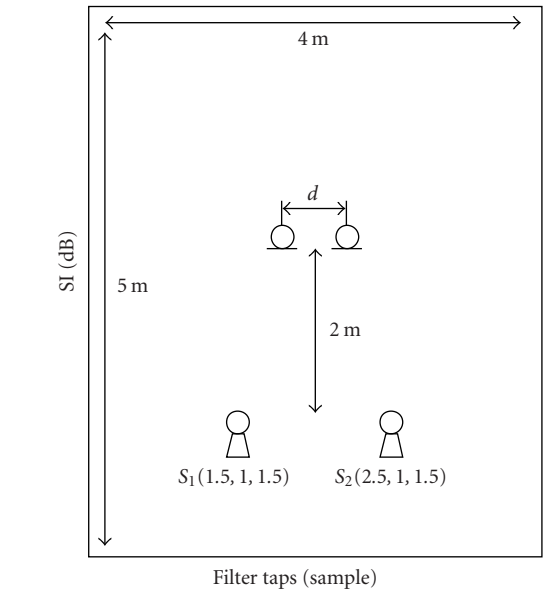
In order to evaluate the separation result with respect to different filter lengths L's under different ASI values, we define the following separation index (SI):

$$\mathrm{SI}_m = \lim_{k\to\infty} 10 \log \frac{\sum_{q=-T}^{T} \left|\tilde{s}'_m[k+q]\right|^2}{\sum_{q=-T}^{T} \left|\tilde{s}'_l[k+q]\right|^2}, \tag{26}$$

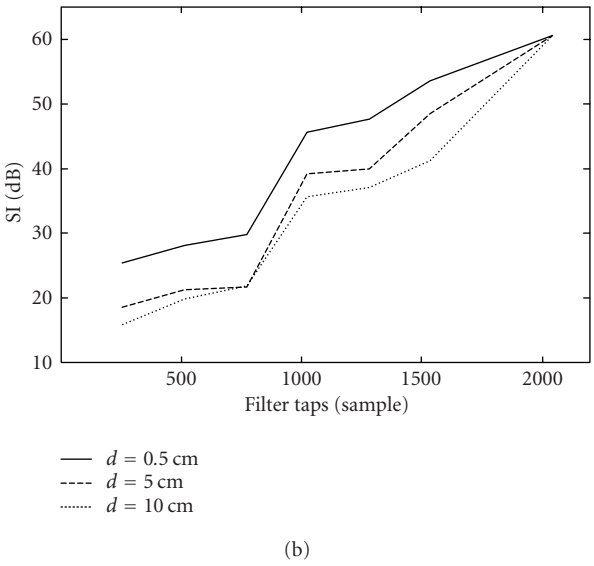
where only s_m is active, $m, l = 1, 2$, $m \neq l$,

$$\mathrm{SI} = \frac{\mathrm{SI}_1 + \mathrm{SI}_2}{2} \tag{27}$$

and T is a proper time period. If a white noise is assumed as an input signal, by using (25), the SI may be also expressed



(a)



(b)

Figure 10: The SI versus the number of taps for DRIR. (a) The simulation environment ($T_{60} = 0.27$ second). (b) The results in three different microphone spacings.

in z domain as

$$\mathrm{SI}_i = 10\log\left[\left(\oint \left|\Lambda^{-1}(z^{-1})\left[z^{-2\tau} - \Delta h_{ml}(z^{-1})\Delta h^{o}_{lm}(z^{-1})\right]\right.\right.$$
$$\left.\times h_{mm}(z^{-1})\right|^2 dz\Big)$$
$$\left.\times\left(\oint \left|\Lambda^{-1}(z^{-1})z^{-\tau}\epsilon_{lm}(z^{-1})h_{mm}(z^{-1})\right|^2 dz\right)^{-1}\right]. \quad (28)$$

In the current case where the sources are known to be alternatively active, one is able to get the DRIRs with only one source present, which makes the separation nonblind, so the SI value in fact indicates the maximum separation effect one can achieve when using the simplified mixing model with some given L and τ. In real BSS, extra modelling errors will exist due to nonoptimal estimation of DRIRs. Thus the SI will no longer reach its maximum. By definition, the SI is equivalent to an SNR or an SIR (signal-to-interference ratio) normally used in literature.

The simulation is done as described in Figure 10a. The reverberation time in the room is set as $T_{60} = 0.27$ second and input signals are white noise sampled in the frequency of 8 kHz. The results are plotted in Figure 10b. For a better comparison, the highest SI values of the three microphone spacings are normalized to be the same. In all cases, the SI decreases with the reduction of the filter taps, which coincides very well with (28). For the same filter length, the SI with $d = 0.5$ cm is higher than that with $d = 5$ cm by more than 5 dB. That is because the DRIR in the former case resembles a single pulse due to the high similarity of the acoustical paths (ASI = 0.85). It makes possible a considerable filter length reduction. The SI with $d = 10$ cm is about 3 dB lower than that with $d = 5$ cm, meaning that the acoustic similarity disappears further.

Two conclusions can be drawn. First, thanks to the similarity between the acoustical paths, the computational load of the audio signal separation can be significantly relieved, while the separation effect stays still reasonably good (above 20 dB). This gives us an opportunity to implement an audio signal separation in real time. Secondly, with large microphone spacings, a satisfying separation can be still acquired if the DRIRs are provided with enough taps. Hence, the proposed mixing model is also suitable for a normal use where microphones are not closely spaced, having the advantage of less filters to estimate.

Notice that the proposed model remains feasible as long as the difference between acoustical paths is distinguishable. A too small microphone spacing ($d < 1$ cm) gives little path difference to low frequency components of sources. A higher accuracy (more bits in digital signal processing) during sampling helps, but will be limited by the background noise level. The small path difference can accumulate after a number of reflections, so it reveals itself strongly in the reverberation part. However, since the signal power decays exponentially, the path difference becomes "invisible" especially at the presence of noise. Obviously, the separation fails when microphones are placed at the same point since no difference can be detected regardless of the given accuracy. In the case of a large spacing ($d > 10$ cm), the time delay of the arrival of the direct sound can help to "build" a path difference to avoid the spatial aliasing of high frequency components.

To be able to cope with a more complicated situation, for example, with moving speakers, in the next section, an online adaptive separation algorithm will be developed specifically for the simplified mixing model.

5. ADAPTIVE BLIND SIGNAL SEPARATION ALGORITHM

In literature, to achieve BSS, a variety of approaches based on different methodologies and theories have been proposed, which in general fall into two categories. If sources are stationary, Gaussian processes, it can be shown that blind separation is impossible in a certain sense. By stipulating non-Gaussianity on source signals, one can apply higher-order statistics (HOS) evaluation to realize separation. The methods in the first category are characterized by computing HOS explicitly [14, 15, 16, 17], or implicitly [18, 19], ML [20], INFOMAX [21], MMI [22], and NM [23] to achieve separation. While in the other category, with the help of some extra constraints, SOS is proved to be sufficient to determine the mixing process, for instance, additional time-delayed correlations [24], sources of different spectra [15, 25, 26], spectral matching and FIR constraints on mixing process [27], and nonstationary sources [28].

In this paper, by taking advantage of the nonstationarity of the audio sources, we develop an adaptive blind audio signal separation (BLASS) algorithm in frequency domain only based on SOS evaluation. Apparently, from the application point of view, an SOS method is preferred due to its less computational complexity and stronger robustness to noise.

Some first theoretical proof of how the BSS problem of nonstationary sources can be solved only using SOS has been given in [28]. We are not going to look at that further because it is out of the scope of this paper. Roughly speaking, the point is on the fact that, with the help of nonstationarity, one is able to do decorrelation through time to eliminate the ambiguity of the mixing (or demixing) model.

5.1. A gradient descent rule

Recalling the proposed mixing model in frequency domain from (18), after having obtained the mixing matrix $\Delta\mathcal{H}(\omega_i)$, the separated signals for each frequency component can be written as

$$\underline{S}'(\omega_i, p) = \Delta\mathcal{H}^{-1}(\omega_i)\underline{X}(\omega_i, p-\tau), \qquad \omega_i = \frac{(i-1)}{N}2\pi,\ i = 1, \ldots, N. \tag{29}$$

We construct an objective function as follows:

$$\begin{aligned} J(p) &= \sum_{i=1}^{N}\sum_{m\neq l} \left|\left(R_{s'}(\omega_i, p)\right)_{ml}\right|^2 \\ &= \sum_{i=1}^{N}\sum_{m\neq l} \left|\left(\langle \underline{S}'(\omega_i, p)\underline{S}'^{H}(\omega_i, p)\rangle\right)_{ml}\right|^2 \\ &= \sum_{i=1}^{N}\sum_{m\neq l} \left|\left(\Delta\mathcal{H}^{-1}(\omega_i)R_x(\omega_i, p-\tau)\Delta\mathcal{H}^{-H}(\omega_i)\right)_{ml}\right|^2, \end{aligned} \tag{30}$$

where

$$R_x(\omega_i, p-\tau) = \langle \underline{X}(\omega_i, p-\tau)\underline{X}^H(\omega_i, p-\tau)\rangle, \tag{31}$$

$(\cdot)^H$ and $(\cdot)^{-H}$ denote the complex conjugate and transpose of a matrix and the inversion of the resulting matrix, respectively, $R_x(\omega_i, p-\tau)$ and $R_{s'}(\omega_i, p)$ represent the power matrix of $\underline{X}$ and $\underline{S}'$ for each frequency component. Basically they are the function of a different p because of the nonstationarity of audio signals. The objective function is in fact the sum of off-diagonal elements in the power matrix of the signal vector $\underline{S}'$ over all the frequency components, reflecting the cross power spectra between the demixed signals. Due to the mutual independency of the sources, $J(p)$ should reach its minimum. Therefore, the mixing matrix may be learned by means of a gradient descent algorithm

$$\Delta\hat{\mathcal{H}}(\omega_i, p+1) = \Delta\hat{\mathcal{H}}(\omega_i, p) + \mu\left[-\frac{\partial J(n)}{\partial\Delta\mathcal{H}^*(\omega_i)}\bigg|_{\Delta\mathcal{H}(\omega_i)=\Delta\hat{\mathcal{H}}(\omega_i, p)}\right], \tag{32}$$

where $\partial J(n)/\partial\Delta\mathcal{H}^*(\omega_i) = (\partial J(n)/\partial\Delta\mathcal{H}(\omega_i))^*$ and $(\cdot)$ denotes the complex conjugate of a matrix, and μ is a positive factor that determines the rate of updating. The gradients in (32) are given by

$$\frac{\partial J(p)}{\partial\Delta\mathcal{H}(\omega_i)} = -2\{\Delta\mathcal{H}^{-H}(\omega_i)[R_{s'}(\omega_i, p) - \mathrm{diag}\{R_{s'}(\omega_i, p)\}]R_{s'}(\omega_i, p)\}^*, \tag{33}$$

$$R_{s'}(\omega_i, p) = \Delta\mathcal{H}^{-1}(\omega_i)R_x(\omega_i, p)\Delta\mathcal{H}^{-H}(\omega_i), \tag{34}$$

where $\mathrm{diag}\{\cdot\}$ denotes taking the diagonal elements of a matrix and putting them at the corresponding positions of a zero matrix.

5.2. Constraints on the gradients

There are two constraints on the gradients during parameter updating. The diagonal elements in each gradient matrix $\partial J(p)/\partial\Delta\mathcal{H}(\omega_i)$ must be zero because the parameters have been known as a constant $e^{-j\omega_i\tau}$ or 1 ($\tau = 0$) on the diagonal of the mixing matrix $\Delta\hat{\mathcal{H}}(\omega_i, p)$, so the first constraint is

$$\mathcal{C}^{(1)}\left\{\frac{\partial J(p)}{\partial\Delta\mathcal{H}(\omega_i)}\right\} := \frac{\partial J(p)}{\partial\Delta\mathcal{H}(\omega_i)} - \mathrm{diag}\left\{\frac{\partial J(p)}{\partial\Delta\mathcal{H}(\omega_i)}\right\}. \tag{35}$$

Secondly, the gradients have to be constrained in order to make the time domain solutions satisfying $\Delta h[k] = 0$ for $k > L$. This is important for the expression in (18) to have a good approximation. Thus we use the following constraint [29]:

$$\mathcal{C}^{(2)}\left\{\frac{\partial J(p)}{\partial\underline{\Delta H}_{ml}(\omega)}\right\} := FZF^{-1}\frac{\partial J(p)}{\partial\underline{\Delta H}_{ml}(\omega)}, \tag{36}$$

where

$$\frac{\partial J(p)}{\partial\underline{\Delta H}_{ml}(\omega)} = \left(\frac{\partial J(p)}{\partial\Delta H_{ml}(\omega_1)}, \frac{\partial J(p)}{\partial\Delta H_{ml}(\omega_2)}, \ldots, \frac{\partial J(p)}{\partial\Delta H_{ml}(\omega_N)}\right)^T, \tag{37}$$

Z is an $N \times N$ diagonal matrix with $Z_{ii} = 1$ for $i \leq L$ and $Z_{ii} = 0$ for $i > L$, F denotes the Fourier matrix operating DFT, and accordingly F^{-1} operates IDFT.

As well known, the side effect of the frequency domain separation is that one cannot guarantee that the frequency components used to reconstruct the time domain output come from the same source because any permutation of the coordinates at every frequency will lead to exactly the same $J(p)$. If the permutation appears, generally the spectra of the estimated filters will become nonsmooth. Forcing zero coefficients for $k > L$ in time domain equivalently smooths their spectra through a convolution with a sinc function in frequency domain. Therefore, the permutation problem can be effectively removed by applying the constraint $\mathcal{C}^{(2)}$.

In addition, there is another point which may have not been realized in previous literature. If one of the sources does not have any power at a certain frequency component ω_i, the separation fails because $\Delta\widehat{\mathcal{H}}(\omega_i)$ is singular and therefore $\Delta\widehat{\mathcal{H}}^{-1}(\omega_i)$ does not exist. The smoothing of the spectrum by $\mathcal{C}^{(2)}$ helps to remove the zeros so as to relieve the problem to some extent.

5.3. Practical approximation

Although in general audio signals are nonstationary, it is still said that human voice has temporally stationary structure within a few 10 milliseconds [30]. This means it is possible to use time averaging to approximate the SOS needed in our algorithm:

$$R_x(\omega_i, p) \approx \frac{1}{2T} \sum_{q=-T}^{T} \underline{X}(\omega_i, p+q)\underline{X}^H(\omega_i, p+q), \tag{38}$$

where T is properly chosen to make the averaging within the stationary period. Formula (38) is only needed for the initialization of the power matrices. During the adaptation the power estimates are updated using

$$\begin{aligned} R_x(\omega_i, p+1) &= \alpha R_x(\omega_i, p) \\ &\quad + (1-\alpha)\underline{X}(\omega_i, p+1)\underline{X}^H(\omega_i, p+1). \end{aligned} \tag{39}$$

The forgetting factor α may vary from zero to one depending on the degree of nonstationarity. Additionally, to reduce the computational complexity further, the inversion of the matrix $\Delta\widehat{\mathcal{H}}(\omega_i, p)$ may be approximated by

$$\begin{aligned} &\Delta\widehat{\mathcal{H}}^{-1}(\omega_i, p+1) \\ &\approx \Delta\widehat{\mathcal{H}}^{-1}(\omega_i, p) \\ &\quad - \Delta\widehat{\mathcal{H}}^{-1}(\omega_i, p)[\Delta\widehat{\mathcal{H}}(\omega_i, p+1) - \Delta\widehat{\mathcal{H}}(\omega_i, p)]\Delta\widehat{\mathcal{H}}^{-1}(\omega_i, p) \end{aligned} \tag{40}$$

when it does not vary greatly for each update, for instance, already close to the optimum.

5.4. Causality issue

Some extra attention must be paid to the separation done in (29). The inversion of the mixing matrix $\Delta\mathcal{H}(\omega_i)$ concerns the causality because its equivalent operation in time domain needs the inversion of an FIR filter, which is probably of nonminimum phase. The postprocessing filter $\Lambda^{-1}[k]$ mentioned in Section 4 is an example in the 2×2 case. In particular, $\Lambda[k]$ is very likely to be nonminimum phase with a nonzero τ. Since we can express $\Delta\mathcal{H}^{-1}(\omega_i)$ as

$$\Delta\mathcal{H}^{-1}(\omega_i) = \frac{\operatorname{adj}\Delta\mathcal{H}(\omega_i)}{\det\Delta\mathcal{H}(\omega_i)}, \tag{41}$$

if we simply omit the denominator that has nothing to do with the separation, which is equivalent to skipping the filtering by Λ^{-1} in time domain, the noncausal problem can be avoided. However, this will cause some signal quality loss, and being more serious, make the dependence of the gradients on the mixing matrix much more complicated than in (33), which leads to high computational complexity. To keep the expression of (33) concise and systematic, we therefore introduce another time delay τ_c and let

$$\Delta\mathcal{H}_c(\omega_i) = \Delta\mathcal{H}(\omega_i)e^{j\omega_i\tau_c}, \tag{42}$$

where $\tau_c \ll N$ still holds. Accordingly, the separation done in (29) is replaced with

$$\begin{aligned} \underline{S}'(\omega_i, p) &= \Delta\mathcal{H}_c^{-1}(\omega_i)\underline{X}(\omega_i, p-\tau) \\ &= \Delta\mathcal{H}^{-1}e^{-j\omega_i\tau_c}(\omega_i)\underline{X}(\omega_i, p-\tau). \end{aligned} \tag{43}$$

It can be easily proved that this modification has no impact on the objective function $J(p)$, and thus the expression in (33) will remain the same.

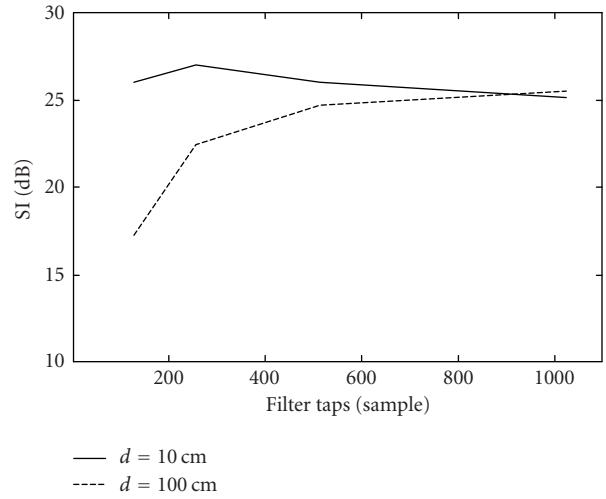


Figure 11: The SI versus the tap number L given to the DRIRs ($T_{60} = 0.27$ second, $\tau = \tau_c = 0$).

5.5. Experimental results

In this section, the BLASS algorithm based on the proposed simplified mixing model is applied on both synthetic and real-world audio signal separation. First a blind separation experiment with synthetic signals is done. A piece of human speech and a piece of music, both lasting 15 seconds, are mixed artificially with four RIRs generated in "Room."

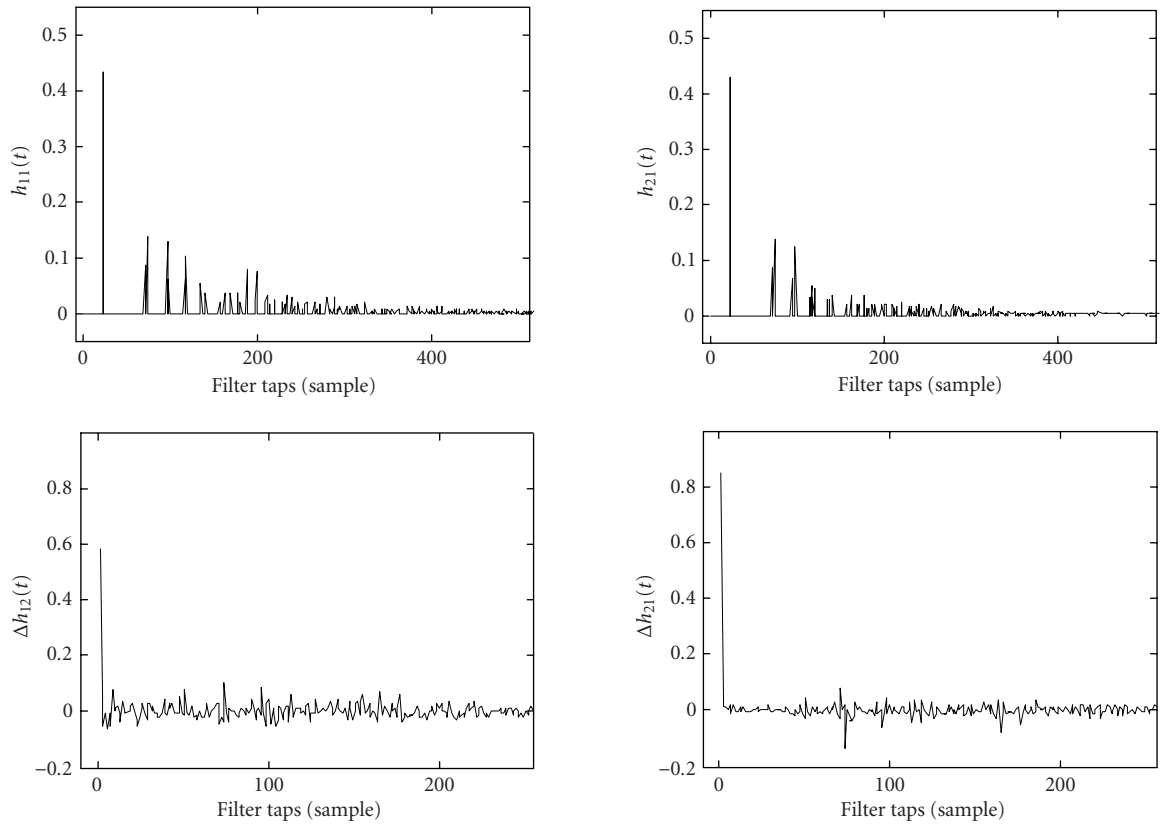
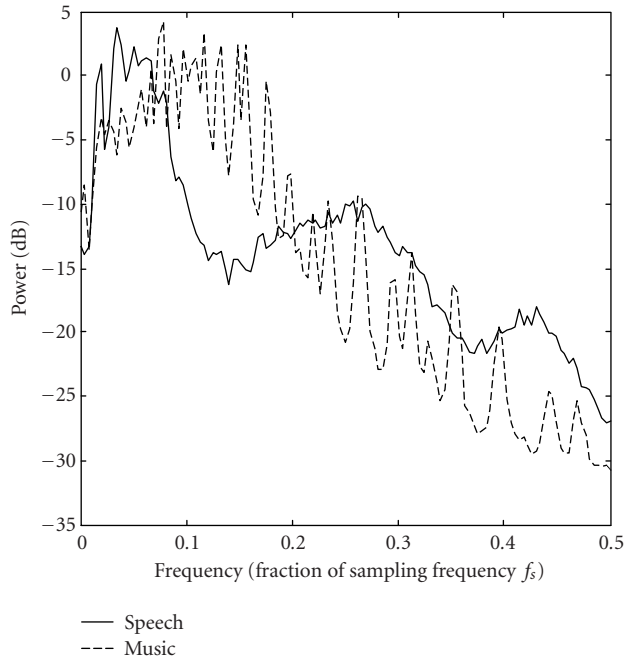


Figure 12: The artificially generated RIRs and the DRIRs obtained by the BLASS.

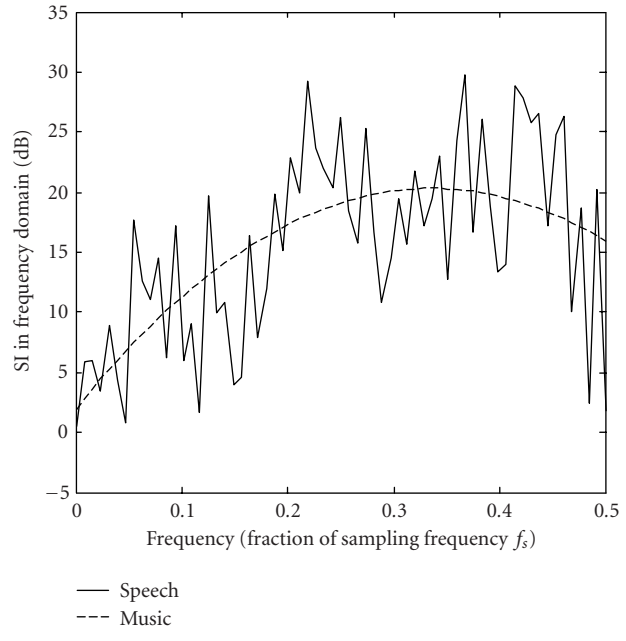
The same setup as described in Figure 10a is used except that the source-to-microphone distance w reduces to 1 m. As seen from Figure 7, this increases the ASI from below 0.05 to above 0.6 in the case of $T_{60} = 0.270$ second and $d = 10$ cm. The sampling frequency f_s is 8 kHz and the four RIRs have 1024 taps each. The signal-to-signal ratio of the sources is set to almost 0 dB. After the BLASS finishes the separation, we take the DRIRs that have been identified and use (26) and (27) to calculate the SI. According to the definition, two sources are switched on alternatively for the SI calculations. We repeat the separation with various number L of taps given to the DRIRs. The results are plotted in Figure 11.

One can observe that in the case of a small microphone spacing, the SI value almost stays the same as L varies from more than 1000 to around 100, while with a large microphone spacing, the SI declines by about 8 dB. The curves manifest that the length L of the DRIRs needed for signal separation can be considerably shortened because of the existence of acoustic similarity between sound propagating paths. From the upper curve, we also see the dependence of the separation performance on the length L of the DRIRs. Too few parameters are not sufficient to perform the separation, while too many parameters bring larger misadjustment in estimation, which only exchanges with a limited amount of increase in adaptability. In Figure 12, the artificially generated RIRs and the two DRIRs obtained by the BLASS are shown in the case of $d = 10$ cm. The RIRs are truncated in order to show fine structures of the early reflection parts. Because of the symmetry of the setup in Figure 10a, we just plot two RIRs. The difference between the spectra of the speech and the music (shown in Figure 13a) causes different shapes of resulting DRIRs, which ideally should be identical. Nevertheless, with a single-pulse-like shape, the DRIRs still imply a strong acoustic path similarity.

Also plotted in Figure 13b is the SI values calculated for each frequency component, where we take the resulting DRIRs in the case of $d = 10$ cm and $L = 1024$ and use white noise as sources. The dashed line describes the average separation level for different frequencies. From there, one can see that with $d = 10$ cm, high frequency components of the sources are separated better than low frequency ones. One reason is that from around 1.4 kHz down to 800 Hz, the music signal is quite dominant, and below 800 Hz the speech prevails, which is not good for separation. The other reason could be that the 10 cm microphone spacing provides acoustic path differences for separation to the low frequency components (greater than 25 cm in this particular case) not as sufficient as to the high-frequency components. This has been briefly discussed also at the end of Section 4. Therefore, depending on different acoustic environments, the microphone spacing to be chosen will be always a trade-off between generating path similarities and differences.



(a)



(b)

Figure 13: (a) The spectra of the sources. (b) The frequencywise SI values.

To show the effectiveness of the BLASS algorithm for real-world data, we design the following experiment with recorded audio signals. The signals are recorded in a room illustrated in Figure 14 with dimensions of about 3 m × 4 m × 3 m (width × depth × height). The same pieces of speech and

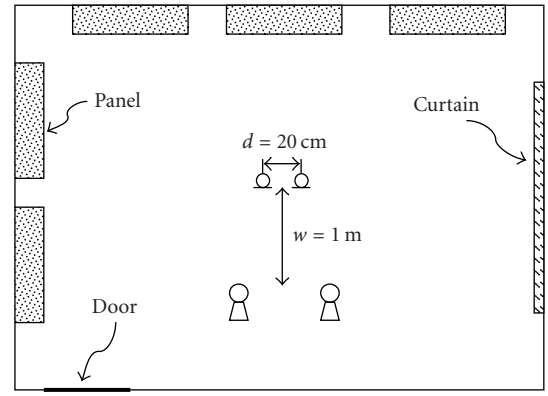


Figure 14: The room for real-world audio signal recording.

music as in the synthetic data experiment are played by two loudspeakers simultaneously and recorded with two microphones. The microphones and loudspeakers are placed as depicted in the figure. The panels hanging on the walls and the curtain can be flipped to choose the smooth or the coarse side. Different combinations lead to different acoustic environments. In this experiment, we use the coarse sides of the panels and the curtain in order to acquire a mild reverberation. The reverberation time, being averaged over various positions, is $T_{60} = 0.335$ second. With different tap numbers L given to DRIRs, the BLASS algorithm is applied to the recorded microphone signals and the results are shown in Figure 15.

One can see that the SI curve in Figure 15a reveals the pattern similar to that in Figure 11 with $d = 10$ cm. The SI value stays around 20 dB as L decreases from 1024 to about 300, and the best separation appears with $L \approx 380$. This is understood when we look at the DRIRs acquired by the BLASS in Figure 15b. They have stronger and more dense side taps than the DRIRs obtained with synthetic data in Figure 12, but they still look quite single-pulse like, which indicates significant similarities existing in acoustic paths and allow the reduction of filter taps. The best SI is 5 dB lower than that with the synthetic data because of the more complex acoustic environment in the real world, while the corresponding number of taps is about 150 more since the acoustic similarities reduce due to a larger microphone spacing ($d = 20$ cm).

Finally, we apply the BLASS to one of the recordings provided in [31] that are normally considered as benchmarks. Two speakers have been recorded speaking simultaneously. Speaker 1 says the digits from one to ten in English and speaker 2 counts at the same time the digits in Spanish (uno, dos, ...). The recording has been done in a normal office room. The distance between the speakers and the microphones is about 60 cm in a square ordering. The sampling frequency is 16 kHz. Take $L = 128$ and $\tau = \tau_c = 0$ because the RIRs are likely to be minimum phase in such a source-to-microphone distance. One piece (around 1.5 second) of the mixtures and the corresponding separated signals are shown

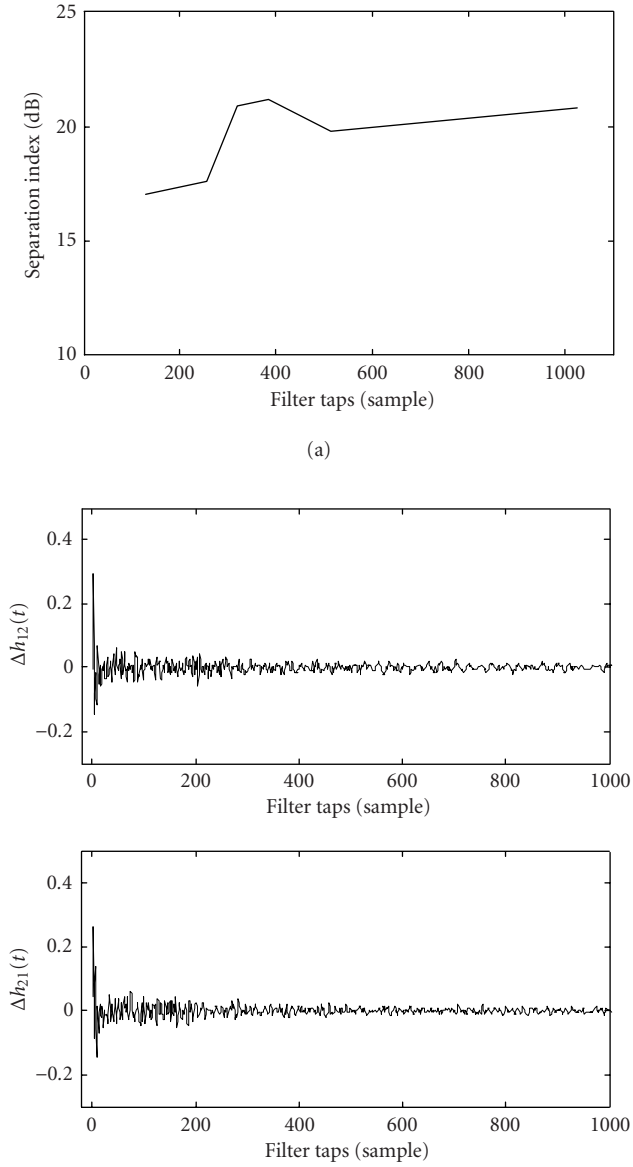


FIGURE 15: (a) The SI versus the tap number L given to DRIRs, $\tau = \tau_c = 0$. (b) The DRIRs acquired by the BLASS.

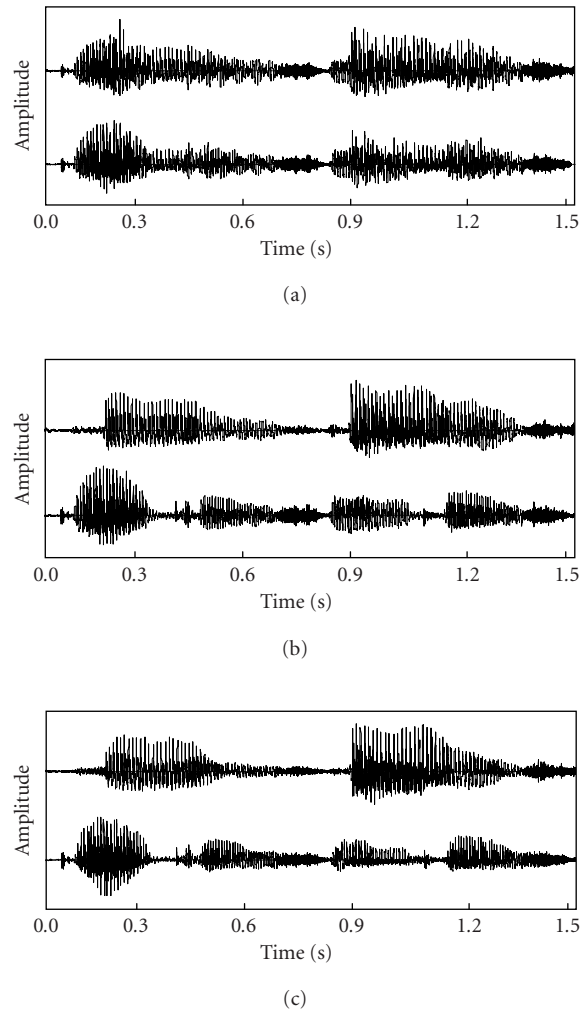


FIGURE 16: The separation results of the real-world data. (a) Microphone recordings. (b) Separated signals obtained in [31]. (c) Separated signals with BLASS.

in Figure 16. In Figure 17, we plot the two DRIRs obtained by the BLASS. Inspecting by sight or by hearing, one can hardly find any separation performance loss with the reduced filter number and shortened filter length. This confirms the statement in Section 4 that the proposed model is in general also applicable to the case of large microphone spacings provided that the filter length L is properly chosen.

An overview of the experiment results is given in Table 1, where N_{filter} and N_{coef} denote the number of filters and the number of filter taps to be adjusted, respectively. SI loss means the separation degradation due to the use of the proposed model instead of a conventional model. The total number of filter coefficients needed to achieve a comparable performance is considerably reduced. In Figure 18, one can see more generally the reduction of the needed filter taps as the number of sources/microphones n increases, where

$$\frac{N_{\text{coef,pro}}}{N_{\text{coef,con}}} = \beta - \frac{\beta}{n}, \tag{44}$$

$N_{\text{coef,pro}}$ and $N_{\text{coef,con}}$ represent the number of filter taps needed for the proposed model and a conventional model, respectively, and β is the ratio between the filter lengths used in these two models provided that the separation performances are comparable. With a large number of sources, the ratio $N_{\text{coef,pro}}/N_{\text{coef,con}}$ approaches β.

Other experimental aspects with real-world data can be found in [32], where the BLASS has been implemented on a TI TMS320C6701 DSP system and is able to realize audio signal separation in real time.

Table 1: An overview of the experiment results regarding the reduction of the total filter taps.

Data type		Mixing model	N_{filter}	L	N_{coef}	SL loss
Synthetic		normal	4	1024	4096	negligible
		proposed	2	∼250	∼500	
Real world	Own recorded	normal	4	1024 (expected)	4096 (expected)	negligible
		proposed	2	∼400	∼800	
	Benchmark	normal	4	1024 (expected)	4096 (expected)	Hardly recognized by hearing
		proposed	2	128	256	

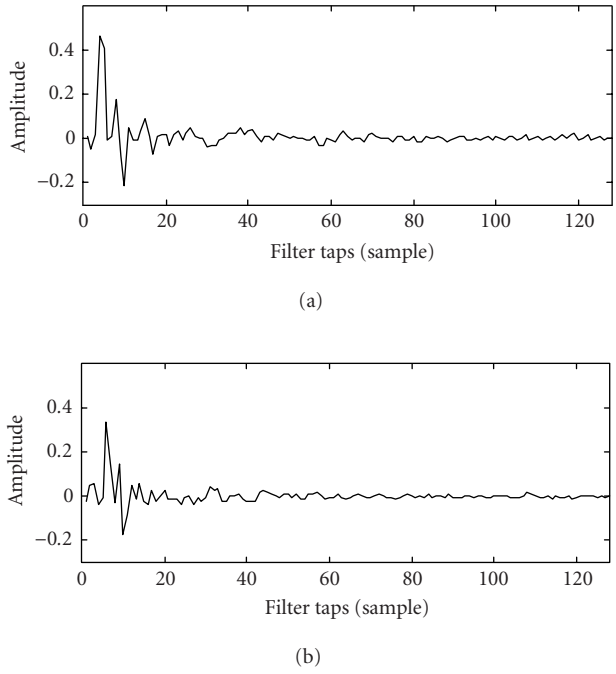


Figure 17: The DRIRs acquired by the BLASS. (a) $\Delta h_{12}[k]$. (b) $\Delta h_{21}[k]$.

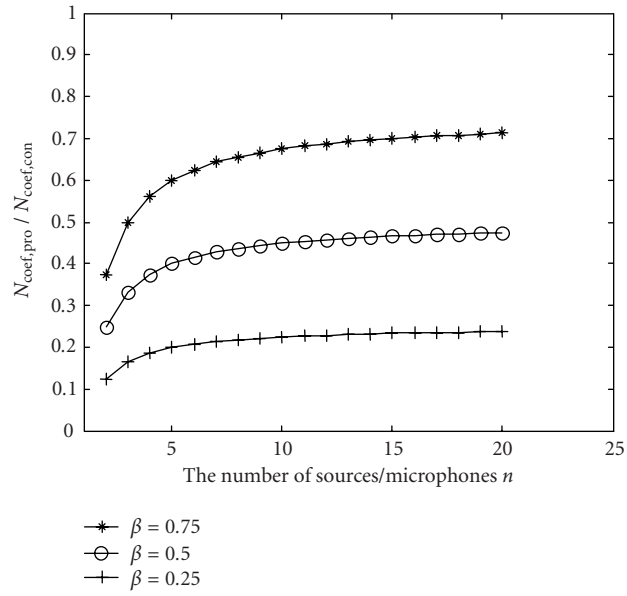


Figure 18: Reduction of the total filter taps using the proposed model.

6. CONCLUSIONS

In this paper, the concept of acoustic similarities existing between two propagating paths of a sound is presented. In order to quantitatively describe the similarity, an acoustic similarity index (ASI) is defined and studied. There are three ways to increase the ASI value, which are (1) reducing the microphone spacing, (2) making the environment less reverberant, and (3) decreasing the source-to-microphone distance. Then a new mixing model of a multispeaker-multimicrophone setup is proposed. The model is proved to be feasible in practice and can be applied to simplify an audio signal separation problem. With a reasonably high ASI, for example, by means of closely spacing microphones (within 5–10 cm), the model can relieve the computational load of the separation algorithm by considerably reducing the number and length of the filters to be adjusted. It is also applicable in the normal microphone spacings if the filters are provided with enough taps. Therefore, the implementation of a blind audio signal separation (BLASS) is used specifically for the proposed algorithm.

In principle, various BSS algorithms can be designed on the proposed model. As an example, in this paper, we have developed a BLASS in order to cope with real and more complicated situations. BLASS only uses the second-order statistics and performs efficiently in frequency domain. Its effectiveness is shown by the separation results of both synthetic and real-world signals.

REFERENCES

[1] J. Herault and C. Jutten, "Space or time adaptive signal processing by neural network models," in *Neural Networks for Computing: AIP Conference Proceedings*, J. S. Denker, Ed., vol. 151, pp. 206–211, American Institute for Physics, Snowbird, Utah, USA, April 1986.

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

[3] K. Torkkola, "Blind separation for audio signals—are we there yet?" in *Proc. 1st International Workshop on Independent Component Analysis and Blind Signal Separation (ICA '99)*, pp. 239–244, Aussois, France, January 1999.

[4] H.-L. N. Thi and C. Jutten, "Blind source separation for convolutive mixtures," *Signal Processing*, vol. 45, no. 2, pp. 209–229, 1995.

[5] S. Van Gerven and D. Van Compernolle, "Signal separation by symmetric adaptive decorrelation: Stability, convergence, and uniqueness," *IEEE Trans. Signal Processing*, vol. 43, no. 7, pp. 1602–1612, 1995.

[6] D. Yellin and E. Weinstein, "Multichannel signal separation: Methods and analysis," *IEEE Trans. Signal Processing*, vol. 44, no. 1, pp. 106–118, 1996.

[7] J. T. Ngo and N. A. Bhadkamkar, "Adaptive blind separation of audio sources by a physically compact device using second-order statistics," in *Proc. 1st International Workshop on Independent Component Analysis and Signal Separation (ICA '99)*, pp. 257–260, Aussois, France, January 1999.

[8] J. Garas, *Room Impulse Response*, Version 2.1, http://www.dspalgorithms.com/room/room25.html.

[9] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[10] J. Garas, *Adaptive 3D sound systems*, Ph.D. thesis, Eindhoven University of Technology, Eindhoven, The Netherlands, 1999.

[11] V. Capdevielle, C. Serviere, and J. Lacoume, "Blind separation of wide-band sources in the frequency domain," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '95)*, pp. 2080–2083, Detroit, Mich, USA, May 1995.

[12] F. Ehlers and H. G. Schuster, "Blind separation of convolutive mixtures and an application in automatic speech recognition in noisy environment," *IEEE Trans. Signal Processing*, vol. 45, no. 10, pp. 2608–2612, 1997.

[13] P. He, P. C. W. Sommen, and B. Yin, "A realtime DSP blind signal separation experimental system based on a new simplified mixing model," in *Proc. International Conference on Trends in Communications (EUROCON '01)*, pp. 467–470, Bratislava, Slovak Republic, July 2001.

[14] P. Comon, "Independent component analysis," in *Proc. International Workshop on Higher-Order Statistics (HOS '91)*, pp. 111–120, Chamrousse, France, July 1991.

[15] L. Tong, R. Liu, V. Soon, and Y. Huang, "Indeterminacy and identifiability of blind identification," *IEEE Trans. Circuits and Systems*, vol. 38, no. 5, pp. 499–509, 1991.

[16] J.-F. Cardoso, "Iterative techniques for blind source separation using only fourth order cumulants," in *Proc. 6th European Signal Processing Conference (EUSIPCO '92)*, pp. 739–742, Brussels, Belgium, August 1992.

[17] D. Yellin and E. Weinstein, "Criteria for multichannel signal separation," *IEEE Trans. Signal Processing*, vol. 42, no. 8, pp. 2158–2168, 1994.

[18] C. Jutten and J. Herault, "Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture," *Signal Processing*, vol. 24, no. 1, pp. 1–10, 1991.

[19] J. Karhunen, L. Wang, and R. Vigario, "Nonlinear PCA type approaches for source separation and independent component analysis," in *Proc. IEEE International Conference on Neural Networks (ICNN '95)*, pp. 995–1000, Perth, Western Australia, Australia, November–December 1995.

[20] B. A. Pearlmutter and L. C. Parra, "Maximum likelihood blind source separation: A context-sensitive generalization of ICA," in *Proc. Advances in Neural Information Processing Systems (NIPS '96)*, vol. 9, pp. 613–619, MIT Press, Denver, Colo, USA, December 1996.

[21] A. J. Bell and T. J. Sejnowski, "An information maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[22] S.-I. Amari, A. Cichocki, and H. H. Yang, "A new learning algorithm for blind signal separation," in *Proc. Advances in Neural Information Processing Systems (NIPS '96)*, vol. 8, pp. 757–763, MIT Press, Denver, Colo, USA, December 1996.

[23] M. Girolami and C. Fyfe, "Extraction of independent signal sources using a deflationary exploratory projection pursuit network with lateral inhibition," *IEE Proceedings on Vision, Image and Signal Processing Journal*, vol. 14, no. 5, pp. 299–306, 1997.

[24] L. Molgedey and H. Schuster, "Separation of independent signals using time-delayed correlations," *Physical Review Letters*, vol. 72, no. 23, pp. 3634–3637, 1994.

[25] K. Abed-Meraim, A. Belouchrani, J.-F. Cardoso, and E. Moulines, "Asymptotic performance of second order blind source separation," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '94)*, vol. 4, pp. 277–280, Adelaide, Australia, April 1994.

[26] S. Van Gerven and D. Van Compernolle, "On the use of decorrelation in scalar signal separation," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '94)*, pp. 57–60, Adelaide, Australia, April 1994.

[27] E. Weinstein, M. Feder, and A. V. Oppenheim, "Multi-channel signal separation by decorrelation," *IEEE Trans. Speech, and Audio Processing*, vol. 1, no. 4, pp. 405–413, 1993.

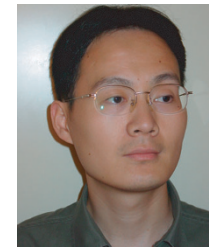
[28] K. Matsuoka, M. Ohya, and M. Kawamoto, "A neural net for blind separation of nonstationary signals," *Neural Networks*, vol. 8, no. 3, pp. 411–419, 1995.

[29] L. Parra, C. Spence, and B. De Vries, "Convolutive blind source separation based on multiple decorrelation," in *Proc. IEEE Workshop on Neural Networks for Signal Processing (NNSP '98)*, pp. 23–32, Cambridge, UK, September 1998.

[30] H. Kawahara and T. Irino, "Exploring temporal feature representations of speech using neural networks," Tech. Rep. SP88-31, IEICE, Tokyo, Japan, July 1988.

[31] T.-W. Lee, A. Ziehe, R. Orglmeister, and T. J. Sejnowski, "Combining time-delayed decorrelation and ICA: towards solving the cocktail party problem," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '98)*, vol. 2, pp. 1249–1252, Seattle, Wash, USA, May 1998, Authors provide signals and results at http://www.cnl.salk.edu/~tewon/.

[32] J. van de Laar, E. A. P. Habets, J. D. P. A. Peters, and P. A. M. Lokkart, "Adaptive blind audio signal separation on a DSP," in *Proc. 12th Annual Workshop on Circuits, Systems and Signal Processing (ProRISC '01)*, pp. 475–479, Veldhoven, The Netherlands, November 2001.



Bin Yin received his B.S., M.S., and Ph.D. degrees in electrical engineering from Southeast University (SEU), Nanjing, China, in 1992, 1995, and 1998, respectively. From 1999 to 2001, as a Postdoctor, he was with the Faculty of Electrical Engineering at Eindhoven University of Technology, the Netherlands. Currently he is a Senior Scientist in Philips Research Laboratories (Nat.Lab.), Eindhoven, the Netherlands. His main interests of research involve model identification, adaptive filtering, and adaptive array signal processing, with applications in adaptive control systems, blind audio signal separation, and, currently, the signal processing in optical storage.

Piet C. W. Sommen received the Ingenieur degree in electrical engineering from Delft University of Technology in 1981 and his Ph.D. degree from Eindhoven University of Technology, the Netherlands, in 1992. From 1981 to 1989 he was with Philips Research Laboratories, Eindhoven, and since 1989, he has been with the Faculty of Electrical Engineering at Eindhoven University of Technology, where he is currently an Associate Professor. Dr. Sommen is involved in internal and external courses, all dealing with different basic and advanced signal processing topics. His main field of research is in adaptive array signal processing, with applications in acoustic communication systems. Dr. Sommen is the Editor of EURASIP Newsletter.

Peiyu He received her B.S. degree from Tsinghua University, Beijing, China and her M.S. degree from Sichuan University, Chengdu, China, in 1986 and 1989, respectively. From 2000 to 2001, she was with the Faculty of Electrical Engineering at Eindhoven University of Technology, the Netherlands, as a visiting Researcher. Currently she is Professor in Sichuan University, Chengdu, China. Her main field of research is adaptive signal processing, with applications in telecommunications, such as channel equalization, acoustic echo cancellation, and blind signal separation.