PHASE-BASED INFORMED SOURCE SEPARATION OF MUSIC

Nicolas Sturmel, *

GIPSA-Lab, Speech and Cognition Department
CNRS - Grenoble INP
nicolas.sturmel@gipsa-lab.fr

Laurent Daudet,

Institut Langevin, ESPCI
CNRS - Université Paris Diderot
laurent.daudet@espci.fr

Laurent Girin,

GIPSA-Lab, Speech and Cognition Department
CNRS - Grenoble INP
laurent.girin@gipsa-lab.fr

ABSTRACT

This paper presents an informed source separation technique of monophonic mixtures. Although the vast majority of the separation methods are based on the time-frequency energy of each source, we introduce a new approach using solely phase information to perform the separation. The sources are iteratively reconstructed using an adaptation of the Multiple Input Spectrogram Inversion (MISI) algorithm from Gunawan and Sen. The proposed method is then tested against conventional MISI and Wiener filtering on monophonic signals and oracle conditions. Results show that at the cost of a larger computation time, our method outperforms both MISI and Wiener filtering in oracle conditions with much higher objective quality even with phase quantization.

1. INTRODUCTION

Audio source separation has recently drawn a lot of interest, mainly because of its potential multimedia applications. In particular, audio source separation can be a vital component of active listening: the ability for a listener to modify the way music is mixed by changing the volume or spatial location of its composing tracks, the so-called sources. In order to perform active re-mixing, the user has to have access to the separated tracks of the mixture, which is problematic if they are lost, or not delivered to the general public. In this case, blind source separation (BSS) is used in order to recover the tracks. However, while blind source separation techniques (e.g. [1]) have made tremendous progress, they still cannot guarantee a sufficient quality or fast-enough implementation.

The recent paradigm of Informed Source Separation (ISS) addresses these limitations in specific conditions. This approach uses compacted (quantized [2]) information about the priorly known separated tracks and embeds this information in an easy to read and popular stereo format such as WAV PCM or MPEG2/AAC. This way, side information can be used to enhance the quality and accelerate the separation process [2, 3, 4, 5, 6]. Actually, the challenge of ISS is to find the best balance between the final quality of the separated tracks and the amount of side information, so that it can easily be transmitted alongside the mix, or even watermarked into it. All these techniques are based on the same principle: coding energy information about each source in order to facilitate the posterior separation. Sources are then recovered by adaptive filtering of the mixture which leads to limited reconstruction quality, even in the so-called oracle condition, when the prior information (in general, the energy of each source) is perfectly known. Other methods such as [7] are not bounded, but rely on joint posterior coding of both phase and magnitude relative to the mixture.

The motivation of this article is to propose a dual approach of ISS on monophonic mixtures, using solely the phase information of each source to perform the separation and to reach better quality than any separation method based on power spectrogram only. In order to perform such phase-based ISS, an iterative reconstruction technique adapted to source separation known as Multiple Input Spectrogram Inversion (MISI - [8]) is modified to use the phase information. The coding approach of ISS will not be addressed and methods will be compared considering infinite side-info bit rate. However, phase quantization will be considered.

The article is structured as follows. First, the principles of magnitude-based ISS and signal reconstruction are presented on Section 2. Then, the proposed method of phase-based ISS (P-ISS) is detailed in Section 3. Section 4 presents the experimental setup used to evaluate the method and Section 5 discusses the results. Finally, Section 6 concludes the paper.

2. MAGNITUDE-BASED ISS

In this study, we only consider monophonic mixtures where J sources $s_j(t)$, $j = 1 \dots J$, are linearly mixed into the mix signal $m(t) = \sum_j s_j(t)$.

2.1. Wiener filtering

If the signals are stationary and the local time-frequency (TF) variances of all sources are known, then the individual source $s_j(t)$ can be estimated from the mix $m(t)$ of the Short-Time Fourier Transform (STFT) M using a generalized time-frequency Wiener filter. Take, for instance, the energy of the STFT, noted $|S_j(t,f)|^2$, $j = 1 \dots J$ as the variance. Computing the Wiener filter is then equivalent to computing the relative energy contribution of the source j with respect to the total energy of the sources. At a given time-frequency bin (t, f), the estimated source $\hat{s}_j(t)$ is then:

$$\hat{s}_j(t) = STFT^{-1}\left[M(t,f)\frac{|S_j(t,f)|^2}{\sum_k |S_k(t,f)|^2}\right]. \qquad (1)$$

* This work was supported by the french ANR DReaM project

2.2. Iterative signal reconstruction

By construction the Wiener filtering framework satisfies the so-called re-mixing constraint $\sum_j \tilde{s}_j(t) = m(t)$. However, the phase is not estimated: only the magnitude $|\hat{S}_j(t,f)|$ of each source is estimated by this adaptive Wiener filter, and the reconstruction uses the phase of the mixture. This limits the separation quality.

To this extent, iterative reconstruction frameworks such as the Griffin and Lim technique (G&L [9]) have been proposed [10, 11]. By iteratively constraining the magnitude of a TF redundant representation, one can partially recover the phase of a signal (see [12] for a complete review). The Multiple Input Spectrogram Inversion (MISI, [8]) has especially been proposed to allow iterative reconstruction of signals in a source separation framework when knowing the amplitude of each source STFT $|S_j(t,f)|$. At each iteration k, source reconstruction "à la G&L" generates a remix error $E_k = M - \sum \hat{S}_j^{(k)}$. This error is distributed on each source to correct the phase pattern so that :

$$\hat{s}_j^{(k+1)} = STFT^{-1}\left[|S_j| e^{i\angle(\hat{S}_j^{(k)} + E_k/J)}\right]$$

Results obtained with MISI are very good, but require the perfect knowledge of the source magnitude STFT $|S_j|$, which is problematic when such data has to be quantized or factorized [2] to be embedded with or watermarked to the mix signal.

However, it has also been shown in [13] that using G&L with phase information led to overall better reconstruction quality than with magnitude only in a similar way to [14] but applied to compact time-frequency representations. Since to the best of our knowledge, no existing method performs phase-based ISS, we propose in the next section a modified MISI algorithm that uses solely the phase information of each source to separate them from the mixture. In order to take embedding constraints into account, we also consider possible quantization of the phase.

3. PHASE-BASED ISS

3.1. Adapting MISI

MISI uses a G&L iteration framework but adds a supplementary phase correction step using the remix error E_k. When dealing with phase-based ISS (PB-ISS), the phase only can be constrained and therefore the reconstruction error has to be distributed on the magnitude. We choose to initialize the estimated source magnitude STFT with the magnitude STFT from the mixture. Note that because we do not code the energy of the sources, we are totally immune to the energy relations between them (the mixing coefficients) provided that they are scalar.

The dual algorithm then differs on two points :

1. The phase is constrained and the magnitude is reconstructed
2. The re-mix error E_k is distributed on the magnitude of each estimated source.

Note that the proposed method uses an offline implementation based on the original G&L algorithm, it should however be easily implemented in real time using fast algorithms doing similar iterative constrained reconstruction (e.g [10, 11]).

3.2. Compacting the phase information

Phase information can be compacted using various techniques, and notably using phase delay and group delay. In order to simplify this study, we only consider uniform quantization of the phase with a step Q_Φ designated by the operator $U_{Q_\Phi}[.]$. The transmitted phase used for initialization is then $\hat{\Phi}_j^{(0)} = U_{Q_\Phi}[\Phi_j]$.

When the phase information is quantized, direct application of MISI is problematic because the signal is constrained to wrong values. Therefore, one can not directly constrain the phase to the exact initial quantized value. At each iteration step the goal is rather to shift the estimated phase so that it remains in the close neighborhood of the original quantized value. This can be simply implemented by adding the quantization residual to the original phase quantization value, that is :

$$\hat{\Phi}_j^{(k+1)} = \hat{\Phi}_j^{(k)} - U_{Q_\Phi}[\hat{\Phi}_j^{(k)}] + \hat{\Phi}_j^{(0)} \quad (2)$$

3.3. Proposed algorithm

The proposed algorithm of PB-ISS consists in the following steps:

1. Consider the quantized phase information $\Phi_j^{(0)}$ of each source s_j using step Q_Φ and the monophonic mixture m of STFT M.
2. Initialize the estimated source spectrograms with the amplitude $|M|$ of the mix and the quantized phase $\hat{\Phi}_j^{(0)}$ so that:
$$\hat{S}_j^{(0)} = |M| e^{\hat{\Phi}_j^{(0)}}.$$
3. Perform G&L : $\dot{S}_j^{(k+1)} = STFT[STFT^{-1}[\hat{S}_j^{(k)}]]$
4. Constrain the phase $\hat{\Phi}_j^{(k+1)}$ of $\dot{S}_j^{(k+1)}$ within the quantization step Q_Φ as given by equation (2).
5. Distributes the reconstruction error $E_{k+1} = M - \sum_j \dot{S}_j^{(k+1)}$ to each estimated STFT amplitude:
$$\hat{S}_j^{(k+1)} = |\dot{S}_j^{(k+1)} + \frac{E_{k+1}}{J}| e^{\hat{\Phi}_j^{(k+1)}}$$
6. Finally, when the number of iteration has reached an arbitrary maximum K, output $\hat{s}_j = STFT^{-1}[\hat{S}_j^{(K)}]$

4. EXPERIMENTS

The experiments are performed on a database of 2 different monophonic linear instantaneous stationary (LIS) music mixtures under creative commons. The first mixture is composed of 5 sources: lead voice, drums, bass, guitar and synthesizer. The second mixture is composed of 7 sources : lead voice, snare track (sn), bass, bass drum (bd), backing vocals, first guitar and second guitar. The proposed method is tested against the Wiener filter, the MISI technique and PB-ISS without error distribution (independent phase only G&L reconstruction) all in oracle conditions. All these methods are based on the same original TF distribution (STFT). Therefore, other iterative reconstruction algorithm based on different phase distributions (e.g. [14]) will not be tested.

Six different phase quantization steps are used, as power of two from 2 to 64, but a seventh testing conditions with no quantization is also used (designated as a "NO" quantization step). We used a 2048 point STFT with half sinus analysis window and 50% overlap. The results are assessed using the three objective criteria of the BSS Eval toolbox [15] : Source to Distortion Ratio (SDR), Source to Interference Ratio (SIR) and Source to Artifact Ratio (SAR).

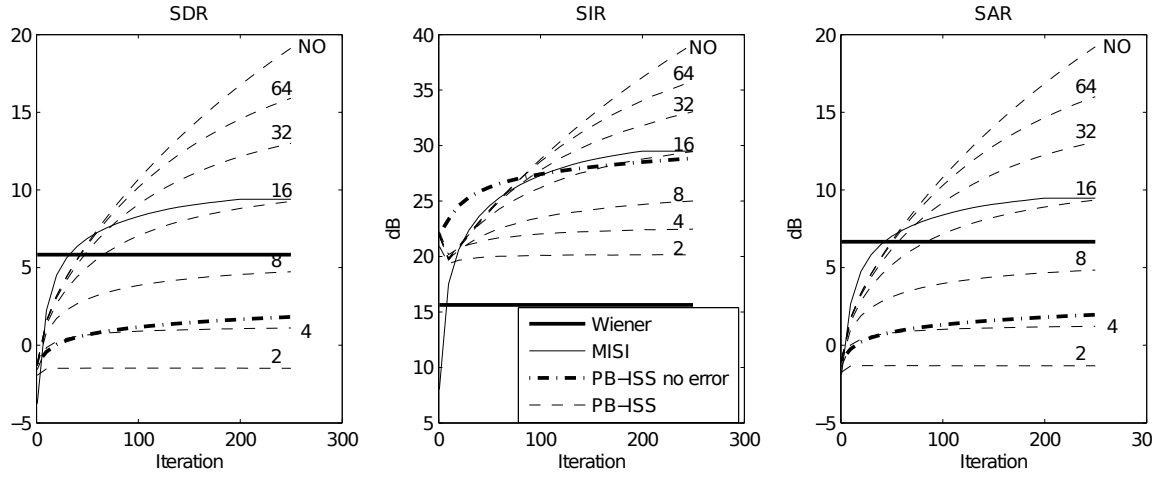

Figure 1: *Averaged results. SDR, SIR and SAR are given as a function of the iteration number (one point every ten iteration) for each of the four methods : Wiener filter, MISI and PB-ISS without error distribution, PB-ISS at oracle conditions and for 6 levels of quantization.*

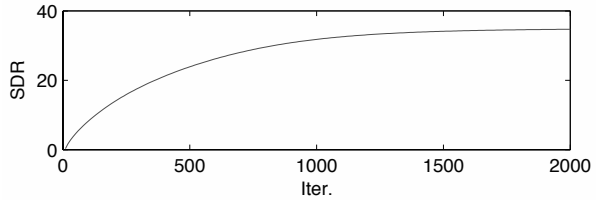

Figure 2: *SDR on oracle PB-ISS for 2000 iterations on M2.*

Results are presented in Figure 1. SDR, SIR and SAR are presented as a function of the iteration number of the considered method. In the case of Wiener filter, only one value is given. Wiener filter is represented as a bold black line, MISI as a thin black line and the proposed method is presented using annotated dashed line for each testing condition with quantization steps.

From an overall distortion (SDR) or artifact (SAR) point of view, the MISI reconstruction outperforms the Wiener filter starting from iteration 25, but PB-ISS becomes better from iteration 40. The oracle PB-ISS outperforms the Wiener filter by up to 12dB of SDR at 250 iterations, and outperforms MISI by 7dB approximately. In terms of interference (SIR), the Wiener filter is much faster outperformed by MISI or PB-ISS. Results of MISI are 15dB higher and results of PB-ISS in oracle conditions are up to 20dB higher than the oracle Wiener filter. PB-ISS outperforms MISI only from the 120th iteration. This is expected because PB-ISS is initialized with the mixture amplitude, producing a lot of interferences especially when the original source is muted. Moreover, G&L reconstructions tend to reconstruct interferences signals out of phase with the originals, which are then seen as artifacts by BSS Eval. PB-ISS without error distribution proposes poor results.

As visible in Figure 1, quantization lowers the performances of PB-ISS. With 2 quantization steps (a maximum error of π), the information used for the separation is equivalent to only knowing the phase sign of each TF bin. In those conditions the results are poor and not improved by increasing the number of iterations beyond 20, it is however sufficient to outperform the Wiener filter in terms of SIR. Considering SDR and SAR, 16 quantization steps are needed to outperform the Wiener filter whereas 32 quantization steps are needed to outperform MISI. Increasing the number of quantization steps also increases the reconstruction quality.

5. DISCUSSION

In oracle conditions, PB-ISS allows a better quality reconstruction than both Wiener filter and MISI, but at the cost of a slower convergence, therefore needing more computation power. More importantly, especially in Figure 2, we can see that PB-ISS presents a ceiling in reconstruction which is much higher than the two other methods. For M2, the ceil is reached at approx. 1700 iterations and 35dB of SDR. This shows the potential of phase information for signal separation or reconstruction should it be properly modeled. The results of the oracle PB-ISS without error distribution show the paramount importance of step 5 of the algorithm.

Audio examples are available online[1]. Original and reconstructed sources of the second mixture are given for the Wiener filter and the two iterative methods for three iteration values (10, 40 and 200) all in Oracle conditions. On top of that, the 2000 iteration point for PB-ISS is also given. Audio samples for the 6 levels of phase quantization PB-ISS at 40 iteration are also given. Contrary to spectrogram based ISS, PB-ISS tends to increase SIR by removing the interferences using the phase constraint. At each iteration, the target source is enhanced and interfering sources are removed. This effect leads to much less "subjective" musical noises but a decreased aspect of "cleanness" of the separation: the source are harder isolated. This should however not be a problem in an active listening framework where the sources are not intended to be listened to separately but rather remixed. The audible saturation shows that the reconstruction is not gain controlled, but the gain eventually converges to the original value.

Detailed results in oracle conditions are shown in Figure 3 for two iteration numbers: the approximate SDR crossing point between the three methods (40 iterations) and the higher 200 iterations point. At 40 iterations, PB-ISS does not always outperform the other two methods, but leads to more consistent results across sources. For instance, the Synthesizer of Mixture 1 is hardly well separated using magnitude-based ISS whereas PB-ISS easily gives

[1] http://nicolas.sturmel.com/DAFx12

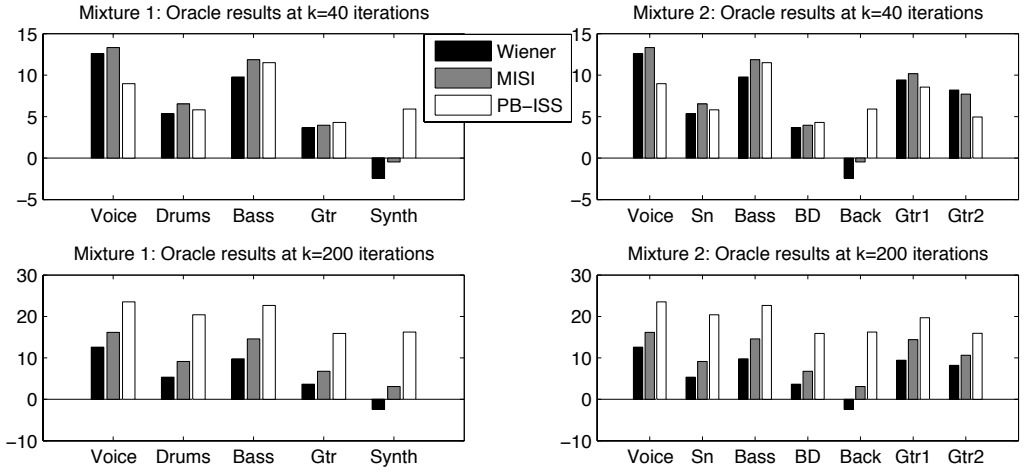

Figure 3: *SDR Results per source and per mixture at 40 iterations and at 200 iterations for the three tested methods in oracle conditions.*

more than 5dB SDR. This is also the case for the backing vocals of Mixture 2. Therefore, on the average, PB-ISS is better. On the contrary, predominant sources in the mixture such as the lead voice, do not immediately benefit from PB-ISS. The phase of the mixture is generally very close to the phase of the predominant source. In this case, magnitude adaptive filtering already gives a very high quality separation. As expected, at 200 iteration PB-ISS outperforms other methods by at least 5dB.

The proposed method is much slower than classical Wiener filter. In Matlab, while Wiener filtering needs approx. 0.1 second of CPU per second of processed signal (s/s), our method needs approx. 4 s/s at 40 iterations. However, optimization of the reconstruction algorithm should allow real time computation.

At the exception of PB-ISS, all method are in oracle condition. For such condition, the size of the side-information is too big to be considered, therefore no bit rate will be given. Very small bit rates can be achieved on spectrograms using image compression of matrix factorization (e.g. 3 to 10kbit/source, see [2]), but such compression on phase distributions remains to be investigated.

6. CONCLUSION

In this paper we have presented a new approach to informed source separation. Whereas existing ISS methods addressing monophonic mixtures use the magnitude of the STFT to perform the separation, we presented an algorithm that uses solely the phase information. For this, we adapted the MISI technique to suit our needs and tested the proposed Phase-based ISS (PB-ISS) against MISI and the Wiener filter both in oracle condition (i.e. knowing perfectly the source spectrogram). We showed that our approach outperforms both methods up to 32 quantization steps but at higher computation cost. In oracle conditions, our method presents a much higher ceiling than other spectrogram only based ISS method do. This indicates that phase is a valuable information of the source signal TF representation in a source separation framework. Therefore, further work should focus on modeling and coding this phase information in order to develop a full ISS system. Aside from the existing posterior coding implementation of [7], hybrid iterative reconstruction models using both phase and spectrogram information should also be investigated.

7. REFERENCES

[1] A. Ozerov and C. Fevotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE TASLP*, vol. 18, no. 3, pp. 550 –563, march 2010.

[2] A. Liutkus, J. Pinel, R. Badeau, L. Girin, and G. Richard, "Informed source separation through spectrogram coding and data embedding," *Signal Processing*, vol. 92, no. 8, pp. 1937–1949, 2011.

[3] J. Engdegard and al., "MPEG spatial audio object coding, the ISO/MPEG standard for efficient coding of interactive audio scenes," in *AES Convention 129*, november 2010.

[4] C. Faller, A. Favrot, Y.-W. Jung, and H.-O. Oh, "Enhancing stereo audio with remix capability," in *AES Convention 129*, november 2010.

[5] M. Parvaix and L. Girin, "Informed source separation of linear instantaneous under-determined audio mixtures by source index embedding," *IEEE TASLP*, vol. 19, no. 6, pp. 1721 –1733, aug. 2011.

[6] S. Gorlow and S. Marchand, "Informed source separation: Underdetermined source signal recovery from an instantaneous stereo mixture," in *Proc. IEEE WASPAA*, 2011, pp. 309–312.

[7] A. Ozerov, A. Liutkus, R. Badeau, and G. Richard, "Informed source separation: source coding meets source separation," in *Proc. IEEE WASPAA*, Mohonk, Oct 2011, pp. 257–260.

[8] D. Gunawan and D. Sen, "Iterative phase estimation for the synthesis of separated sources from single-channel mixtures," *IEEE Signal Processing Letters*, vol. 17, no. 5, pp. 421 –424, may 2010.

[9] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32(2), pp. 236–243, 1984.

[10] J. Le Roux, H. Kameoka, N. Ono, and S. Sagayama, "Fast signal reconstruction from magnitude STFT spectrogram based on spectrogram consistency," *Proc. of DAFx '10*, 2010.

[11] V. Gnann and M. Spiertz, "Multiresolution STFT phase estimation with frame-wise posterior window length decision," in *Proc. of DAFx '11*, Sep. 2011, pp. 101–106.

[12] N. Sturmel and L. Daudet, "Signal reconstruction from its STFT amplitude: a state of the art," in *Proc. of DAFx '11*, 2011.

[13] L. D. Alsteris and K. K. Paliwal, "Iterative reconstruction of speech from short-time fourier transform phase and magnitude spectra," *Computer Speech & Language*, vol. 21, no. 1, pp. 174 – 186, 2007.

[14] M. Hayes, J. Lim, and A. Oppenheim, "Signal reconstruction from phase or magnitude," *IEEE ASSP*, vol. 28, pp. 672 – 680, 1980.

[15] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE TASLP*, vol. 14, no. 4, pp. 1462 –1469, july 2006.