Comparison of the Wavelet and Short Time Fourier Transforms for Spectral Analysis of Speech Signals

Mohammad A. Tinati Behzad Mozaffary
Faculty of Electrical and Computer Engineering
Univercity of Tabriz
29 Bahman Blvd., Tabriz, East Azerbaijan
IRAN

Abstract— In mixtures of speech signals the energy content of the components of the mixture is important and determine the structure of the mixture. Energy contents of signals are better shown when time-frequency or time-scale planes are used. In this paper we present a comparison of wavelet transform (WT) and short time Fourier Transform (STFT) in spectral analysis of speech signals. We will show in wavelet domain, speech signals are very uncorrelated and sparsity of signal is increased.

Keywords—STFT, WT, uncorrelated, sparsity, ICA

1 Introduction

Blind source separation problem is relatively new and an important signal processing issue. It involves recovering unknown sources by using only mixtures of them [1]. Generally it is assumed that sources are statistically independent from each other and at most one of them could be Gaussian [2]. Recently, time-frequency representation (TFR) algorithms have been developed by many researchers [3] which in many cases could be considered as very powerful signal processing tools. In [1] and [4] wigner-ville representation is used to separate up to three speech signals from single observed mixture. They assumed that the time-frequency signatures of sources are disjoint. In [5] it is assumed that speech signals are windowed disjoint orthogonal in time-frequency and can separate speech sources from two mixtures of speech signals. In [6] and [7] it is assumed that speech and music representations are sparse and using frequency domain analysis, speech signal is separated from music. In [8] a solution for the blind source separation problem by shifting the problem to time-frequency domain and applying independent component analysis (ICA) algorithm is presented. In [9] using STFT, an algorithm is proposed for separation of heart beat cycles. Other time-frequency methods have been developed during the past decades applicable to different fields. One can find most of them with detailed references in [10], [11], [12], [13].

2 Backgrounds

2.1 Time-Frequency

In many applications such as speech processing, we are interested in the frequency content of a signal localized in time. The reason is that the signal parameters such as frequency content change over time. In other words these signals are non-stationary. For a non-stationary signal, $s(t)$, the standard Fourier Transform is not useful for analyzing the signal. Information which is localized in time such as spikes and high frequency bursts cannot easily be detected from Fourier Transform. Time localization can be achieved by first windowing the signal so as to cut off only a well-localized slice of $s(t)$ and then taking its Fourier Transform. This gives rise to the short time Fourier Transform or windowed Fourier Transform. The magnitude of the STFT is called spectrogram. The Short Time Fourier Transform of a signal $s(t)$ using a window function $w(t)$ is defined as :

$$\boldsymbol{S}(\tau,\omega) = STFT(s(t)) = \int_{-\infty}^{\infty} s(t) w(t-\tau) e^{-j\omega t} dt \qquad (1)$$

As the window $w(t)$ slides along the signal $s(t)$, for each shift τ, the usual Fourier Transform of the product function $s(t)w(t-\tau)$ is calculated. In two dimensional plots of the spectrogram is made with time on the horizontal axis, frequency on the vertical axis and amplitude given by a gray-scale colors. Therefore three dimensional plots are made with the amplitude on the third axis.

2.2 Wavelets

Wavelets are a set of basis functions generated by dilation and translation of a compactly supported scaling function $\psi_{j,k}(t)$, and basis function $\varphi(t)$, associated with an r-regular multi-resolution analysis of $L^2(R)$. Many types of functions encountered in practice can be sparsely and uniquely represented in terms of a wavelet series [14].

Wavelet transform method has received great deal of attention over the past several years. The wavelet transform is a time-scale representation method that decomposes signals into basis functions of time and scale, which makes it useful in applications such as signal de-noising, wave detection, data compression, feature extraction, etc. There are many techniques based on wavelet theory, such as wavelet packets, wavelet approximation and decomposition, discrete and continuous wavelet transform, etc. Wavelets are generated according to the following equation from a mother wavelet as [14]:

$$\psi_{j,k}(t)=\sum_{j}2^{j/2}\psi(2^{j}t-k) \qquad (2)$$

A wavelet system is a set of building blocks to construct or represent a signal or function. It is a two dimensional expansion set whose linear expansion would be:

$$s(t)=\sum_{k=-\infty}^{+\infty}c_k\varphi(t-k)+\sum_{k=-\infty}^{+\infty}\sum_{j=0}^{+\infty}d_{j,k}\psi(2^{j}t-k) \qquad (3)$$

Most of the results of wavelet theory are developed using filter banks and in applications one never has to deal directly with the scaling functions or wavelets, only the coefficients of the filters in the filter bank are needed. The wavelet decomposition for three scales is shown in Fig. (1), where LP and HP denote low-pass and high-pass filters respectively.

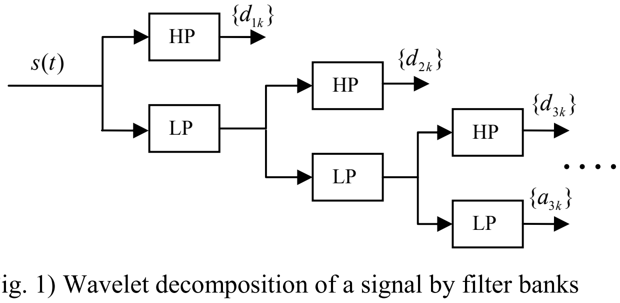

Fig. 1) Wavelet decomposition of a signal by filter banks

3 Discussions and Results

In this paper we compare the time-frequency and time-scale features of speech signals by using short time Fourier Transform and wavelet transform. We will show that speech signals are more orthogonal in WT domain than STFT domain and sparsity of signal increases and therefore it is a better domain for speech separation. We used about 100 speech signal from TIMIT databases. Each signal is two seconds long in time. Each signal is normalized for unit energy and their averages are removed, then spectrogram of signals are computed by utilizing a hamming window that has 256 samples. Let $s(t)$ be the signal and $w(t)$ be the hamming window, then the spectrogram of the windowed signal is given as:

$$\boldsymbol{S}(\tau,\omega)=\left|FFT\{s(t).w(t-\tau)\}\right|^{2} \qquad (4)$$

The cross-energy of the windowed $s_i(t)$ and $s_j(t)$ in time-frequency domain is defined as :

$$\mathcal{E}_{ij}=\iint_{\tau,\omega}E_{i,j}(\tau,\omega)d\tau d\omega \qquad (5)$$

where

$$\begin{aligned}\boldsymbol{E}_{i,j}(\tau,\omega)&=\boldsymbol{S}_{i}(\tau,\omega)\times\boldsymbol{S}_{j}(\tau,\omega)\\&=\left[s_{i,kl}\right]_{N\times N}\times\left[s_{j,kl}\right]_{N\times N} \qquad (6)\\&=\left[s_{i,kl}\times s_{j,kl}\right]_{N\times N}\end{aligned}$$

With $k=1,2,\ldots,N$ and $l=1,2,\ldots,N$.

In equation (6), $\boldsymbol{S}_p(\tau,\omega)$ is N dimensional matrix and $s_{p,kl}$ is its k^{th} row and l^{th} column entry. Any non-zero entries in matrix $\boldsymbol{E}_{ij}(\tau,\omega)$ means that both signals $s_i(t)$ and $s_j(t)$ have considerable energies in their corresponding TFR planes, and therefore we consider this location in the time-frequency plane as a common energy location. The speech signals shown in Fig. (2) are used for time-frequency domain analysis as well. The spectrogram of the speech signals are calculated according to equation (4) and are shown in Fig. (3).

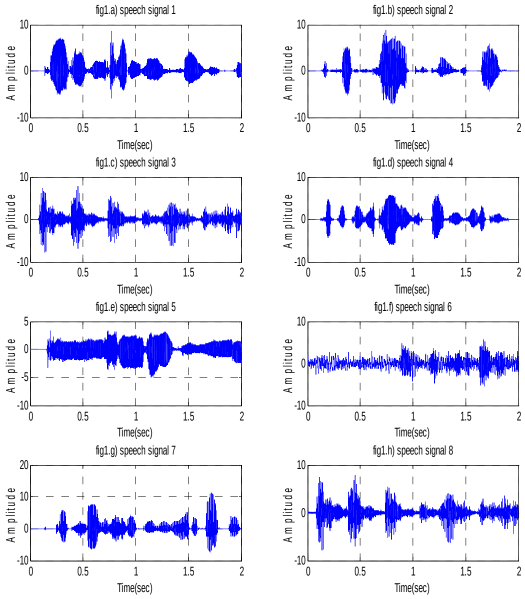

Fig. 2) Eight of the speech signals used for analysis

Using equation (5), the cross-energies of each signal in Fig. (2) with rest of the signals are calculated, and results are shown in table (1) and four of the cross-spectrogram of signals are shown in Fig. (4). In table (1) every entry shows percentage of cross-energies of two speech signals. For example, $\varepsilon_{i,j}$=1.2176 for i=2 , j=4 means that $s_2(t)$ and $s_4(t)$ have %1.2176 cross-energy in time-frequency plane. Note that table (1) would be symmetrical about its diagonal, where for sake of simplicity the lower part of the table is omitted.

Table 1) Percentage of cross-energy of speech signals in TF domain

$\varepsilon_{i,j}$	j=2	j=3	j=4	j=5	j=6	j=7	j=8
i=1	0.297	0.369	0.940	0.287	0.369	0.128	0.102
i=2	-	1.188	1.218	0.103	1.188	0.130	0.137
i=3	-	-	0.694	0.198	8.598	0.245	0.260
i=4	-	-	-	0.254	0.684	0.129	0.061
i=5	-	-	-	-	0.195	0.290	0.331
i=6	-	-	-	-	-	0.245	0.259
i=7	-	-	-	-	-	-	0.317

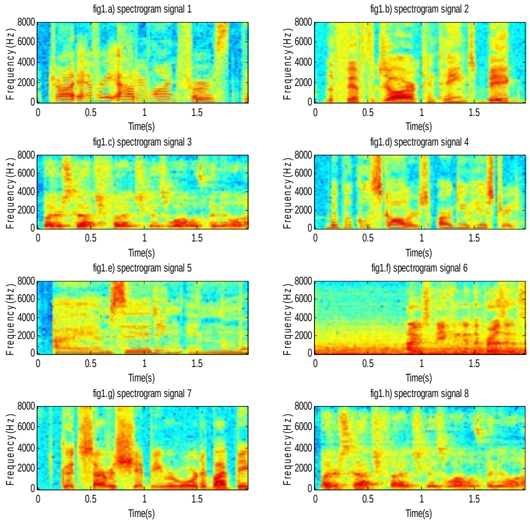

Fig. 3) spectrogram of speech signals shown in Fig. (2)

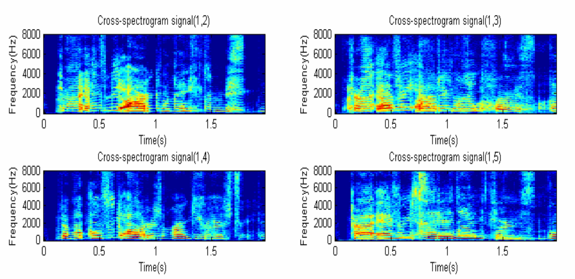

Fig. 4) Cross-energy of signals shown in Fig. (2) in STFT domain

We decompose speech signals in space-time by WT and define energy of signal in scale j as:

$$E_j = \sum_k |d_{jk}|^2 \qquad (7)$$

$$d_{jk} = \int_{-\infty}^{\infty} s(t)\psi_{j,k}(t)dt \qquad (8)$$

The energy distribution in WT domain could be calculated as:

$$\boldsymbol{E}(j,k) = |d_{jk}|^2 \qquad (9)$$

Using Parsaval's theorem, energy of the signal could be computed using wavelet coefficients according to equation (7). Time-scale distribution of the energy of signals shown in Fig. (2) are plotted in Fig. (5). We define cross-energy of $S_n(t)$ and $S_m(t)$ in time-scale as:

$$\mathcal{E}_{n,m} = \sum_j \sum_k E_{n,m}(j,k) \quad (10)$$

Where

$$\begin{aligned} \boldsymbol{E}_{n,m}(j,k) &= \boldsymbol{E}_n(j,k) \times \boldsymbol{E}_m(j,k) \\ &= \left[e_{n,kl}\right]_{N\times N} \times \left[e_{m,kl}\right]_{N\times N} \qquad (11) \\ &= \left[e_{n,kl} \times e_{m,kl}\right]_{N\times N} \end{aligned}$$

With $k=1,2,\ldots,N$ and $l=1,2,\ldots,N.$

In equation (11), $\boldsymbol{E}_p(j,k)$ is N dimensional matrix and $e_{p,kl}$ is its k^{th} row and l^{th} column entry. Any non-zero entries in matrix $\boldsymbol{E}_{n,m}(j,k)$ means that both signals $s_n(t)$ and $s_m(t)$ have considerable energies in their corresponding time-scale planes, and therefore we consider this location in the time-scale plane as a common energy location. The speech signals shown in Fig.(2) are used for wavelet domain analysis as well. The scalogram of the speech signals are calculated according to equation (7) , (8) and are shown in Fig. (5).

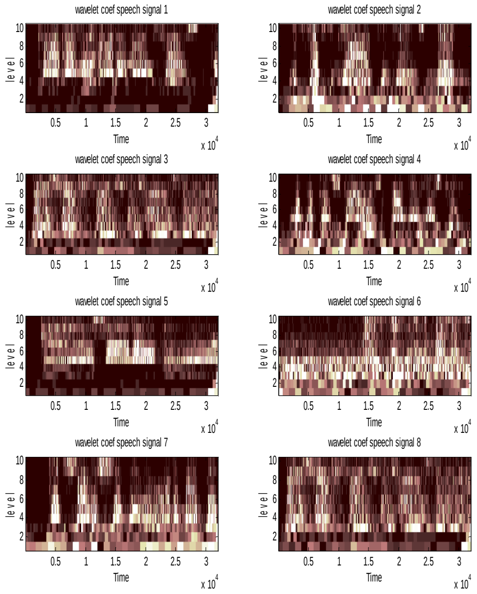

Fig. 5) scalogram of speech signals shown in Fig. (2)

Using equation (10), the cross-energies of each signal in Fig. (1) with rest of the signals are calculated, and results are shown in table (2) and four of the cross-scalograms are shown in Fig. (6). In this table each entry shows percentage of cross-energies of two speech signals. For example, $\varepsilon_{m,n}$=0.1080 for m=2, n=4 means that $s_2(t)$ and $s_4(t)$ have %0.1080 cross-energy in time-scale plane. Again, the lower part of the table (2) is not shown.

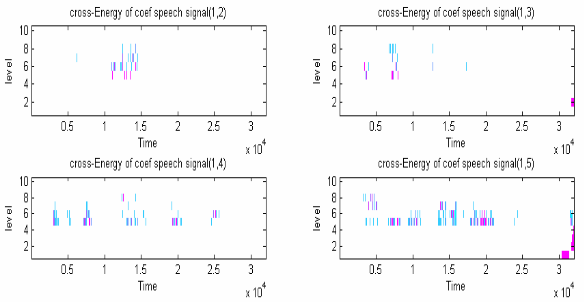

Fig. 6) Cross-energy of speech signals in wavelet domain

Table 2) percentage of cross-energy of speech signals in wavelet domain

$\varepsilon_{m,n}$	n=2	n=3	n=4	n=5	n=6	n=7	n=8
m=1	0.022	0.021	0.027	0.018	0.021	0.008	0.009
m=2	-	0.035	0.108	0.009	0.035	0.007	0.007
m=3	-	-	0.022	0.013	0.296	0.005	0.009
m=4	-	-	-	0.019	0.022	0.005	0.004
m=5	-	-	-	-	0.013	0.015	0.126
m=6	-	-	-	-	-	0.005	0.009
m=7	-	-	-	-	-	-	0.013

4 Conclusions

As is stated, it is well knew that sparsity of signals increase in wavelet domain. We have used this characteristic in speech source separation. Both in time-frequency and time-scale domains, cross-energies of speech signals are calculated. We have shown that cross-energies of speech signals have less common regions in time-scale plane than STFT time-frequency planes. The cross-energies of both domains are summarized in tables (1) and (2) and plotted in figures (4) and (6). By comparing these tables and figures the difference are revealed. As we see in table (2) the cross-energies have decreased in wavelet domain which is direct consequences of wavelet transform properties. We can therefore state that orthogonality of speech signals used in wavelet domain has increased and sparsity of signals are better than STFT domain. Therefore if wavelet domain is used in separation problem, we can separate sources from mixtures much better in wavelet domain than STFT.

5 References

1) A. Mansour, A. Kardec Barros, and N. Ohnishi, “Blind separation of sources: Methods , assumptions and applications.” IEICE Transactions on Fundamentals of Electronics,

Communications and Computer Sciences , Vol. E83-A, no.8, pp. 1498-1512, August 2000.

2) P. Comon, “Independent component analysis, a new concept ?,” Signal processing, Vol. 36, no.3, pp. 287-314, April 1994.

3) P. Flandrin, “Time-frequency/Time-scale analysis,” volume 10 of *Wavelet Analysis and its Applications*, Academic Press, Paris, 1999.

4) A. Mansour, M. Kawamato, C. Puntonent, “A time-frequency approach to blind separation of under-determined mixture of sources,” Proceeding of the IASTED International Conference APPLIDE SIMULATION AND MODELLING September 3-5, 2003, Marbela, Spain

5) O. Yilmaz , S. Rickard , “Blind Separation of Speech Mixtures via Time – Frequency Masking,” IEEE Transaction on signal processing , November 4 , 2002

6) Bofill P. , “Underdetermined Blind Separation of Delayed Sound Sources in the frequency Domain ”, submitted to Neurocomputing , special issue ICA and BSS, 2 march 2001.

7) Bofill P. and Zibulevsky M. , “Underdetermined Blind Source Separation using Sparse Representations” , submitted to Signal Processing, 2000 http://www.ac.upc.es/homes/pau/

8) Dr. S. Jayaraman , G. Sitaraman , R. Seshadri, “ Blind source separation of acoustic mixtures using time-frequency domain independent component analysis,” IEEE conference, ICCS2002, Nov. 2002, Vol.3, pp. 1383- 1387

9) M.A. Tinati , A. Bouzerdoum , J. Mazumdar, L.J. Mahar, “Time-Frequency analysis of heart sounds befor and after angioplasty,”13th int. conf. on digital signal processing proceedings, DSP97, santorini, Greece ,1997

10) J.K. Hammond and P.R. White, “The analysis of non-stationary signals using time-frequency methods,” journal of sound and vibrations, pp. 419-447, 1996.

11) F. Hlawatsch and G.F. Boudreax-Bartels, “linear and quadratic time-frequency signal representations,” IEEE Signal Processing Magazine, Vol. 9, pp. 21-67, April 1992.

12) L. Cohen, “Time-frequency analysis,” Prentice hall PTR, Englewood Cliffs, New Jersy,1995

13) L. Cohen, “Time-frequency distributions – a review,” in proceedings of the IEEE , July 1989, Vol. 77, No.7, pp. 941-979

14) C. S. Burrus, R. A. Gopinath, H. Guo, “Introduction to Wavelets and Wavelet Transforms, a primer” Prentice Hall New jersey, 1998.