MULTIRESOLUTION SPECTRAL CONVERSION FOR MULTICHANNEL AUDIO RESYNTHESIS

Athanasios Mouchtaris, Shrikanth S. Narayanan, and Chris Kyriakakis

Integrated Media Systems Center
University of Southern California
3740 McClintock Ave., EEB 432
Los Angeles, CA 90089-2564, USA
mouchtar@sipi.usc.edu, shri@sipi.usc.edu, ckyriak@imsc.usc.edu

ABSTRACT

Multichannel audio is attracting rapidly increasing popularity in audio reproduction. In most cases, however, its transmission requirements are extremely demanding compared to the available bandwidth. One possible solution to this problem could be to transmit a reference channel and recreate the remaining channels at the receiving end. In this paper such a method is proposed by taking advantage of spectral conversion techniques that have been successfully applied to speech processing. Applications of the proposed system include transmission of multichannel audio over the current Internet infrastructure and, as an extension of the methods proposed here, remastering of existing monophonic and stereophonic recordings for multichannel rendering.

1. INTRODUCTION

Multichannel audio is one of the most demanding media types in terms of transmission requirements. A novel architecture allowing delivery of uncompressed multichannel audio over high-bandwidth communications networks has been presented in [1]. As suggested there, for applications in which bandwidth limitations prohibit transmission of multiple audio channels, an alternative is to transmit only one or two channels (denoted as *reference* channels or recordings here, *e.g.* the left and right signals in a traditional stereo recording) and reconstruct the remaining channels at the receiving end. The system proposed there partially provided a solution for reconstructing the channels of a specific recording from the reference channels and was particularly suitable for live concert hall performances. The proposed algorithm was based on information of the acoustics of the specific concert hall and the microphone locations with respect to the orchestra, information that was extracted from the specific multichannel recording.

In this paper the methods for recreating the channels of a multichannel recording proposed in [1] are extended. Before proceeding to the description of the algorithm proposed here, a brief outline of the previously published analysis is given. The reader is asked to examine Fig. 1, where an example is given of how microphones may be arranged in a recording venue in a multichannel recording. A number of microphones are used to capture several characteristics of the venue, resulting in an equal number of *stem recordings*. These are then mixed and played back through a multichannel audio system that recreates the spatial realism of the recording venue. Our objective is to design a system based on available stem recordings that is able to recreate them from the reference channels at the receiving end (thus, these are referred to as *target* recordings here). The result would be a significant reduction of transmission requirements, while at the same time mixing could take place at the receiving end. Consequently, such a system would be suitable for resynthesizing any number of channels in the initial recording. This is different than what commercial systems accomplish today. By examining the acoustical characteristics of the various microphone recordings, the distinction of microphones is made into reverberant and spot microphones.

Reverberant microphones are the microphones placed far from the sound source, for example C and D in Fig. 1. These microphones are treated separately as one category because they mainly capture reverberant information (that can be reproduced by the surround channels in a multichannel playback system). The recordings captured by these microphones can be synthesized by passing the reference recordings through a linear time-invariant (LTI) filter, designed as in [1].

Spot microphones are the microphones that are placed close to the sound source (*e.g.* G in Fig. 1). These microphones introduce an even more challenging situation. 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 acoustics of the hall. Synthesizing the recordings of these microphones, therefore, involves enhancing certain instruments and diminishing others, which in most cases overlap both in the time and frequency domains. The algorithm described here focuses on this problem and is based on spectral conversion (SC).

2. SPECTRAL CONVERSION

Based on the analysis given in the previous paragraph, the goal is to modify the short-term spectral properties of the reference audio signal in order to recreate the desired one. The short-term spectral properties are extracted by using a short sliding window with overlapping (resulting in a sequence of signal segments or frames). Each frame is modeled as an autoregressive (AR) filter excited by a residual signal. The AR filter coefficients are found by means of linear predictive analysis (LPC, [2]) and the residual signal is the result of inverse filtering the audio signal of the current frame

This research has been funded by the Integrated Media Systems Center, a National Science Foundation Engineering Research Center, Cooperative Agreement No. EEC-9529152.

by the AR filter. The LP coefficients are modified in a way to be described later in this section and the residual is filtered with the designed AR filter to produce the desired signal of the current frame. Finally, the desired response is synthesized from the designed frames using overlap-add techniques.

In order to obtain the desired response for each frame, an algorithm is required for converting the LP coefficients into the desired ones. Although the target coefficients in the application examined can be found by applying the same residual/LP analysis described (assuming that the reference and target waveforms are time-aligned), our intention is to design a mapping function based on the reference and target responses whose parameters will remain constant. The result will be a significant reduction of information as the target response can be reconstructed using the reference signal and this function.

Such a mapping function can be designed by following the approach of voice conversion algorithms [3, 4, 5]. The objective of voice conversion is to modify a speech waveform so that the context remains as is but appears to be spoken from a specific speaker. Although the application is completely different, the approach followed is very suitable for our system. In voice conversion pitch and time-scaling need to be considered, while in the application examined here this is not necessary since the reference and target waveforms come from the same excitation recorded with different microphones and the need is not to modify but to *enhance* the reference waveform. However, in both cases, there is the need to modify the short-term spectral properties of the waveform. The method to do that is briefly described next.

Assuming that a sequence $[\boldsymbol{x}_1\boldsymbol{x}_2\ldots\boldsymbol{x}_n]$ of reference spectral vectors (*e.g.* line spectral frequencies (LSF's), cepstral coefficients, *etc.*) is given, as well as the corresponding sequence of target spectral vectors $[\boldsymbol{y}_1\boldsymbol{y}_2\ldots\boldsymbol{y}_n]$ (training data from the reference and target recordings respectively), a function $\mathcal{F}(\cdot)$ can be designed which, when applied to vector $\boldsymbol{x}_k$, produces a vector close in some sense to vector $\boldsymbol{y}_k$. Many algorithms have been described for designing this function (see [3, 4, 5] and the references therein). Here the algorithms based on vector quantization (VQ, [3]) and Gaussian mixture models (GMM, [4, 5]) were implemented and compared.

2.1. Spectral Conversion based on VQ

Under this approach, the spectral vectors of the reference and target signals (training data) are vector quantized respectively using the well-known modified K-means clustering algorithm (see for example [6] for details). Then, a histogram is created indicating the correspondences between the reference and target centroids. Finally, the function $\mathcal{F}$ is defined as the linear combination of the target centroids using the designed histogram as a weighting function. It is important to mention that in this case the spectral vectors were chosen to be the cepstral coefficients so that the distance measure used in clustering is the truncated cepstral distance.

2.2. Spectral Conversion based on GMM

In this case, the assumption made is that the sequence of spectral vectors $\boldsymbol{x}_k$ is a realization of a random vector $\boldsymbol{x}$ with a probability density function (pdf) that can be modeled as a mixture of M multivariate Gaussian pdf's. Thus, the pdf of $\boldsymbol{x}$, $\mathrm{g}(\boldsymbol{x})$, can be written

$$\mathrm{g}(\boldsymbol{x}) = \sum_{i=1}^{M} p(\omega_i)\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu}_i^x,\boldsymbol{\Sigma}_i^{xx}) \qquad (1)$$

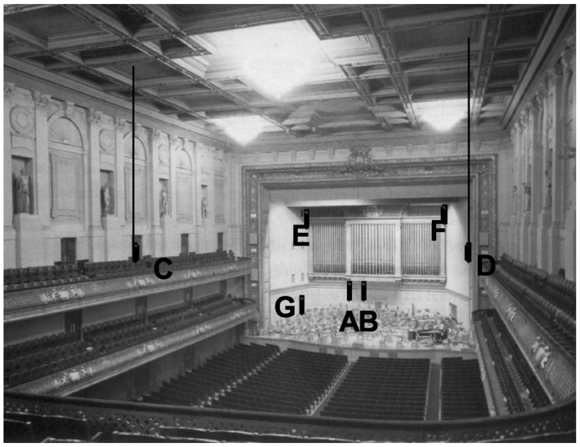

Figure 1: An example of how microphones may be arranged in a recording venue for a multichannel recording.

where, $\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$ is the normal multivariate distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ and $p(\omega_i)$ is the prior probability of class ω_i. The parameters of the GMM, *i.e.* the mean vectors, covariance matrices and priors, can be estimated by applying the expectation maximization (EM) algorithm [7] to the training data.

As already mentioned, the function $\mathcal{F}$ is designed such that the spectral vectors $\boldsymbol{y}_k$ and $\mathcal{F}(\boldsymbol{x}_k)$ are close in some sense. In [4], the function $\mathcal{F}$ is designed such that the error

$$\mathcal{E} = \sum_{k=1}^{n} \|\boldsymbol{y}_k - \mathcal{F}(\boldsymbol{x}_k)\|^2 \qquad (2)$$

is minimized. Since this method is based on least-squares estimation, it will be denoted as the LSE method. This problem becomes possible to solve under the constraint that $\mathcal{F}$ is piecewise linear,

$$\mathcal{F}(\boldsymbol{x}_k) = \sum_{i=1}^{M} p(\omega_i|\boldsymbol{x}_k)\left[\boldsymbol{v}_i + \boldsymbol{\Gamma}_i{\boldsymbol{\Sigma}_i^{xx}}^{-1}(\boldsymbol{x}_k - \boldsymbol{\mu}_i^x)\right] \qquad (3)$$

where the conditional probability that a given vector $\boldsymbol{x}_k$ belongs to class ω_i, $p(\omega_i|\boldsymbol{x}_k)$ can be computed by applying Bayes' theorem

$$p(\omega_i|\boldsymbol{x}_k) = \frac{p(\omega_i)\mathcal{N}(\boldsymbol{x}_k;\boldsymbol{\mu}_i^x,\boldsymbol{\Sigma}_i^{xx})}{\sum_{j=1}^{M} p(\omega_j)\mathcal{N}(\boldsymbol{x}_k;\boldsymbol{\mu}_j^x,\boldsymbol{\Sigma}_j^{xx})} \qquad (4)$$

The unknown parameters ($\boldsymbol{v}_i$ and $\boldsymbol{\Gamma}_i$, $i = 1,\ldots,M$) can be found by minimizing (2) which reduces to a least-squares equation.

A different solution for function $\mathcal{F}$ results when a different function than (2) is minimized [5]. Assuming that $\boldsymbol{x}$ and $\boldsymbol{y}$ are jointly Gaussian for each class ω_i, then, in mean-squared sense, the optimal choice for the function $\mathcal{F}$ is

$$\begin{aligned}\mathcal{F}(\boldsymbol{x}_k) &= \mathrm{E}(\boldsymbol{y}|\boldsymbol{x}_k) \qquad (5)\\ &= \sum_{i=1}^{M} p(\omega_i|\boldsymbol{x}_k)\left[\boldsymbol{\mu}_i^y + \boldsymbol{\Sigma}_i^{yx}{\boldsymbol{\Sigma}_i^{xx}}^{-1}(\boldsymbol{x}_k - \boldsymbol{\mu}_i^x)\right]\end{aligned}$$

where $\mathrm{E}(\cdot)$ denotes the expectation operator and the conditional probabilities $p(\omega_i|\boldsymbol{x}_k)$ are given again from (4). If the source and

Band	Frequency Range		LPC	GMM
Nr.	Low (kHz)	High (kHz)	Order	Centroids
1	0.0000	0.1723	4	4
2	0.1723	0.3446	4	4
3	0.3446	0.6891	8	8
4	0.6891	1.3782	16	16
5	1.3782	2.7563	32	16
6	2.7563	5.5125	32	16
7	5.5125	11.0250	32	16
8	11.0250	22.0500	32	16

Table 1: Parameters for chorus microphone example.

target vectors are concatenated, creating a new sequence of vectors $\boldsymbol{z}_k$ that are the realizations of the random vector $\boldsymbol{z} = [\boldsymbol{x}^T \boldsymbol{y}^T]^T$ (where ^{T} denotes transposition), then all the required parameters in the above equations can be found by estimating the GMM parameters of $\boldsymbol{z}$. Then,

$$\boldsymbol{\Sigma}_i^{zz} = \begin{bmatrix} \boldsymbol{\Sigma}_i^{xx} & \boldsymbol{\Sigma}_i^{xy} \\ \boldsymbol{\Sigma}_i^{yx} & \boldsymbol{\Sigma}_i^{yy} \end{bmatrix}, \boldsymbol{\mu}_i^z = \begin{bmatrix} \boldsymbol{\mu}_i^x \\ \boldsymbol{\mu}_i^y \end{bmatrix} \qquad (6)$$

Once again, these parameters are estimated by the EM algorithm. Since this method estimates the desired function based on the joint density of $\boldsymbol{x}$ and $\boldsymbol{y}$, it will be referred to as the JDE method.

2.3. Subband Processing

Audio signals contain more information than speech signals. The sampling rate for audio signals is usually 44.1 or 48 kHz compared to 16 kHz for speech. Moreover, high acoustical quality for audio applications is essential. For these reasons, the decision to follow an analysis in subbands seems natural. Instead of warping the frequency spectrum using the Bark scale as is usual in speech analysis, the frequency spectrum was divided in subbands and each one was treated separately under the analysis of the previous paragraphs. Perfect reconstruction filter banks, based on wavelets [8], provide a solution with acceptable computational complexity as well as the appropriate, for audio signals, octave frequency division. The choice of filter bank was not a subject of investigation but steep transition is a desirable property.

2.4. Residual Processing for Percussive Sounds

The SC methods described earlier will not produce the desired result in all cases. One such case of particular importance is the case of percussive drum-like sounds. It is usual in multichannel recordings to place a microphone close to the tympani as drum-like sounds are considered perceptually important in recreating the acoustical environment of the recording venue. For percussive sounds, a similar model to the residual/LP model described here can be used [9], but for the enhancement purposes investigated in this paper, the emphasis is given to the residual instead of the LP parameters. It is proposed to extract the residual of an instance of the particular percussive instrument from the recording of the microphone that captures this instrument and then recreate this channel from the reference channel by simply substituting the residual of all instances of this instrument with the extracted residual. As explained in [9], this residual corresponds to the interaction between the exciter and the resonating body of the instrument and lasts until the structure reaches a steady vibration. This signal characterizes the attack part of the sound and is independent of the frequencies and amplitudes of the harmonics of the produced sound (after the instrument has reached a steady vibration). Thus, it can be used for synthesizing different sounds by using an appropriate all-pole filter. This method was successfully tested and more details are given in the next section. The drawback of this approach is that a robust algorithm is required for identifying the particular instrument instances in the reference recording.

SC	Cepstral Distance		Centroids
Method	Train	Test	per Band
LSE	0.6451	0.7144	Table 1
JDE	0.6629	0.7445	Table 1
VQ	1.2903	1.3338	1024

Table 2: Normalized distances for LSE-, JDE- and VQ-based methods.

3. IMPLEMENTATION DETAILS

The three spectral conversion methods outlined in Section 2 were implemented and tested using a multichannel recording, as described in the first section of this paper. The objective was to recreate the recording that mainly captured the chorus of the orchestra (residual processing for percussive sound resynthesis is also considered at the last paragraph of this section). Acoustically, therefore, the emphasis was on the male and female voices. At the same time, it was clear that some instruments, inaudible in the target recording but particularly audible in the reference recording, needed to be attenuated. A database of about 10,000 spectral vectors for each band was carefully created (so that only parts of the recording where the chorus is present are used) with the choice of spectral vectors being the cepstral coefficients. Parts of the chorus recording were selected so that there were no segments of silence included. Results were evaluated through informal listening tests and through objective performance criteria. The SC methods were shown to provide promising enhancement results. Formal listening tests are currently underway and will be available in the near future. The experimental conditions are given in Table 1. The number of octave bands used was 8, a choice that gives particular emphasis on the frequency band 0-5 kHz and at the same time does not impose excessive computational demands. The frequency range 0-5 kHz is particularly important for the specific case of chorus recording resynthesis since this is the frequency range where the human voice is mostly concentrated. For producing good quality results, the entire frequency range 0-20 kHz must be considered. The order of the LPC filter varied depending on the frequency detail of each band and for the same reason the number of centroids for each band was different.

In Table 2, the average quadratic cepstral distance (averaged over all vectors and all 8 bands) is given for each method, for the training data as well as for the data used for testing (9 sec. of music from the same recording). The cepstral distance is normalized with the average quadratic distance between the reference and the target waveforms (*i.e.* without any conversion of the LPC parameters). The improvement is large for both the GMM-based algorithms, with the LSE algorithm being slightly better, for both the training and testing data. The VQ-based algorithm, in contrast, produced a deterioration in performance which was audible

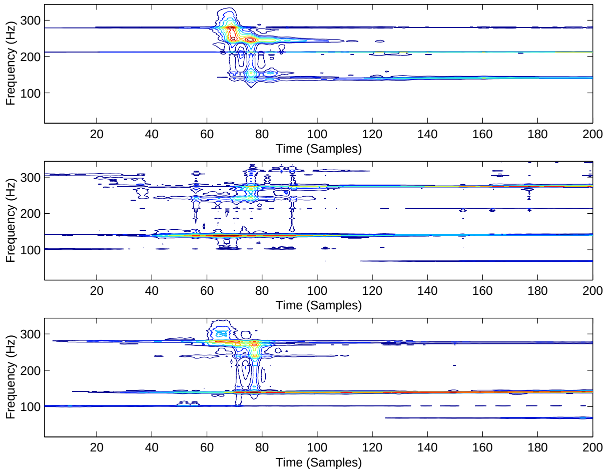

Figure 2: Choi-Williams distribution of the desired (top), reference (middle) and synthesized (bottom) waveforms at the time points during a tympani strike (samples 60-80).

as well. This can be explained based on the fact that the GMM-based methods result in a conversion function which is continuous with respect to the spectral vectors. The VQ-based method, on the other hand, produces audible artifacts introduced by spectral discontinuities because the conversion is based on a limited number of existing spectral vectors. This is the reason why a large number of centroids was used for the VQ-based algorithm as seen in Table 2 compared to the number of centroids used for the GMM-based algorithms. However, the results were still unacceptable both from the objective and subjective perspectives.

The algorithm described in Section 2 considering the special case of percussive sound resynthesis was tested as well. Fig. 2 shows the time-frequency evolution of a tympani instance using the Choi-Williams distribution [10], a distribution that achieves the high resolution needed in such cases of impulsive nature. Fig. 2 clearly demonstrates the improvement in drum-like sound resynthesis. The impulsiveness of the signal at around samples 60-80 is observed in the desired response and verified in the synthesized waveform. The attack part is clearly enhanced, significantly adding naturalness in the audio signal, as our informal listening tests demonstrated. Formal listening tests demonstrating the perceptual benefits of this method will be described in a future publication.

4. FUTURE RESEARCH DIRECTIONS

Multichannel audio resynthesis is a new and important application that allows transmission of only one or two channels of multichannel audio and resynthesis of the remaining channels at the receiving end. Spectral conversion algorithms that have been used successfully for voice conversion can be adopted for the task of multichannel audio resynthesis quite favorably. Three of the most common spectral conversion methods have been compared and our objective results, in accordance with our informal listening tests, have indicated that GMM-based spectral conversion can produce extremely successful results. Residual signal enhancement was also found to be essential for the case of percussive sounds.

Our current research is focused on audio quality improvement for the proposed methods. Towards this direction, one possible alternative to the residual/LP model proposed here, is the sinusoidal model of [11]. This model follows the short-term analysis of the residual/LP model, modeling each frame of the audio signal as a summation of r sinusoids with additive noise. The additive noise corresponds to the modeling error and can be represented with simple models quite effectively. A multiresolution sinusoidal model has also been proposed [12], analogous to the multiresolution residual/LP model described here. Other possible directions for future research include extensions of this research for the purpose of remastering existing monophonic and stereophonic recordings for multichannel rendering.

5. REFERENCES

[1] A. Mouchtaris, Z. Zhu, and C. Kyriakakis, "High-quality multichannel audio over the Internet," in *Conf. Record of the Thirty-Third Assilomar Conf. Signals, Systems and Computers*, Pacific Grove, CA, October 1999, vol. 1, pp. 347–351.

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

[3] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *IEEE Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, New York, NY, April 1988, pp. 655–658.

[4] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, March 1998.

[5] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *IEEE Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Seattle, WA, May 1998, pp. 285–289.

[6] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, NJ, 1993.

[7] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, January 1995.

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

[9] J. Laroche and J.-L. Meillier, "Multichannel excitation/filter modeling of percussive sounds with application to the piano," *IEEE Trans. Speech and Audio Processing*, vol. 2, pp. 329–344, 1994.

[10] H.-I. Choi and J. Williams, "Improved time-frequency representation of multicomponent signals using exponential kernels," *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. 37, no. 6, pp. 862–871, June 1989.

[11] X. Serra and J. O. Smith III, "Spectral modeling sythesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Computer Music Journal*, vol. 14, no. 4, pp. 12–24, Winter 1990.

[12] S. N. Levine, T. S. Verma, and J. O. Smith III, "Multiresolution sinusoidal modeling for wideband audio with modifications," in *IEEE Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Seattle, WA, May 1998, pp. 3585–3588.