Separation of Speech Signals Using Trigonometric Transforms

Hossam Hammam*, Atef E. Abu El-Azm**, and Mohamed E. Elhalawany**
*Telecom Egypt, Alexandria, Egypt, hossammha@yahoo.com
**Department of Electronics and Electrical Communications, Faculty of Electronic Engineering, Menoufia University, Menouf 32952, Egypt.
E-mails: elazm40@hotmail.com, and halawany_m85@yahoo.com

Abstract—This paper deals with the problem of blind separation of signals from noisy mixtures. It proposes the application of a blind separation algorithm on the discrete cosine transform (DCT) or the discrete sine transform (DST) of the mixed signals, instead of performing the separation on the mixtures in the time domain. Both the DCT and the DST have an energy compaction property, which concentrates most of the signal energy in a few coefficients in the transform domain, leaving most of the transform domain coefficients close to zero. As a result, the separation is performed on a few coefficients in the transform domain. Another advantage of signal separation in transform domains is that the effect of noise on the signals in the transform domains is smaller than that in the time domain. As an application, the separation of audio signals in a noisy environment is studied in time and transform domains. The simulation results confirm the superiority of transform domain separation to time domain separation.

Index Terms—DCT, DST and Signal separation.

I. INTRODUCTION

Blind signal separation is an important branch of signal processing. It deals mainly with mixed signals, which are frequently encountered in real life. Real life signals are frequently mixed with undesired signals. This fact has motivated the evolution of blind signal separation algorithms. The word blind means that there is no apriori information about the mixed signals and their sources. Several approaches have been proposed for blind signal separation [1-8]. Some of these approaches depend on independent component analysis [1]. Others depend on higher order statistics [2]. There is also a category of correlation based separation algorithms [3]. Adaptive signal separation algorithms also exist [4].

Generally, most of the separation algorithms deal with the problem in the presence of noise. The existence of noise causes some deterioration in the quality of the separated signals. As a result, there is a bad need for a technique to reduce the effect of noise in the separation process. The DCT and the DST are techniques for converting a signal into elementary coefficients in different domains. They are widely used in signal and image compression.

This paper proposes the application of the blind signal separation algorithm after converting the mixtures into the DCT domain or the DST domain. The rest of the paper is organized as follows. Sections II and III present the blind signal separation algorithm. Section IV gives the trigonometric transforms. Section V gives the objective quality metrics for audio signals. Section VI gives the simulation results. Finally, the concluding remarks are given in section VII.

II. PROBLEM FORMULATION

Blind signal separation deals with mixtures of signals in the presence of noise. If there are two signal sources \( s_1(k) \) and \( s_2(k) \), and two observations \( x_1(k) \) and \( x_2(k) \), these observations can be represented as follows [9]:

\[
\begin{align*}
    x_1(k) &= \sum_{i=0}^{p} h_{11}(i)s_1(k-i) + \sum_{i=0}^{p} h_{12}(i)s_2(k-i) + v_1(k) \\
    x_2(k) &= \sum_{i=0}^{p} h_{21}(i)s_1(k-i) + \sum_{i=0}^{p} h_{22}(i)s_2(k-i) + v_2(k)
\end{align*}
\]

or in matrix form as follows:

\[
\begin{pmatrix}
    x_1(k) \\ x_2(k)
\end{pmatrix} =
\begin{pmatrix}
    h_{11} & h_{12} \\ h_{21} & h_{22}
\end{pmatrix}
\begin{pmatrix}
    s_1(k) \\ s_2(k)
\end{pmatrix} +
\begin{pmatrix}
    v_1(k) \\ v_2(k)
\end{pmatrix}
\]

where \( \begin{pmatrix}
    h_{ij}(0) & \ldots & h_{ij}(p)
\end{pmatrix} \) is the impulse response from source \( j \) to sensor \( i \) and \( p \) is the order of the filter. For simplicity, the source signals are assumed to be statistically independent with zero mean. From (1) and (2), it is clear that the mixtures are convolutive sums of sources in the presence of noise.

The problem is simplified by assuming that the signals arrive at the sensors unfiltered, which is equivalent to setting \( h_{11}=h_{22}=1 \). Taking Z-transform of (2) and neglecting the effect of noise lead to:

\[
\begin{pmatrix}
    X_1(z) \\ X_2(z)
\end{pmatrix} =
\begin{pmatrix}
    H_{11}(z) & H_{12}(z) \\ H_{21}(z) & H_{22}(z)
\end{pmatrix}
\begin{pmatrix}
    S_1(z) \\ S_2(z)
\end{pmatrix}
\]

\( v_1(k) \) and \( v_2(k) \) are due to noise, \( h_{ij} \) is the impulse response from source \( j \) to sensor \( i \) and \( p \) is the order of the filter. For simplicity, the source signals are assumed to be statistically independent with zero mean. From (1) and (2), it is clear that the mixtures are convolutive sums of sources in the presence of noise.

The problem is simplified by assuming that the signals arrive at the sensors unfiltered, which is equivalent to setting \( h_{11}=h_{22}=1 \). Taking Z-transform of (2) and neglecting the effect of noise lead to:

\[
\begin{pmatrix}
    X_1(z) \\ X_2(z)
\end{pmatrix} =
\begin{pmatrix}
    H_{11}(z) & H_{12}(z) \\ H_{21}(z) & H_{22}(z)
\end{pmatrix}
\begin{pmatrix}
    S_1(z) \\ S_2(z)
\end{pmatrix}
\]
where $y_j$ is the objective of blind signal separation is to get the signals $s_j(k)$ from the mixed signals $x_j(k)$. Substituting (7) into (8) leads to

$$\begin{pmatrix} Y_1(z) \\ Y_2(z) \end{pmatrix} = \begin{pmatrix} 1 + W_1(z)H_{12}(z) \\ W_2(z) + H_{12}(z) \end{pmatrix} \begin{pmatrix} S_1(z) \\ S_2(z) \end{pmatrix}$$

(9)

For $H_2(z)=1$, which is the case of interest, (5) simplifies to:

$$\begin{pmatrix} Y_1(z) \\ Y_2(z) \end{pmatrix} = \begin{pmatrix} 1 & H_{12}(z) \\ H_{21}(z) & 1 \end{pmatrix} \begin{pmatrix} S_1(z) \\ S_2(z) \end{pmatrix}$$

(10)

where the cost function $C$ is defined as the sum of the squares of the cross-correlations between the two inputs as follows [9]:

$$C = \sum_{i=1}^{l_1} r_{x_1 y_2}^i$$

(11)

where $r_{x_1 y_2}^i = E[x_1(k+i)x_2(k)]$ is a $(q+1)^2$ matrix, which is a function of the cross-correlations of $s_1$ and $s_2$. The cost function can be also expressed as:

$$C = \sum_{i=1}^{l_1} r_{x_1 y_2}^i = \sum_{i=1}^{l_1} r_{x_1 y_2}^i = \sum_{i=1}^{l_1} r_{x_1 y_2}^i$$

(12)

Substituting (8) into (12) gives:

$$r_{x_1 y_2}(l) = r_{x_1 y_2}(l) + \mathbf{w}_1^T \begin{pmatrix} r_{x_1 y_2}(l) \\ \vdots \\ r_{x_1 y_2}(l+q) \end{pmatrix}$$

(13)

where $\mathbf{R}_{x_1 y_2}(l) = E[\mathbf{x}_1(k)\mathbf{x}_2(k+l)]$, is a $(q+1)^2$ matrix, which is a function of the cross-correlations of $x_1$ and $x_2$. The cost function $C$ is defined as the sum of the squares of the cross-correlations between the two inputs as follows [9]:

$$C = \sum_{i=1}^{l_1} r_{x_1 y_2}^i$$

(14)

Thus:

$$r_{x_1 y_2} = r_{x_1 y_2} + [\mathbf{Q}_{x_1 y_2} - \mathbf{w}_1^T \mathbf{R}_{x_1 y_2}(l)] \mathbf{w}_2$$

(15)
signal separation in these transform domains can give better results than in the time domain. In the following subsections, a brief explanation of these transforms is presented.

### A. DISCRETE COSINE TRANSFORM

The DCT expresses a sequence of data points of finite length, in terms of a sum of cosine functions oscillating at different frequencies. The DCT is important for numerous applications in science and engineering, from the lossy compression of audio signals and images, where small high-frequency components can be discarded, to spectral methods for the numerical solution of partial differential equations. The use of cosine rather than sine functions is critical in these applications.

In particular, the DCT is a Fourier-related transform, similar to the discrete Fourier transform (DFT), but using only real numbers. DCTs are equivalent to DFTs of roughly twice the length, operating on real data with even symmetry, since the Fourier transform of a real and even function is real and even, where in some variants the input and/or output data are shifted by half a sample. There are eight standard DCT variants, of which four are common.

The most common variant of the DCT is the type-II DCT, which is often called simply the DCT; its inverse, the type-III DCT, is correspondingly often called simply the inverse DCT or the IDCT. Two related transforms are the DST, which is equivalent to the DFT of real and odd functions, and the modified discrete cosine transform (MDCT), which is based on the DCT of overlapping data.

There are several variants of the DCT with slightly modified definitions. The K real numbers x(0), ..., x(K-1) are transformed into the K real numbers X(0), ..., X(K-1), according to one of the formulas:

**DCT-I**

\[
X(m) = \frac{1}{2} x(0) + (-1)^m x(K-1) + \sum_{k=1}^{K-2} x(k) \cos \left[ \frac{\pi}{K-1} km \right]
\]

\[m = 0, \cdots, K-1\]  \hspace{1cm} (25)

Some authors further multiply the x(0) and x(K-1) terms by \(\sqrt{\frac{2}{K}}\), and correspondingly multiply the X(0) and X(K-1) terms by \(1/\sqrt{2}\). This makes the DCT-I matrix orthogonal.

**DCT-II**

\[
X(m) = \sum_{k=0}^{K-1} x(k) \cos \left[ \frac{\pi}{K} (k + \frac{1}{2})m \right]
\]

\[m = 0, \cdots, K-1\]  \hspace{1cm} (26)

The DCT-II is the most commonly used form, and is often simply referred to as the DCT. This transform is exactly equivalent, up to an overall scale factor of 2, to the DFT of 4K real inputs of even symmetry where the even-indexed elements are zero. Some authors further multiply the X(0) term by \(1/\sqrt{2}\). This makes the DCT-II matrix orthogonal.

**DCT-III**

\[
X(m) = \frac{1}{2} x(0) + \sum_{k=1}^{K-1} x(k) \cos \left[ \frac{\pi}{K} (k + \frac{1}{2})m \right]
\]

\[m = 0, \cdots, K-1\]  \hspace{1cm} (27)

The use of cosine rather than sine functions is critical in these applications.

---

**IV. TRIGONOMETRIC TRANSFORMS**

Trigonometric transforms include the DCT and the DST. It is known that these algorithms have an energy compaction property, which maps a signal of correlated samples to a group of coefficients having a lower level of correlation. The effect of noise is reduced in these transform domains. As a result, signal separation in these transform domains can give better
Because it is the inverse of DCT-II, this form is sometimes simply referred to as the IDCT. Some authors further multiply the $x(0)$ term by $\sqrt{2}$. This makes the DCT-III matrix orthogonal.

**B. DISCRETE SINE TRANSFORM**

In mathematics, the DST is a Fourier-related transform similar to the DFT, but using a purely real matrix. It is equivalent to the imaginary parts of a DFT of roughly twice the length, operating on real data with odd symmetry. There are several variants of the DST with slightly modified definitions. The $K$ real numbers $x(0), \ldots, x(K-1)$ are transformed into the $K$ real numbers $X(0), \ldots, X(K-1)$, according to one of the formulas:

**DST-I**

$$X(m) = \sum_{k=0}^{K-1} x(k) \sin \left[ \frac{\pi}{K+1} (k+1)(m+1) \right]$$  \hspace{1cm}  m = 0, \ldots, K-1  \hspace{1cm} (28)

The DST-I matrix is orthogonal up to a scale factor.

**DST-II**

$$x(m) = \sum_{k=0}^{K-1} X(k) \sin \left[ \frac{\pi}{K} (k+\frac{1}{2})(m+1) \right]$$  \hspace{1cm}  m = 0, \ldots, K-1  \hspace{1cm} (29)

Some authors further multiply the $X(K-1)$ term by $1/\sqrt{2}$. This makes the DST-II matrix orthogonal up to a scale factor.

**DST-III**

$$x(m) = \frac{(-1)^m}{2} x(K-1) + \sum_{k=0}^{K-2} x_n \sin \left[ \frac{\pi}{K} (k+1)(m+\frac{1}{2}) \right]$$  \hspace{1cm}  m = 0, \ldots, K-1  \hspace{1cm} (30)

Some authors further multiply the $x(K-1)$ term by $\sqrt{2}$. This makes the DST-III matrix orthogonal up to a scale factor.

**V. OBJECTIVE QUALITY METRICS FOR AUDIO SIGNALS**

There is a bad need of some metrics to assess the perceptual quality of the audio signals at the output of the proposed technique. Several approaches, based on subjective and objective metrics, have been adopted in the literature for this purpose [16-19]. Concentration in this paper will be on the objective metrics.

Objective metrics are generally divided into intrusive and non-intrusive metrics. Intrusive metrics can be classified into three main groups. The first group includes time domain metrics such as the traditional signal-to-noise ratio (SNR) and segmental signal-to-noise ratio (SNRseg). The second group includes linear predictive coefficients (LPCs) metrics, which are based on the LPCs of the audio signal and its derivative parameters, such as the linear reflection coefficients (LRCs), the log likelihood ratio (LLR), and the cepstral distance (CD). The third group includes the spectral domain metrics, which are based on the comparison between the power spectrum of the original signal and the processed signal. An example of such metrics is the spectral distortion (SD) [16-19].

**A. THE SIGNAL-TO-NOISE RATIO**

The SNR is defined as follows [16-19]:

$$\text{SNR} = 10 \log_{10} \frac{\sum_{i=1}^{N} x^2(i)}{\sum_{i=1}^{N} (x(i) - y(i))^2}$$ \hspace{1cm} (31)

where $x(i)$ is the original audio signal, $y(i)$ is the output audio signal, $i$ is the sample index, and $N$ is the number of samples of the output audio signal.

**B. THE SEGMENTAL SIGNAL-TO-NOISE RATIO**

The most popular one of the time-domain metrics is the segmental signal-to-noise ratio (SNRseg). SNRseg is defined as the average of the SNR values of short segments of the output signal. It is a good estimator for audio signal quality. It is defined as follows [16-19]:

$$\text{SNRseg} = \frac{10}{L} \sum_{m=0}^{M-1} \log_{10} \left( \frac{\sum_{i=mN}^{(m+1)N-1} x^2(i)}{\sum_{i=mN}^{(m+1)N-1} (x(i) - y(i))^2} \right)$$ \hspace{1cm} (32)

where $L$ is the number of segments in the output audio signal.

**C. THE LOG LIKELIHOOD RATIO**

The LLR metric for an audio segment is based on the assumption that the segment can be represented by a p-th order all-pole linear predictive coding model of the form [16-19]:

$$x(n) = \sum_{m=1}^{p} a_m x(n-m) + G_y u(n)$$ \hspace{1cm} (33)

where $x(n)$ is the $n$th audio sample, $a_m$ (for $m=1, 2, \ldots, p$) are the coefficients of an all-pole filter, $G_y$ is the gain of the filter and $u(n)$ is an appropriate excitation source for the filter. The audio signal is windowed to form frames of 15 to 30 ms length. The LLR metric is then defined as [19]:

$$\text{LLR} = \log \frac{\tilde{\alpha}_x \tilde{R}_y \tilde{\alpha}_y^T}{\tilde{\alpha}_y \tilde{R}_y \tilde{\alpha}_x^T}$$ \hspace{1cm} (34)

where $\tilde{\alpha}_x$ is the LPCs coefficient vector $[1, a_x(1), a_x(2), \ldots, a_x(p)]$ for the original audio signal $x(n)$, $\tilde{\alpha}_y$ is the LPCs coefficient vector $[1, a_y(1), a_y(2), \ldots, a_y(p)]$ for the output audio signal $y(n)$, and $\tilde{R}_y$ is the autocorrelation matrix of the output audio signal. The closer the LLR to zero, the higher is the quality of the output audio signal.

**D. THE SPECTRAL DISTORTION**

The SD is a form of metrics that is implemented in frequency domain on the frequency spectra of the original and the output signals. It is calculated in dB to show how far is the
spectrum of the output signal from that of the original signal. The SD can be calculated as follows [16-19]:

$$SD = \frac{1}{M-1} \sum_{m=0}^{M-1} \sum_{i=0}^{N-1} \left| V_x(i) - V_y(i) \right|$$

(35)

where $V_x(i)$ is the spectrum of the original audio signal in dB for a certain segment and $V_y(i)$ is the spectrum of the output audio signal in dB for the same segment. The smaller the SD, the better is the quality of the audio output signal.

VI. RESULTS AND DISCUSSION

In this section, three experiments are conducted to test the effect of using the DCT based separation, the DST based separation and the normal blind separation. The experiments are performed on two mixtures of a music signal and a speech signal contaminated by white Gaussian noise. The results of these experiments are shown in Figs. 3 to 10.

These figures give a comparison between the speech and music quality at the output of the separating algorithms using the different separation algorithms, revealing the effect of the proposed techniques. The speech and music quality is measured by four metrics; the SNR, and the SNRseg, which are time domain metrics. The LPCs are also used to determine the LLR. The SD is also used as an evaluation metric. Results reveal that signal separation in the DCT and DST domains give music and speech signals with better quality than separation in time domain.

Fig. 3. Output SNR vs. input SNR for all techniques for the music signal.

Fig. 4. Output SNR vs. input SNR for all techniques for the speech signal.

Fig. 5. Output SNR$_{seg}$ vs. input SNR for all techniques for the music signal.

Fig. 6. Output SNR$_{seg}$ vs. input SNR for all techniques for the speech signal.

Fig. 7. Output LLR vs. input SNR for all techniques for the music signal.

Fig. 8. Output LLR vs. input SNR for all techniques for the speech signal.
In this paper, trigonometric transforms are introduced as new techniques to reduce the effect of the noise and to achieve a better performance of blind signal separation algorithms. In order to show the effectiveness of the proposed techniques, three experiments have been presented. The experimental results showed better quality of signals separated in the transform domains for speech as well as music signals.

ACKNOWLEDGMENT

We would like to thank Dr. Fathi E. Abd El-Samie from the Faculty of Electronic Engineering for his efforts during the course of this work and the arrangement of this paper.

REFERENCES


Hossam Hammam received his B.Sc., in electrical engineering from Department of Electronics and Electrical Communications, Faculty of Electronic Engineering, Menoufia University, Egypt, in 2000. He is currently a central office manager in Telecom Egypt. His areas of interests are signal separation and signal processing.

Atef Abou El-Azm was born in 1954, Egypt. He has the B.Sc. in electronic engineering and M.Sc. in antennas from Faculty of Electronic Engineering, Menoufia University in 1977, 1984 respectively. He has the Ph.D in communications from Warsaw University of Technology, Poland in 1990. Dr. Abou El-Azm's research is in the area of digital communications, with special emphasis on coding theory, information theory, error control coding, and coded modulation. Topics of current interest include the use of convolutional, trellis codes and Turbo codes in the development of coding standards for high speed data modems, digital satellite communication, and deep space channels. He is the author of many papers in the field of line codes, channel codes and signal processing.

Mohamed Elhalawany received the M.Sc. degree in electrical engineering from Menoufia University in Egypt in 1982. He received the Ph.D. degree from Wroclaw University in 1989. He is an associate professor in the Department of Electronics and Electrical Communications, Faculty of Electronic Engineering, Menoufia University. His areas of interests are signal processing, image processing, and digital communications.