Available online at www.elixirpublishers.com (Elixir International Journal)

Computer Science and Engineering

Elixir Comp. Sci. & Engg. 66 (2014) 20442-20445

ISSN: 2229-712X

DSP algorithm for music-less audio stream generation

Munaza Razzaq*, Hafiz Adnan Habib and M.Usman Khan
Department of Computer Engineering. University of Engineering & Technology Taxila, Pakistan.

ARTICLE INFO

Article history:
Received: 21 May 2013;
Received in revised form:
19 December 2013;
Accepted: 31 December 2013;

Keywords
Discrete Fourier transform,
Masking, time frequency filters,
Overlap-add,
Stereophonic audio.

ABSTRACT

In this paper we investigate the problem of separation of human voice from a mixture of voice and different music instruments. The human voice may be a part of singing voice in a song or it may be a part of some news broadcasted by a channel and it contains background music. The final outcome of this work is a file containing only vocals. In our approach we consider stereo audio for separation. We process the signal in time frequency domain. In our method of blind source separation we processed the input stereo audio file in the form of frames, windowed them and then applied discrete Fourier transform (DFT) on signal. Then the signal is masked for de-mixing purpose using time frequency filters and non-zero DFT coefficients that are estimated as a part vocals are selected and signal is reconstructed by overlap add (OLA) method to get the final output signal containing only vocals.

Introduction

Blind source separation is a process of separating a source of interest from a mixture of different sources without prior knowledge of mixing process, number of sources and mixing parameters. Presently most of the audio is recorded, saved and processed in the form of stereophonic audio. Many algorithms are there that processed stereophonic audio for different purposes like singing voice detection, melody extraction, instruments detection, speech recognition, singer identification, instruments classification and voice separation from the mixture. The work is being carried out on stereophonic audio to get clear separation results [2,6,10]. Similar attempt is made in this paper; we collected dataset of 26 stereo audio files and get promising separation results for almost all of them. Section II presents our algorithm overview. In section III we discuss filtering and signal reconstruction process. In section IV we discuss experimental results. Section V concludes our paper. Section VI presents future work.

GORITHM OVERVIEW

We perform steps shown in figure 1; to carryout separation process.

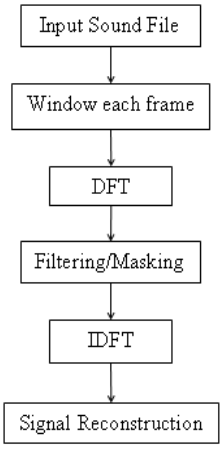


Fig. 1: Algorithm block diagram

Input audio file is processed in form of overlapped time frames. Each frame is windowed. DFT [1,3,4] is applied to each frame; it gives DFT coefficients of the input file. Some of these coefficients are kept making others zero by the help of masking or in other words filtering. Then IDFT is applied to signal to get a real signal which is then reconstructed by overlap add using triangular window with 50% overlap.

SIGNAL FILTERING AND RECONSTRUCTION

In this step we apply independent layers of time frequency filters (TFF) [5] on our signal to build our DFT frames keeping some of the DFT coefficients and setting others to zero. Pan TFF and Inter-channel Phase Difference TFF are utilized for filtering and also for non-zero DFT coefficient selection. We define a different mask in each filtering technique. The mask is defined manually based upon the filtering method used here.

Pan TFF

Mono signals are panned in both channels to form a stereo mixture. Non reverberated tracks do not show significant overlapping and it is easier to define a range to select their DFT coefficients but if the tracks are overlapping their coefficients may change in time and cannot be estimated correctly as belonging to one source or the other in the mix. We are not considering those files in which stereo reverberation is added to one mono track to form a stereo file.

Voice is a pure mono track present in a file so we can define a mask in pan TFF to select DFT coefficients of this mono signal as our desired output. In this paper, pan TFF mask range is taken from 0.4 to 0.6 which is selecting center panned coefficients which can be estimated as belonging to voice as voice as a mono signal is found in center of audio file. This mask selects DFT coefficients that are within its range setting others to zero.

IPD TFF

Some of the noise or interference that cannot be completely extracted by pan TFF is filtered out by Inter-channel Phase Difference (IPD) TFF. It is found that DFT phase spectrum of mirrored pure mono tracks is same for both channels [7] as in equation 1.

Tele:
E-mail addresses: engr.munaza_razzaq@yahoo.com

$$\left|Arg\left(DCT_p\left(s_i^L\right)[f]\right) - Arg\left(DCT_p\left(s_i^R\right)[f]\right)\right| = 0$$

$$\forall f \in 0 \ldots . \frac{N}{2} \qquad (1)$$

Mono tracks with artificial reverberation exhibit opposite criteria as shown in equation 2.

$$\left|Arg\left(DCT_p\left(s_i^L\right)[f]\right) - Arg\left(DCT_p\left(s_i^R\right)[f]\right)\right| > 0$$
$$\forall f \in 0 \ldots . \frac{N}{2} \qquad (2)$$

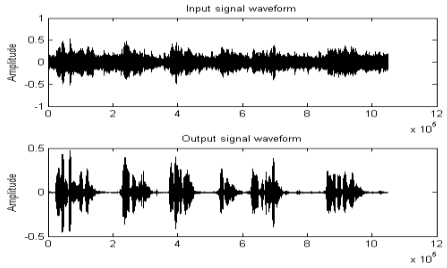
However IPD TFF [7] helps to differentiate between non-reverberated mono and artificially reverberated/stereo tracks. The energy of DFT coefficients possessing same phase is standing around zero IPD value. It is the case for pure mono tracks mirrored in two channels. DFT coefficients with different phase exhibit the presence of artificial reverberation in tracks. As our source of interest is voice, so based upon previous criteria we define a mask to perform IPD filtering on signal. Its range is from -π and +π. We set our IPD TFF mask in range [-0.2 0.2]. Application of this mask makes DFT coefficients zero that are outside this range. This process can also be termed as binary masking.

Finally we overlap and add [8] these processed frames to reconstruct our final output signal. We use triangular window to perform overlap; as triangular window obey constant overlap-add constraint for a wide variety of hop sizes and yields perfect reconstruction due to this property.

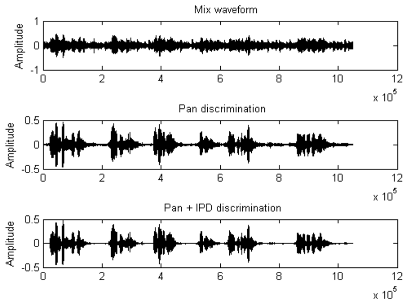
PERIMENTAL RESULTS

In our experiments we choose frames of size N=8192 for a better perceived quality of output sound. We set hop size M= N/4. MATLAB is used as a simulation platform. Blackman Harris -92dB window is used as it show good performance with reduced spectral leakage that's why it is selected. We apply our algorithm upon dataset of 26 songs we collected. Some of the experimental results are presented in the following figures.

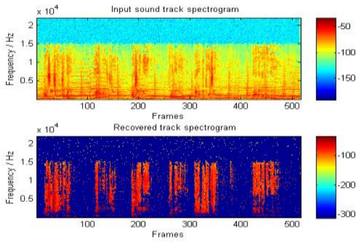
We tested our algorithm on a song named 'wonderful world'. Its results are in figure 2 (a-c).Figure 2(a) shows waveform of original signal and signal after separation for this particular song. It is clearly seen from this plot that vocals are separated. Figure 2(b) shows waveform of signal after filtering the signal with pan and IPD TFF.



(a) Original and recovered signal



(b) Pan and IPD TFF application



(c) Spectrogram

Fig. 2: Original and extracted vocals signal

It can be observed from this plot that some residual noise left after pan TFF is removed by IPD TFF. Figure 2(c) show spectrogram of input and processed sound file containing only vocals.

Some more experimental results that we performed on some popular songs from our dataset are shown in figure 3, 4 and 5. We further evaluate our algorithm separation performance by the help of one of the blind source separation evaluation (bss_eval) parameter that is signal to interference ratio (SIR).

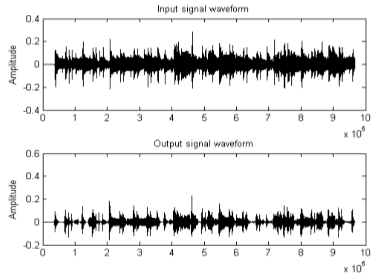
AGl To Interference Ratio (Sir)

Bss_Eval is a toolbox [9] used for measuring the separation quality of a separation algorithm. It tests separation performance by means of different parameters like signal to interference ratio (SIR), signal to distortion ratio (SDR) and signal to artifacts ratio (SAR). We make use of one of its parameters i.e. SIR for evaluation of our algorithm's performance.

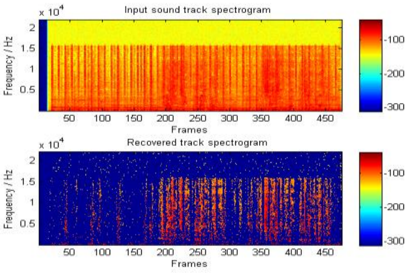
The reason of measuring SIR is to evaluate our results in terms of finding out that whether the interfering source (music) is removed or not. The SIR gives a measure of level of interference present in a source of interest that is estimated in a separation algorithm. Let $\mathbf{D_{interf}}$ be the distortion due to interference. Then in mathematical terms SIR can be written as in equation 3.

$$\mathbf{SIR \triangleq 10 \log_{10} D_{inter\,f}^{-1}} \qquad (3)$$

We calculated SIR values for our separation method. We tested all 26 songs in our dataset. Table 1 enlists SIR results for separation performed on these audio files. Large and positive SIR values represent good separation performance. Table 1 has large and positive values of SIR which yield excellent separation performed by our algorithm. These results are also presented graphically in figure 4.

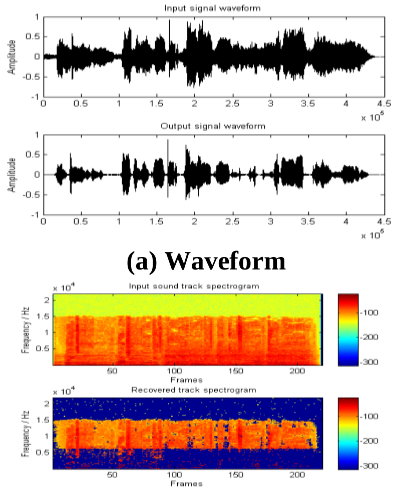


(a) Waveform



(b) Spectrogram

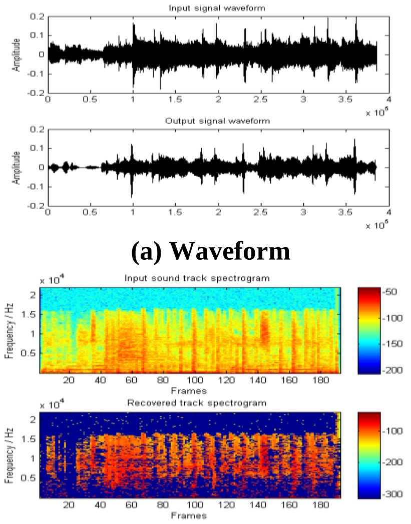
Fig. 3 Suspended.wav



(a) Waveform

(b) Spectrogram

Fig. 4: Waiting.wav



(a) Waveform

(b) Spectrogram

Fig 5: Sun rise.wav

Note that all file in table 1 have sampling resolution of 16bits and sampling frequency of 44.1 kHz and SIR values are in decibels.

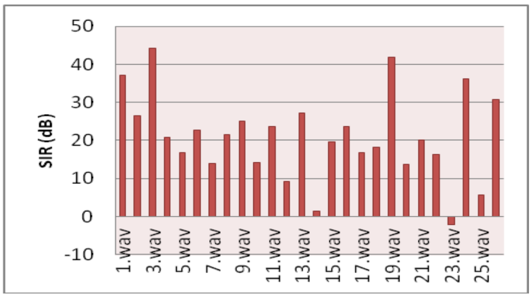


Fig 4: SIR values

Shortcomings

There are some shortcomings in our algorithm. From SIR values in table 1; it can be seen that a good separation performance is achieved but there is a wav file ‘wondering.wav’ in which our algorithm has shown degraded performance as its negative SIR value illustrates poor performance of our algorithm for this particular file. The reason for this is that most of the music instruments used in this particular song is bass drum and bass drum is mostly panned near the center where voice also resides so it interferes with the voice and can’t be removed completely as if an attempt is made to completely extract this drum then it will remove voice as well. Due to this reason our algorithm could perform only partial separation in this case. Other than this file all other files have shown large positive values of SIR yielding excellent separation performance.

CONCLUSION

In this paper we deal with blind audio source separation problem in order to remove background music leaving only foreground vocals in stereophonic audio. Signals are observed in time frequency domain and well known techniques like DFT, binary masking/time frequency filtering and overlap-add methods are used for this purpose. The results are analyzed and discussed in the form of waveforms, spectrograms and bss_eval metric SIR. The SIR values are acceptable and envision qualitative separation results.

FUTURE WORK

Our algorithm could not perform fully if there are bass heavy instruments panned near the center interfering with the voice. In future this shortcoming could be worked upon to get better results.

Table 1: Bss_Eval Results

Title	Duration (sec)	SIR
Saturation.wav	**13**	**37.09**
Long.wav	8	26.50
Geo news.wav	15	44.35
Magic candel	8	20.83
Sun rise.wav	8	16.76
Waiting.wav	10	22.78
Wonderful world.wav	23	13.91
Words of wisdom.wav	25	21.45
Leave me.wav	13	24.98
My way.wav	9	14.23
Exactly.wav	15	23.64
Want to.wav	21	9.254
Akomo heo.wav	23	27.29
Getting bore.wav	15	1.432
Trying to see you.wav	21	19.59
Matter.wav	9	23.58
Story.wav	20	16.89
Fortminer.wav	12	18.23
Ready.wav	22	41.94
Suspended.wav	21	13.60
Something new.wav	18	20.04
Ready.wav	24	16.41
Wondering.wav	25	-2.13
Kuandu.wav	13	36.12
Rainy.wav	14	5.618
Help.wav	20	30.80

References

Wang Chao, Fang Yong, Feng Jiuchao, “A weighted general discrete Fourier transform for the frequency domain blind source separation of convolutive mixtures”, *in Journal of Electronics*, China, vol. 25. No. 6, November, 2008.

Maximo Cobos, Jaume Segura and Jose J. Lopez, “Magnitude Ratio Modelling of Instantaneous Stereo Audio Mixtures in the Time-Frequency Domain”, in *2nd International Conference on Computer Design and Engineering (ICCDE 2012)*, vol. 49. Singapore, 2012.

D. Sundararajan, *The Discrete Fourier Transform, Theory Algorithms and Applications*, World Scientific, 2001, ISBN: 981-02-4521-1

Babu Ram, *Engineering Mathematics*, Volume II, Second Edition, May 14, 2012, Print ISBN**:** 978-81-317-8503-4

Vaninirappuputhenpurayil Gopalan Reju, Soo Ngee Koh and Ing Yann Soon, "Underdetermined convolutive blind source separation via time–frequency masking", *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, No. 1, January 2010.

Sofianos, Stratis; Ariyaeeinia, A.; Polfreman, R., "Towards effective singing voice extraction from stereophonic recordings*", in IEEE Int. Conf on Acoustics Speech and Signal Processing*, ICASSP, pp. 233-236 , 2010 .

MarC Vinyes, Jordi Bonada, Alex Loscos, "Demixing commercial music productions via human-assisted time-frequency masking", *Audio Engineering Society (AES), 120th Convention*, Paris, France, 2006 May 20–23

S. Araki, S. Makino, H. Sawada, and R. Mukai, "Reducing musical noise by a fine-shift overlap-add method applied to source separation using a time frequency mask*", in IEEE Int. Conf on Acoustics Speech and Signal Processing ICASSP* 05, vol. III, pp. 81–84, 2005.

C. Févotte, R. Gribonval and E. Vincent, BSS_EVAL toolbox user guide-Revision 2.0, Technical Report 1706, IRISA, April 2005.

E. Vincent, H. Sawada, P. Bofill, S. Makino and J.P. Rosca, "First stereo audio source separation evaluation campaign: Data, algorithms and results", in *Proc. 7th Int. Conf. on Independent Component Analysis and Signal Separation (ICA)*, pp 552-559, 2007.