STFT based Blind Separation of Underdetermined Speech Mixtures

Prasanna Kumar M K
VCET, Puttur
Nehru Nagar
Puttur, DK, Karnataka, India

Padmavathi K
NMAMIT, Nitte
Karkala
DK, Karnataka, India

ABSTRACT
Analysis of non stationary signals like audio, speech and biomedical signals require good resolution both in time and frequency as their spectral components are not fixed. There are many applications of time-frequency analysis in non stationary signals like source separation, signal denoising etc. This paper presents an application of time frequency analysis using STFT, Short Time Fourier Transform in speech separation. The method is blind since the information about the sources and mixing type is not available. The method uses relative amplitude information of speech mixtures in time frequency domain and ideal binary mask of source signals. The speech mixture used is underdetermined where number of sources are more than number of sensors. A mixture of male and female speech with a musical note is considered for the separation first with a strong mixing matrix and next with a weak mixing matrix. The performance parameter like SNR, signal to noise ratio obtained with this approach proves that time-frequency analysis using STFT can be useful to identify the tracks for separation out of determined speech mixtures. Short time spectrum representation of speech signal requires on the order of two to four times as many samples as required to represent the waveform. However in return a very flexible representation of the signal can be obtained from which extensive modifications in both time and frequency domains can be made.

General Terms
Signal processing, speech processing, time frequency analysis, speech separation

Keywords
STFT, SNR, ISTFT, ABS, TFM, ASR

1. INTRODUCTION
There are many instances where it is necessary to separate signals from a mixture to obtain the original sources, by observing only the mixture i.e. without prior knowledge of the original sources. This is a typical requirement especially in speech enhancement applications. Blind Source Separation (BSS) is one way of solving this kind of problems. In this proposed model, an underdetermined speech and audio mixtures with strong and weak mixing matrices are considered at multiple observations. Speech separation is done in Time- Frequency domain by using relative amplitude information and time frequency ratios of Short Time Fourier Transform of individual observations. Performance evaluation can be done in three ways. (1) Visual Observation (2) Auditory Observation (3) Mathematical Observation. Visual observations can be done through spectrograms and histograms. Auditory observations can be done by hearing original and recovered speech signals. Finally mathematical analysis can be done by calculating and comparing Signal to Noise Ratio (SNR) at various observations. In the proposed model Blind separation is done in Time-Frequency domain using STFT (Short Time Fourier Transform). Since audio and speech signals are non stationary in nature it is not adequate to use Fourier Transform. BSS methodology is dependent on the type of input signals. If the input signal is stereo audio track then both left and right channel signals can be analyzed separately in Time Frequency domain for the elimination of unwanted signals. If the input signal is a mono track like most of the recorded speech signals then the mixture of multiple sources has to be analyzed at various sensors or observations in Time frequency domain to eliminate the unwanted signal.

2. MIXING MODEL
The speech mixture used is under determined mixture where number of source signals are greater than number of sensors

$$S=3 \text{ and } X=2 \quad (1)$$

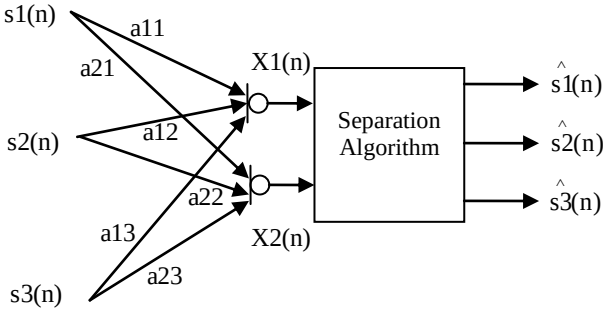

Fig 1: Instantaneous underdetermined mixing model

Where S is number of sources and X is number of sensors or observations. The mixing model shown above can be modeled with following equations.

$$X=A.S \quad (2)$$

Where X is the observation matrix, S is the source matrix, and A is the mixing matrix.

$$X = [X1(n), X2(n)]^T \quad (3)$$

Where X1 (n) and X2 (n) are the observation vector at sensor1 and sensor 2 respectively.

$$S=[S1(n),S2(n),S3(n)] \quad (4)$$

Where S1(n), S2(n) and S3(n) are the source vector1 source vector2 and source vector3 respectively.

$$A = \begin{bmatrix} a11 & a12 & a13 \\ a21 & a22 & a23 \end{bmatrix} \quad (5)$$

$$X1(n) = a11.S1(n)+a12.S2(n)+a13.S3(n) \quad (6)$$

$$X2(n)=a21.S1(n)+a22.S2(n)+a23.S3(n) \quad (7)$$

Where a11, a12, a13, a21, a22, a23 are the coefficients of mixing matrix. Separation algorithm uses STFT and relative amplitudes of each time frequency cluster. Separation algorithm is discussed in next section. s1^, s2^ and s3^ are the recovered source 1, source 2 and source3 respectively.

3. SEPARATION ALGORITHM

Figure 2 shows various stages in separation algorithm. Separation algorithm involves finding STFT of speech mixtures individually .Once the individual STFT is obtained in the matrix format next step is to find STFT ratio. STFT ratio is the ratio of STFT1 and STFT2 element by element. Where STFT1 is the STFT of mix1 at observation 1 and similarly STFT2 is the STFT of mix2 at observation 2.It is the element by element division denoted by STFT1./STFT2

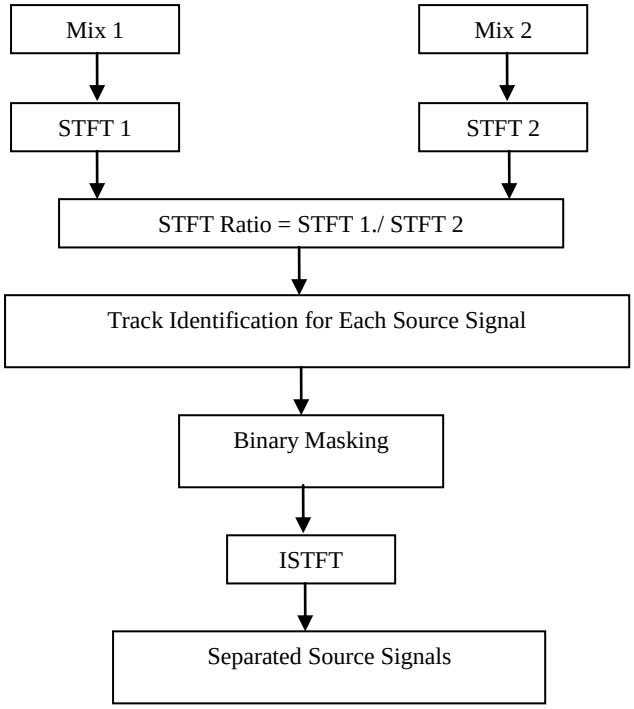

Fig 2: Various stages in separation algorithm

The figure 3 shows frequency of STFT ratios between observation 1 and observation 2. Both mixtures at observation 1 and observation 2 consist of all the three source signals with varying amplitude levels depending on the attenuation level. The first smaller peak represents the first source signal, middle larger peak represents second source which is assumed to be at equal distance from both observations. In the proposed model it is assumed that source 2 is at equal distance from both observations. Hence it is observed that larger peak corresponding to second source signal is obtained at a ratio of '1'. It indicates that source 2 is equally sensed by observation 1 and observation 2. The last smaller peak represents the third source signal. The first and last peaks are smaller compared to the middle peak since source 1 and source 3 are not sensed equally due to varying attenuation levels at observation 1 and observation 2. Once the tracks are identified, proceed to substitute them by zeros. This method is usually known as binary masking because the coefficients are multiplied by either one or by zero. This can be seen mathematically as follows

$$M = \begin{cases} 0 & \text{if} \quad a < \text{Ratio} < b \\ 1, & \text{otherwise} \end{cases} \quad (8)$$

Where M is the binary mask. From the figure 4 for first peak a=0.7 and b=0.8, for the second peak a=0.9 b=1.1 and for the last peak a=1.3 and b=1.4. Binary masking is a method of retaining required values by multiplying those values by one and discarding other values which are not required by multiplying those values by zeros. Since we use '1' for retaining the required samples and '0' for discarding the other samples this method is known as binary masking. For the proposed model suppose we want to separate only first source out of the mixture then we have to track the source one and multiply them by ones and multiply the other tracks by zeros. Similar process has to be applied if we want to separate other two sources.. After the binary mask has been applied, the signal has to be transformed back to the time domain. In order to do this, the ISTFT is used on each one of the observations.

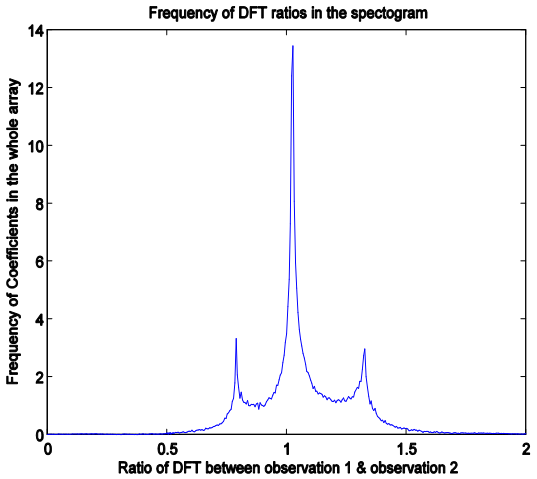

Fig 3: Frequency of STFT ratios

4. TIME FREQUENCY ANALYSIS

The separation algorithm discussed in section 3 is successfully applied to a mixture with a male speech, a female speech and a musical tone. The entire analysis and synthesis results are shown in the following spectrograms. Since the mixture considered here is under determined it is required to consider lesser observations than source signals. Figure 4 shows steps involved in separation of under determined speech mixtures with musical note for S=3 and X=2 in detail. Here S is the number of source signals and X is the number of observations or sensors. In figure 5 (a) represents the spectrogram of original male speech (b) spectrogram of original female speech (c) spectrogram of original saxophone tune (d) spectrogram of speech mixture (e) & (f) spectrogram of mixture after masking saxaphone (g) spectrogram of mixture after masking female speech (h) spectrogram of mixture after masking both saxaphone and female speech (i) spectrogram of speech mixture after masking both saxaphone and male speech (j) spectrogram of speech mixture after masking both female and male speech (k) recovered male speech from the mixture (l) recovered female speech from the mixture (m) recovered saxaphone tune from the mixture.

5. PERFORMANCE EVALUATION

In order to analyze the results of source separation, a method for performance evaluation should be used based on distortion measures. These distortion measures take into account interference from other sources, signal to noise ratio (SNR)

Table 1. SNR (in db) at various observations

Source	Observation1	Observation2
Male	26.758	25.848
Female	22.301	38.548

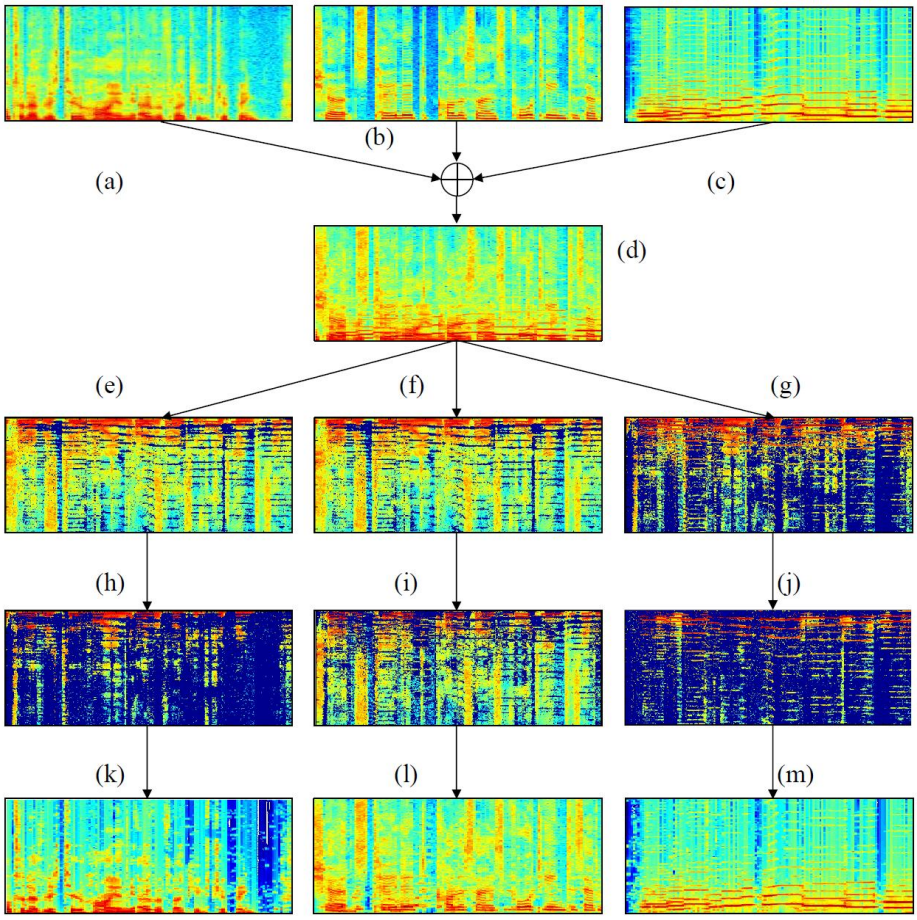

Fig 4: Spectrograms at various stages of speech separation

and artifacts introduced by the algorithm of source separation. Performance of separation algorithm was evaluated using SNR. Performance was evaluated for a strong mixing matrix and also for a weak mixing matrix. The results obtained are tabulated.

$$\mathrm{SNR} = 20\log_{10}\frac{\|\mathrm{output}\|}{\|\mathrm{output} - \mathrm{input}\|} \qquad (9)$$

Where output is the normalized array of output after source separation and input is the normalized array of original input speech signal. The above formula is applied for both male and female speech signal. The mixing matrix used for under determined mixture with S=3 and X=2 is given by

$$\mathrm{Mixing\ matrix} = \begin{bmatrix} 0.9 & 0.8 & 0.7 \\ 0.7 & 0.8 & 0.9 \end{bmatrix} \qquad (10)$$

ble 1. SNR of the male speech at observation 1 is greater than at observation 2. This is due to the fact that male speaker is closer to observation 1 compared to observation 2 .Similarly SNR of female speech at observation 2 is greater than at observation 1 since female speaker is closer to observation 2 than observation 1. However SNR at observation1 and observation2 does not vary much indicating that sources can be recovered from any of the mixtures quite efficiently. The values of SNR shows good separation accuracy.

6. CONCLUSION

In the proposed model initially an instantaneous mixture of speech and audio signals were analyzed (with 20 kHz sampling rate), with different gains in relation to each source and sensor position. An ideal binary mask was obtained for the speech signals. Objective is to analyze the speech separation using an ideal mask and using only information of the relative amplitude, without the phase information (relative delay). Time- Frequency analysis using STFT gives greater flexibility in choosing the masking region. This is due to the Fact that large number of frequency components can be

extracted for the better selectivity. There is no unique solution to Blind Source Separation Applications. The separation algorithm should be modified according to the application. Further research can be done on development of unique algorithms for various applications of speech separation .Since audio and speech signals are non stationary in nature FT (Fourier Transform) is inadequate for the separation of multiple signals. Hence transforms like STFT and Wavelets has to be used which are more complex. Further wavelet transforms can be applied in order to achieve good resolution both in time and frequency.

7. REFERENCES

[1] E. Vincent, R. Gribonval, and C. F'evotte, "Performance measurement in blind audio source separation," IEEE Transactions on Speech and Audio Processing, vol. 14, no. 4, pp.1462– 1469, 2006.

[2] M. Babaie-Zadeh, C. Jutten, and A. Mansour, "Sparse ica via clusterwise pca," Neurocomputing, vol. 69, pp. 1458–1466, 2006.

[3] F.Abrard and Y.Deville, Atime frequency blind signal separation method applicable to underdetermined mixtures of dependent sources,Signal Processing,Vol 85,Issue 7,pp 1389- 1403,July 2005.

[4] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," IEEE Transactions on Signal Processing, vol. 52, no. 7, pp. 1830–1847, July 2004.

[5] S. Araki, S. Makino, H. Sawada, and R. Mukai, "Underdetermined blind separation of convolutive mixtures of speech with directivity pattern based mask and ica," Fifth International Conference on Independent Component Analysis and Blind Signal Separation, pp. 898–905, 2004.

[6] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 5, pp. 2737–2740, 2001.

[7] A. Hyv¨arinen, J. Karhunen, and E. Oja. Independent Component Analysis. John Wiley & Sons, 2001.

[8] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures," In IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP2000),vol. 5, pp. 2985–2988, June 2000.

[9]. K. Torkkola, "Blind separation of convolved sources based on information maximization," IEEE Worshop on Neural Networks for Signal Processing, Kyoto, pp. 423–432, september 1996.

[10] 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.