FREQUENCY-DOMAIN SOURCE IDENTIFICATION AND MANIPULATION IN STEREO MIXES FOR ENHANCEMENT, SUPPRESSION AND RE-PANNING APPLICATIONS

Carlos Avendano

Creative Advanced Technology Center
1500 Green Hills Road
Scotts Valley, CA 95065, USA
carlosa@atc.creative.com

ABSTRACT

In this paper we describe a frequency-domain framework for source identification, separation and manipulation in stereo music recordings. Based on a simplified model of the stereo mix, we describe how a similarity measure between the Short-Time Fourier Transforms (STFT) of the input signals is used to identify time-frequency regions occupied by each source based on the panning coefficient assigned to it during the mix. Individual sources are identified and manipulated by clustering the time-frequency components with a given panning coefficient. After modification, an inverse STFT is used to synthesize a time-domain processed signal. We describe applications of the technique in source suppression, enhancement and re-panning.

1. INTRODUCTION

There have been recent developments in the area of frequency-domain processing for multi-channel audio processing and compression [1, 2, 3]. These methods are based on analyzing the multi-channel input signal into a time-frequency transform and use cross-channel metrics to derive a number of parameters useful in identifying individual sources or other components in the mix. One basic assumption made by these methods is that in the time-frequency transform domain, signal components corresponding to different sources do not overlap significantly. This non-overlapping requirement, called by some authors W-disjoint orthogonallity [4], is hardly met in real audio material. However, some studies have shown that with a limited number of speech sources, the condition is closely met [3]. In practice, the non-overlapping nature of the sources in the audio signal will introduce an error in the parameter estimates that assume no overlap. The effect of this error will be different depending on the type of parameter, and on the particular application. For example, in blind source separation of five linearly-mixed speech sources, only 14 dB SNR improvement was achieved in [4]. Another interesting property of these methods is that they can be applied when the number of observations (at least two) is smaller than the number of sources. For the applications described in this paper we are interested in stereo input signals with any number of sources in the mix.

We first describe a cross-channel metric, known as the *panning index* [1], that identifies the different sources based on their panning coefficients in the mix. The metric is shown to be robust and its estimation error increases predictably when the amount of overlap increases. Given the behavior of the panning index error, we then propose an adaptive mapping or window function to separate and/or manipulate the individual sources in the mix. Finally we show applications of this technique to several problems such as source suppression, enhancement and re-panning.

2. FRAMEWORK

We start by presenting a simplified model of the stereo signal. Stereo recordings can be roughly categorized into two main classes: *studio* or artificial, and *live* or natural [5]. In this paper we focus on the *studio* recording, where the different sources are individually recorded and then mixed into a single stereo signal by amplitude panning (we discuss delay-panned mixes in the last section). Stereo reverberation is then added artificially to the mix. In general, the left and right impulse responses of the reverberation processor have equal direct paths and different tails, to increase the spaciousness of the stereo presentation.

A model for this signal is as follows: assume that there are N amplitude-panned sources $s_j(t), j = 1, ..., N$ convolved with reverberation impulse responses $r_i(t)$ to generate the left ($i = 1$) and right ($i = 2$) stereo channels respectively. The stereo signal can be written as:

$$x_i(t) = \left[\sum_{j=1}^{N} \alpha_{ij} s_j(t)\right] * r_i(t), \tag{1}$$

where α_{ij} are amplitude-panning coefficients. For amplitude-panned sources we assume the sinusoidal energy-preserving panning law where $\alpha_{2j} = \sqrt{1 - \alpha_{1j}^2}$.

2.1. Panning Index

The source identification technique described here has been applied in the context of multi-channel upmix [1]. The basic idea is to compare the left and right signals in the time-frequency plane to derive a two-dimensional map that identifies the different source components based on the panning gains assigned to them during the mix. For instance, if we are looking for a source panned to the center, we select time-frequency bins in the map whose values correspond to $\alpha_{1j} = \alpha_{2j} = \frac{1}{\sqrt{2}}$. Once identified, these components can be modified (e.g. attenuated) or separated to create a new signal.

To formalize let us first denote the STFT's of the channel signals $x_i(t)$ as $X_i(m, k)$, where m is the time index and k is the frequency index, and $i = 1, 2$. We define the following similarity measure:

0-7803-7850-4/03/$17.00 © 2003 IEEE

$$\psi(m,k) = 2\frac{|X_1(m,k)X_2^*(m,k)|}{|X_1(m,k)|^2 + |X_2(m,k)|^2}, \qquad (2)$$

where $*$ denotes complex conjugation. The properties of this function are very useful to our purposes as shown next. If we assume that only one amplitude-panned source $s_j(t)$ is present in the mix (assuming no reverberation), from the signal model in (1) we can write the left and right signals as $x_1(t) = \sqrt{1-\alpha^2}s_j(t)$ and $x_2(t) = \alpha s_j(t)$ respectively. The similarity function (2) will have a value proportional to the panning coefficient α in those time-frequency regions where the source has energy (in [1] the non energy-preserving panning law was used), i.e.

$$\psi(m,k) = 2\alpha\sqrt{1-\alpha^2}.$$

If the source is panned to the center (i.e. $\alpha = 0.7071$), then the function will attain its maximum value of one, and if the source is panned completely to either side, the function will attain its minimum value of zero. In other words the function is bounded (unlike other metrics in [4, 2, 7]). Notice, however, that given the quadratic dependence on α, the function (2) is multi-valued and there exists ambiguity with regards to the lateral direction of the source. The ambiguity can easily be resolved by using the following partial similarity measures:

$$\psi_i(m,k) = \frac{|X_i(m,k)X_j^*(m,k)|}{|X_i(m,k)|^2}, \quad i \neq j, \qquad (3)$$

and their difference

$$\Delta(m,k) = \psi_1(m,k) - \psi_2(m,k), \qquad (4)$$

where we notice that time-frequency regions with positive values of $\Delta(m,k)$ correspond to signals panned towards the left, and negative values correspond to signals panned towards the right. A value of $\Delta(m,k)$ equal to zero corresponds to non-overlapping time-frequency regions of signals panned to the center. Thus we can define an ambiguity-resolving function as

$$\widehat{\Delta}(m,k) = \begin{cases} 1 & \text{if} \quad \Delta(m,k) > 0 \\ 0 & \text{if} \quad \Delta(m,k) = 0 \\ -1 & \text{if} \quad \Delta(m,k) < 0 \end{cases} \qquad (5)$$

Shifting and multiplying the similarity function by $\widehat{\Delta}(m,k)$ we obtain the *panning index* $\Psi(m,k)$ as,

$$\Psi(m,k) = [1 - \psi(m,k)]\,\widehat{\Delta}(m,k), \qquad (6)$$

which is bounded but whose values now vary from minus one to one as a function of the panning coefficient as shown in Figure 1.

Notice that the panning index will uniquely identify the time-frequency components of the sources in the stereo mix only when they are all panned to different locations and do not overlap significantly in the transform domain. This is unfortunately rarely the case, so many times there will be an estimation error. However, given the properties of (6), the estimation error is bounded. The upper bound is attained when a source panned to either side overlaps in the time-frequency plane with an overlapping source panned to the opposite side. To see this, assume that in (1) source $s_1(t)$ is the desired source and there is no reverberation. Given the linearity of the STFT we can write this in the frequency domain as:

$$X_i(m,k) = \alpha_{i1}S_1(m,k) + \sum_{j=2}^{N} \alpha_{ij}S_j(m,k) \qquad (7)$$

where $S_1(m,k)$ is the STFT of $s_1(t)$ and $S_j(m,k)$ are the STFTs of the overlapping signals $s_j(t)$. At any given point (m_0, k_0) in the time-frequency plane we can assume, without loss of generality, that the contribution of the overlapping signals can be reduced to a single term corresponding to an equivalent interfering component. Using this model we can compute the estimation error for a source component with panning index Ψ_0 and magnitude g_0, and an interference component with panning index Ψ_e and magnitude g_e. This is illustrated in Figure 2, where the error is shown as a parametric function of signal-to-interference ratio (SIR = g_0^2/g_e^2) and Ψ_e for three different values of Ψ_0.

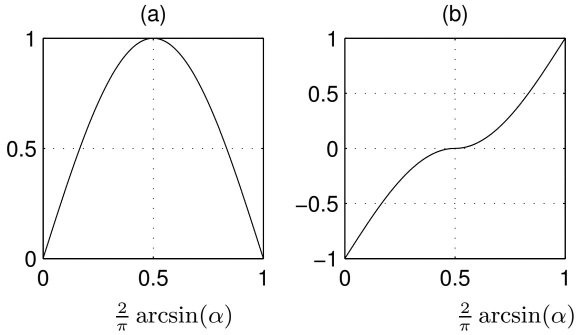

Figure 1: *(a) Similarity and (b) panning index. The absissa has been warped according to the energy-preserving panning law to illustrate the symmetry in this domain.*

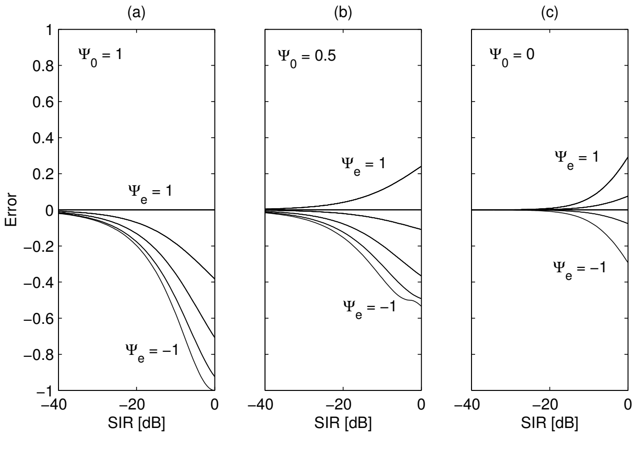

Figure 2: *Panning Index error as a function of SIR and panning index of the interfering component Ψ_e.*

For a fixed SIR, the maximum error is introduced by the source panned to the most distant location. For instance, when the source is panned to the right (i.e. $\Psi_0 = 1$) the largest error will be caused by sources panned to the left (i.e. $\Psi_e = 1$). In general, as the magnitude of the interference increases and exceeds the magnitude of the source, the range of the error will increase asymptotically to a maximum equal to $E_{max} = \Psi_0 - \Psi_e$. In Figure 2 we also notice that for a given SIR, the magnitude of the error decreases as $\|\Psi_0\|$ decreases. The error with the smallest magnitude corresponds to sources panned to the center.

2.2. Panning Index Window

For the applications in this paper we are interested in identifying, selecting and processing a source or sources panned to a particular direction. For this we need to select time-frequency bins with panning indices equal to Ψ_0. Due to the overlap with other sources, selecting only these bins will exclude bins where the source might still have significant energy but whose panning index has been altered by the presence of the interference. Thus, selecting bins in a window around Ψ_0 will help to reduce distortion (at the price of increased interference). Using the properties of the error we can design a panning index window as follows.

The desired behavior of the window function is to let components with values equal to Ψ_0 pass unmodified, weigh and pass components with panning indices near Ψ_0, and reject the rest. In this paper we propose a symmetrical tapering window function centered around Ψ_0. The width of the window will determine the trade-off between distortion and interference, and will vary in width depending on Ψ_0 and the maximum level of interference allowed. A useful function for these purposes is a Gaussian window function, i.e.:

$$\Theta(m,k) = \nu + (1-\nu)e^{-\frac{1}{2\xi}(\Psi(m,k)-\Psi_0)^2} \quad (8)$$

where Ψ_0 is the desired panning index value, ξ controls the width of the window, and ν is a floor value necessary to avoid setting STFT values to zero, which might result in musical-noise artifacts. Since the Gaussian function reaches zero asymptotically, the value of ξ is obtained by assuming that the window will effectively reject values beyond a certain point Ψ_c where the function reaches a small value of A (e.g. $A_{\mathrm{dB}} = -60$ dB). The rejection point Ψ_c is calculated as the maximum panning index error introduced by the interference for a given SIR value. From (8) with $\nu = 0$ we compute the value of ξ as

$$\xi = -\frac{(\Psi_c - \Psi_0)^2}{2\log A}. \quad (9)$$

3. APPLICATIONS

We have previously shown applications of the panning index to multi-channel audio upmix (e.g. center-channel synthesis [1]). In this paper we focus on applications to stereo recordings. The idea is to use the panning index to identify and manipulate the signals in the STFT domain by computing and applying a time-frequency mask to modify the STFT magnitude, and reconstructing a processed signal using a least-squares optimal reconstruction STFT synthesis [6].

3.1. Source Suppression

Techniques capable of removing the lead vocals or lead instruments from a commercial musical recording have been of interest to Karaoke enthusiasts, music students and professional musicians. While play-along and sing-along recordings are widely available for these purposes, their selection and quality cannot always meet the demands of singers and instrumentalists. Thus, an automatic (or semi-automatic) method to remove lead vocals and instruments is highly appealing to these users.

The problem of lead vocal elimination is extremely difficult since most of the times there is no a priori knowledge about how the different instruments and vocals were recorded and mixed into a stereo signal. A well-known vocal elimination technique is the left-right (L-R) technique that subtracts the left minus the right channels assuming a simplified model of the mix in which the lead source is panned in amplitude (and phase) to the center. While this assumption is valid for the vast majority of popular music recordings, the lead vocal or instrument is not always the only source panned to the center, thus the L-R technique will remove these other center-panned sources as well. Another problem with this technique is that the resulting signal is monaural. Some techniques try to overcome this limitation by applying pseudo-stereophony processing to the resulting monaural signal. A refinement of the L-R approach is to perform the signal subtraction only in a frequency band in the range of interest (e.g. for vocals roughly 100 Hz to 8 kHz). The resulting signal of this partial L-R is stereo outside the elimination band. However, the soundstage image in this frequency band will be compromised. Other refinements include the suppression of sources when they are off-center, for example by doing a weighed L-R subtraction.

The technique proposed in this paper is similar to a frequency-domain vocal suppression method proposed in [7], which overcomes many of the limitations of the L-R approach. The technique uses the ratio of the left and right STFT's to identify components that are panned to the center (i.e. ratio values near unity) and applies magnitude modification to suppress these components. In our case we use the panning index (6) and the window (8) to identify and suppress the source. The idea is to multiply the input STFT's by (8) and subtract the result from the input signals to obtain a new STFT as:

$$Y_i(m,k) = X_i(m,k)[1 - \Theta(m,k)], \;\; \forall k \in K,$$

where K is the frequency range of interest [1], and we finally apply an inverse STFT to the new transform $Y_i(m,k)$ to obtain the time domain stereo signal where the center-panned components in the range K have been suppressed. The width of the window is adjusted according to the desired trade-off between distortion and suppression.

3.1.1. Simulation

To validate the performance of the vocal suppression method we have performed a series of simulations where known vocal signals were panned to the center of instrumental play-along stereo tracks. The performance of the method varied depending on the material used. A typical result is shown in Figure 3, where waveforms and spectrograms of the original and processed versions of the instrumental track, the vocal signal and the stereo mix are shown. In this example, a window width of $\xi = 0.006$ corresponding to an SIR of -60 dB gave the best results in terms of the overall Itakura-Saito (IS) distance [2] between the original instrumental track and the processed stereo mix with the vocal suppressed (IS$= 0.24$ in this case). Notice that the center-panned components of the instrumental track have been suppressed along with the vocal.

[1] Notice that this is equivalent to a brick wall filter. In practice, a less aggressive filter is used to reduce time-domain aliasing.

[2] This metric was chosen due to its perceptual relevance [8]. Values of this measure larger than approximately 0.1 indicate audible distortion.

3.2. Source Enhancement

In some applications one might be interested in accentuating the lead instrument or vocal. Using the panning index in (6) it is straightforward to design a source enhancement algorithm. To enhance the source we simply multiply the input STFT's by (8), apply a gain and add to the input signal STFTs as:

$$Y_i(m,k) = X(m,k)[1+\beta\Theta(m,k)],\ \forall k \in K,$$

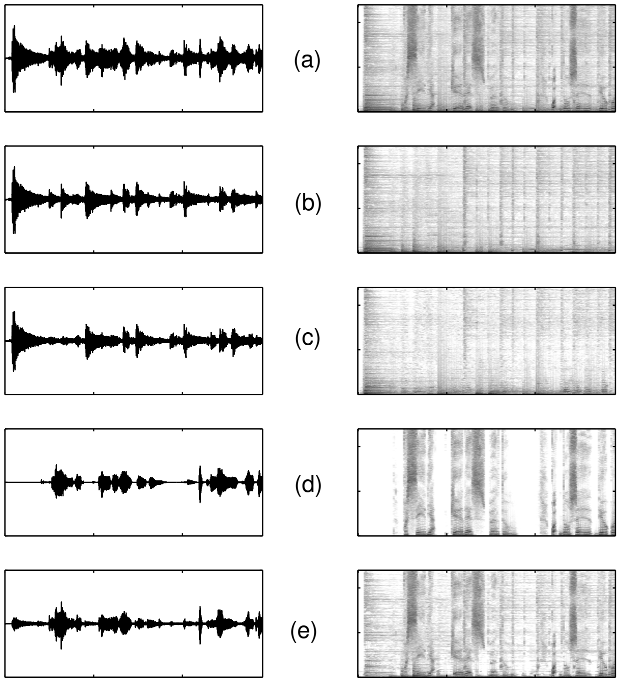

Figure 3: *Vocal suppression example: waveforms and spectrograms of (a) left channel of mixture, (b) original left channel, (c) processed left channel, (d) original vocal, and (e) extracted vocal. The absissa is time (*0-5 *s) and the ordinate is amplitude for the waveforms (*-1-1*) and frequency for the spectrograms (*0-12 *kHz).*

3.3. Source Re-Panning

Another interesting application is to modify the direction of a source in the stereo image (panorama). To do this we identify the source using the panning index and multiply the input STFT components by $\Theta(m,k)$ and a gain factor that is the ratio of the actual panning gains and the desired panning gains, i.e.:

$$Y_i(m,k) = X_i(m,k)\left[1+\Theta(m,k)(\rho_i - 1)\right],\ \forall k \in K,$$

where ρ_i is a gain factor calculated as:

$$\rho_i = \frac{\gamma_i}{\alpha_i}$$

where γ_i is the desired panning gain. Notice that the re-panning gain ρ can be made dependent on frequency. This is useful when repanning wideband sources, where the apparent direction at low and high frequencies will deviate according to the panning law [9].

The results with this algorithm vary depending on the music material used. In an informal listening test, where stereo mixtures with two musical instruments and a vocalist were artificially generated, the vocal signal was re-panned to multiple directions. In all cases the source was identified as being in the correct direction. However, depending on the amount of overlap with other sources, the re-panned source suffered some amount of spatial smearing.

4. DISCUSSION

So far we have illustrated some of the capabilities of the panning index approach. It is worth noting that since we are dealing with commercial music recordings, there is little information about the mixing process and the results will vary according to how much the actual signal deviates from the simplified model considered in this study. One area of future research is to derive a robust metric for the case of *live* recordings, where non-coincident microphone techniques will result in a stereo mix panned in delay. While some techniques have been proposed to deal with this problem in simplified scenarios, such as mixtures of a few speech signals or communication signals [4], it seems that a more robust metric is needed to handle the case of musical recordings.

5. REFERENCES

[1] C. Avendano and J.M. Jot, "Frequency-Domain Techniques for Stereo to Multichannel Upmix." In Proc. AES 22nd International Conference on Virtual, Synthetic and Entertainment Audio, pp. 121-130, Espoo, Finland 2002.

[2] C. Faller and F. Baumgarte, "Binaural Cue Coding: A Novel and Efficient Representation of Spatial Audio." In Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'02, Vol. 2, pp. 1841-1844, Orlando, Florida, May 2002.

[3] A. Radke and S. Richard, "Audio Interpolation for Virtual Audio Synthesis" In Proc. AES 22nd International Conference on Virtual, Synthetic and Entertainment Audio, pp. 51-57, Espoo, Finland 2002.

[4] A. Jourjine, S. Richard, and O. Yilmaz, "Blind Separation of Disjoint Orthogonal Signals: Demixing N Sources from 2 Mixures." In Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'00, Vol. 5, pp. 2985-2988, Turkey, April 2000.

[5] F. Rumsey, "Controlled Subjective Assessment of Two-to-Five Channel Surround Processing Algorithms." *Journal of the Audio Engineering Society*, Vol. 47, No. 7/8, pp. 563-582, 1999.

[6] D. W. Griffin and J.S. Lim, "Signal Estimation from Modified Short-Time Fourier Transform." IEEE TASSP, Vol. 32, No. 2, pp. 236-243, April 1984.

[7] J. Laroche, "Process for Removing Voice from Stereo Recordings." US Patent US6405163.

[8] J.R. Deller, J.H.L. Hansen and J.G. Proakis, "Discrete-Time Processing of Speech Signals," IEEE Press, New York, 1993.

[9] J.M. Jot, V. Larcher and J.M. Pernaux, "A Comparative Study of 3-D Audio Encoding and Rendering Techniques." *AES 16th International Conference on Spatial Sound Reproduction*, Rovaniemi, Finland 1999.