Audio Demixing with Decorrelation, Cross Cancellation, Normalization, and Regularization

Sean Webster

August 17, 2011

Abstract

Many methods for demixing audio exist, but their recovered demixing coefficients may not be very close to the actual values. These methods then do not accurately describe the environment the sources were recorded in. By using multiple methods and enforcing a decaying model, the coefficients can come closer to the actual values. These coefficients can then give us a better picture of the environment and better demixing results.

1 Introduction

The basic equation used in this problem is as follows:

\[ x = A \ast s \] (1.1)

Where \( s \) contains two demixed signals, \( x \) contains two different mixtures of the two signals in \( s \), \( A \) is the matrix used to mix the signals, and \( \ast \) denotes the convolution function. \( A \) is a \( 2 \times 2 \) matrix of vectors \( a_{xy} \) of length \( q \). (1.1) can then be written out as

\[
\begin{pmatrix}
  x_1 \\
  x_2
\end{pmatrix} =
\begin{pmatrix}
  a_{11} \ast & a_{12} \ast \\
  a_{21} \ast & a_{22} \ast
\end{pmatrix}
\begin{pmatrix}
  s_1 \\
  s_2
\end{pmatrix}
\] (1.2)

The goal of the audio demixing problem is to get \( s_1 \) and \( s_2 \) from \( x_1 \) and \( x_2 \) by way of recovering \( A \). As described later, there is a better way of writing out this problem, but the basic idea remains the same: given \( x_1 \) and \( x_2 \), estimate \( A \) and then recover \( s_1 \) and \( s_2 \).
The method used in this attempt to solve the audio demixing problem is a combination of the decorrelation method discussed in [2], which is discussed further in section 2, and a cross cancellation method similar to one discussed in [3], which is discussed further in section 3. These methods are combined with extra terms that take out the trivial case where every entry in $A$ is 0 (discussed in section 4), and enforce a decaying model not unlike the one discussed in [1] (discussed in section 5).

2 Decorrelation

With our current definition of our problem, solving for both $A$ and $s$ is quite difficult. This can be alleviated, however, with partial inversion. If we use

$$A^{-1} = \frac{1}{\det A} \begin{pmatrix} a_{22}^* & -a_{12}^* \\ -a_{21}^* & a_{11}^* \end{pmatrix}$$ (2.1)

and multiply both sides of (1.1) with it, we come up with

$$A^{-1}x = s$$ (2.2)

which we can rewrite as

$$\det As = \begin{pmatrix} a_{22}^* & -a_{12}^* \\ -a_{21}^* & a_{11}^* \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$ (2.3)

If we define $v$ as $\det As$, then we’ve come from

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} a_{11}^* & a_{12}^* \\ a_{21}^* & a_{22}^* \end{pmatrix} \begin{pmatrix} s_1 \\ s_2 \end{pmatrix}$$ (2.4)

to

$$\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} a_{22}^* & -a_{12}^* \\ -a_{21}^* & a_{11}^* \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$ (2.5)

The first assumption this method makes is that $s_1$ and $s_2$ are independent. Furthermore, it is expected
that the two signals are independent even if one of the signals is shifted by a small number. With these two assumptions, we expect \( E[s_1(t)s_2(t - n)] = 0 \). Since \( v_1 \) and \( v_2 \) are dependent on only \( s_1 \) and \( s_2 \), we also expect

\[
E[v_1(t)v_2(t - n)] = 0
\]

(2.6)

to be true. If we substitute the values of \( v_1 \) and \( v_2 \) derived from (2.5) into (2.6), we get

\[
-a_{22}a_{21}E[x_1(t)x_1(t - n)] + a_{12}a_{21}E[x_2(t)x_1(t - n)] + a_{22}a_{11}E[x_1(t)x_2(t - n)] - a_{12}a_{11}E[x_2(t)x_2(t - n)] = 0
\]

(2.7)

If we define \( C_{ij}^n = E[x_i(t)x_j(t - n)] \), then (2.7) becomes

\[
-a_{22}a_{21}C_{11}^n + a_{12}a_{21}C_{21}^n + a_{22}a_{11}C_{12}^n - a_{12}a_{11}C_{22}^n = 0
\]

(2.8)

which can then be written in matrix notation to become

\[
\begin{pmatrix}
a_{22} & a_{21} \\
a_{12} & a_{11}
\end{pmatrix}
\begin{pmatrix}
-C_{11}^n & C_{12}^n \\
C_{21}^n & -C_{22}^n
\end{pmatrix}
\begin{pmatrix}
a_{12} \\
a_{11}
\end{pmatrix} = 0
\]

(2.9)

3 Cross Cancellation

Let us assume there exist two mixtures each of the source signals \( s_1 \) and \( s_2 \) mixed with 0 such that:

\[
\begin{pmatrix}
x_{11} \\
x_{12}
\end{pmatrix} = \begin{pmatrix}
a_{11} & a_{12} \\
a_{21} & a_{22}
\end{pmatrix}
\begin{pmatrix}
s_1 \\
0
\end{pmatrix}
\]

(3.1)

and

\[
\begin{pmatrix}
x_{21} \\
x_{22}
\end{pmatrix} = \begin{pmatrix}
a_{11} & a_{12} \\
a_{21} & a_{22}
\end{pmatrix}
\begin{pmatrix}
0 \\
s_2
\end{pmatrix}
\]

(3.2)
Where \( x_{xn} \) represents the \( n \)th mixture of source \( s_x \) and 0. From (3.1) and (3.2), we can get all of the \( x \) values:

\[
\begin{align*}
    x_{11} &= a_{11} \cdot s_1 \\
    x_{12} &= a_{21} \cdot s_1 \\
    x_{21} &= a_{12} \cdot s_2 \\
    x_{22} &= a_{22} \cdot s_2
\end{align*}
\]

We do not know \( s_1 \), so we must remove them from (3.3) and (3.4).

\[
\begin{align*}
    x_{11} &= a_{11} \cdot s_1 \\
    a_{21} \cdot x_{11} &= a_{21} \cdot a_{11} \cdot s_1 \\
    a_{21} \cdot x_{11} - a_{21} \cdot a_{11} \cdot s_1 &= 0 \quad (3.7)
\end{align*}
\]

and

\[
\begin{align*}
    x_{12} &= a_{21} \cdot s_1 \\
    a_{11} \cdot x_{12} &= a_{11} \cdot a_{21} \cdot s_1 \\
    a_{11} \cdot x_{12} - a_{11} \cdot a_{21} \cdot s_1 &= 0 \quad (3.8)
\end{align*}
\]

We can now set (3.7) and (3.8) equal to each other, giving us

\[
\begin{align*}
    a_{21} \cdot x_{11} - a_{21} \cdot a_{11} \cdot s_1 &= a_{11} \cdot x_{12} - a_{11} \cdot a_{21} \cdot s_1 \\
    a_{21} \cdot x_{11} &= a_{11} \cdot x_{12}
\end{align*}
\]

which is more useful to us in the form:

\[
a_{21} \cdot x_{11} - a_{11} \cdot x_{12} = 0 \quad (3.9)
\]
The same method is used on (3.5) and (3.6) to remove $s_2$:

$$a_{22} * x_{21} - a_{12} * x_{22} = 0$$  \hspace{1cm} (3.10)

These two equations should then be minimized, which should ultimately give you values for $A$, which can in turn be used with $x_1$ and $x_2$ to find $v_1$ and $v_2$.

4 Normalization

Both the decorrelation and cross cancellation methods have a trivial case where every entry in $A$ is 0. To remove this case, we used the $l_1$ normalization contraint applied to each half of the $A$ matrix. We define these halves as

$$u = \begin{pmatrix} a_{22} \\ a_{21} \end{pmatrix} \quad w = \begin{pmatrix} a_{12} \\ a_{11} \end{pmatrix}$$

which we can substitute into (2.9) along with

$$C_n = \begin{pmatrix} -C_{11}^n & C_{12}^n \\ C_{21}^n & -C_{22}^n \end{pmatrix}$$

to create

$$u^T C_n w = 0$$  \hspace{1cm} (4.1)

As discussed in [2], taking the $l_1$ norm of both $u$ and $w$ and forcing them to be near 1 removes the trivial case and provides some benefits over a higher norm, notably the ability to distinguish peaks in $A$. The constraint then looks like

$$\sigma^2(||u||^2_{l_1} - 1)^2 = 0$$  \hspace{1cm} (4.2)

$$\sigma^2(||w||^2_{l_1} - 1)^2 = 0$$  \hspace{1cm} (4.3)

Combined with (4.1), (4.2) and (4.3) create a function described in [2] that can be minimized with the least squares method.
5 Regularization

Early experiments with the decorrelation and normalization method recovered an $A$ whose components did not decay fast enough in the instantaneous case. The nature of the demixing coefficients is a decaying one. As per a representation of the room impulse response as modelled by [1], $A$ should represent direct sounds around $t = 0$, early reflections of slightly smaller magnitude after the direct sound, and late reverberations of much smaller magnitude after that.

The solution is to add terms that penalize large values as we get further into each $a_{xy}$ vector. We can accomplish this by multiplying each vector by an exponential function, like:

$$a_{xy} \ast \sqrt{\alpha} e^{c[0:q-1]} \quad (5.1)$$

Where $q$ indicates the length of $a_{xy}$. This vector is then added to the function that is put into the least squares method.

6 Results

![Figure 6.1: Actual $A$ values used in experiments.](image)

Through the use of the methods mentioned in the previous sections, we hope to recover $A$, which can then be used with (2.5) to find $v_1$ and $v_2$. Figures 6.1a and 6.1b represent the two $A$ matrices used to mix $x_1$, $x_2$, and the mixtures needed to use the cross cancellation method, where every $q = 40$ points represents a different $a_{xy}$. Since the decorrelation approach uses the $a_{xy}$ vectors in reverse order, these plots and the
ones that follow show the vectors in the order $a_{22}$, $a_{21}$, $a_{12}$, and then $a_{11}$. The goal is for this method to get similar plots.

![Figure 6.2: A values recovered for the instantaneous case.](image)

Cross cancellation and normalization alone shown in figure 6.2a works well to get the magnitudes of the first and third peaks, but has no such luck otherwise. Adding regularization smoothed out the rest of $a_{22}$ and $a_{12}$ and added peaks to $a_{21}$ and $a_{11}$, but the peaks of $a_{21}$ and $a_{11}$ were far smaller than the ones in figure 6.1a.

Adding decorrelation smoothed out each vector and made the peaks of $a_{21}$ and $a_{11}$ more distinct, as seen in figure 6.2b. $a_{22}$ and $a_{12}$ are just about the same as those in figure 6.1a, but $a_{21}$ and $a_{11}$ are scaled down by roughly $\frac{1}{2}$.

![Figure 6.3: A values recovered for the convolutive case.](image)

Figure 6.3a represents the only good result on the convolutive case. It’s accurate for $a_{22}$ and most of
$a_{12}$, but is not even close for the rest of $A$. Not many tests were done on different weights on each part of this method, however. The regularized case in particular decayed far too quickly, as did the case with decorrelation.

7 Conclusion

As shown in section 6, the instantaneous case works out pretty well save for scaling issues on $a_{21}$ and $a_{11}$. The convolutive case did not work out so well, and its best results did not work at all to find $a_{21}$ and $a_{11}$.

Since both are having issues with $a_{21}$ and $a_{11}$, it would be a good subject to look in to. The best thing to look in to in the short-term, however, would be the weights set on all the different terms put in to the least squares method. The constant values that work for the instantaneous case have not worked on the convolutive case, so finding values that work should be a priority.

8 Acknowledgements

I would like to thank my mentor Ernie Esser for all the help and guidance he’s given over the course of this research. I would also like to thank all of the people responsible for the ICAMP program for making this opportunity possible, and the participants of the program for making the shared experience fun.

References

