EXPLOITING ACOUSTIC SIMILARITY OF PROPAGATING PATHS FOR AUDIO SIGNAL SEPARATION

Bin Yin , Piet Sommen and Peiyu He

Eindhoven University of Technology
P.O. Box 513, 5600 MB Eindhoven, the Netherlands
Email: P.C.W.Sommen@tue.nl

ABSTRACT

Blind signal separation can easily find its position in audio applications where mutually independent audio sources need to separate 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 computation load is mainly caused by either direct or indirect estimation of thousands of acoustic parameters. Aiming at the complexity reduction, in this paper the similarity of acoustic paths is first investigated in 1-speaker-2-microphone case. Then a simplified mixing model is proposed and furthermore the simulation results show the effectiveness of the model in audio signal separation.

1. INTRODUCTION

Blind Signal Separation (BSS) deals with the problem of separating mutually independent sources from their mixtures while both the mixing process and the sources are unknown. For acoustical applications, it can be used to extract individual audio sources from microphone signals when the sources are simultaneously active. Therefore, it becomes possible, e.g. in a teleconferencing system, to pick up one desired speech under a relatively low signal-to-noise ratio.

In a reverberant environment, usually the mixing process can be modelled as

$$\underline{x}[k] = H[k] * \underline{s}[k] \,, \tag{1}$$

where $\underline{s}[k] = (s_1[k], \cdots, s_n[k])^T$ and $\underline{x}[k] = (x_1[k], \cdots, x_n[k])^T$ denote the signal 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 matrix of filters whose element h_{ij} expresses the room impulse response (RIR) from the jth source to the ith microphone, which can be approximately modelled as an FIR filter. The symbol $*$ denotes linear convolution. We assume the numbers of sources and microphones are the same. The sources can be separated either by inverting the mixing process (forward model method), or by finding a demixing process directly (backward model method)[1]. A RIR is normally a filter of considerably long length, e.g., having $1000 - 2000$ taps with $8kHz$ sampling frequency in a usual office. Hence, in both methods 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.

Aiming at the feasibility of realtime applications, we proposed a BSS algorithm with a simplified mixing model which takes advantage of acoustic propagation similarities [2]. In this paper, the investigation is focused on the simplified mixing model where more insight is specifically given into the acoustic similarity. First an acoustic similarity index (ASI) is defined. Then we concentrate on an 1-speaker-2-microphones system to study the relationship between microphone spacing and ASI. Finally a setup is designed to measure the separation effect. Several advantages are discussed as well which appear to benefit a real time implementation.

2. SIMILARITY OF ACOUSTIC PATHS

Consider an 1-speaker-2-microphones setup. The RIRs from the speaker to two microphones are described by $h_{11}[k]$ and $h_{21}[k]$, each of length L_0, respectively. A difference room impulse response $\Delta h_{21}[k]$ (DRIR) can be defined as

$$\Delta h_{21}[k] = (h_{21} * h_{11}^{-1})[k] \,. \tag{3}$$

$\Delta h_{21}[k]$ often has to be a non-causal infinite double sided filter due to the inversion of h_{11} which shows non-minimum phase characteristic in most acoustic conditions. However, for practical applications we shift it by a delay of τ samples and then truncate it such that

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

$\Delta h_{21}[k]$ from here on denotes a causal FIR filter of length L, and $\mathbf{Tr}\{\cdot\}$ denotes the truncation.

Now we are in the position of defining an acoustic similarity index (ASI). Suppose $\Delta h_{21}[k]$ has coefficients $\underline{c} = [c_1, \cdots, c_L]^T$. An ASI can be expressed as the following:

$$\text{ASI}_{[h_{11},h_{21}]} = \exp\{-\frac{\| \underline{c} - E_m \underline{c} \|_2}{\| E_m \underline{c} \|_2}\} , \qquad (5)$$

where E_m represents a matrix with 1 at the mmth position and zeros otherwise, and $E_m \underline{c} = [0, \cdots, c_m, \cdots, 0]^T$ where $c_m = \max_i\{c_i^2\}$.

When two microphones are exactly of the same location, Δh_{21} becomes a single pulse. According to the definition (5), ASI equals one. Assuming that the coefficients of RIRs vary continuously within a small spatial range, from (3), we have the following satisfied in z domain:

$$\Delta h_{21}(z^{-1}) = 1 + \frac{d(z^{-1})}{h_{11}(z^{-1})} , \qquad (6)$$

where

$$d(z^{-1}) = h_{21}(z^{-1}) - h_{11}(z^{-1}) = \sum_{i=0}^{L-1} (a_i - b_i) z^{-i} . \quad (7)$$

a_i's and b_i's represent the coefficients of two RIRs respectively. Due to the continuity assumption, the magnitude of $a_i - b_i$ will be small in accordance with the close microphone spacing. Thus $\frac{d(z^{-1})}{h_{11}(z^{-1})}$ only has a small contribution to $\Delta h_{21}(z^{-1})$. Visually, $\Delta h_{21}[k]$ in this case consists of not only a main pulse but also a one-side or double-side tail, which makes ASI smaller than one. As the microphone spacing increases, the tail will become longer in time and higher in amplitude because of the increased $a_i - b_i$. Consequently, ASI decreases furthermore towards zero. Therefore, ASI reflects the similarity of two acoustic paths. In particular ASI$= 1$ indicates the situation of highest similarity where two microphones are in the same location.

In practice, the acoustic conditions in a room are often so complicated because of, e.g., other moving objects and irregularly arranged furniture, that the continuity assumption can not hold especially when microphones are moving further away from each other. However, that ASI varies from 1 to 0 as the distance between microphones is enlarged is very likely to be the truth in general. This fact is demonstrated by the simulation results shown below.

In the simulations we use the software 'Room' which can generate RIRs for certain acoustic conditions. The dimension of the virtual room is $5\text{m}\times 4\text{m}\times 3\text{m}$ ($l \times w \times h$). 1 speaker and 2 microphones are set up as a triangle in the center of the room and the speaker is about 1m from the microphones. We apply the efficient block frequency domain adaptive filter (BFDAF) algorithm for DRIR estimation. A mean square error (MSE) is defined as usual to evaluate the estimation results.

The simulation results about how ASI changes regarding microphone spacing d are shown in Fig.1. For a certain reverberation level, ASI decreases as d increases. ASI keeps very close to one within 50cms in a weakly echoic environment (solid line), while in a more reverberant case ASI declines drastically at first several cms (dotted line), and after that ASI stays almost the same over several 10cms. This implies that one should expect a more *Delta*-like DRIR within around 10cm microphone spacing in normal room acoustics. In all cases we take $L = 512$.

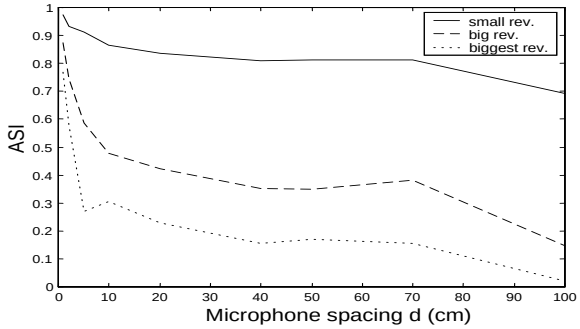

Fig. 1. The relationship between microphone spacing d and ASI for different reverberations.

A possibility of filter tap reduction is reflected by the simulation results in Fig.2. For a certain MSE requirement (corresponding to some separation effect), fewer taps are needed with a smaller microphone spacing. It is because in this case the DRIR is more like a pure pulse as demonstrated above in Fig. 1. In both experiments $\tau = 0$ is taken for simplicity.

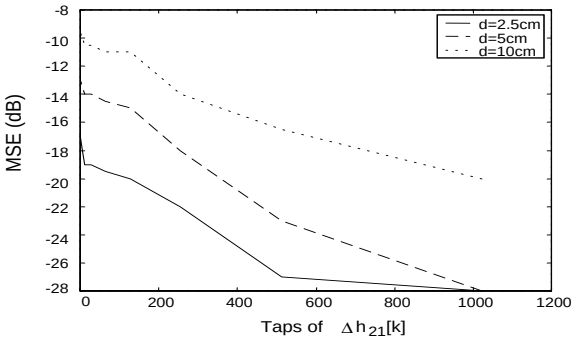

Fig. 2. The relationship between filter taps and MSE for different microphone spacing.

3. A SIMPLIFIED MIXING MODEL

In the case of high ASI, introducing a simplified mixing model becomes very attractive to audio signal separation

which normally suffers from high complexity. Applying the relationship (4), we can rewrite (1) as

$$\underline{x}[k-\tau] = \Delta H[k] * \underline{\tilde{s}}[k] \,, \tag{8}$$

where $\underline{x}[k-\tau] = (x_1[k-\tau], \cdots, x_n[k-\tau])^T$, $\underline{\tilde{s}}[k] = ((h_{11} * s_1)[k], \cdots, (h_{nn} * s_n)[k])^T$ and

$$\Delta H[k] = \begin{pmatrix} \delta[k-\tau] & \Delta h_{12}[k] & \cdots & \Delta h_{1n}[k] \\ \vdots & \vdots & \ddots & \vdots \\ \Delta h_{n1}[k] & \Delta h_{n2}[k] & \cdots & \delta[k-\tau] \end{pmatrix} . \tag{9}$$

For more efficient implementation, a frequency domain counterpart of the mixing model 8 is also given as follows,

$$\underline{X}(\omega, p-\tau) \approx \Delta\mathcal{H}(\omega)\underline{\widetilde{S}}(\omega, p) \,, \omega = 0, \cdots, \frac{(N-1)}{N} 2\pi \,, \tag{10}$$

where ω denotes the frequency and N denotes the number of points in the discrete Fourier transform (DFT). $\underline{X}(\omega, p-\tau) = (X_1(\omega, p-\tau), \cdots, X_n(\omega, p-\tau))^T$ represents the DFT of the microphone signals where $X_i(\omega, p-\tau)$ comes from the DFT of the ith microphone signal vector $\underline{x}_i[p-\tau] = (x_i[p-\tau] \cdots, x_i[p-\tau+N-1])^T$, starting at $p-\tau$ and of length N, which is given by

$$X_i(\omega, p-\tau) = \Sigma_{\kappa=0}^{N-1} e^{-j\omega\kappa} x_i[p-\tau+\kappa] \,. \tag{11}$$

The approximation in (10) is due to expressing linear convolution by circular convolution and linear time shift by circular time shift. $\underline{\widetilde{S}}(\omega, p-\tau)$ has a similar expression as above. $\Delta\mathcal{H}(\omega)$ denotes the Fourier transform of the filter matrix $\Delta H[k]$ and can be expressed as

$$\Delta\mathcal{H}(\omega) = \begin{pmatrix} e^{-j\omega\tau} & \cdots & \Delta h_{1n}(\omega) \\ \vdots & \ddots & \vdots \\ \Delta h_{n1}(\omega) & \cdots & e^{-j\omega\tau} \end{pmatrix} . \tag{12}$$

Since the components in the vector $\underline{\tilde{s}}$ are mutually independent, the signal separation can be actually 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) The number of filters to be estimated is reduced from n^2 to $n(n-1)$.

2) Instead of the sources themselves we recover the signals just in front of the microphones (i.e. the sources convolved by RIRs) which often sound more natural. In a strongly reverberant environment, for instance, a large conference hall, the RIR h_{ii}'s still need to be inverted to improve the intelligibility of separated signals.

3) It is shown in the previous section that the ASI will get closer to one with the distance between microphones decreasing to a small range (Fig.1). This implies that the DRIR Δh_{ij}'s appear more *Delta*-function-like so that fewer coefficients are required (Fig.2). As a result, the computational load for the mixing model estimation can be significantly reduced.

Hence, a closely spaced microphone array provides a realtime audio signal separation with a big possibility.

4. SIGNAL SEPARATION EXPERIMENTS

An experimental scheme is designed to demonstrate the effectiveness of the proposed mixing model.

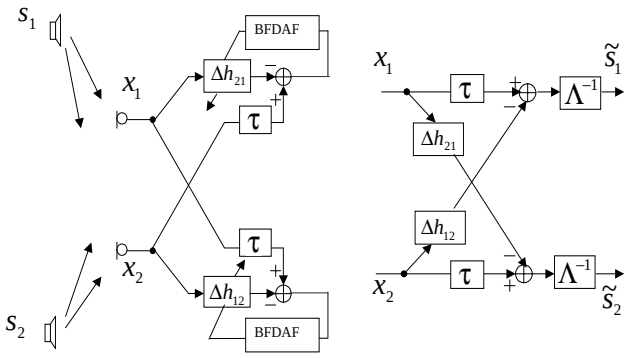

Fig. 3. The signal separation method (2×2 case).

Fig.3 shows a setup of the separation in 2×2 case . The left diagram expresses the parameter estimation part where two BFDAF algorithms are used parallelly, and the right describes the separation part. With the postprocessing filter $\Lambda[k] = \delta[k-2\tau] - (\Delta h_{21} * \Delta h_{12})[k]$, this part functions exactly as an inversion of the mixing matrix $\Delta H[k]$. Δh_{ij} is measured only when s_j is present. The measurement may be first done with alternatively active sources and then the separation is switched on after convergence of the parameters. Like BFDAF, the separation may be implemented efficiently in frequency domain as well. Its equivalent frequency domain structure is shown below where the input $\underline{x}_i$ to FFT is the buffered microphone signal vector of length N. One can see that the separation is independently operated for each frequency bin ω_i, which converts the convolutively mixing problem into an instantaneous one.

In principle, the structure in Fig.4 is only an approximation of its time domain counterpart due to the quasi equality in Equation (10). Nevertheless, a good approximation can be always obtained if the conditions $L \ll N$ and $\tau \ll N$ are satisfied.

The time delay τ is introduced for a causal stable inversion of non-minimum phase filters. The proper choice of τ depends on different factors, e.g., the wall refection and the distance between sources and microphones. In particular, if the reverberation is present quite weakly and the audio source is located closely to its microphone (say within several $10cm$s) τ may be chosen as zero since h_{ii} in this case

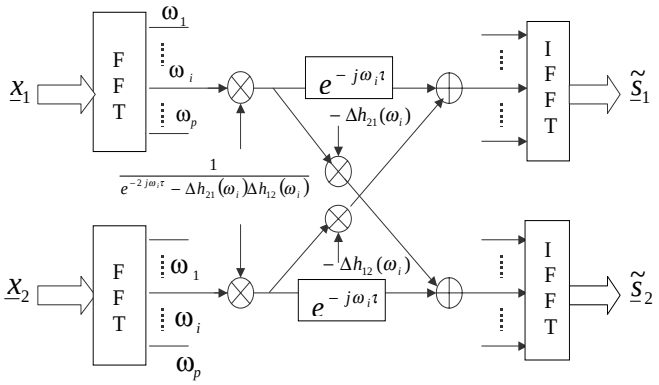

Fig. 4. The separation in frequency domain (2×2 case).

shows minimum phase characteristics. The detailed experimental results can be found in [3].

One can see that in Fig.3 the filtering Λ^{-1} is again concerning the inversion of a non-minimum phase filter. To solve the problem, one possibility is simply moving it away because it has nothing to do with the separation (correspondingly omitting the item $\frac{1}{e^{-2j\omega\tau}-\Delta h_{12}(\omega)\Delta h_{21}(\omega)}$ in Fig. 4); another possibility is keeping it there to improve the sound quality at the cost of another extra time delay.

In order to evaluate the separation results, we define the following separation index (SI):

$$SI_i = 10\log\frac{\Sigma_{t=-T}^{T}\tilde{s}_i^2[k+t]/x_i^2[k+t]}{\Sigma_{t=-T}^{T}\tilde{s}_j^2[k+t]/x_j^2[k+t]}, \quad (13)$$

only s_i is active, $i,j=1,2, i\neq j$,

$$SI = (SI_1+SI_2)/2. \quad (14)$$

T is a proper time period. The separating outputs are normalized by the corresponding microphone signals so that the influence on separation caused by the variation of signal levels can be reduced as much as possible. SI may be also expressed via z domain as:

$$SI_i = 10\log\oint|\frac{1-\Delta h_{ji}(z^{-1})\Delta\hat{h}_{ij}(z^{-1})}{\Delta h_{ji}(z^{-1})-\Delta\hat{h}_{ji}(z^{-1})}|^2\mathrm{d}z\ , \quad (15)$$

where $\Delta\hat{h}_{ij}$ denotes the estimate of Δh_{ij} and for simplicity τ is assumed to be zero.

Still using the virtual experimental setup in Section 2, we study the relationship of the microphone spacing d and SI. The results, which are obtained for the most reverberant case, are shown in Fig.5. In general, SI is increasing as the microphone spacing is getting small. The corresponding increasing of MSE is also presented there. However, when d becomes extremely small (around 1cm), instead of further increasing SI comes across a huge drop. This phenomenon may be explained by (15). With a very close microphone spacing, the denominator $\Delta h_{ji}-\Delta\hat{h}_{ji}$ approaches zero due to the decreased MSE. Meanwhile, the numerator approaches zero as well because Δh_{ji} and $\Delta\hat{h}_{ij}$ all become a nearly unit impulse. Unfortunately, with a limited computing accuracy the competition result is that the denominator fails! The limit of this trend can be more easily understood: it is impossible to do any separation (with unlimited computing accuracy) when two microphones are placed at the same point.

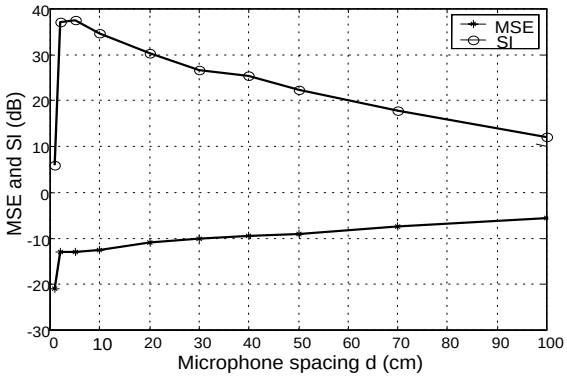

Fig. 5. Relationship of the microphone spacing d and SI.

5. CONCLUSIONS

A simplified mixing model is proposed to make a realtime audio signal separation possible. By placing multiple microphones close to each other, a significant computation reduction can be gained. The proper microphone spacing for 2×2 case is around 10 cms (but never close to 1cm). To cope with a more complicated situation, e.g., moving speakers, a blind separation algorithm which only uses cross-correlation information of the output signals can be developed, one possibility of which is shown in [2].

6. REFERENCES

[1] L.Parra and C.Spence, "Convolutive blind source separation based on multiple decorrelation", in *Proc. of NNSP98*, Cambridge, UK, September, 1998.

[2] B.Yin and P.C.W.Sommen, "A new convolutive blind signal separation algorithm based on second order statistics using a simplified mixing model", in *Proc. EUSIPCO 2000*, Tampere, Finland, Sept., 2000, pp2049-2052.

[3] 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. of EUROCON'2001*, Bratislava, Slovak Republic, July, 2001.