ESTIMATION OF THE ENERGY RATIO BETWEEN PRIMARY AND AMBIENCE COMPONENTS IN STEREO AUDIO DATA

Aki Härmä

Philips Research
High Tech Campus 36, 5656AE, Eindhoven, The Netherlands
email: aki.harma@philips.com
web: www.research.philips.com

ABSTRACT

Stereo audio signal is often modeled as a mixture of instantaneously mixed primary components and uncorrelated ambience components. This paper focuses on the estimation of the primary-to-ambience energy ratio, PAR. This measure is useful for signal decomposition in stereo and multichannel audio coding, format conversion, and spatial audio enhancement. The conventional approaches for the estimation of the ratio are based on the ratio of eigenvalues which requires equal energies of the ambience signals. This often leads to an inaccurate estimate of PAR. An alternative measure is proposed which reduces those estimation errors but requires a priori information about the primary component signal. The performance of the method is demonstrated with synthetic signals and a large collection of stereo audio data.

1. INTRODUCTION

Stereo audio content such as music, movies, or radio or TV program material typically consist of a time-varying mixture of signals from multiple sound sources. In various audio reproduction applications it is beneficial to be able to modify the content such that the desired spatial stereo image is preserved in a different reproduction system. For example, in *upmixing* from stereo audio to five-channel surround the goal is to route the sources from the center of the stereo image to the center channels of the surround setup while some other spectrum parts of the mixture are routed to the surround channels. The logic in the modification is based on some assumed *signal model* and the estimation of the parameters and component signals of that model.

The modification of stereo audio data typically consists of two steps: decomposition and remixing. The decomposition part can usually be further divided into estimation and filtering. The topic of this paper is the estimation of signal parameters in the primary-ambience (PA) model (2) which is one of the most common signal models for stereo audio manipulation [3, 1, 2, 7, 8, 4]. In this paper we study computationally efficient methods to estimate the energy ratio between primary and ambient components in individual time-frequency regions of the input stereo signals.

The properties of the PA signal model are discussed in the following sections and it is shown how the signal characteristics are influence the statistical metrics such as correlation coefficients. It is noted that many metrics introduced for audio applications based on the PA model use the ratio of eigenvalues of the cross-correlation matrix. It is demonstrated that this gives a biased estimate for the energy ratio. An alternative measure is introduced and it is shown that it gives more accurate results especially in the cases where the energy of the ambience signal is not equal in the two input channels. However, the alternative method requires *a priori* information about the panning direction of the primary component signal.

2. SIGNAL MODEL

A stereo audio signal can be represented by the following signal model

$$\begin{aligned} X_1(n) &= aP(n)+U(n) \\ X_2(n) &= bP(n)+V(n). \end{aligned} \qquad (1)$$

where n is an index of short-time Fourier transform coefficients in a small time-frequency region. It is assumed that $P(n)$, $U(n)$, and $V(n)$ are mutually uncorrelated and the model is applied separately to small time-frequency regions. The source signals can be seen *sparse* such that only one source at the time dominates a small time-frequency region in the signal [6]. Therefore, the model applies also to a mixture of multiple simultaneous primary sources. The *panning coefficients* of the primary component of the model are defined so that $a^2+b^2=1$. The model has been used in different applications and it is often called primary/ambience, PA signal model, or non-diffuse/diffuse model [7]. The model contains three independent component signals in a two-channel mixture, which makes the blind source separation problem ill-posed [9]. The same applies to the problem of the estimation of the energy ratios between signals and therefore some regularization of the problem, or additional information about the signal is needed.

The model can be used for many typical signal configurations. For example, the case of a filtered source signal $X_1(n) = A(n)P(n)+U(n)$ can also be represented by the model. For example, we may write $X_1(n) = (a+\tilde{A}(n))P(n)+U(n)$, where a is a constant and $\tilde{A}(n)$ is a zero-mean sequence. In this case the cross correlation coefficient $< aP(n), \tilde{A}(n)P(n) >= 0$, which means that the *instantaneous* component with the scalar gain $aP(n)$ is orthogonal to the *ambience* component $\tilde{A}(n)P(n)$.

3. PARAMETER ESTIMATION

The most common measure to compare two signals is the cross-correlation coefficient

$$\sigma_{12}^2 =< X_1, \bar{X}_2 > \qquad (2)$$

where $\bar{X}_2$ is the complex conjugate of X_2 and $< \cdot >$ represent expectation (e.g., the mean over the data points). The signal energies are represented by the variances $\sigma_k =< X_k, \bar{X}_k >$

σ_1^2	$a^2\sigma_p^2 + \sigma_u^2 +$ cross-terms
σ_2^2	$b^2\sigma_p^2 + \sigma_v^2 +$ cross-terms
σ_{12}^2	$ab\sigma_p^2 +$ cross-terms
d_{dif}	$(a-b)^2\sigma_p^2 + \sigma_u^2 + \sigma_v^2 +$ cross-terms
d_{sum}	$(a+b)^2\sigma_p^2 + \sigma_u^2 + \sigma_v^2 +$ cross-terms
d_{E}	$\lvert\sigma_u^2 - \sigma_v^2 + (a^2-b^2)\sigma_p^2 +$ cross-terms
d_{S}	$(a^2+b^2)\sigma_p^2 + \sigma_u^2 + \sigma_v^2 +$ cross-terms

Table 1: Expressions for some of the elementary measures for the PA signal model.

$, k = 1, 2$. Different combinations of these can be used to formulate a number of useful measures such as the difference and sum of energies given by

$$d_{\mathrm{E}} = |\sigma_1^2 - \sigma_2^2| \text{ and } d_{\mathrm{S}} = \sigma_1^2 + \sigma_2^2, \qquad (3)$$

respectively, the energies of sum and difference signals $d_{\mathrm{sum}} =< |X_1 + X_2|^2 >$, and $d_{\mathrm{dif}} < |X_1 - X_2|^2 >$.

One may also try to search for coefficients $\mathbf{z} = [z_1, z_2]$ such that $< |z_1 X_1 + z_2 X_2| >$ is minimized under some constraint. In matrix notation, this becomes the quadratic form

$$\mathbf{z}^T \bar{X} (\mathbf{z}^T \bar{X})^T = \mathbf{z}^T \bar{X} \bar{X}^T \mathbf{z} = \mathbf{z}^T \mathbf{C} \mathbf{z}, \qquad (4)$$

where $\mathbf{C}$ is the correlation matrix. With the constraint $z_1^2 + z_2^2 = 1$ the maximum and minimum of the quadratic form are obtained when $\mathbf{z}$ has the eigenvectors corresponding to the largest or smallest eigenvalues of $\mathbf{C}$, respectively. The eigenvalues are given by

$$\lambda_{1,2} = \frac{1}{2}(\sigma_1^2 + \sigma_2^2 \pm \sqrt{4\sigma_{12}^2 \bar{\sigma}_{12}^2 + (\sigma_1^2 - \sigma_2^2)^2}). \qquad (5)$$

The ratio of eigenvalues

$$\delta_{eig} = \frac{2\lambda_1}{\lambda_1 + \lambda_2} = 1 - \frac{\sqrt{4\sigma_{12}^2 \bar{\sigma}_{12}^2 + (d_{\mathrm{E}})^2}}{d_{\mathrm{S}}}, \qquad (6)$$

is often used as a measure characterizing the energy difference between primary and ambient sounds [5]. The denominator of δ_{eig} can be seen as a normalization term. The square root term is an Euclidean distance from the origin in plane spanned by the values of σ_{12}^2 and d_E.

The expressions in terms of the variances σ_p^2, σ_u^2, and σ_v^2, of the signals $P(n)$, $U(n)$, and $V(n)$, respectively, can be derived using the expressions above and they are shown in Table 1. The cross-terms in several measures are terms that contain inner products of component signals. If the component signals are uncorrelated all cross-terms vanish.

4. PRIMARY-TO-AMBIENCE RATIO

The *Primary-to-Ambience Ratio* is defined by

$$\mathrm{PAR} = \frac{\sigma_p^2}{\sigma_u^2 + \sigma_v^2} \qquad (7)$$

For the PA signal the ratio of eigenvalues (6) becomes

$$\delta_{eig} = 1 - \frac{\sqrt{4a^2b^2\sigma_p^4 + \left((a^2-b^2)\,\sigma_p^2 + \sigma_u^2 - \sigma_v^2\right)^2}}{(a^2+b^2)\,\sigma_p^2 + \sigma_u^2 + \sigma_v^2} \qquad (8)$$

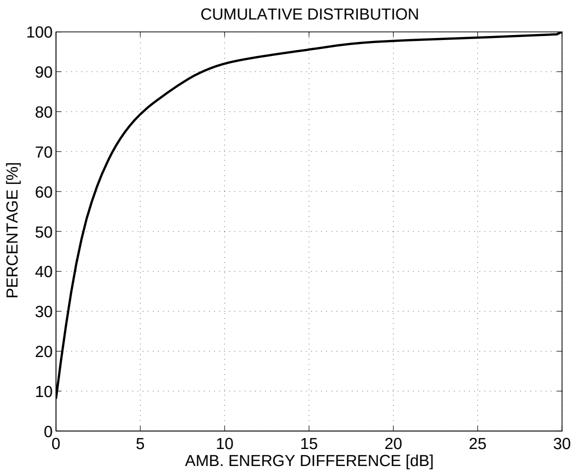

Figure 1: CDF of level difference estimates computed over a large database of stereo ambience data.

This is a function of the panning coefficients a and b, but also depends on the level difference between the ambience components $|\sigma_u^2 - \sigma_v^2|$. With the convention $a^2 + b^2 = 1$, and additional assumption that the ambience signal has the same level in both channels, that is, $\sigma_u^2 - \sigma_v^2 = 0$, we obtain

$$\delta_{eig} = \frac{\sigma_u^2 + \sigma_v^2}{\sigma_p^2 + \sigma_u^2 + \sigma_v^2}. \qquad (9)$$

which is a representation that is independent of the panning coefficients a and b. Moreover, combining (9) and (7) gives the estimator

$$\mathrm{PAR}_e = \frac{1}{\delta_{eig}} - 1 \qquad (10)$$

$$= \frac{d_S}{d_S - \sqrt{d_E^2 + 4\sigma_{12}^4}} - 1 \qquad (11)$$

The requirement that ambience signal has the same energy in both channels is a common simplifying assumption [2, 7] and its main benefit is to eliminate the panning coefficients a and b from the estimate. Various different formulations proposed in the literature lead essentially to the result shown in (11). The equal-energy assumption is sometimes linked to the room acoustic concept of *diffuse sound*, i.e., diffuse room reverberation is basically at the same level in both ears of a listener. This interpretation may be debatable when the methods are applied at separate frequency bands. Audio recordings with unequal levels of *non-primary* sound are not unusual. For example, the cumulative distribution function (CDF) in Fig. 1 represents the amplitude differences between individual time-frequency regions in a database of typical ambience stereo audio data (see Sect. 5). Although the amplitude difference in half of the cases is less than 3dB, there is still a significant number of regions where level difference is larger, and the PAR estimate based on (9) then becomes inaccurate.

If the energies are not equal, the value of the PAR estimate based on (11) yields

$$\mathrm{PAR}_e = \frac{\sigma_p^2 + \sigma_u^2 + \sigma_v^2}{\sigma_p^2 + \sigma_u^2 + \sigma_v^2 - \chi} - 1 \qquad (12)$$

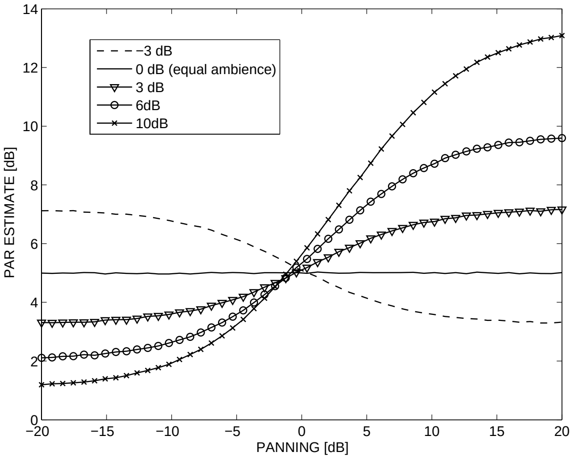

Figure 2: The values of (12) as a function of the panning direction of the primary component in PA model as a function of the level difference $\sigma_u^2 - \sigma_v^2$ of the ambience component.

where

$$\chi = \sqrt{\sigma_p^4 + (4a^2-2)\,\sigma_p^2(\sigma_u^2-\sigma_v^2) + (\sigma_u^2-\sigma_v^2)^2}$$

.

The values of (12) as a function of the panning direction of the primary component for five values of the level difference $\sigma_u^2 - \sigma_v^2$ of the ambience component are shown in Fig. 2. In the plotted example the true PAR is 5dB, which is obtained when the energies of the two ambience signals $u(n)$ and $v(n)$ are equal (solid curve). The error in the PAR estimate is large for the case where the energies of $u(n)$ and $v(n)$ differ and this difference also depends on the direction of the primary component, i.e., the panning coefficients a and b. If the true panning coefficients and ambience energy difference were known, one could possibly derive a correction term (or a lookup table) to eliminate the bias. However, the correction term would then depend on the panning coefficients, energy difference between the ambience components, and also the energy of the primary component.

The expressions from Table 1 can be collected into a matrix equation given by

$$\begin{pmatrix} \sigma_1^2 \\ \sigma_2^2 \\ \sigma_{12}^2 \end{pmatrix} = \begin{pmatrix} a^2 & 1 & 0 \\ b^2 & 0 & 1 \\ ab & 0 & 0 \end{pmatrix} \begin{pmatrix} \sigma_p^2 \\ \sigma_u^2 \\ \sigma_v^2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix} \quad (13)$$

or, equivalently,

$$\bar{d} = \mathbf{A}\bar{s} + \bar{e}. \quad (14)$$

Ignoring the cross-terms in $\bar{e}$ an estimate for the variances can be computed from $\tilde{s} = \mathbf{A}^{-1}\bar{d}$. The PAR is then obtained by insertion of the values of $\tilde{s}$ into (7). In fact, it turns out that this solution will be the same as

$$\mathrm{PAR}_a = \frac{\sigma_{12}^2}{\sigma_{12}^2 - abd_\mathrm{S}} \quad (15)$$

This PAR estimate is independent of the level difference of the ambient components. However, it is a function of a and b. The estimation of ab will be discussed below. When the value of a or b is close to zero, the estimator becomes highly sensitive to the non-zero cross-terms in values of σ_{12}^2 and d_S. This problem can be partly alleviated by using a least squares solution of (14) and Tikhonov regularization. In this case the estimate for the variances is obtained from

$$\tilde{s} = (\tilde{\sigma}_p^2, \tilde{\sigma}_u^2, \tilde{\sigma}_v^2) = (\mathbf{A}^T\mathbf{A} + \Gamma^T\Gamma)^{-1}\mathbf{A}^T\bar{d} \quad (16)$$

where Γ is the regularization matrix. In the following experiments we use $\Gamma = \lambda \max[a,b]\mathbf{I}$ where λ is a small constant and $\max[a,b]$ operator adds more regularization to side-panned primary components. The PAR estimate is then given by

$$\mathrm{PAR}_a = \frac{\tilde{\sigma}_p^2}{\tilde{\sigma}_u^2 + \tilde{\sigma}_v^2} \quad (17)$$

4.1 Estimation of *ab*

The estimation of the panning coefficients is typically based on eigenvectors of (4). In particular, the coefficients are computed from

$$a = \frac{\nu}{\sqrt{1-\nu^2}} \text{ and } b = \sqrt{1-a^2} \quad (18)$$

where

$$\nu = \frac{d_\mathrm{E} \pm \sqrt{4\sigma_{12}^2\bar{\sigma}_{12}^2 + (d_\mathrm{E})^2}}{2\sigma_{12}^2}, \quad (19)$$

where ± depends on which eigenvalue in (5) is larger. A direct application of the computation of a and b from the eigenvectors and substitution to (15) gives, after arithmetic manipulations, exactly the formula in (12).

In the case of a typical stereo audio signal the panning position of a certain instrument signal, for example, remains constant often over the entire duration of the recording. Therefore, it is possible to get more accurate estimates of the panning coefficients when the information is integrated over time. In the experiments reported in the current paper the values of a and b were estimated from long signal segments in parts where the ratio of eigenvalues is close to maximum. An alternative to this is adaptive estimation of the panning coefficients, e.g., using various formulations of adaptive eigenvalue decomposition.

5. EXPERIMENTS

First, let us compare the performance of the two PAR estimation methods for synthetic mixtures of uncorrelated white noise signals. The signals were mixed such that the real PAR value was set to 10dB and the amplitude difference between ambience components and the panning coefficients were varied. The results for the method of (11) shown in the top panel of Fig. 3 are similar to the analytic results of Fig. 2. The PAR estimates based on (11) were computed with an accurate *a priori* estimates of a and b. The results are close to the original PAR value independently of the panning coefficients or the energy difference between the ambience components. The histograms of the differences between the true PAR and the estimates shown in the top panel of Fig. 4.

The true PAR value is not available in the case of real stereo audio data and this makes the comparison difficult. In this paper, artificial mixtures of signals from a database of clean multi-track recordings were used as test material.

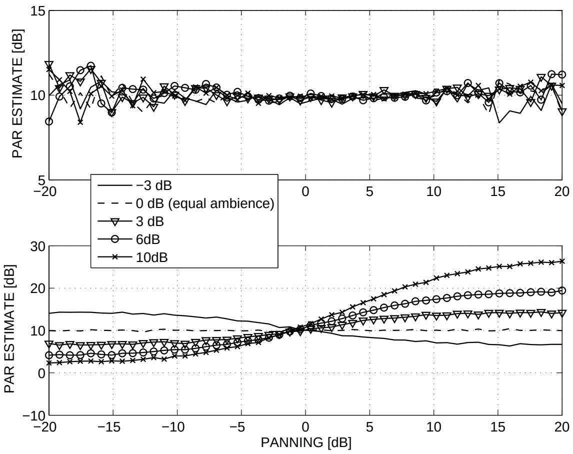

Figure 3: The PAR estimates computed from synthetic mixtures of white noise signals **top)** using (11) **bottom)** using (15). The true PAR was 10dB.

Since the original signals are available, it is possible to compute the true PAR value separately for each time-frequency region prior to the mixing. The experimental data consisted of 1000 artificially mixed songs where the primary component was either vocal or music instrument signal, and an ambience signal. The ambience signals were selected stereo signals of wide orchestral recordings, environmental sounds, or movie audio backgrounds characterized by a low Pearson correlation coefficient between the left and right channels. The sample rate of the signals was 44.1kHz and the duration of each song was 60s. The signals were converted to the frequency domain using the Fast Fourier Transform with the frame size of 1024 samples and 50% overlap in consecutive frames. The obtained spectrogram was split into uniform 16×16 sample time-frequency tiles and the estimates were computed from those tiles. In a typical application the frequency division would follow some perceptually relevant frequency scale such as the ERB rate scale.

The histograms of differences between the true PAR value and the estimate given by the methods proposed above are shown in Fig. 4 (bottom). The PAR measure based on (17) gives a more accurate estimate of PAR also in the case of realistic stereo signals, however, the difference is now smaller than with mixtures of white noise signals, which can also be predicted from Fig. 1.

6. CONCLUSIONS

A stereo audio signal is often modeled consisting of an amplitude panned primary component and two uncorrelated ambience signal components in the two channels. Primary-to-Ambience Ratio, PAR, is a useful measure that characterizes the energy ratio between the primary signal and stereo ambience data. The limitations of a conventional method based on the ratio of eigenvalues were discussed and a new estimator for PAR was proposed. The measure is independent of the levels of the ambience components but depends on the amplitude panning coefficients of the primary signal. The possibilities for the estimation of the coefficients was discussed and it was demonstrated that the method gives more accurate results for synthetic and realistic stereo audio data.

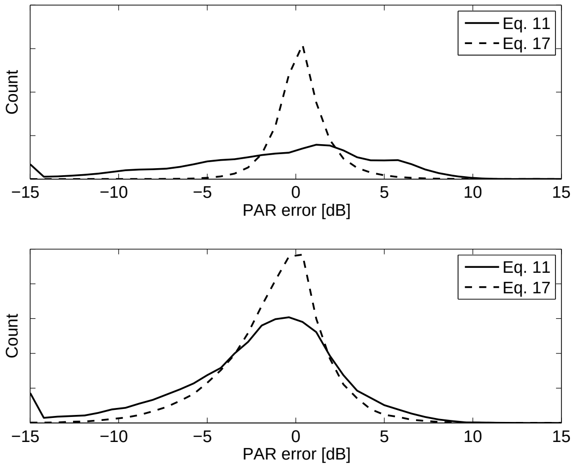

Figure 4: The difference between true PAR value and the different estimates for 100 mixes of **top)** white noise sequences **bottom)** synthetic songs.

Potential topics for the future work are the development of the methods for estimation of the panning coefficients based on signal history and evaluation of the proposed new measure in a complete audio format conversion application.

REFERENCES

[1] C. Avendano and J. M. Jot. A frequency-domain approach to multichannel upmix. *J. Audio Eng. Soc.*, 52(7/8):740–749, July 2004.

[2] J. Breebaart and C. Faller. *Spatial Audio Processing: MPEG Surround and Other Applications*. Wiley, 2007.

[3] R. Irwan and R. M. Aarts. Two-to-five channel sound processing. *J. Audio Eng. Soc.*, 50(11), November 2002.

[4] K. Kinoshita, T. Nakatani, and M. Miyoshi. Blind upmix of stereo music signals using multi-step linear prediction based reverberation extraction. In *Proc. ICASSP'2010*, pages 49–52, March 2010.

[5] J. Merimaa, M. Goodwin, and J. M. Jot. Correlation-Based Ambience Extraction from Stereo Recordings. In *AES 123rd Convention Preprint 7282*, NY, USA, October 2007.

[6] M. D. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M. E. Davies. Sparse representations in audio and music: from coding to source separation. *Proc. of the IEEE*, 2009.

[7] V. Pulkki. Spatial Sound Reproduction with Directional Audio Coding. *J. Audio Eng. Soc*, 55(6):503–516, 2007.

[8] J. Usher and J. Benesty. Enhancement of Spatial Sound Quality: A New Reverberation-Extraction Audio Upmixer. *IEEE Transactions on Audio, Speech and Language Processing*, 15(7):2141–2150, September 2007.

[9] D. Yellin and E. Weinstein. Criteria for multichannel source separation. *IEEE Trans. Signal Processing*, 42(8):2158–2168, August 1994.