TOWARDS EFFECTIVE SINGING VOICE EXTRACTION FROM STEREOPHONIC RECORDINGS

Stratis Sofianos, Aladdin Ariyaeeinia, and Richard Polfreman

University of Hertfordshire, Hatfield, UK

ABSTRACT

Extracting a singing voice from its music accompaniment can significantly facilitate certain applications of Music Information Retrieval including singer identification and singing melody extraction. In this paper, we present a hybrid approach for this purpose, which combines properties of the Azimuth Discrimination and Resynthesis (ADRess) method with Independent Component Analysis (ICA). Our proposed approach is developed specifically for the case of singing voice separation from stereophonic recordings. The paper presents the characteristics of the proposed method and details an objective evaluation of its effectiveness.

Index Terms— audio source separation, singing voice separation

1. INTRODUCTION

Singing Voice Separation (SVS) can be defined as the process of extracting the vocal element from a given song recording. The impetus for research in this area is mainly that of facilitating certain important applications of Music Information Retrieval (MIR) such as lyrics recognition, singer identification, and singer melody extraction [1], [2].

The field of SVS can be seen as a subset of audio source separation, which has mainly taken two paths (mono and stereo) depending on the nature of the observed mixture. Mono approaches utilize such diverse methods as pitch detection and amplitude modulation [3], source-adapted models [4], and normalized cuts [5]. On the other hand, most stereo methods make use of the Interchannel Intensity Difference (IID) in addition to the Interchannel Time Difference (ITD) or the Interchannel Phase Difference (IPD) [1], [6-8].

The concern in this paper is that of unsupervised singing voice separation from stereophonic studio recordings. As a base for our method, the *Azimuth Discrimination and Resynthesis* (ADRess) algorithm [7] is used. A main attraction of ADRess is that it aims to isolate individual music sources, by exploiting the IID between the two channels of the stereophonic mix. However, ADRess is not optimized for the vocal element of a song, requires user supervision, and is ineffective when two or more music sources share the same panning position [7]. The motivation for incorporating ADRess is that, with appropriate modifications, the method can provide an automated means for isolating the central panning subspace of the stereo field. This central subspace is further processed in our system with the aim of enhancing the effectiveness of SVS.

Our method exploits assumptions that are valid in most commercial song recordings, such as the traditional placement of the lead vocal element at the center of the stereo field, and its amplitude dominance over the coexisting music sources. In addition, a novel approach based on combining the modified version of ADRess with Independent Component Analysis (ICA) [9] is proposed and investigated. We term our novel approach "Singing Extraction through Modified Adress and Non-vocal Independent Component Subtraction (SEMANICS)".

The rest of the paper is organized as follows. Section II describes the ADRess algorithm [7]. Section III gives a detailed description of the proposed system. Section IV presents the experimental investigations. Finally, Section V provides overall conclusions and suggestions for future work.

2. ADRESS

According to the theoretical concept that ADRess incorporates, every music source in a stereo recording has a panoramic position that can be expressed as an intensity ratio between the two channels of the stereo mixture [7]. Therefore, by applying a scaling factor on the STFT (short-term Fourier transform) of one channel and subtracting it from the other channel, the target source can be cancelled. Subsequently, the minima of the minuend STFT are located and the target source is reconstructed. The process can be further described as follows. Let

$$x_i(t) = \sum_{j=1}^{J} a_{ij} s_j(t), \qquad i = 1, 2 \tag{1}$$

represent the stereo mix, where s_j are the J independent sources, a_{ij} are panning coefficients for the jth source, and i is the index of the channel. The jth source can be cancelled out by:

$$x_1(t) - g_j \times x_2(t) \tag{2}$$

where g_j *is* the intensity ratio between $x_1(t)$ *and* $x_2(t)$ for the jth source[1]. The estimation of the value of g_j and the reconstruction of the jth source can be detailed as follows. Initially, (2) is rewritten in the time-frequency domain and different values of g (gains) are

[1] x_1 *and* x_2 represent the left and right channel of a stereo recording but not necessarily in that order. A condition for avoiding possible distortions is that of $g_j \leq 1$ [7]. Therefore, x_2 is always the channel that contains the target source more prominently.

applied:

$$AZ(k,l) = \left| X_1(k) - g(l) \times X_2(k) \right|, \text{ for } l = 1,2,3 \ldots \beta \quad (3)$$

where $X_1(k)$ and $X_2(k)$ are the Hann-windowed frequency transforms of the two channels, and β is the total number of gains that are applied. *AZ* is termed *Azimugram* in [7]. The value of each gain is $g(l)=l/\beta$. The value of β in [7] (termed *azimuth resolution*) is equal to 100. This is retained in this paper. The resulting minima are first located and then substituted with peaks as follows:

$$AR(k,l) = \begin{cases} AZ(k)_{\max} - AZ(k)_{\min}, \text{ if } AZ(k,l) = AZ(k)_{\min} \\ 0, \qquad \text{otherwise} \end{cases} \quad (4)$$

$\forall\ k,l$, where $AZ(k)_{\max}$ and $AZ(k)_{\min}$ are the maximum and minimum magnitude values over a given frequency bin (i.e. *k*) respectively.

In an ideal scenario, where all sources are panned to different locations and there is no frequency overlap between them, the target source spectrogram is obtained by a single value of *l* in (4). However, as this is usually not the case, a *range* of values of *l* are needed to be included in the target source spectrogram which is obtained by:

$$W(k) = \sum_{l=d-H/2}^{d+H/2} AR(k,l), \quad \forall\ k \quad (5)$$

where *H* is the number of values in the extracted range of *l*, and *d* is the midpoint of the range. In [7], *H* and *d* are termed *subspace width* and *discrimination index* respectively. Depending on the value of *H,* a trade-off applies: a large value includes more vocal components but results in poor separation, whereas a small value provides better separation, but excludes some of the target bins. Finally, the phase information from the original mixture is used together with *W(k)*, in order to transfer the recovered source to the time domain.

As mentioned before, ADRess on its own will only help to isolate the central panoramic subspace. This subspace will include the voice mixed with all the instruments that exist in the center. Therefore, in order to increase voice isolation, further processing is required.

3. PROPOSED APPROACH

SEMANICS is a hybrid singing voice separation system that could be better described as a fusion between ADRess and ICA. The ADRess part of the system has fixed parameters (i.e. can run unsupervised), and is enhanced by an approach introduced in this paper, termed "amplitude discrimination". The ICA part of SEMANICS utilizes the application of ICA to stereophonic mixtures. A main attraction of SEMANICS is that it can also be effectively applied on stereophonic mixtures that are subjected to typical commercial mixing and mastering processes (e.g. equalization, compression, artificial reverb). This is unlike ADRess that can only operate on pure stereo mixtures (see (1)). In addition, SEMANICS is not hindered by the use of stereo microphone techniques during studio recording (e.g. stereo recording of a piano). The only requirements are that the voice is panned at the center of the mix, and that it is the dominant source at this panning position. These assumptions, arguably, cover a wide range of commercial recordings.

3.1. Amplitude Discrimination

The proposed extensions take place after Equation (4) of original ADRess. Equation (4) provides a matrix *AR*, whose rows *k* are frequency bins that contain peaks at specific azimuth values (i.e. columns *l*). The peaks that are near the end of the Azimuth (i.e. ~β) contain the music sources that are at the center of the original mix. This panning location is usually the most occupied, as it traditionally hosts the lead vocal part together with other instruments (e.g. bass, bass drum). Hence, ADRess is unable to separate them from the voice. The purpose of the proposed "amplitude discrimination" approach is to enhance the capability for the isolation of voice.

The proposed modification is based on the assumption that the vocal component will be more dominant than the other music components in the estimated magnitude[2] spectrogram *AR*. By more dominant, it is implied that the magnitude of each of the individual bins that contain the vocal frequencies is generally higher than the mean of the frequency bins within designated frequency bands. This holds because the estimated *AR* contains essentially only the center location of the stereo mix and mixing engineers usually process the vocal part (e.g. with filtering or compression) so that is dominant over the whole frequency spectrum at the same panning position, in order to avoid "masking" phenomena (i.e. increase intelligibility)[10]. Furthermore, it is common that during the mixing process, the voice is subjected to highpass filtering in order to avoid saturation of the low frequencies. Based on these assumptions, we define amplitude discrimination subbands based on the equal division of the full mel-scale. The mean magnitude is then calculated for each of the subbands and only the individual bins that exceed the mean within their corresponding subband are extracted. An additional subband is defined ($\mathbf{b_0}$), such that it accounts for the highpass filtering of the voice in the mixing process.

Initially, the matrix *AR* from (4) is used in order to calculate the mean of the magnitudes for *M* number of subbands:

$$\mu_m = \frac{1}{Q} \sum_{l=\beta-H}^{l=\beta} \sum_{k \in \mathbf{b}_m} AR(k,l) \quad (6)$$

for m=1, 2, … *M,* where μ_m is a scalar, *Q* is the number of elements that are summed, β is the azimuth resolution, **b** is the subband, and *H* is the subspace width. The amplitude discrimination is then applied as follows:

[2] As in the original ADRess, in our system we use magnitude spectra instead of power spectra.

$$AD_{(k,l)} = \begin{cases} AR(k,l), & \text{if } AR(k,l) > \mu_m \text{ and } k \in \mathbf{b}_m \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

for $(\beta\text{-}H) \le l \le \beta$, and m=1, 2, … M. This algorithm functions as a brick wall, allowing only the bins with magnitude higher than their respective subband thresholds to pass.

3.2. Non-vocal Independent Component (NIC) Subtraction

The process of amplitude discrimination helps significantly towards the voice isolation, but it is not able to completely filter out all the frequency components originating from other sources. In this section, we describe how *SEMANICS* uses the Fast ICA algorithm [9] in order to achieve further voice isolation after the amplitude discrimination is applied.

Initially, we exploit the properties of ICA in order to obtain a mixture of music instruments from the unprocessed song. Subsequently, this mixture of music instruments is "subtracted", in the frequency-domain, from the summed result of the amplitude discrimination process.

In general, when the ICA algorithm is applied to an arbitrary mixture, it separates the mixture into subspaces (in the case of stereo mixture they are two) that are as independent as possible [11]. Some of the source signals will be in the first output while the other sources will find place in the second output [12]. Hence, one of the outputs will contain the vocal element mixed together with some of the sources, while the other will contain only a mixture of the remaining sources, with little vocal information. We can exploit the latter mixture in order to achieve further voice isolation. For our system we use the Fast ICA algorithm as proposed by [13].

Since the output order of ICA is not known, the NIC determination takes place after Fast ICA is applied to the original mixture. In order to automatically choose which one of the two outputs contains less vocal part, each of the ICA outputs is cross-correlated with the original mixture. For this operation, the Pearson Product Moment Correlation Coefficient (PMCC) is used:

$$\rho = \frac{1}{\mathrm{T}} \sum_{t=1}^{\mathrm{T}} \left(\frac{IC_n(t) - \mu_{ICn}}{\sigma_{ICn}} \right) \left(\frac{x_2(t) - \mu_{x2}}{\sigma_{x2}} \right) \quad (8)$$

where IC_n is the n^{th} (1st or 2nd in this case) ICA output , and T is the number of samples in each of IC_n and x_2. The letters μ and σ represent the sample mean and standard deviation respectively.
The ICA output containing the vocal will give a higher correlation index, whereas the other, i.e. *NIC(t)*, outputs a lower value. The latter is used as follows to enhance the vocal separation process. First, in a similar manner to (5), we add up all the columns of the matrix *AD* to obtain a magnitude spectrogram in one vector.

$$Y(k) = \sum_{i=\beta-H}^{\beta} AD(k,l), \quad \forall\, k \quad (9)$$

Despite the magnitude of the *NIC* being arbitrary (due to ICA limitations [13]), the magnitude ratios *between* the sources that are contained in *NIC* will be similar to that in the original mixture. Hence, we define *FNIC*, Fourier transform of *NIC*, and then scale it to match the sample mean of the magnitude spectrum *Y(k)*. By subtracting the scaled absolute of *FNIC* from *Y(k)*, attempts are made to further *reduce* some of the music sources, i.e.

$$\hat{S}(k) = Y(k) - \frac{\mu_Y}{\mu_{|FNIC|}} \left| FNIC(k) \right| \quad (10)$$

where μ_Y and $\mu_{|FNIC|}$ are scalars. Subsequently, all the negative elements of $\hat{S}(k)$ are set to zero. Finally, we follow the procedures in the original ADRess: we use the phase information from the original mixtures and apply ISTFT on $\hat{S}(k)$, to transfer it to the time domain. The overview of the proposed system can be seen in Fig.1.

4. EXPERIMENTAL INVESTIGATIONS

For our experimental investigations, we used the *bss_eval* metrics [14-15]. The results can be read as Source-to-Distortion Ratio (SDR), source Image to Spatial distortion Ratio (ISR), Source-to-Interference Ratio (SIR), and Source-to-Artifacts Ratio (SAR). For the purpose of testing, a small database comprising a number of songs was created.

The dataset used for the experiments consisted of 10 samples as detailed in Table I. All mixtures in the database were subjected to some form of convolution process during the mastering stage. The last two samples on the list (i.e. "bearlin roads" and "tanto") were taken from [15]. It is worth noting that the results given in Table I for "tanto" are comparable to those reported previously for this sample using other source separation methods [16].

Our system runs in an unsupervised mode and the all its parameters were fixed a priori for the whole of the dataset. These were the number of subbands ($M=4$) as in (6), the $\mathbf{b_0}$ range [0-140Hz], the discrimination index ($d=85$), and azimuth width ($H=30$, as in (5)). These settings were chosen empirically, as they gave satisfactory results during our initial experiments. The window size was 4096 samples long, and the overlap was 75%. The results for our dataset can be seen in Table I, where column "SDR mix" shows the results for the original mix (i.e. SDR results

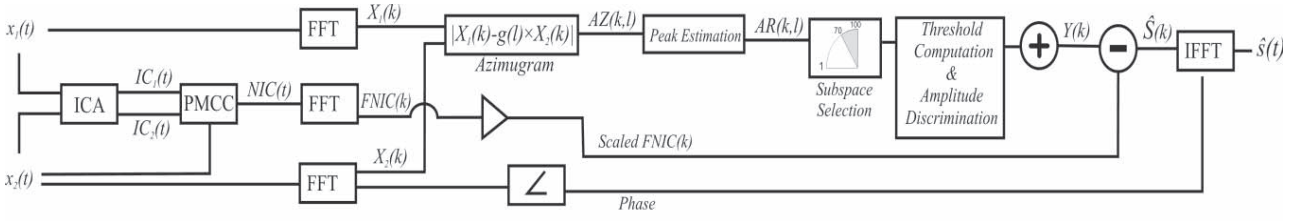

Fig.1 Structure of the proposed SEMANICS approach to singing voice separation.

TABLE I
BSS_EVAL RESULTS

				SEMANICS				ADRESS				
Title	**Dur. (sec)**	**Artist**	**Genre**	**SDR**	**ISR**	**SIR**	**SAR**	**SDR**	**ISR**	**SIR**	**SAR**	**SDR Mix**
Help Me	8.9	Brian. & D. Byrne	Experimental	3.4	14.6	6.6	5.3	1.0	16.9	0.3	12.7	-1.1
Kunlarim	8.0	Sevara Nazarkhan	World	4.8	9.8	7.8	6.2	2.1	7.9	-3.1	14.9	-3.6
L'Americano	7.7	FeelM	Jazz	5.7	12.2	12.8	7.2	5.7	3.9	6.8	10.0	-0.4
Nude	13.8	Radiohead	Alternative Rock	6.8	9.1	19.3	8.0	3.0	6.0	1.9	7.2	-6.0
Only	8.3	Nine Inch Nails	Industrial Rock	5.8	12.4	12.0	6.3	3.0	3.9	-4.0	13.8	-4.6
Resistencia	6.0	Los De Abajo	Latin Ska	5.3	8.7	9.6	6.0	2.3	5.0	-6.6	15.8	-6.5
Salala	8	Angelique Kidjo	Adult Contemporary	8.8	13.6	18.8	10.2	3.6	16.2	3.7	11.9	-0.9
Monkey	4.6	Peter Gabriel	New Wave	3.9	7.9	8.7	4.5	-1.3	14.9	-2.5	21.3	-2.9
Bearlin Roads	14	Brian. & D. Byrne	Experimental	4.1	9.6	9.3	4.7	-0.6	5.1	-5.8	11.4	-5.2
Tanto	23	Sevara Nazarkhan	World	9.4	16.2	19.1	12.3	6.4	16.2	11.2	8.6	-4.6

Results are given in dB. All songs have a sampling frequency of 44.1 kHz and a sample resolution of 16 bits.

if no separation is performed).

As observed in this table, SEMANICS outperformed ADRess in almost every aspect except for SAR. However, this is not considered an issue, since the primary aim has been that of minimizing the interference (i.e. achieving a high SIR) which has proven to facilitate certain important applications of MIR, such as Singer Identification [2]. The resulting files from the separation can be downloaded from: **http://tinyurl.com/nopdyl**

5. CONCLUSION

In this paper, we have introduced SEMANICS, a SVS system that is specifically for commercially produced stereophonic recordings. It is based on the fusion of independent component analysis (ICA) with an extended version of ADRess. The results compare very well with the latest Stereo Source Separation methods [16]. Future work will involve further exploitation of ICA principles in order to improve the vocal separation effectiveness.

6. REFERENCES

[1] J. L. Durrieu, G. Richard, and B. David, "Singer melody extraction in polyphonic signals using source separation methods," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 2008, pp. 169-172.

[2] A. Mesaros, T. Virtanen, and A. Klapuri, "Singer Identification in Polyphonic Music using vocal separation and pattern recognition methods," in *International Conferences on Music Information Retrieval*, pp. 375-378, 2007

[3] Y. Li and D. L. Wang, "Separation of Singing Voice From Music Accompaniment for Monaural Recordings," *Audio, Speech, and Language Processing, IEEE Transactions,* vol. 15, pp. 1475-1487, 2007.

[4] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval, "Adaptation of Bayesian Models for Single-Channel Source Separation and its Application to Voice/Music Separation in Popular Songs," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 15, pp. 1564-1578, 2007.

[5] M. Lagrange, L. G. Martins, J. Murdoch, and G. Tzanetakis, "Normalized Cuts for Predominant Melodic Source Separation," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 16, pp. 278-290, 2008.

[6] C. Avendano, "Frequency-domain source identification and manipulation in stereo mixes for enhancement, suppression and re-panning applications," in *Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on.*, 2003.

[7] D. Barry and B. Lawlor, "Sound Source Separation: Azimuth Discrimination and Resynthesis," in *DAFX* Naples, Italy, 2004.

[8] E. Vincent, "Musical source separation using time-frequency source priors," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 14, pp. 91-98, 2006.

[9] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Networks,* vol. 13, pp. 411-430, 2000.

[10] R. Izhaki, *Mixing Audio: Concepts, Practices and Tools*: Elsevier Science Ltd, 2007.

[11] P. Bofill and M. Zibulevsky, "Underdetermined blind source separation using sparse representations," *Signal Processing,* vol. 81, pp. 2353-2362, 2001.

[12] M. S. Pedersen, D. Wang, J. Larsen, and U. Kjems, "Separating Underdetermined Convolutive Speech Mixtures " in *Independent Component Analysis and Blind Signal Separation*. vol. 3889 Berlin/Heidelberg: Springer, 2006.

[13] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *Neural Networks, IEEE Transactions on,* vol. 10, pp. 626-634, 1999.

[14] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 14, pp. 1462-1469, 2006.

[15] E. Vincent, S. Araki, and P. Bofill, "The 2008 Signal Separation Evaluation Campaign: A Community-Based Approach to Large-Scale Evaluation," in *Independent Component Analysis and Signal Separation*, pp. 734-741, 2009.

[16] E.Vincent, "The 2008 Signal Separation Evaluation Campaign: Professionally produced music recordings, accessed on 24/08/2009 http://www.irisa.fr/metiss/SiSEC08/SiSEC_professional/2008.