14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

# MULTIBAND SOURCE/FILTER REPRESENTATION OF MULTICHANNEL AUDIO FOR REDUCTION OF INTER-CHANNEL REDUNDANCY

*A. Mouchtaris, K. Karadimou, and P. Tsakalides*

Department of Computer Science, University of Crete and
Institute of Computer Science (FORTH-ICS)
Heraklion, Crete, Greece 71110
{mouchtar, karadsun, tsakalid}@ics.forth.gr

## ABSTRACT

*In this paper we propose a model for multichannel audio recordings that can be utilized for revealing the underlying interchannel similarities. This is important for achieving low bitrates for multichannel audio and is especially suitable for applications when there is a large number of microphone signals to be transmitted (such as remote mixing or distributed musicians collaboration). Using this model, we can encode a multichannel audio signal using only one full audio channel and some side information in the order of few KBits/sec per channel, which can be used to decode the multiple channels at the receiving end. We apply objective and subjective measures in order to evaluate the performance of our method.*

## 1. INTRODUCTION

Multichannel audio offers significant advantages regarding music reproduction when compared to stereophonic audio. The use of a large number of channels around the listener results in a more realistic acoustic space, adding more sound directions and thus immersing the listener into the acoustic scene. By using a higher number of channels than in stereo systems, multichannel audio recordings require higher data rates for transmission. For multichannel audio, in addition to reducing the intra-channel redundancies, methods have been explored for reducing the inter-channel redundancies, such as Mid/Side Coding [1], Intensity Stereo Coding [2], and KLT-based methods [3]. This paper focuses on inter-channel redundancies.

Although the multichannel audio coding algorithms mentioned in the previous paragraph result in reduction of the data rates required by the original recording, they still remain highly demanding for many practical applications when the available channel bandwidth is low. This is especially important given the fact that many multichannel audio systems require even more than the 5.1 channels of currently popular standards, and thus even higher data rates. In recent years, the concept of Spatial Audio Coding has been introduced, with the objective of further taking advantage of interchannel redundancies in multichannel audio recordings. Under this approach, the objective is to decode a (stereo or mono) channel of audio using some additional (side) information, so as to recreate the spatial rendering of the original multichannel recording. The side information is extracted during encoding; in the most popular implementation of this approach, Binaural Cue Coding (BCC) [4], the side information contains the interchannel level difference, time difference, and correlation. The resulting signal contains one full channel of audio (downmix), along with the side information with bitrate in the order of few KBits/sec per channel.

Multichannel audio recordings are made using a large number of microphones in a venue, resulting in numerous microphone signals. These are then mixed in order to create the final multichannel audio recording. In many applications it would be desirable to transmit the multiple microphone signals of a performance, before those are mixed into the (usually much smaller number of) channels of the multichannel recording. This would allow for remote mixing of the multichannel recording, which is an important aspect for many applications in the music industry. Remote collaboration of geographically distributed musicians is a field of great significance with extensions to music education and research. Current experiments have shown that high data rates are needed so that musicians can perform and interact with minimal delay [5]. Remote mixing in the client side would also enable the user to interact with the music in an unparalleled fashion, allowing him to create his own music by mixing sounds as he pleases.

In this paper, we present a source/filter representation of multichannel audio that allows for transmission of the multiple microphone signals of a music performance with moderate data-rate requirements. This would allow for transmission through low bandwidth channels such as the current Internet infrastructure or wireless networks for broadcasting. Our method is tailored towards the transmission of the various microphone signals of a performance *before* they are mixed and thus can be applied to applications such as remote mixing and distributed performances. Our innovative approach relaxes the current bandwidth constraints of these demanding applications, enabling their widespread usage and more clearly revealing their value. Our method has the same objective with Spatial Audio Coding, *i.e.* to reduce a multichannel recording into one full audio channel and some side information of the order of few KBits/sec per channel. However, it should be viewed as a generalization of BCC. In BCC, the side information can be used to recreate the spatial rendering of the various channels. In our method, the side information can theoretically (as we explain next) recreate the *exact* microphone signals of the multichannel recording. In addition these microphone signals need not be the actual channels of the recording but rather they can be the microphone recordings (stem recordings) *before* those are mixed into the final multichannel signal.

## 2. RECORDING FOR MULTICHANNEL AUDIO

Before proceeding to the description of the proposed method, a brief description is given of how the multiple microphone signals for multichannel rendering are recorded. In this paper, we mainly focus on live concert hall performances, although this does not result in a loss of generality of our methods. A number of microphones are used to capture several characteristics of the venue, resulting in an equal number of microphone signals (stem recordings). These signals are then mixed and played back through a multichannel audio system. Our objective is to design a system based on available microphone signals, that is able to recreate all of these *target* microphone signals from a smaller set (or even only one) of *reference* microphone signals at the receiving end. The result would be a significant reduction in transmission requirements, while enabling remote mixing at the receiving end. In our previous work [6], we were interested to completely synthesize the target signals using the reference signals, without any additional information. Here, we propose using some additional information for each microphone for achieving high quality resynthesis, with the constraint that this additional information requires minimal data rates for transmission. By examining the acoustical characteristics of the various stem recordings, the distinction of microphones is made into reverberant and spot microphones.

Spot microphones are microphones that are placed close to the sound source. Because the source of sound is not a point source but rather distributed such as in an orchestra, the recordings of these microphones depend largely on the instruments that are near the microphone and not so much on the hall acoustics. Resynthesizing the signals captured by these microphones, therefore, involves enhancing certain instruments and diminishing others, which in most cases overlap in the time and frequency domains. Reverberant microphones are the microphones placed far from the sound source, that mainly capture the reverberation information of the venue. In our previous work [6], we showed that the reverberant recordings can be resynthesized from a reference recording using specially designed LTI filters. Here, we focus on the spot microphone signals. Our objective is to design a system that *recreates* these signals from a smaller subset of the microphone recordings.

In order to achieve low bitrates for audio coding, it is generally considered necessary to introduce some trade-off regarding the quality of the recording. Here we propose, instead, a tradeoff regarding the *accuracy* of the final multichannel recording. We propose that it is possible to achieve low bitrates by substituting some microphone signals with others, which, although differ acoustically, they however retain the “objectives” of the initial recording. Here, the term “objectives” corresponds to the main purpose for the microphone placement in a particular position of the venue. If a microphone was placed, for example, near the chorus of an orchestra, then the main objective of the microphone placement is to capture a recording of the music where the chorus sounds as the most prevailing part with respect to the remaining parts of the orchestra. If this microphone signal is substituted by a different (*i.e.* resynthesized) one, which again contains the same performance and the chorus is the prevailing part of the new signal, this is considered as a signal that retains the “objective” of the initial microphone signal. The term accuracy corresponds to the distance between the two signals. Our methods result in a resynthesized signal that, although retains the “objective” of the initial microphone signal, it introduces a tradeoff between the required bitrates and the accuracy achieved. In our previous example, in the resyntheszied signal the chorus will still be the prevailing part of the orchestra (thus the objective is retained), but the other parts of the orchestra might be now more audible than in the initial signal (*i.e.* loss of accuracy). Subjectively, this will have the effect that the new signal sounds as if it was captured with a microphone that was placed farther from the chorus compared with the microphone placement of the original recording. However, since the objective is retained, the resynthesized signal will still sound as if it was made with a microphone placed close to the chorus. We claim that low data rates can be achieved, without significant sacrifices regarding the accuracy of the multichannel recording.

## 3. MODEL AND MOTIVATION

Our proposed methodology, which is based on a multiband source / filter representation of the multiple microphone signals, consists of the following steps. Each microphone signal is segmented into a series of short-time overlapping frames using a sliding window. For each frame, the audio signal is considered approximately stationary, and the spectral envelope is modeled as a vector of linear predictive (LP) coefficients [7]. Under the source/filter model, the signal $s(n)$ at time $n$ is related with the $p$ previous signal samples by the following autoregressive (AR) equation

$$s(n) = \sum_{i=1}^{p} a(i)s(n-i) + e(n) \tag{1}$$

where $e(n)$ is the modeling error (residual signal), and $p$ is the AR filter order. In the frequency domain, this relation can be written as

$$P_s(\omega) = |A(\omega)|^{-2} P_e(\omega) \tag{2}$$

where $P_x(\omega)$ denotes the power spectrum of signal $x(n)$. $A(\omega)$ denotes the frequency response of the AR filter, *i.e.*

$$A(\omega) = 1 - \sum_{i=1}^{p} a(i)e^{-j\omega i} \tag{3}$$

The $p+1^{th}$-dimensional vector $a^T = [1,\ -a_1,\ -a_2,\ \cdots,\ -a_p]^T$ is the low dimensional representation of the signal spectral properties. If $s(n)$ is an AR process, the noise $e(n)$ is white, thus $a$ completely characterizes the signal spectral properties. In the general case, the error signal will not have white noise statistics and thus cannot be ignored. In this general case, the all-pole model that results from the LP analysis gives only an approximation of the signal spectrum, and more specifically the spectral envelope. For the particular case of audio signals, the spectrum contains only frequency components that correspond to the fundamental frequencies of the recorded instruments, and all their harmonics. The AR filter for an audio frame will capture its spectral envelope. The error signal is the result of the audio frame filtered with the inverse of its spectral envelope. Thus, we conclude that the error signal will contain the same harmonics as the audio frame, but their amplitudes will now have significantly flatter shape in the frequency spectrum.

Consider now two microphone signals of the same music performance, captured by microphones placed close to two different groups of instruments of the orchestra. Each of these microphones mainly captures that particular group of instruments, but also captures all the other instruments of the orchestra. For simplification, consider that the orchestra consists of only two instruments, *e.g.* a violin and a trumpet. Microphone 1 is placed close to the violin and microphone 2 close to the trumpet. It is true in most practical situations, that microphone 1 will also capture the trumpet, in much lower amplitude than the violin, and vice versa for microphone 2. In that case, the signal $s_1$ from microphone 1, and the signal $s_2$ from microphone 2 will contain the fundamentals and corresponding harmonics of both instruments, but they will differ in their spectral amplitudes. Consider a particular frame for these 2 signals, which corresponds to the exact same music part (*i.e.* some time-alignment procedure will be necessary to align the two microphone signals). We model each of the two audio frames with the source/filter model:

$$s_k(n) = \sum_{i=1}^{p} a_k(i) s_k(n-i) + e_k(n), k = 1,2. \tag{4}$$

From the previous discussion it follows that the two residual signals $e_1$ and $e_2$ will contain the same harmonic frequency components. If the envelope modeling was perfect, then it follows that they would also be equal (differences in total gain are of no interest for this application), since they would have flat magnitude with exactly the same frequency components. In that case, we would be able to resynthesize each of the two audio frames using only the AR filter that corresponds to that audio frame, and the residual signal of the other microphone. If we used similarly this model for all the spot microphone signals of a single performance, we would be able to completely resynthesize these signals using their AR vector sequences (one vector for each audio frame) and the residual error of only one microphone signal. This would result in a great reduction of the data rate of the multiple microphone signals.

In practice, the AR filter is not an exact representation of the spectral envelope of the audio frame, and the residual signals for the two microphone signals will not be equal. However, we can improve the modeling performance of the AR filter by using filterbanks. We divide the spectrum of the audio signals and apply LP analysis in each band separately (subband signals are downsampled). A small AR filter order for each band can result in much better estimation of the spectral envelope than a high-order filter for the full frequency band. The multiband source/filter model achieves a flatter frequency response for the residual signals. Then we can use one of them for resynthesizing the other microphone signals, in the manner explained in the previous paragraph. However, the error signals cannot be made exactly equal, thus the resynthesized signals will not sound exactly the same as the originally recorded signals. This corresponds to the loss of “accuracy” for the multichannel recording that was discussed earlier. We claim that the use of the multiband source/filter model results in audio signals of high-quality which retain the “objective” of the initial recordings, in the sense that was introduced here. In other words, the “main” group of instruments that is captured still remains the prominent part of the microphone signal, while other parts of the orchestra might be more audible in the resynthesized signal than in the original microphone signal. Returning to the example of the two microphones and the two instruments, if we use the residual of microphone 1 to resynthesize the signal of microphone 2, then in the result the violin will most likely be more audible than in the original microphone 2 signal. This happens because some information of the first microphone signal remains in the error signal, since the spectral envelope modeling is not perfect. However, the trumpet will still be the prominent of the two instruments in the resynthesized signal for mic 2, since we used the original spectral information of that microphone signal. Equally important is the fact that the accuracy and the final audio quality the multiband source/filter model can be controlled with a variety a parameters: (1) the duration of the audio frames for each band, (2) the AR order for each band, (3) the percentage of frame overlapping, and (4) the total number of bands. By changing these parameters we can achieve various data rates with the corresponding varying audio quality. Thus our system is quality scalable which is a significant property for the applications in mind. The appropriate values can be found experimentally, while the choice of filterbank is a subject which is currently under investigation.

It is easy to verify experimentally that our claims hold for other types of harmonic signals, *e.g.* speech signals. Some types of microphone signals, such as percussive signals and signals from reverberant microphones, present different challenges [6]. Here, we focus on the large class of audio signals that can be modeled using short-time analysis with emphasis on their spectral envelope.

## 4. RESULTS AND DISCUSSION

In this section, we show that the use of the proposed method results in a modeled signal that is objectively and subjectively very close to the original recording. For this purpose, we use two microphone signals of a live concert hall performance. One of the microphones captures mainly the male voices of the chorus of the orchestra, while the other one mainly captures the female voices. These recordings are very easy to distinguish acoustically. The objective is to resynthesize one of these recordings using its corresponding low-dimensional model coefficients along with the residual of the other recording.

From initial listening tests it has been clear that using a number of bands around 8 for our model produced high quality resynthesis without loss of the objective of the initial recording. For example, we have been able to resynthesize the male voices recording based on the residual from the female voices. Without the use of a filterbank, the resulting quality of the resynthesized signal greatly deteriorated with a complete loss of the recording objective. In order to show this objectively, we measured the distance between the residual signals of the two recordings, using the normalized mutual information as a distance measure. As mentioned, the intuitive claim is that decreasing the distance of the two residuals will increase the quality of the resynthe-

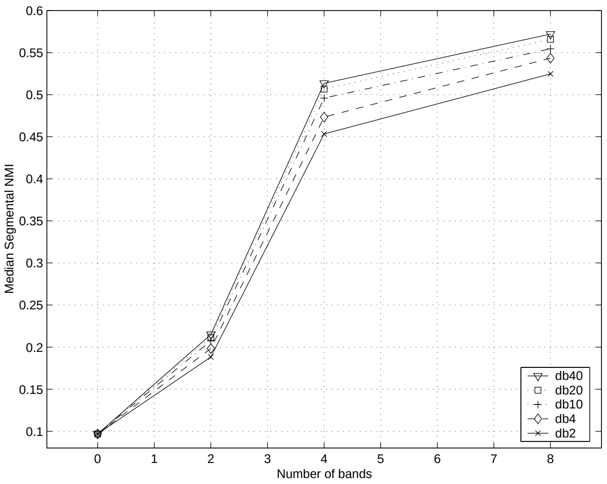

Figure 1: Normalized Mutual Information between the residual signals from the reference and target recordings as a function of the number of bands of the filterbank, for various Daubechies (db) filters.

sized recording. Our listening tests indicated that increasing the number of subbands in our model, and consequently improving the model accuracy, resulted in much better quality of the resynthesized signals. While several measures were tested, the normalized mutual information proved to be very consistent in this sense.

The use of mutual information $I(X;Y)$ as a distance measure between random variables $X$ and $Y$ is very common in pattern comparison. Since our interest is in comparing two vectors $X$ and $Y$ ($Y$ being the desired response), it is useful to use a modified definition for the mutual information, the Normalized Mutual Information (NMI) $I_N(X;Y)$ which is the mutual information normalized by the entropy of $Y$, so that $0 \leq I_N \leq 1$. The NMI obtains its minimum value when $X$ and $Y$ are statistically independent and its maximum value when $X = Y$. The NMI does not constitute a metric since it lacks symmetry, however it is invariant to amplitude differences which is very important when comparing audio waveforms.

In Fig. 1 we plot the NMI between the power spectra of the two residual signals with reference to the number of different subbands used, for different orders of the Daubechies wavelet filters, which were used for our tree-structured filterbank [8]. As a result, our filterbank has the perfect reconstruction property, which is essential for an analysis/synthesis system, and also octave frequency-band division, which is important since the LP algorithm is especially error-prone in lower frequency bands. For our implementation, we used 32$^{\text{nd}}$ order LP filter for a 1024 sample frame (corresponding to about 23 msec. for 44.1 kHz sampling rate) for the full band analysis. For the subband analysis, we used an 8$^{\text{th}}$ order filter for each band, with a constant frame rate of 256 samples for each band (thus varying frame in msec.). The amount of overlapping for best quality was found to be 75% for all cases. These parameters were chosen so that the total number of transmitted coefficients for the resynthesized recording remains the same for both the full band and the subband cases. For the particular number of parameters used, the total number of coefficients used for the resynthesis is eight times less than the total number of audio samples. The coefficients that we intend to code for each microphone signal are the line spectral frequencies (LSF's) given their favorable quantization properties.

|  | ABX-1 | ABX-2 | ABX-3 |
|---|---|---|---|
| Results correct | 86% | 63% | 10% |

Table 1: Results from the ABX listening tests.

The NMI values in Fig. 1 are median values of the segmental NMI between the two residual signals using an analysis window of 6 msec. The residual signals are obtained using an overlap-add procedure so that they can be compared using the same analysis window. Our claim, that using a subband analysis with a small LP order for each band will produce much better modeling results than using a high LP order for the full frequency band, is greatly justified by the results shown. For the full band analysis we obtain a NMI value of 0.0956 while for a 8-band filterbank the median NMI is 0.5720 (40$^{\text{th}}$ order wavelet filters). In Fig.1 we plot the median NMI for different orders of the Daubechies filters. We can see that increasing the filter order results in slightly better results. Intuitively this was expected; an increase in the filter order results in better separation of the different bands, which is important since we model each subband signal independently of the others. In a similar experiment, we compared the residual signals in the time-domain and found that the median NMI doubles when using the 8-band system when compared to the full-band case. The results for both the frequency and time domains are similar regardless of the analysis window length for obtaining the NMI segmental values. When increasing the window size the NMI drops, which is expected since more data are compared. However, the decrease is similar for the various numbers of bands we tested.

In order to test the performance of our method, we also employed subjective (listening) tests, in which a total of 17 listeners participated (individually, using good quality headphones). We used the two concert hall recordings from the same performance as mentioned earlier (one capturing the male voices and one capturing the female voices of the chorus). We chose three parts of the performance (about 10 sec. each, referred to as Signals 1-3 here) where both parts of the chorus are active so that the two different microphone signals can be easily distinguished. For each signal we designed an ABX test, where A and B correspond to the male and female chorus recording (in random order), while each listener was asked to classify X as being closer to A or B regarding as to whether the male or female voices prevail in the recording. We tested 3 different types of wavelet-based filterbanks, namely 8-band with filters db40 (test ABX-1) and db4 (ABX-2), and 2-band with db40 (ABX-3). For each of these 3 tests, we used all three of the chosen signals, thus a total of 9 ABX tests was conducted per listener. The results are given in Table 1. We can conclude that the objective results, as well as the various claims made in the previous sections regarding the model, are verified by the listening tests. It is clear that

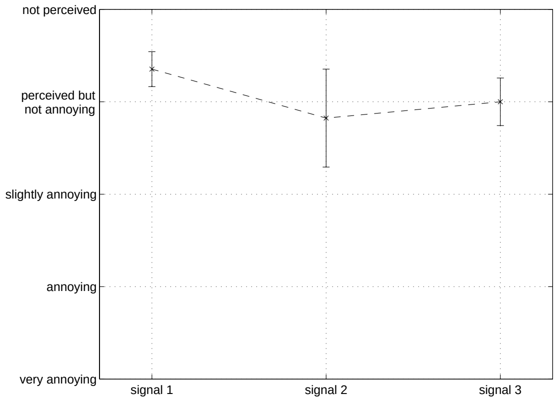

Figure 2: Results from the quality rating listening tests.

the 8-level wavelet-based filterbank produces excellent results when aliasing is limited (*i.e.* db40 case), although there is certainly room for improvement and further enhancement to our model is currently underway. On the other hand, when aliasing is high or when the number of bands (and thus the modeling accuracy) drops, the performance of the proposed method greatly deteriorates, not only in the sense of enhancing the male voices, but also regarding final quality (which most listeners noticed during the experiments). At this point we note that we obtained very similar results (using the NMI as well as informal listening tests) with a Laplacian pyramid filterbank, which is a different type of octave-spaced filterbank [9]. The choice of filterbank and whether octave-spaced filterbanks are indeed better for our model is a subject of our ongoing research.

We also conducted DCR-based (Degradation Category Rating) [10] listening tests for evaluating the quality of the resynthesized signals using a 5-grade scale in reference to the original recording (5 corresponding to being of same quality, and 1 to the lowest quality, when compared with the original male chorus recording). Subjects listened to the three sound clips (Signals 1-3), where the resynthesized signals were obtained using the best modeling parameters (8-level db40 wavelet-based). The results are depicted in Fig. 2, where graphical representations of the 95% confidence interval are shown (the x's mark the mean value and the two horizontal lines indicate the confidence limits). These results show clearly that the resynthesized signals are of high quality and the model does not seem to introduce any serious artifacts. It is important to note that we are currently working on the coding part based on our model. Our initial experiments have shown that bitrates in the order of 10 KBits/sec per channel are possible regarding the side information for good quality audio. However, these are only preliminary results, and we are confident that the bitrates can be further reduced.

## 5. CONCLUSIONS

A multiband source/filter model for multichannel audio was proposed, which can be used for resynthesizing the multiple microphone signals before they are mixed, and is thus tailored towards applications such as remote mixing and distributed musicians collaboration. The main advantage of the model is that it separates each microphone signal into a low-dimensional signal which mainly captures the microphone-specific properties, and a high-dimensional signal which mainly contains the interchannel similarities. Our model can result in a multichannel audio coding scheme where only one audio channel, along with side information of few KBits/sec per channel, can be decoded into the multiple channels of the original recording. The authors wish to thank the listening tests volunteers, and Prof. Kyriakakis of USC for his continued support of the project.

## 6. ACKNOWLEDGMENT

This work has been co-funded by the European Social Fund and Greek Government National Resources, and by a Marie Curie International Reintegration Grant within the 6th European Community Framework Program.

## REFERENCES

[1] J. D. Johnston and A. J. Ferreira, "Sum-difference stereo transform coding," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pp. 569–572, 1992.

[2] J. Herre, K. Brandenburg, and D. Lederer, "Intensity stereo coding," in *Proc.* 96*th Convention of the Audio Engineering Society (AES), preprint No. 3799*, 1994.

[3] D. Yang, H. Ai, C. Kyriakakis, and C.-C. J. Kuo, "High-fidelity multichannel audio coding with Karhunen-Loeve transform," *IEEE Trans. Speech and Audio Processing*, vol. 11, pp. 365–380, July 2003.

[4] F. Baumgarte and C. Faller, "Binaural cue coding - Part I: Psychoacoustic fundamentals and design principles and Part II: Schemes and applications," *IEEE Trans. Speech and Audio Processing*, vol. 11, pp. 509–531, November 2003.

[5] A. Sawchuk, E. Chew, R. Zimmermann, C. Papadopoulos, and C. Kyriakakis, "From remote media immersion to distributed immersive performance," in *Proc. ACM SIGMM Workshop on Experiential Telepresence (ETP)*, 2003.

[6] A. Mouchtaris, S. S. Narayanan, and C. Kyriakakis, "Virtual microphones for multichannel audio resynthesis," *EURASIP Jrnl. on Applied Signal Process., Special Issue on Digital Audio for Multimedia Comm.*, vol. 2003:10, pp. 968–979, 2003.

[7] S. Haykin, *Adaptive Filter Theory*. Prentice Hall, 1996.

[8] G. Strang and T. Nguyen, *Wavelets and Filter Banks*. Wellesley-Cambridge, 1996.

[9] P. J. Burt and E. H. Adelson, "The Laplacian pyramid as a compact image code," *IEEE Trans. Comm.*, pp. 532–540, 1983.

[10] W. B. Kleijn and K. K. Paliwal, eds., *Speech Coding and Synthesis*. Elsevier Science, 1995.