POSITION-BASED ATTENUATION AND AMPLIFICATION FOR STEREO MIXES

Luca Marinelli

Audio Communication Group
Technical University of Berlin
Berlin, Germany
luca.marinelli@campus.tu-berlin.de

Holger Kirchhoff

zplane.development GmbH & Co KG
Berlin, Germany
kirchhoff@zplane.de

ABSTRACT

This paper presents a position-based attenuation and amplification method suitable for source separation and enhancement. Our novel sigmoidal time-frequency mask allows us to directly control the level within a target azimuth range and to exploit a trade-off between the production of musical noise artifacts and separation quality. The algorithm is fully describable in a closed and compact analytical form. The method was evaluated on a multitrack dataset and compared to another position-based source separation algorithm. The results show that although the sigmoidal mask leads to a lower source-to-interference ratio, the overall sound quality measured by the source-to-distortion ratio and the source-to-artifacts ratio is improved.

1. INTRODUCTION

Over the past years, research on sound source separation and upmixing techniques has produced a vast body of literature. Nonnegative matrix factorisation (NMF) [1], independent component analysis (ICA) [2], computational auditory scene analysis (CASA) [3] and time-frequency (TF) masking [4] appear to be the main families of blind audio source separation (BASS) methods. With regard to stereo recordings many different approaches have been proposed to model the mixing process and the nature of the sources. The derived techniques can be divided into blind or informed (guided) source separation [5].

This paper proposes a guided TF masking algorithm, assuming that the direction of the source can be approximately estimated by the user. As with other position-based source separation methods [4, 6], only the interaural intensity difference (IID) between the two channels (left and right) is taken into account to model the position of the sources. Our signal model is similar to the ones in [7] and [4] and assumes mono sources that have been positioned in the stereo image by a panorama potentiometer. Each TF bin is assumed to belong to a single source and we estimate its position as well as its mono magnitude assuming the energy-preserving panning law. Given a target azimuth range, we then compute a sigmoidal TF mask that weights the amplitudes with regard to their distance from the target azimuth range. In addition to source separation, our mask is able to perform source enhancement and attenuation with precise level indications. A binary mask as in [4] produces significant musical noise due to isolated non-zero TF bins. The sigmoidal mask has a smoother transition between the target and the adjacent azimuth ranges which reduces this kind of artifact.

In section 2 we briefly introduce our signal model, while in section 3 our method is presented in a closed analytical form. Finally, in section 4, we confirm the effectiveness of our approach.

2. FRAMEWORK

Commercial recordings are often instantaneous mixes of mono tracks combined through amplitude panning to generate a stereophonic effect [8].

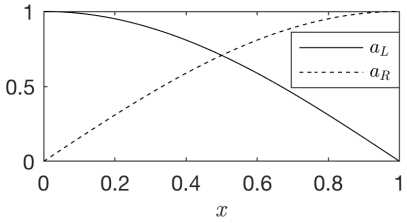

Figure 1: *Energy preserving panning coefficients*

2.1. Mixing Model

Given a set of mono sources $\{S_j\}_{j=1}^{J}$ and the relative amplitude panning gains a_j^L, a_j^R, a stereo mix can be modelled as:

$$\begin{aligned} L &= \textstyle\sum_j a_j^L S_j \\ R &= \textstyle\sum_j a_j^R S_j \end{aligned} \tag{1}$$

where L and R are the left and the right channels, respectively.

As reported in [8], the majority of analog and digital mixers approximate the *energy-preserving panning law* (Fig. 1), where the value of the panorama potentiometer takes on values $x_j \in [0, 1]$ and $(a_j^L)^2 + (a_j^R)^2 = C^2$:

$$\begin{aligned} a_j^L &= C \cdot \cos(x_j \cdot \pi/2) \\ a_j^R &= C \cdot \sin(x_j \cdot \pi/2) \end{aligned} \tag{2}$$

where $C = 1$ satisfies the energy preserving condition.

2.2. W-disjoint orthogonality

Our method is based on the W-disjoint orthogonality assumption, where two or more sources do not overlap in the short-time Fourier transform (STFT) domain. Mathematically, this condition can be expressed as:

$$S_i(k, m) \cdot S_j(k, m) = 0 \quad \forall i \neq j, \ \forall k, m \tag{3}$$

where $S_j(k, m)$ is the STFT of the j-th source at frame m and frequency bin k.

3. METHOD

In a first step, we estimate the panning position of each TF bin (sec. 3.1) as well as its mono magnitude (sec. 3.3). Given a target azimuth range, a sigmoidal mask is computed based on the estimated panning positions (sec. 3.2).

The sigmoidal mask is applied to the mono magnitudes which are then re-panned (sec. 3.4) and recombined with the phase from the original mixture.

3.1. Panning map

Given equation 3 and our assumptions from eq. 1 and 2, it is now possible to estimate the panning position for each element in the spectrograms:

$$x(k,m) = \arctan\left(\frac{|X_R(k,m)|}{|X_L(k,m)|}\right) \cdot 2/\pi \tag{4}$$

where X_L and X_R are the left and the right channel in the STFT domain. A similar estimation of the panning position has been used in [9].

3.2. Sigmoidal mask

The smoothness of sigmoidal functions has been proven useful in the post-processing of signal estimates coming from methods like ICA, CASA or NMF [10, 11, 12]. Those estimates can be then used to compute TF sigmoidal masks that are then applied on the original mixture. In this work we combine two sigmoids with the panning map to create a position-based mask that can control the level in a given azimuth range.

In order to attenuate or amplify the elements inside a target azimuth range, it is necessary to find a function that weights TF bins based on their estimated position. The target range is defined by its center position $T \in [0,1]$ and a width R. We define two complementary sigmoid functions that control the amount of attenuation and amplification both inside and outside the target azimuth range:

$$\begin{aligned}\sigma_L(x) &= \frac{1}{1+e^{-\beta(x-T+\frac{R}{2})}} \\ \sigma_R(x) &= \frac{1}{1+e^{+\beta(x-T-\frac{R}{2})}}\end{aligned} \tag{5}$$

In these equations, β defines the slope of the sigmoids. Choosing $\beta = \infty$ is equivalent to a binary mask as in [4], whereas lower values for β result in smoother transitions. In order to amplify the target azimuth range, we choose $\beta > 0$ and combine the two sigmoids as follows to get the sigmoidal mask:

$$M(x) = \min\left(\sigma_L(x), \sigma_R(x)\right) \tag{6}$$

For attenuation, we choose $\beta < 0$ and obtain the sigmoidal mask:

$$M(x) = \max\left(\sigma_L(x), \sigma_R(x)\right) \tag{7}$$

Finally, to control the level in decibels, one can simply rearrange one of the previous equations as follows:

$$M_{dB}(x) = 10^{(\alpha \cdot M(x) - \alpha)/20} \tag{8}$$

where $\alpha > 0$ is the desired attenuation in decibels (see Fig. 2).

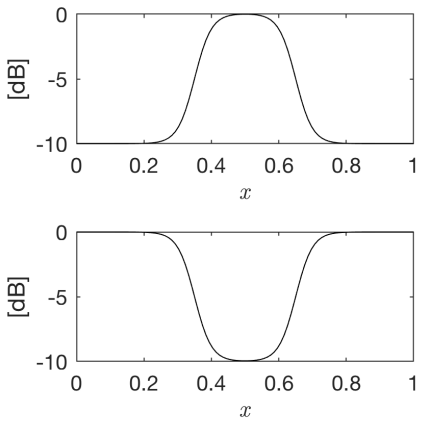

Figure 2: *Sigmoidal masks as in eq. 6 (upper) and eq. 7 (lower).* $\alpha = 10\ dB\ ,\ T = 0.5\ ,\ R = 0.3\ ,\ \beta = \pm 40$

3.3. Pre-panning magnitudes

The assumptions made in eqs. 1, 2 and 3 pose the *ideal* conditions to recover the mono magnitude of each source. Generally, the mono magnitude in the STFT domain can be computed as:

$$|S(k,m)| = \sqrt{|X_L(k,m)|^2 + |X_R(k,m)|^2} \tag{9}$$

3.4. Masking and re-panning

To synthesize the modified signal, the mono magnitudes are masked

$$|S_{out}(k,m)| = |S(k,m)| \cdot M_{dB}(x(k,m)), \tag{10}$$

and each component is re-panned to its original position.

$$\begin{aligned}|Y_L(k,m)| &= |S_{out}(k,m)| \cdot \cos(x(k,m) \cdot \pi/2) \\ |Y_R(k,m)| &= |S_{out}(k,m)| \cdot \sin(x(k,m) \cdot \pi/2)\end{aligned} \tag{11}$$

Finally, the phase from the original mixture has to be recombined:

$$\begin{aligned}Y_L(k,m) &= |Y_L(k,m)| \cdot e^{j \cdot \angle X_L(k,m)} \\ Y_R(k,m) &= |Y_R(k,m)| \cdot e^{j \cdot \angle X_R(k,m)}\end{aligned} \tag{12}$$

4. EVALUATION

4.1. Procedure

To evaluate our proposed method we use MedleyDB [13] a database of 122 royalty free multitrack recordings with a total length of 7:17 hours. The dataset provides stems (i.e. processed individual instrument tracks) for each song. For our purpose we eliminated tracks that were recorded in a live setting, due to their significant amount of spill between sources. We evaluated all tracks with a number

of stems greater than or equal to three and less than or equal to six, resulting in 43 tracks for a total of 199 sources. Due to the lack of metadata about the sources' spatial position, we created separate mixtures by downmixing the stems from stereo to mono and remixing them with random azimuth positions. The azimuth values were chosen from a uniform distribution over the whole azimuth range.

As a baseline, we compare our position-based sigmoidal source separation (PoSiS) method against the ADRess algorithm [4] which uses a different method for azimuth estimation and a binary mask instead of our proposed sigmoidal mask. We use an implementation written for the Csound system by Victor Lazzarini [14]. The ADRess algorithm was parameterized with 600 equally spaced azimuth positions and a target azimuth range of 60 azimuth positions for each source. To make our algorithm comparable, our mask was set as in equation 6 with $\beta = 30$, $R = 0.1$ and without rearranging it as in 8 to effectively emulate an attenuation of $-\infty$ dB for the TF bins outside the target range. For both algorithms, we opted for a 4096 points Hann window with 50% overlap.

To measure the quality of the separations, we used the MATLAB toolbox BSS_EVAL [15] distributed under GNU Public Licence. The computation of the criteria is performed in two steps. First, the estimated source signal is decomposed as:

$$\hat{s} = s_{\text{target}} + e_{\text{interf}} + e_{\text{noise}} + e_{\text{artif}} \tag{13}$$

where s_{target} is a modified version of the source through an allowed distortion (in this case a time invariant filter, with a 512 samples delay) and where e_{interf}, e_{noise} and e_{artif} are respectively the interference, noise and artifacts errors. From these terms, assuming no noise in our model, three numerical performance criteria are computed:

- the source-to-distortion ratio (SDR)
  that can be seen as a global quality assessment
- the source-to-artifacts ratio (SAR)
  in our case mainly related to musical noise
- the source-to-interference ratio (SIR)
  that measures the interference from other sources

4.2. Results

Figure 3 displays box plots of the measurements grouped by the number of sources present in the track.

In general, all quality measures for both algorithms show a decreasing trend when the number of sources increases, which can be attributed to the increased complexity and TF overlap of the sources when more sources are present. It can be observed that ADRess yields higher source-to-interference ratios than our proposed method, particularly when the number of sources increases. The sigmoidal mask provides a smoother transition between azimuth values inside and outside the target range and hence leads to a higher amount of contributions from other sources. On the other hand, however, PoSiS generally yields higher source-to-distortion and source-to-artifacts ratios which both capture the overall sound quality of the separated source signals. Artifacts — mainly musical noise — are reduced by the sigmoidal mask because it leads to less isolated TF bins in comparison with ADRess' binary mask.

The results suggest that the sigmoidal mask trades separation accuracy against artifacts, which can be controlled by the slope of the sigmoidal mask. With higher slopes, the mask approaches the binary mask, resulting in more artifacts and a better separation accuracy, whereas sigmoidal masks with lower slopes reduce musical noise artifacts but lead to more interference from other sources.

Δ_{SDR}	Δ_{SIR}	Δ_{SAR}
$\mu \simeq 3.3$ dB	$\mu \simeq 0.6$ dB	$\mu \simeq 3.0$ dB
$p \simeq 0.00$	$p \simeq 0.12$	$p \simeq 0.00$

Table 1: Paired difference t-test: μ is the average and p the p-value.

Assuming an underlying normal distribution of the source-wise differences of the performance measurements, where:

$$\begin{aligned} \Delta_{SDR} &= SDR_{PoSiS} - SDR_{ADRess} \\ \Delta_{SIR} &= SIR_{PoSiS} - SIR_{ADRess} \\ \Delta_{SAR} &= SAR_{PoSiS} - SAR_{ADRess} \end{aligned} \tag{14}$$

We then checked the statistical significance of our results by performing a paired t-test. With the resulting p-values in Table 1 we can, for the SDR and SAR, safely reject the null-hypothesis, while there is no statistically significant difference between the two methods in the SIR measurements.

5. CONCLUSION

We presented a system for position-based source separation from a stereo mixture. The algorithm first estimates a panning position and mono magnitude for each TF bin based on the energy-preserving panning law, assuming W-disjoint orthogonality. Given a target azimuth range, a sigmoidal mask is computed that enables attenuation and amplification of the audio within the target range. The attenuation/amplification level can be specified in dB. The mask is applied to the estimated mono magnitudes of each TF bin and the bins are re-panned to their estimated azimuth position. A resynthesis combining the magnitudes with the mixture phases yields the separated source signal.

We could confirm that using a sigmoidal mask, that is, a smoother transition between the target azimuth range and adjacent azimuth ranges, significantly reduces musical noise artifacts that occur in position-based algorithms that rely on binary masking. Binary masking often leads to isolated TF bins which cause perceptually disturbing musical noise. The sigmoidal mask smoothes the spectrogram of the separated source thereby trading musical noise artifacts against separation accuracy.

For certain use cases such as amplifying an instrument for the purpose of transcribing its musical performance, it is often not necessary to have a sharp separation and a complete suppression of interfering sources, but rather to provide a limited amplification that allows users to better listen to what has been played by the performer. In these cases an improved overall sound quality with less artifacts might be preferred.

Future work on position-based source separation will have to consider methods that do not assume W-disjoint orthogonality, which does not hold in general for professionally produced music mixtures. Even though it is possible to isolate sources under this assumption, a significant improvement in separation accuracy *and* sound quality will only be achieved if the TF contributions of each individual source can be estimated and reassigned to the corresponding source. Therefore monaural source separation methods

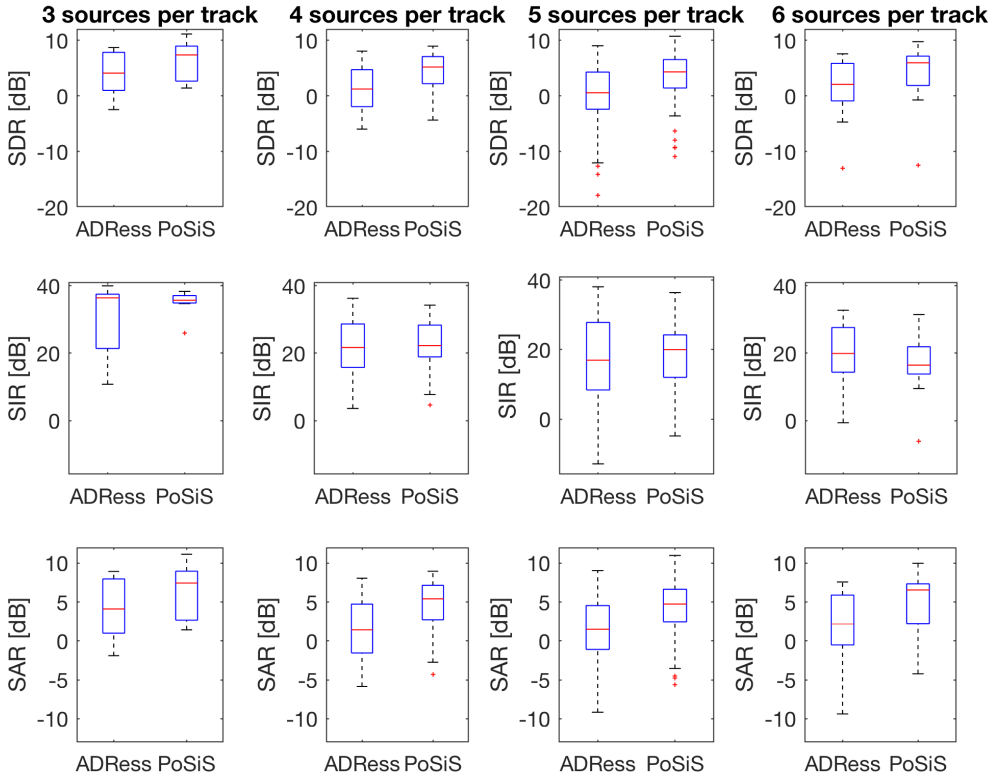

Figure 3: *SDR, SIR, SAR of ADRess and PoSiS grouped by number of sources in the mixture*

will have to be combined with position-based algorithms in order to improve sound source separation from stereo mixtures.

6. REFERENCES

[1] P. Smaragdis and J.C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, Oct. 19-22, 2003.

[2] J.-F. Cardoso, "Blind source separation: statistical principles," in *Proceedings of the IEEE, vol. 9, no. 10*, Oct. 1998, pp. 2009–2025.

[3] D. Ellis, *Prediction-driven computational auditory scene analysis*, Ph.D. thesis, MIT, June 1996.

[4] E. Coyle D. Barry, B. Lawlor, "Sound source separation: Azimuth discrimination and resynthesis," in *7th Conference on Digital Audio Effects (DAFX 04)*, Neaples, IT, Oct. 5-8, 2004.

[5] Emmanuel Vincent, Nancy Bertin, Rémi Gribonval, and Frédéric Bimbot, "From blind to guided audio source separation: How models and side information can improve the separation of sound," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 107–115, May 2014.

[6] Maximo Cobos and J.Lopez, "Stereo audio source separation based on time-frequency masking and multilevel thresholding," *Digital Signal Processing*, vol. 18, no. 6, pp. 960 – 976, 2008.

[7] C. Avendano, "Frequency-domain source identification and manipulation in stereo mixes for enhancement, suppression and re-panning applications," in *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct 2003, pp. 55–58.

[8] J. Bonada. A. Loscos M. Vinyes, "Demixing commercial music productions via human-assisted time-frequency masking," in *120th AES Convention*, Paris, FR, May 20-23, 2006.

[9] Y. Mitsufuji and A. Roebel, "Sound source separation based on non-negative tensor factorization incorporating spatial cue as prior knowledge," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 71–75.

[10] Toby Stokes, Christopher Hummersone, Tim Brookes, and Andrew Mason, "Perceptual quality of audio separated using sigmoidal masks," in *137th Audio Engineering Society Convention 2014*, Oct. 2014.

[11] Dorothea Kolossa, Ramon Fernandez Astudillo, Eugen Hoffmann, and Reinhold Orglmeister, "Independent component analysis and time-frequency masking for speech recognition

in multitalker conditions," *EURASIP J. Audio Speech Music Process.*, vol. 2010, no. 1, Dec. 2010.

[12] Dorothea Kolossa and Reinhold Haeb-Umbach, *Robust Speech Recognition of Uncertain or Missing Data: Theory and Applications*, Springer Publishing Company, Incorporated, 1st edition, 2011.

[13] R. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam and J. P. Bello, "MedleyDB: A Multitrack Dataset for Annotation-Intensive MIR Research," in *15th International Society for Music Information Retrieval Conference*, Taipei, Taiwan, Oct. 2014.

[14] Victor Lazzarini, "pvsdemix - spectral azimuth-based de-mixing of stereo sources," Available at http://www.csounds.com/manualOLPC/pvsdemix.html, accessed March 19, 2018.

[15] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.