AMBIENCE EXTRACTION AND SYNTHESIS FROM STEREO SIGNALS FOR MULTI-CHANNEL AUDIO UP-MIX.

Carlos Avendano and Jean-Marc Jot

Creative Advanced Technology Center
1500 Green Hills Road, Scotts Valley, CA 95066
{carlosa,jmj}@atc.creative.com

ABSTRACT

In this paper we propose a frequency-domain technique to identify and extract the ambience information in stereo audio signals. The method is based on the computation of an inter-channel coherence index and a non-linear mapping function that allow us to determine time-frequency regions that consist mostly of ambience components in the two-channel signal. Ambience signals are then synthesized and used to feed the surround channels of a multi-channel playback system. Simulation results demonstrate the effectiveness of the technique in extracting ambience information and up-mix tests on real audio reveal the various advantages and disadvantages of the system compared to previous up-mix strategies.

1. INTRODUCTION

While surround multi-speaker systems are already popular in home and desktop theater settings, the number of multi-channel audio recordings available to the public is still limited. Recent movie soundtracks and a few musical recordings are available in discrete multi-channel format (e.g. 5.1 surround), but most legacy audio recordings are available only in stereo. Thus, a system that can enhance stereo recordings for reproduction over a multi-speaker surround setup is desirable. However, playback of stereo material over a multi-channel system poses a fundamental problem: stereo audio is mixed with a very particular listening setup in mind, which consists of a pair of loudspeakers placed symmetrically in front of the listener. Due to the increased number of loudspeakers, it is not obvious what signal or signals should be sent through the additional channels. To solve this problem several factors need to be considered, including the number and locations of the loudspeakers, the desired degree of "preservation" of the original material, the subjective attributes of the overall immersion experience, etc. Thus, no single objective criterion exists and the solution will depend strongly on individual preferences [7].

In this paper we describe an up-mix strategy inspired by the direct/ambient approach for mixing multi-channel audio [5]. In this approach, multiple microphones are placed at different spatial locations inside the studio or venue. Some are placed near the individual instruments or sound sources to capture the "primary signals" and others are appropriately placed to capture the surrounding "ambience" of the performance (the latter can include room reverberation, applause, or various types of background sounds). In the mixing phase, the primary signals are typically panned among the front channels and the rear channels are typically fed only ambience signals. This method can for instance create the impression that the listener is in the audience of a concert hall, in front of the stage (best seat in the house) [1]. Accordingly, the up-mix system should be capable of extracting ambience information from two-channel material and of synthesizing the rear channel signals using this information.

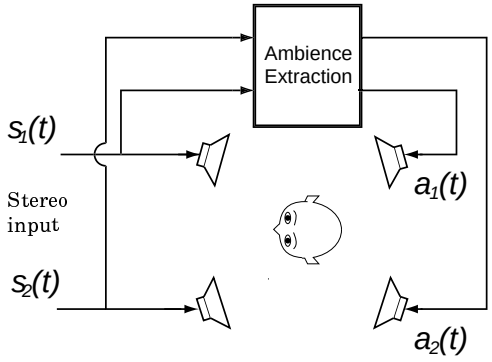

Fig. 1. *Example: stereo to four-channel upmix system.*

The up-mix system is shown in Figure 1, where we observe that the front channels are unmodified. There are many alternative approaches to deriving the front channels, some including the addition of a center loudspeaker that will provide a more stable frontal image for off-axis listeners. In this study, we do not discuss how to derive these front channel signals, but we assume that for a listener located at the "sweet spot", the frontal stereo image is identical to the original stereo recording. It is out of the scope of this paper to assess the subjective attributes of the up-mix. Our goal is to demonstrate, through objective measurements and simulations, the effectiveness of the proposed ambience extraction algorithm and some of its potential applications.

2. BACKGROUND

The existing up-mixing algorithms rely on combinations of the following methods for deriving the rear channels:

a) Applying artificial reverberation to the original stereo signal. The resulting impression is essentially of lis-

Submitted to IEEE ICASSP 2002.

[1] A second approach is the "in-the-band" approach, where both the primary signals and the ambience signals are panned among all the loudspeakers, creating the impression that the musicians surround the listener.

tening to the original recording in a virtual listening room (see e. g. [2]). This artificial ambience information does not match the conditions in which the original recording was produced.

b) Computing the difference of the original left and right signals. This provides a monaural signal whose content includes the desired ambience information and excludes any primary signal panned in the center of the original stereo image. However, the resulting ambience signal also contains unwanted "leakage" from any non-centered primary signals. This leakage can be partially reduced by use of "logic steering" techniques (see e. g. [4]).

c) Deriving a stereo ambience signal from a monaural signal (pseudo-stereophony). Two weakly correlated signals can be obtained by applying a pair of all-pass filters to a single audio signal [8].

d) Applying a small delay (typically 5 to 20 ms) on the rear-channel signals to alleviate unwanted localization artifacts caused by any leakage of primary signals into the rear channels [4, 7]. This is an effective method for better preserving the frontal stereo image of the original recording, but it cannot correct the ambience information itself.

In [6], room responses corresponding to "virtual microphone" positions are derived and then used to process a stereo recording that was captured in the same room. This improves on method (a) by synthesizing rear-channel reverberation signals that match the acoustics of the original venue. However, the application of this method is in principle restricted to "live" acoustical recordings for which detailed additional historical information is available on the original recording conditions and techniques. This method is also limited in that the only form of ambience information that it handles is the acoustical reverberation of the recording venue. It cannot be used to up-mix the ambient noises that may be present in the original recording.

3. SIGNAL MODEL

Stereo recordings can be roughly categorized into two main classes: "studio" or artificial, and "live" or natural [7]. In studio recording, the different sources (or instruments) are individually recorded and then mixed into a single stereo signal. Stereo reverberation is added artificially to the mix. In general, the left and right impulse responses of the reverberation processor are different (weakly correlated) to increase the perception of spaciousness [3]. Live recording involves a number of spatially distributed microphones to capture all sound sources. With stereo microphone techniques, the ambience is naturally included in the recording and presents a weak correlation between the left and right channels [7] [5]. Notice that the ambience is not only produced by reverberation, but can be introduced by other spatially distributed sources such as wind, audience noise, etc.

A signal model can be defined as follows: assume that there are N sources $c_j(t), N = 1, ..., j$ convolved with room impulse responses (or artificial reverberation reponses) $h_{ij}(t)$ to generate the left ($i = 1$) and right ($i = 2$) stereo channels respectively. The responses $h_{ij}(t)$ can be split into a direct-path or primary signal component and a reverberation component as $h_{ij}(t) = d_{ij}(t) + r_{ij}(t)$, and the stereo signal can be written as

$$s_i(t) = \left[\sum_{j=1}^{N} c_j(t) * d_{ij}(t)\right] + \left[\sum_{j=1}^{N} c_j(t) * r_{ij}(t) + n_i(t)\right], \quad (1)$$

where $n_j(t)$ are background noise signals contributing to the surrounding ambience (e.g. audience noise.). In (1) the left term corresponds to the primary signals and the right term corresponds to the ambience signals. Notice that by definition the ambience signals will have comparable levels, i.e. $||r_{1j}(t)||^2 \simeq ||r_{2j}(t)||^2$ and $||n_1(t)||^2 \simeq ||n_2(t)||^2$.

Notice that if there is a source $c_0(t)$ panned to the center (i.e. $d_{10}(t) = d_{20}(t)$) the difference of the left and right signals will eliminate its direct-path component and yield a signal that contains only the reverberation infornation corresponding to this source, i.e. $s_1(t) - s_2(t) = c_0(t) * (r_{10}(t) - r_{20}(t))$. However, notice that only the ambience of signals panned to the center can be extracted with this method, and the ambience component is monaural and consists of the difference between the stereo ambience signals. In fact this approach is commonly used to extract ambience information from stereo recordings, as described in Section 2, and we readily see the limitations. None of techniques reviewed can remove all of the primary signals from the surround channels to present ambience information only (including both reverberation and ambient noise), without relying on additional knowledge of the recording conditions.

4. FREQUENCY-DOMAIN AMBIENCE EXTRACTION

In this section, we describe the proposed technique for extracting the ambience information of a stereo signal. The method is based on the assumption that the left and right ambience signals are weakly correlated. This assumption is in general valid for stereo recordings as we discussed above. The proposed technique essentially attempts to separate the parts of the signals that are uncorrelated between left and right channels from the primary signal components (i.e. those that are maximally correlated), and generates two signals which contain most of the ambience information from each channel and exclude most of the primary signal information. An additional criterion for recognizing ambience components is that the left and signals must have comparable energies.

The basic idea is inspired by the binaural processing of the hearing mechanism and the well established fact that binaural hearing and localization involve the computation, in each critical band, of the cross-correlation between the left and right channel signals [3]. Higher processes then use across-frequency coherence information and other inter and intra-channel features to determine position, distance, environment, etc. The proposed technique exploits a process similar to the psycho-acoustic phenomena on which the auditory perception of ambience relies.

Our signal processing front-end consists of a discrete short-time Fourier transform (STFT). In the time-frequency plane, the time correlation between the stereo channels at each frequency band will be high in regions where the direct component is dominant, and low in regions dominated by the reverberation or ambience. A similar rationale is used by the two-microphone speech de-reverberation algorithm described in [1], where the regions of low correlation are attenuated by a factor directly proportional to the amount of correlation, and regions of high correllation are co-phased and added to produce a new STFT. This new STFT is then

inverted to yield a monaural time domain signal, which consists mainly of the direct path component (de-reverberated speech).

The ambience extraction algorithm is similar in structure to [1] although the goal and metrics are different. Let us first denote the STFTs of the channel signals $s_i(t)$ as $S_i(m,k)$, where m is the time index and k is the frequency index. We define the following statistical quantities

$$\Phi_{ij}(k) = \mathrm{E}\{S_i(m,k)S_j^*(m,k)\}, \quad (2)$$

where E is the expectation operator with respect to time m and $*$ denotes complex conjugation. Audio signals are in general non-stationary. For this reason the statistics will change with time. To track the changes of the signal we introduce a forgetting factor λ in the computation of the cross-correlation functions. Thus in practice the statistics in (2) are computed as short-time functions:

$$\Phi_{ij}(m,k) = \lambda\Phi_{ij}(m-1,k) + (1-\lambda)S_i(m,k)S_j^*(m,k). \quad (3)$$

Notice that different values of λ can be used in different frequency bands. Using these statistical quantities we define the inter-channel short-time coherence function as

$$\Phi(m,k) = \frac{\Phi_{12}(m,k)}{[\Phi_{11}(m,k)\Phi_{22}(m,k)]^{\frac{1}{2}}}. \quad (4)$$

The coherence function $\Phi(m,k)$ is real and will have values close to one in time-frequency regions where the primary signal is dominant, and will be close to zero in regions dominated by the reverberation tails and surrounding noise (ambience). Given the properties of the coherence function (4), one way of extracting the ambience of the stereo recording would be to multiply the left and right channel STFTs by $1-\Phi(m,k)$ and reconstructing the two time domain ambience signals $a_i(t)$ from these modified transforms. A more general form is to weigh the channel short-time transforms with a non-linear function of the short-time coherence:

$$A_i(m,k) = S_i(m,k)\Gamma\left[\Phi(m,k)\right] \quad (5)$$

where $A_i(m,k)$ are the modified, or ambience transforms. The behavior of the non-linear function Γ that we desire is one in which low-coherence regions are not modified and high-coherence regions are heavily attenuated to remove the direct-path component. Additionally, the function should be smooth to avoid artifacts. One function that presents this behavior is the hyperbolic tangent. Thus we define Γ as:

$$\Gamma\left[\Phi\right] = \left(\frac{\mu_1-\mu_0}{2}\right)\tanh\{\sigma\pi(\Phi_0-\Phi)\} + \left(\frac{\mu_1+\mu_0}{2}\right) \quad (6)$$

where the parameters μ_1 and μ_0 define the range of the output, Φ_0 is the threshold and σ controls the slope of the function. In Figure 2 we show the function Γ for different values of the parameters. In general the value of μ_1 is set to one since we do not wish to enhance the non-coherent regions (though this could be useful in other contexts). The value of μ_0 determines the floor of the function and it is important that this parameter be set to a small value greater than zero to minimize spectral-subtraction-like artifacts at the output. An additional criterion to extract ambience components is that the left and right signals have to have comparable energies. That is because the coherence will be low (or zero) for primary signals panned completely to one side.

A block diagram of the ambience signal extraction algorithm is shown in Figure 3. The inputs to the system are the left and

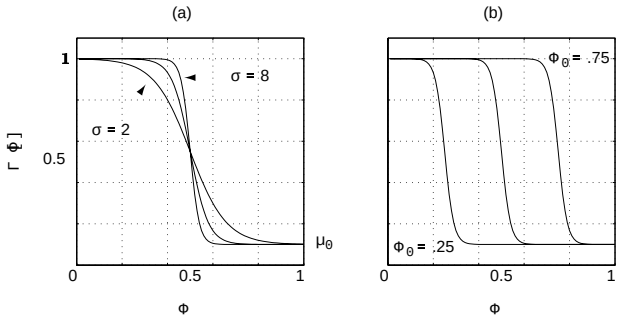

Fig. 2. *Mapping function for the coherence index. (a) As a function of the parameter σ with $\Phi_0 = 0.5$. (b) As a function of the offset Φ_0, with $\sigma = 4$. In both cases $\mu_0 = 0.1$.*

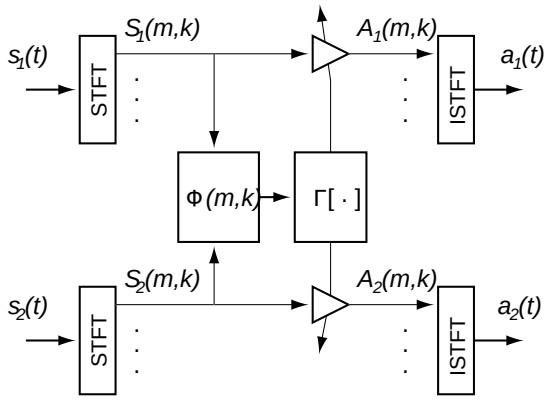

Fig. 3. *Block diagram of the ambience extraction system. Only the processing for one subband is shown.*

right channel signals of the stereo recording, which are first transformed into the short-time frequency domain. The parameters of the STFT are the window length N, the transform size K and the stride length L. The coherence function is estimated and mapped to generate the multiplication coefficients that modify the short-time transforms. After modification, the time domain ambience signals are synthesized by applying the inverse short-time transform via the overlap-and-add (OLA) method. Typical parameter values were obtained based on simulation results and adjusted according to informal listening tests. The analysis parameters for $44.1kHz$ -sampled audio were set to $N = 1024$, $K = 2048$, $L = 256$, with a Hamming window. The coherence function was estimated with a forgetting factor of $\lambda = 0.85$ for all frequencies, and the parameters of the mapping function Γ were set as $\sigma = 8$ and $\Phi_0 = 0.15$. In the following section we illustrate the operation of the algorithm with a simulation.

5. SIMULATION

The ambience extraction system has been implemented and tested on real audio signals. However, to illustrate its operation and performance we carried out a simulation using a simplified stereo signal model. The primary signal was a train of pulses panned

between the left and right channels with gains 0.25 and 0.75. The ambience was generated by convolving the resulting train of pulses with two artificial room impulse responses. These room responses consisted of a direct path component with unit amplitude and a reverberation tail generated by multiplying a white noise sequence with a decaying exponential function. The tail was delayed with respect to the direct path a by a few milliseconds. A portion of the input signal is shown in the top panel of Figure 4.

The results of applying the algorithm are shown in the lower panels of Figure 4. Observe that the ambience signals contain most of the reverberation tail and almost no direct path (approximately -30 dB). The residual of the ambience extraction algorithm is shown in the bottom panels, where we observe that it contains the direct path component plus some very small amount of residual reverberation. This simulation shows the effectiveness of the algorithm for a very simple stereo signal. More complex and realistic examples have been processed as well with similar outcomes. It is worth noting that the ambience components that overlap with direct path components in the time-frequency plane cannot be isolated and extracted effectively. However, when those regions are dominated by the direct path components, the ambience components will be masked and will not contribute significantly [2].

6. SYSTEM IMPLEMENTATION

For the example up-mix system shown in Figure 1 it is useful to apply all-pass filters to the synthetic ambience signals in order to de-correlate them from the ambience information already embedded in the front channels. As mentioned earlier, the rear channels can also be delayed by a few milliseconds in order to minimize the unwanted effects of any residual leakage of primary information into the rear channels.

The computational complexity of the ambience extraction algorithm is dominated by the analysis/synthesis stages. If we assume an efficient FFT algorithm and a table lookup for the mapping function Γ then the overall complexity is about $O = \frac{K}{14L} + \frac{6}{J} K \log_2(K)$ multiply-adds per output sample. For example, at $44.1kHz$ sampling rate the algorithm requires 28 MIPS.

7. DISCUSSION AND CONCLUSION

We have proposed a stereo to multi-channel up-mix strategy based on the direct/ambient approach for mixing multi-channel audio. For this we derived a novel ambience extraction and synthesis algorithm that operates in the short-time frequency domain. Simulations and tests with real audio indicate that the technique is effective. We have noticed some minor artifacts associated with the frequency domain processing. However, these artifacts are not easily perceived in the context of the up-mix application, where all channels are played simultaneously. Other up-mix possibilities can be explored using this ambience extraction technique, for example reducing or enhancing the ambience of the front channels. Other applications might include simply enhancing the natural ambience of the original stereo recording.

The proposed up-mix system has the advantage that it can extract the ambience alone for all sources in the original stereo mix, and does not require complex logic or steering methods to derive the surround channels. Accordingly, it has been informally observed that the perceived front image of the original stereo mix is not significantly modified. However, future work should include the assessment of the perceptual attributes of the ambience signals and of the overall up-mix.

[2] Audio demonstrations will be played at the conference.

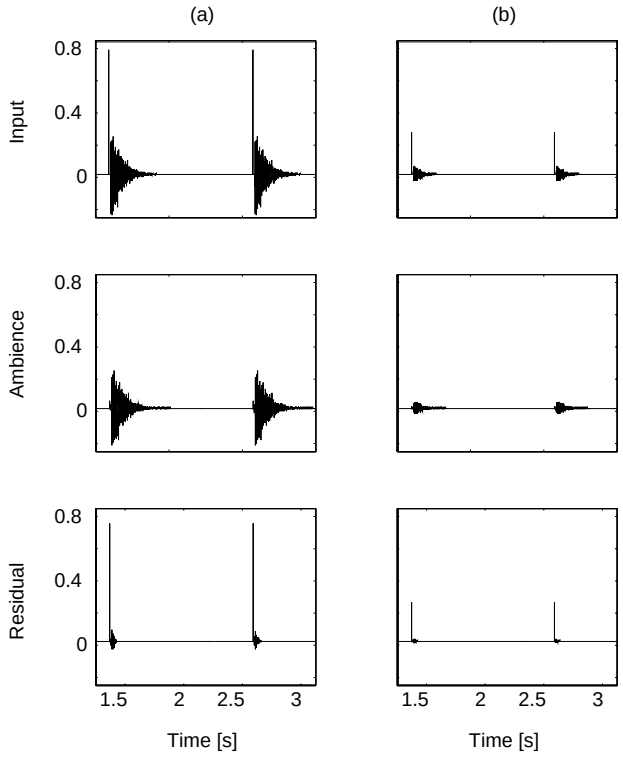

Fig. 4. *Simulation results.*

8. REFERENCES

[1] J. Allen, D.A. Berkeley and J. Blauert, "Multi-microphone Signal-Processing Technique to Remove Room Reverberation from Speech Signals." Journal of the Acoustical Society of America, Vol. 62, No.4, pp. 912-915, October 1977.

[2] D.R. Begault, "3-D Sound for Virtual Reality and Multimedia." pp. 226-229, Academic Press, Cambridge, 1994.

[3] J. Blauert, "Spatial Hearing." MIT Press, Cambridge, 1983.

[4] R. Dressler, "Dolby Surround Pro Logic II Decoder - Principles of Operation."
URL: http://www.dolby.com/tech/l.wh.0007.PLIIops.pdf.

[5] T. Holman, "Mixing the Sound." Surround Magazine, pp. 35-37, June 2001.

[6] C. Kyriakakis and A. Mouchtaris, "Virtual Microphones for Multi-channel Audio Applications." In Proc. IEEE ICME 2000, Vol. 1, pp. 11 - 14, August 2000.

[7] 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.

[8] M. Schroeder, "An Artificial Stereophonic Effect Obtained from Single Audio Signal." Journal of the Audio Engineering Society, Vol. 6, pp. 74-79, 1958.