Evaluation of Techniques for Blind Sources Separation in the Identification of Musical Instruments

Jorge Costa Pires F., Mariane Rembold Petraglia, Diego Barreto Haddad
Laboratory of Analog and Digital Signal Processing
Federal University of Rio de Janeiro, PEE/COPPE
Rio de Janeiro, RJ, Brazil
jcpfilho@gmail.com, mariane@pads.ufrj.br, diego@pads.ufrj.br

Abstract **-** This work makes a comparative analysis of some methods of blind source separation and their respective capabilities of serving as a tool accessory to a system of automatic recognition of musical instruments from polyphonic signals. For such, several methods were used, such as Sparse Component Analysis, Fast Independent Component Analysis and Independent Component Analysis, and an algorithm was elaborated in the present work.

Keywords - Classification, Musical Instruments, Blind Source Separation

I. INTRODUCTION

This work has as main objective to undertake an analysis of the use of methods of Blind Source Separation (BSS) as an accessory tool for a System of Automatic Recognition of Musical Instruments (SARMI) for polyphonic signals. The tested BSS methods include Sparse Component Analysis (SCA) [1], Independent Components Analysis (ICA) [2] and FastICA [3]. The use of these algorithms in conjunction with a SARMI results in a gain of performance when compared to strategies that aim to estimate directly the notes. We search for an optimal supervised separation algorithm that provides minimal disturbance of the classifier, since our main objective in this work is to assess the impact of BSS techniques when used as a pre-processing stage. In this paper we use algorithms that do not require any prior knowledge about the sources (in our case, musical instruments), although some statistical knowledge, even though inaccurate, about the samples distribution is frequently used. This hypothesis is not too restrictive, since it is known that speech and audio signals have a super-Gaussian distribution, which can be modeled in the context of BSS by a Laplacian probability density function.

This work also intends to compare the performance of some BSS methods when used for separating signals in instantaneous determined mixtures (where the number of sources is equal to the number of mixtures) of speech signals or of monophonic sequences of musical notes (from instruments).

Underdetermined mixtures (a difficult configuration where the number of sources exceeds the mixtures) are also treated in this paper, using Sparse Component Analysis techniques.

At the end, we propose an algorithm for the separation of certain mixtures making use of the sparsity property (and not only of independence), which henceforth will be called SCAm.

II. SCAm ALGORITHM

This algorithm consists of a robust method able to estimate directly the coefficients of the mixture matrix. We emphasize this difference, since the methods of Independent Component Analysis (ICA) conduct a search in the space of separation matrices. The proposed method uses two principles:

1. The sources are sparse (when they are not, a sparse transformation technique, such as wavelet and STFT transforms or Matching Pursuit, can be applied);
2. The sources are statistically independent.

To estimate the mixture system coefficients, we use the estimator called "Zibulevsky"[1] [5, 6], which calculates a histogram of the angles of the coordinates, with typical format shown in Figure 1. Histogram peaks are located on the angles whose tangents correspond to the estimated coefficients of each row of the mixing matrix - see [5] for details. It should be noted that the concept of sparsity employed here is weaker than that used in linear algebra, since we only need that the samples of the sources (or the coefficients of the transformed sources) are for the most part close to zero. In other words, a few samples of sources have most of their energy.

[1] Another method could be used, such as the classic clustering algorithm called K-means.

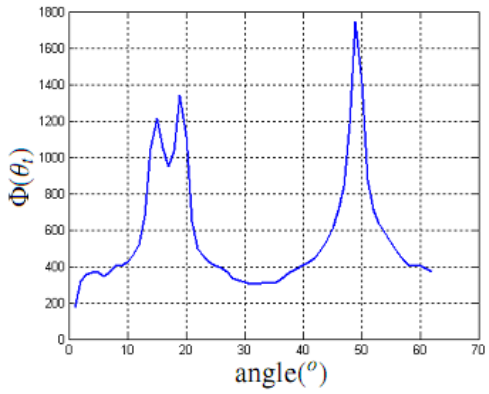

Figure 1. Typical histogram obtained by Zibulevsky estimator.

Thus, the tested algorithm contains the following steps:

1. Estimate all coefficients of each row of the mixture matrix (via Zibulevsky estimator);
2. For each row, get all the possible permutations for the estimated coefficients;
3. Generate all combinations among the distinct rows (candidates);
4. Generate all candidate matrices by combining Steps 2 and 3;
5. Calculate the inverse of all mixture matrix candidates, generating several separation matrices;
6. Determine the sources estimates for each separation matrix candidate;
7. Determine the optimal separation matrix maximizing some measure of the independence of the estimates.

III. SARMI

The employed SARMI uses the classifier bank shown in Figure 2 [4]. The performance of such classifier is superior to any of the ensemble (if used in an isolated manner). This SARMI was designed for single notes recognition, using only a segment of each note for coding.

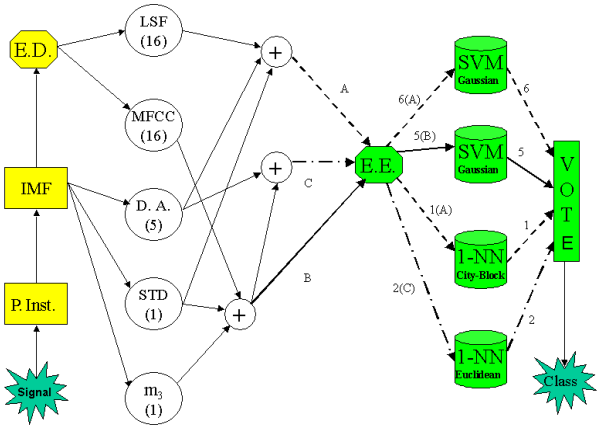

Figure 2. SARMI

As illustrated in Figure 2, the note feature vector consists basically of two types of encoders: coefficients - Line Spectral Frequencies [7] - and MFCC - Mel-frequencies cepstral coefficients [8, 9, and 10]. Both encoders contain 16 coefficients, extracted from a segment of each note. Such segment was obtained by applying a threshold (equal to 90% of the average instantaneous power of the musical note) to determine the initial and final segment samples. In addition to the encoders, the vector of elements also contains the standard deviation, the moment of third order and five audio descriptors (the rate of zero crossing, the spectral flux, the RMS value of the frame, the spectral centroid and the bandwidth of the spectral centroid). These measures are combined in 4 different ways, resulting in 4 distinct feature vectors, which are specific of each classifier. We used as classifiers: 2 SVMs - Support Vector Machine [11, 12, and 13] - and 2 *K*-NNs - *K*-nearest neighbor [14]. At the end, the predicted class is the most voted one from among the four classifiers (in case of a tie, there is a draw among the predictions). The SARMI was trained with a subset[2] of notes from three databases: RWC[3] [15], MIS[4] [16] and MUMS[5] [17]. The adopted system is capable to classify 20 different instruments. The set used in the training phase does not contain notes from the monophonic sequences of test, with 10% of the notes originated from of the RWC database.

IV. BSS

In the implementation of the FastICA algorithm, maximization of kurtosis is employed, while in the ICA algorithm the steepest descent optimization method is used. For both algorithms, we arbitrated the maximum number of iterations (equal to 300) as stop parameter in the process of convergence.

Due to the permutation problem, after we get the estimates from the process of separation, we still need to identify which source is associated with a given estimate. This is not always an easy process of identification because in many cases the estimates still suffer contamination from other sources. To solve such problem, the following metric of comparison of the estimate with the original source was used: the signal-distortion ratio (SDR), defined as

$$SDR = 10\log_{10}\left[\frac{\sum_{i=1}^{N}|s(i)|}{\sum_{i=1}^{N}|s(i)-\hat{s}(i)|}\right]. \quad (.1.)$$

Thus, each estimate will have a SDR measure of similarity with respect to each one of the original sources.

[2] Subset with 90% of the notes originating from the 3 used databases.
[3] *RWCReal World Computing.*
[4] *MISMusical Instruments Samples.*
[5] *MUMSMcGill University Master Samples.*

When the separation process does not get a good performance, it is possible that more than one estimate are associated to the same source or that the same estimate is the best representation for all the original sources. Therefore, the use of metrics to measure the degree of similarity of the estimate with the original source may not yield a good representation of the separation system. When the separation is satisfactory, there will be a distinct association between the estimates and sources. To circumvent this possible distortion, we use the following criterion to obtain a measure of separation (MS) for the algorithm:

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N}\left[\max(SDR(S_j, \hat{S}_i))\right] \quad (2)$$

$$\overline{Y} = \frac{1}{N}\sum_{j=1}^{N}\left[\max(SDR(S_j, \hat{S}_i))\right] \quad (3)$$

$$MS = \min\left(\overline{X}, \overline{Y}\right) \quad (4)$$

where S_j is the j-th source and S_i represents the i-th estimate. The resulting estimator is optimal (that is $\overline{X} = \overline{Y}$) when the separation is satisfactory or when the above cases do not occur

V. BSS+SARMI

This experiment aims at determining the class of each musical instrument in the original sources (formed by monophonic sequences) from the observed instantaneous mixtures. The procedure was developed to achieve this goal through the following stages:

1. Separation of the *M* mixtures in *N* monophonics sequences;

2. Extraction of the notes of each estimated monophonic sequence;

3. Classification of the extracted notes by a previously elaborated classifier.

In this procedure, a BSS algorithm (FastICA, ICA, SCAm or SCA) was employed, followed by a note extractor which used the mean and standard deviation of a previously established window to determine the beginning and the end of each note [4]. All notes are from the database RWC.

Monophonic sequences were constructed from notes of the test sequence (obtained from an established percentage and drawn from the central range of the instrument). These notes are drawn (forming a smaller subset) and separated by intervals (gaps) randomly chosen from 0.045 ms and 0.3 ms. After the formation of the monophonic sequences, the polyphonic signal was constructed using a random 3×3 mixture matrix, resulting in four triple audio signals. Each audio signal of a given triplet is mixed instantaneously through a matrix of mixtures drawn randomly, 10 random matrices of mixtures were elaborated for each triple of signals. The four triples are formed by 3 speech signals [18] (case A) or by 3 monophonic sequences of notes (cases B, C and D). In case B, mixtures of reed instruments, such as oboe and saxophone, are considered. In case C, the sequences are formed by notes of percussion instruments, such as xylophone, glockenspiel and vibraphone. Case D corresponds to brass instruments, such as trumpet and trombone.

We tested the ICA, FastICA, SCA[6], SCAm, algorithms (without the sparse transform and resorting to sparse transformed wavelet packet db32).

The MS results for case A are presented in Table 1.

TABLE I. MS RESULTS OBTAINED FOR MIXTURES OF VOICE SIGNALS

	FASTICA	ICA	SCAm	SCAm + IDCT	SCAm + Db32	SCA - 1,3
1	2,58	30,82	18,52	1,83	3,88	4,28
2	4,64	12,35	2,28	1,30	19,14	4,05
3	2,89	30,82	14,96	3,03	7,46	2,72
4	2,51	17,33	0,74	-0,07	4,98	0,66
5	3,03	17,07	1,14	1,26	2,72	2,90
6	6,39	2,37	0,92	0,94	4,86	0,50
7	6,22	11,84	19,30	0,90	15,19	2,18
8	4,80	30,82	4,70	1,47	3,24	1,93
9	2,84	14,86	22,95	0,56	10,44	1,74
10	3,69	31,38	3,82	0,51	0,92	1,49

As expected, the average results for the ICA, SCAm and FastICA were higher than the results of the SCA, since the SCA only used two of the three mixtures to perform the separation. In the case of percussion instruments, in addition to Table 2 that contains the measures MS, the number of wrong estimates about the class of musical instrument predicted by the classifier is shown in Table 3. The number of errors ranges from zero to three, since each mixture has three notes from distinct musical instruments.

TABLE II. MS RESULTS OBTAINED FOR MIXTURES OF SIGNALS OF PERCUSSION INSTRUMENTS

	FASTICA	ICA	SCAm	SCAm + IDCT	SCAm + Db32	SCA - 1,3
1	3,39	47,66	0,50	-1,32	0,58	2,57
2	16,69	47,62	32,94	1,30	10,78	3,01
3	1,62	47,13	1,04	2,10	2,66	5,10
4	1,89	47,16	27,16	0,43	14,52	2,64
5	0,23	47,51	38,41	-1,20	11,01	0,62
6	1,25	48,01	31,37	-1,23	1,12	15,19
7	6,14	48,01	30,82	2,04	3,17	14,16
8	3,39	47,56	31,85	0,95	2,08	1,34
9	1,24	47,20	28,12	-1,69	6,47	2,58
10	2,87	47,39	36,04	4,03	11,40	1,09

TABLE III. ERRORS OBTAINED FOR MIXTURES OF SIGNALS OF PERCUSSION INSTRUMENTS

	FASTICA	ICA	SCAm	SCAm + IDCT	SCAm + Db32	SCA - 1,3
1	2	0	1	2	2	0
2	0	0	0	2	1	1
3	2	0	2	2	2	1
4	2	0	0	1	1	0
5	2	0	0	2	2	0
6	2	0	0	2	2	0
7	2	0	0	1	2	0
8	2	0	0	1	2	0
9	2	0	0	2	2	0
10	2	0	0	2	2	2

[6]For both mixtures 1 and 2 and mixtures 1 and 3, the wavelet packet db32 transform was used [19].

As can be observed, the best separation results were obtained with the ICA and SCAm algorithms. However, the performance obtained by the SARMI with the SCA algorithm was equivalent to that achieved with the SCAm.

The following tables contain the combination of the results of cases C and D. Table 4 shows the average MS for 10 different of instantaneous mixtures, while Table 5 shows the average accuracy rate obtained by the classification system associated to each of the blind separation algorithms of Table 4.

TABLE IV. AVERAGE RESULTS

mean (MS)	FASTICA	ICA	SCAm	SCAm + IDCT	SCA - 1,2	SCA - 1,3
Percussion	3,87	47,53	25,83	0,54	7,52	4,83
Reeds	5,46	37,96	5,99	1,47	3,00	2,89
Brass	3,42	39,50	16,47	1,59	2,82	2,53
Voice	3,96	19,97	8,93	1,17	2,14	2,24
Mean	**4,18**	**36,24**	**14,31**	**1,19**	**3,87**	**3,12**

TABLE V. AVERAGE OF RATES OF SUCCESS

	FASTICA	ICA	SCAm	SCAm + IDCT	SCA - 1,2	SCA - 1,3
Percussion	40,00%	100,00%	90,00%	43,33%	93,33%	86,67%
Reeds	13,33%	86,67%	20,00%	13,33%	10,00%	10,00%
Brass	80,00%	100,00%	90,00%	70,00%	80,00%	83,33%
Mean	44,44%	95,56%	66,67%	42,22%	61,11%	60,00%

VI. CONCLUSION

This paper proposes a modified SCA method (SCAm), based on the direct estimate of the mixing matrix. The proposed method presented better performance than the FastICA in our experiments. However, it is a combinatorial algorithm, its high computational complexity may turn it prohibitive for cases of a large number of sources.

It was observed that the SCA approach can result in good identification of musical instruments for instantaneous underdetermined mixtures. Although there is a correlation between the SDR matrix and the classification error matrix, the SCA resulted in good classification rates for cases in which the SDR presented poor results, implying that the distortions introduced in the separation process by the SCA approach did not disturb the performance of the classifier.

The results obtained with the classification system using SCA or SCAm were similar, in the average, to those obtained by ICA, except in case B. As it can be observed, the performance of SCA for underdetermined mixtures, in which even a correct estimate of the mixing matrix is not sufficient to recover the sources, is still poor. It was observed that the FastICA and SCAm methods were more robust when compared to ICA, which diverged several times, requiring frequent adjustment of its parameters. It was also observed that the performance of the separation methods were dependent on the mixing matrix, as well as on the sources.

Finally, it is concluded that the use of blind separation techniques in SARMI improves its performance considerably, with excellent results obtained with the ICA method.

REFERENCES

[1] Andrzej Cichocki and Rafal Zdunek, "Regularized alternating least squares algorithms for non-negative matrix/tensor factorization," pp. 793–802, 2007.

[2] Pierre Comon, "Independent component analysis: a new concept? ," 36(3):287–314, 1994.

[3] Aapo Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," 10(3):626-634, 1999.

[4] Jorge Costa Pires Filho, *Classificação de Instrumentos Musicais em Configurações Monfônicas e Polifônicas*, Dissertação de Mestrado, COPPE/UFRJ, Setembro 2009.

[5] P. Bofill and M. Zibulevsky, "Underdetermined blind source separation using sparse representations, signal processing," 81: 2353-2362, 2001.

[6] M. Zibulevsky P. Kisilev and Y. Y. Zeevi, "Multiscale framework for blind separation of linearly mixed signals," 2003.

[7] [P. Kabal and R. P. Ramachandran, "The computation of line spectral frequencies using chebyshev polynomials," vol. 34, no. 6, pp. 1419-1426, December 1986.

[8] P. Mermelstein, "Distance measures for speech recognition, psychological and instrumental," 1976.

[9] S.B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," 1980.

[10] J. S. Bridle and M. D. Brown, "An experimental automatic word-recognition system," 1974.

[11] Corinna Cortes and V. Vapnik, "Support-vector networks," 20, 1995.

[12] I. M. Guyon B. E. Boser and V. N. Vapnik, "A training algorithm for optimal margin classifiers," pp. 144-152, Pittsburgh, PA, 1992.

[13] E. Braverman M. Aizerman and L. Rozonoer, "Theoretical foundations of the potential function method in pattern recognition learning," 1964.

[14] Thomas M. Cover and Peter Hart, "Nearest neighbor pattern classification," Vol. 13 (1) pp. 21-27, 1967.

[15] Masataka Goto and Takuichi Nishimura, "Rwc music database: Music genre database and musical instrument sound database," via http://sta.aist.go.jp/m.goto/RWC-MDB/m, ISMIR, pp. 229-230, 2003.

[16] Lawrence Fritts, "Musical instruments samples of iowa university, mis," via http://theremin.music.uiowa.edu/ MIS.html, 1997.

[17] Frank Opolko and Joel Wapnick, "Mcgill university master samples," via http://www.music.mcgill.ca/ resources/mums/html/mums.html, 1987.

[18] Emanuel Vincent and Hiroshi Sawada, "Stereo audio source separation evaluation campaign," via http://www.irisa.fr/metiss/SASSEC07/dev.zip, January 2010.

[19] Ingrid Daubechies, "Ten lectures on wavelets," ISBN 0-89871-274-2, 1992.