SINGING VOICE SEPARATION BASED ON NON-VOCAL INDEPENDENT COMPONENT SUBTRACTION AND AMPLITUDE DISCRIMINATION

Stratis Sofianos, Aladdin Ariyaeeinia, and Richard Polfreman

University of Hertfordshire, Hatfield, UK

(e.sofianos, a.m.ariyaeeinia,
r.p.polfreman)@herts.ac.uk

ABSTRACT

Many applications of Music Information Retrieval can benefit from effective isolation of the music sources. Earlier work by the authors led to the development of a system that is based on *Azimuth Discrimination and Resynthesis* (ADRess) and can extract the singing voice from reverberant stereophonic mixtures. We propose an extension to our previous method that is not based on ADRess and exploits both channels of the stereo mix more effectively. For the evaluation of the system we use a dataset that contains songs convolved during mastering as well as the mixing process (i.e. "real-world" conditions). The metrics for objective evaluation are based on *bss_eval.*

1. INTRODUCTION

Humans are able to derive a semantic understanding of audio. This remarkable human skill has been extensively studied in Bregman's seminal work [1]. Efforts to replicate this skill using machines have been included in the broad research field of Computational Auditory Scene Analysis (CASA) [2].

However, machines are not yet able to derive comprehensive high-level information from multi-sourced (i.e. polyphonic) audio streams, including the particularly challenging example of recorded music. In this domain, the main obstacle for machines is the lack of a mechanism to "focus" on individual sources of a polyphonic stream in the way that humans do [3]. This is supported by [4] where the authors have found a "glass ceiling" (i.e. limited success) when they tried to extract information such as timbre similarity and music genre from a music track without prior separation of sources. Therefore, in order to facilitate the aforementioned, as well as other applications of Music Information Retrieval (MIR), source separation is needed. Intuitively, the most information-rich source in music is the human voice/singing, as it can provide high-level information, such as the melody, the lyrics, and performing artist, even out of its accompanying context. However, in order to facilitate the automated extraction of this information, the singing voice needs to be isolated [5-6]. The field of research that deals with this separation/isolation has been named Singing Voice Separation (SVS).

Singing Voice Separation from single-channel recordings has been the subject of systematic study, focusing on techniques such as pitch detection and amplitude modulation [7] and source-adapted models [8]. On the other hand, source separation from stereo has enjoyed little attention [9-10], although commercially distributed music recordings today are predominantly produced in this format.

In a recent study [11], the authors have introduced a new method for singing voice separation from stereo recordings. The method (termed SEMANICS: Singing Extraction through Modified Adress and Non-vocal Independent Component Subtraction) is based on the use of *Azimuth Discrimination and Resynthesis* (ADRess) algorithm [9]. It operates by exploiting the Interchannel Intensity Difference (IID) that naturally occurs in stereophonic studio recordings. However, it is well known that ADRess and most other methods utilizing either IID (or Interchannel Phase Difference: IPD) have difficulties when processing reverberant mixtures [12].

In this paper, a modified version of SEMANICS [11] is proposed, which involves removing the ADRess part. In addition to running unsupervised and utilizing the novel approach of Non-vocal Independent Component (NIC) Subtraction that was introduced in [11], the modified system presented here exploits both channels much more effectively. The new algorithm is referred to as Singing Extraction through Multiband Amplitude eNhanced Thresholding and Independent Component Subtraction (*SEMANTICS*).

The rest of the paper is organized as follows. Section 2 describes the use of Independent Component Analysis (ICA) in the proposed system. This process is referred to as NIC Subtraction. Section 3 gives a detailed description of the approach introduced by the authors termed Amplitude Discrimination. Section 4 presents the experimental investigations. Finally, Section 5 provides overall conclusions and suggestions for future work.

2. NON-VOCAL INDEPENDENT COMPONENT (NIC) SUBTRACTION

The first stage of SEMANTICS incorporates the application of ICA to the two channels of the original stereo mixture in order to acquire a time signal that contains less vocal part than either of the original channels. In this section, the basic model of ICA is briefly presented, and its use in the proposed method is discussed.

2.1. ICA Principles

ICA is a general purpose statistical technique that is closely related to Blind Source Separation (BSS), because only weak assumptions of the original sources are made. These assumptions

include the non-Gaussian nature of the sources and their statistical independence. The most concise definition of ICA would be that of a statistical "latent variables" model: we observe U random variables x_u … x_U, which are modeled as linear combinations of J random variables s_j, i.e.

$$x_u = \alpha_{u1}s_1 + \alpha_{u2}s_2 + ... + \alpha_{uJ}s_J, \quad \text{for } u = 1,...,U \tag{1}$$

where α_{uj}, *u, j,* are scalars. By definition s_j are *statistically independent* and non-Gaussian. In vector/matrix notation, the above can be simplified:

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{2}$$

where **A** is the mixing matrix. For the mixing matrix **A,** there is also a de-mixing matrix **W**, such as:

$$\mathbf{s} = \mathbf{W}\mathbf{x} \tag{3}$$

To estimate **W**, the ICA algorithm runs multiple iterations, and converges when the value of **W** gives sources that are maximally non-Gaussian. However, one of the basic limitations of ICA is that it needs as many observations **x** as sources **s**. When the sources are more than the observations, the case is described as underdetermined.

When the ICA algorithm is applied to underdetermined mixtures, it separates the mixtures into subspaces (in the case of stereo mixture they are two) that are as independent as possible [13]. Some of the source signals will be mainly in the first output while the other sources will find a place in the second output [14]. 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 much less vocal. In the case of this study, the latter is referred to as the Non-vocal Independent Component "NIC".

We can exploit the latter mixture in order to suppress components of the accompanying instruments. For our system we used the Fast ICA algorithm as proposed by [13].

2.2. Using NIC to suppress music sources

The NIC determination takes place after Fast ICA is applied on the original mixture. Because of the statistical independence of the components as well as the dominance of the vocal part over its accompaniment (an assumption we make here which has proven valid in testing thus far), one of the components will correlate very well with the original mixture, while the other will have poor correlation (i.e. the one that contains less vocal and more music/accompaniment). Unfortunately, one of the weaknesses of ICA is its ambiguity regarding the order of the independent components. In order to automatically determine 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 absolute value of the Pearson Product Moment Correlation Coefficient (PMCC) is used:

$$\rho = \left| \frac{1}{\mathrm{T}} \sum_{t=1}^{\mathrm{T}} \left(\frac{IC_l(t) - \mu_{ICl}}{\sigma_{ICl}} \right) \left(\frac{V(t) - \mu_V}{\sigma_V} \right) \right| \tag{4}$$

where IC_l is the lth (1st or 2nd in this case) ICA output, $V(t)$ are the samples of the summed channels of the stereo mix, and T is the number of samples in each of IC_l and V. μ_{ICl} and μ_V are the sample means of IC_l and V respectively, and σ_{ICl}, σ_V are their respective standard deviations. The benefit of using the absolute value of PMCC is that the correlation index has fixed boundaries, i.e. $\rho \in \{0,1\}$, where the upper limit indicates strong correlation (in our case it is not significant if it is positive or negative). The ICA output containing the vocal will give a higher correlation index, whereas the other (i.e. *NIC*) outputs a lower value. The latter is used as follows to suppress some of the musical instruments from the vocal independent component.

Initially, the right, i.e. $x_1(t,)$ and left, i.e. $x_2(t)$ channel of the original mixture are subjected to a high-pass filter, for reasons outlined below. Subsequently, $x_1(t)$, $x_2(t)$, and $NIC(t)$,are segmented and transferred to the frequency domain, using a Hann window of 4096-point length before being subjected to a fast Fourier transform (FFT) process at 512-point intervals (i.e. 87.5% Overlap).

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*, modulus of the Fourier transform of *NIC*, and then scale it to match the sample mean of the magnitude spectrum $X_i(k)$. By subtracting the scaled *FNIC* from $X_i(k)$, attempts are made to *reduce* some of the music sources, i.e.

$$Y_i(k) = X_i(k) - \frac{\mu_{Xi}}{\mu_{FNIC}} FNIC(k) \tag{5}$$

for $i=1,2,$ where μ_{Xi} and μ_{FNIC} are scalars. Subsequently, all the negative elements of $Y_i(k)$ are set to zero.

The pre-processing with the high-pass filter provides a more successful result in scaling in (5). This is because in audio mixtures, the lower region of the frequency spectrum usually carries the most energy in the mix, due the sensitivity of the human ear at different frequency ranges (i.e. lower sensitivity to lower frequencies [15]). Due to the way that ICA works on complex mixtures, it will usually cancel out most low frequency components in the output that correlates poorly with the original audio mixture (i.e. NIC). As a consequence, *FNIC* will not contain as much of the bass frequency range. Scaling as in (5) without processing $x_1(t,)$ and $x_2(t)$ with a high-pass filter would bias the scaling factor towards a lesser value. In addition, this filtering is not considered to cause significant loss of the vocal component, as vocal parts are typically high-pass filtered during the mixing process in commercial recordings. During our initial investigations, the cutoff frequency that provided a good compromise between correct scaling and voice loss was around 140Hz. The high-pass filter that was chosen was an IIR first order maximally flat magnitude filter (i.e. Butterworth). The outputs of the NIC subtraction are further processed as described below.

3. AMPLITUDE DISCRIMINATION

The existence of a singer in a music track often implies that the singing part is the leading music source of the mix. Moreover, the lyrics that are sung usually need to be intelligible, even when the singing voice overlaps tonally with other music sources. Therefore, mixing engineers tend to process the vocal part, such that it is not masked by the accompanying instruments. The process often includes the enhancement of the frequency ranges

of the voice where significant overlap occurs [16]. Our proposed method, referred to as "Amplitude Discrimination" is motivated by this phenomenon.

The amplitude dominance of the voice is evident in the obtained magnitude spectrogram following the NIC subtraction, especially since many of the music sources are reduced by the aforementioned process. Hence, it is assumed 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. Based on this assumption, we define M amplitude discrimination subbands. Optimum effectiveness is achieved, when the number of subbands, as well as their crossing points, are set manually (by means of trial and error). However, this runs against the aim of the proposed system, which is unsupervised SVS.

Preliminary investigations suggested that an acceptable trade-off is the selection of crossing points such that each subband spans an equal number of mels. In this case, $3 \leq M \leq 5$ is found empirically to lead to satisfactory results. Formally, the thresholds are computed based on both spectrograms that are obtained after the NIC subtraction:

$$Z_m = \frac{1}{Q} \sum_{i=1}^{2} \sum_{k \in \mathbf{b}_m} Y_i(k,l) \tag{6}$$

for m=1, 2, … M, where Z_m is a scalar, Q is the number of elements that are summed, $\mathbf{b}_m$ is the mth-subband, and i is the channel index. It is noteworthy that the threshold is calculated across both channels and not individually. In addition to the M number of subbands, a subband $\mathbf{b_0}$ is defined, such that it matches the frequencies that were attenuated during the high-pass filtering, but were re-introduced due to STFT errors.

Subsequently, the Amplitude Discrimination is applied. This process functions as a binary mask, allowing only the bins with magnitude higher than their respective subband thresholds to pass. The resulting spectrogram $\hat{S}$ contains the averaged bins from the two channels as follows:

$$\hat{S}(k,l) = \begin{cases} \frac{1}{2}\sum_{i=1}^{2} Y_i(k,l) \;\; if \begin{cases} Y_1(k,l) > Z_m \\ Y_2(k,l) > Z_m \\ k \in \mathbf{b}_m \end{cases} \\ 0, \qquad \text{otherwise} \end{cases} \quad m \in \{1,2,..M\} \tag{7}$$

Fig. 1 shows a magnitude spectrum obtained after the NIC subtraction, for a Hann window of 4096 samples.

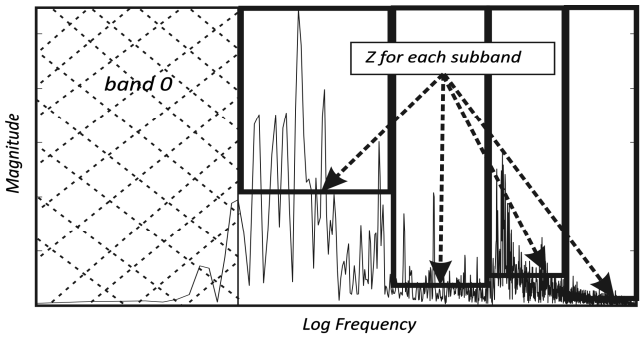

Figure 1: Amplitude Discrimination on the FFT of a windowed frame for M=4. Sampling frequency is 44.1 kHz and sample resolution is 16 bits.

The rectangles show the discrimination between the bins that have a magnitude higher than the mean for each of the subbands and are thus included in the estimation of the target source. For each subband, the computed mean value operates as a threshold. Each bin that is not included in the rectangles is zeroed.

Finally, we use the phase information from the original mixtures and perform ISTFT on $\hat{S}(k)$ to transfer it to the time domain. The overview of the proposed system can be seen in Fig. 2.

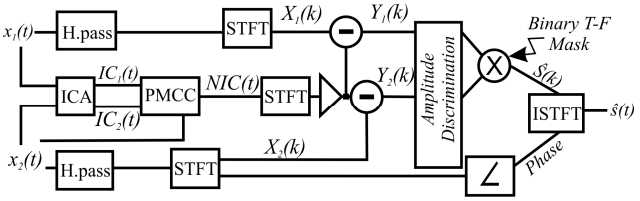

Figure 2: Structure of the proposed SEMANTICS approach to singing voice separation.

4. EXPERIMENTAL INVESTIGATIONS

For our experiments we used the *bss_eval* evaluation system as proposed in [17]. The *bss_eval* metrics system takes the estimated source $\hat{s}_j$, the acapella and the instrumental as input, and decomposes $\hat{s}_j$ as the estimated source (s_{target}), the interference (e_{interf}), and error term (e_{artef}).

In our experiments, the only allowed deformation of s_{target} is a time-invariant gain. The measures (expressed in dB) that are subsequently used to evaluate the SVS performance are as follows.

Source to Distortion Ratio (SDR)

$$\text{SDR} = 10\log\frac{\left\|s_{\text{target}}\right\|^2}{\left\|e_{\text{interf}} + e_{\text{artef}}\right\|^2} \tag{8}$$

Source to Interferences Ratio (SIR)

$$\text{SIR} = 10\log\frac{\left\|s_{\text{target}}\right\|^2}{\left\|e_{\text{interf}}\right\|^2} \tag{9}$$

Source to Artefacts Ratio (SAR)

$$\text{SAR} = 10\log\frac{\left\|s_{\text{target}} + e_{\text{interf}}\right\|^2}{\left\|e_{\text{artef}}\right\|^2} \tag{10}$$

The SIR and SAR can be regarded as valid performance measures with regards to two different goals, namely the rejection of interferences and the absence of "burbling" artefacts (also known as "musical noise") respectively. The SDR can be seen as a global performance measure [18]. It should be noted that, recently, a modified version of *bss_eval*, called *bss_eval_images* [19] includes an additional factor, which is the source Image to Spatial distortion Ratio (ISR). The ISR is of little significance to separation, but is important for applications that use phase cancellation (e.g. Karaoke) [19]. Furthermore, the gain of the estimated output plays a significant role on the results. Therefore, it has not been used in this study.

The testing of the system was performed on a customised dataset comprising 12 songs. During the mixing and mastering process, typical types of convoluted reverberation, as well as equalization and compression, were applied. Two of the songs (i.e. *Roads* and *Tanto*) were taken from [19] while the rest are licensed under Creative Commons 2.5 and can be acquired from the Internet. The results (in dB) for a Hann window of 4096 samples, with 87.5% overlapping and *M=3* are shown in Table 1.

Table 1: *bss_eval Results*

	SEMANTICS			***SEMANICS***		
Title	**SDR**	**SIR**	**SAR**	**SDR**	**SIR**	**SAR**
Salala	8.09	28.98	8.14	8.19	23.35	8.34
Nude	7.58	29.81	7.61	6.24	24.58	6.32
Kunlarim	1.14	4.74	4.88	-1.57	2.84	2.19
Help me	5.12	12.01	6.38	2.19	8.47	3.94
Only	1.35	7.40	3.31	0.12	6.45	2.15
Resistencia	0.35	8.95	1.52	-1.13	7.04	0.36
Americano	4.24	15.55	4.69	4.85	14.73	5.46
Monkey	4.95	13.00	5.90	1.66	11.40	2.45
Roads	2.97	18.77	3.15	2.22	15.64	2.53
Tanto (ex.1)	7.94	18.09	8.44	6.00	11.8	7.58
Tanto(ex. 2)	5.72	17.20	6.12	4.60	14.38	5.24
Don't Know	-2.83	9.10	-2.04	-6.08	2.98	-3.74

Based on this dataset, SEMANTICS provides generally better results, especially with respect to SIR. Compared to [11], where a less demanding dataset was used, SEMANTICS is able to target mixtures that resemble more appropriately the "real-world" conditions. It should be stressed that SDR and SIR are much more important to the objective of this study, which is that of audio separation [20]. The original mixtures, the description of the mixing processing, as well as the resulting audio files from separation are available in **http://tinyurl.com/y9tte98**

5. CONCLUSION

In this paper we have presented SEMANTICS, an extension of our previous method presented in [11]. Our new method does not rely on ADRess and makes more efficient use of both channels of the stereo mix. The dataset that was created simulates "real world" conditions, as the music sources as well as the voice are processed with typical methods such as reverberation, compression, and equalization both during the mixing and the mastering stage. The results indicate significant improvement over our previous method, especially in the area of SIR, and which already improved on ADRess results. Future work will address the issue of artefacts that are introduced during NIC subtraction and Amplitude Discrimination.

6. REFERENCES

[1] A. S. Bregman, *Auditory scene analysis : the perceptual organization of sound*. Cambridge, Mass. ; London: MIT Press, 1990.

[2] S. Yang, *et al.*, "A computational auditory scene analysis system for speech segregation and robust speech recognition," *Comput. Speech Lang.,* vol. 24, pp. 77-93.

[3] D. Byrd and T. Crawford, "Problems of music information retrieval in the real world," *Information Processing & Management,* vol. 38, pp. 249-272, 2002.

[4] J.-J. Aucouturier and F. Pachet, "Improving Timbre Similarity: How high's the sky?," *Journal on Negative Results in Speech Research,* 2004.

[5] J. L. Durrieu, *et al.*, "Singer melody extraction in polyphonic signals using source separation methods," in *ICASSP 2008*, 2008, pp. 169-172.

[6] A. Mesaros, *et al.*, "Singer Identification in Polyphonic Music using vocal separation and pattern recognition methods," presented at the ISMIR, Vienna, Austria, 2007.

[7] Y. Li and D. Wang, "Singing Voice Separation from Monaural Recordings," presented at the ISMIR, Victoria, Canada, 2006.

[8] A. Ozerov, *et al.*, "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.

[9] D. Barry and B. Lawlor, "Sound Source Separation: Azimuth Discrimination and Resynthesis," presented at the DAFx, Naples, Italy, 2004.

[10] 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, pp. 55-58.

[11] S. Sofianos, *et al.*, "Towards Effective Singing Voice Separation from Stereophonic Recordings," in *ICASSP 2010*, Dallas, Texas.

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

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

[14] M. S. Pedersen, *et al.*, "Separating Underdetermined Convolutive Speech Mixtures " in *Independent Component Analysis and Blind Signal Separation*. vol. 3889, ed Berlin/Heidelberg: Springer, 2006.

[15] C. A. Champlin, "Hearing An Introduction to Psychological and Physiological Acoustics (3rd edition)," *Ear and Hearing,* vol. 20, p. 439, 1999.

[16] R. Izhaki, *Mixing Audio: Concepts, Practices and Tools*: Focal, 2007.

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

[18] M. Cobos, *et al.*, "Stereo to Wave-Field Synthesis music up-mixing: An objective and subjective evaluation," in *ISCCSP 2008*, pp. 1279-1284.

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

[20] J. Kornycky, *et al.*, "Comparison of Subjective and Objective Evaluation Methods for Audio Source Separation," *Proceedings of Meetings on Acoustics,* vol. 4, 2008.