Maximum Likelihood Constrained Adaptation for Multichannel Audio Synthesis

A. Mouchtaris, S. S. Narayanan, and C. Kyriakakis
Integrated Media Systems Center
University of Southern California
Los Angeles, CA 90089-2564

Abstract

Multichannel audio can immerse a group of listeners in a seamless aural environment. Previously, we proposed a system capable of synthesizing the multiple channels of a virtual multichannel recording from a smaller set of reference recordings. This problem was termed multichannel audio resynthesis and the application was to reduce the excessive transmission requirements of multichannel audio. In this paper, we address the more general problem of multichannel audio synthesis, i.e. how to completely synthesize a multichannel audio recording from a specific stereophonic or monophonic recording, significantly enhancing the recording's quality. We approach this problem by extending the model employed for the resynthesis problem.

1 Introduction

Multichannel audio can enhance the sense of immersion for a group of listeners by reproducing the sounds that would originate from several directions around the listeners, thus simulating the way we perceive sound in a real acoustical space. However, several key issues must be addressed. Multichannel audio imposes excessive requirements to the transmission medium. A system we previously proposed [1, 2], attempted to address this issue by offering the alternative to synthesize the multiple channels of a multichannel recording from a smaller set of signals (*e.g.* the left and right ORTF microphone signals in a traditional stereophonic recording). The solution provided, termed multichannel audio *resynthesis*, concentrated on the problem of enhancing a concert hall recording and divided the problem in two different parts, depending on the characteristics of the recording to be synthesized. Given the microphone recordings from several locations of the venue (stem recordings, see Fig. 1 for an example of how microphones may be arranged in a recording venue for a multichannel recording), our objective was to design a system that can resynthesize these recordings from the reference recordings. These resynthesized stem recordings are then mixed in order to produce the final multichannel audio recording. The distinction of the recordings was made depending on the location of the microphone in the venue, thus resulting in two different categories, namely reverberant and spot microphone recordings. For simulating recordings of microphones placed far from the orchestra (reverberant microphones, *e.g.* C and D in Fig. 1), infinite impulse response (IIR) filters were designed from existing multichannel recordings made in a particular concert hall. The IIR filters designed were shown to be capable of recreating the acoustical properties of the venue at specific locations. In order to simulate virtual microphones in several locations close and around the orchestra (spot microphones, *e.g.* G in Fig. 1), it is important to design time-varying filters that can track and enhance particular musical instruments and diminish others.

In this paper, we address the more general problem of multichannel audio synthesis. The goal is to convert existing stereophonic or monophonic recordings into multichannel, given that to-date only a handful of music recordings have been made with multiple channels. The same approach is followed as in the resynthesis problem. Based on existing multichannel recordings, we decide which microphone locations must be synthesized. For reverberant microphones, the filters designed in the resynthesis problem can be readily applied to arbitrary recordings. Their time-invariant nature offers the advantage that these filters can be applied to various recordings while having been designed based on a given recording. In contrast, the time-varying nature of the methods designed for spot microphone resynthesis, prohibits us from applying them in an arbitrary recording. This is the problem that we focus on in this paper.

The block diagram of Fig. 2 can serve as a guide to the methods examined in this paper. The part of the diagram to the left of the dotted line corresponds to an existing multimicrophone recording. Multichannel audio resynthesis allows us to reconstruct the stem recordings (target channels) from the reference channel. The part of the diagram to the right of the dotted line, corresponds to multichannel audio synthesis, which is used to fully synthesize stem recordings from the reference channel of a stereo recording. Our approach is to take advantage of the resynthesis parameters that have been derived based on an existing target channel. In order to achieve that, the stereo and multimicrophone recordings are related with the GMM constrained estimation method that is analyzed later in this paper. The adaptation assumption is also needed that relates the (unavailable) target response of the stereo recording with the target response of the multimicrophone recording.

2 Spectral Conversion

The approach followed for spot microphone resynthesis is based on spectral conversion methods that have been successfully employed in speech synthesis applications [3, 4, 5]. A training data set is created from the existing reference and the target recordings by applying a short sliding window and extracting the parameters that model the short-time spectral envelope (in this paper we use the cepstral coefficients).This set is created based on the parts of the target recording that must be enhanced in the reference recording. If for example the emphasis is on enhancing the chorus of the orchestra, then the training set is created by choosing parts of the recording where the chorus is present. This procedure results in two vector sequences, $[\boldsymbol{x}_1\boldsymbol{x}_2\dots\boldsymbol{x}_n]$ of reference spectral vectors and $[\boldsymbol{y}_1\boldsymbol{y}_2\dots\boldsymbol{y}_n]$, as the corresponding sequence of target spectral vectors. 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). In [2] the algorithms based on Gaussian mixture models (GMM, [4, 5]) were found to be very suitable for the resynthesis problem.

According to GMM-based algorithms, a sequence of spectral vectors $\boldsymbol{x}_k$ as above, can be considered as a realization of a random vector $\boldsymbol{x}$ with probability density function (pdf) as GMM

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

where, $\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$ is the multivariate normal 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 using the expectation maximization (EM) algorithm [6].

The analysis that follows focuses on the conversion of [5], which offers great insight as to what the conversion parameters represent. 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

$$\begin{aligned}\mathcal{F}(\boldsymbol{x}_k) &= \mathrm{E}(\boldsymbol{y}|\boldsymbol{x}_k) \qquad (2)\\ &= \sum_{i=1}^{M} p(\omega_i|\boldsymbol{x}_k)\left[\boldsymbol{\mu}_i^y + \boldsymbol{\Sigma}_i^{yx}{\boldsymbol{\Sigma}_i^{xx}}^{-1}(\boldsymbol{x}_k-\boldsymbol{\mu}_i^x)\right]\end{aligned}$$

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

$$p(\omega_i|\boldsymbol{x}_k) = \frac{p(\omega_i)\mathcal{N}(\boldsymbol{x}_k;\boldsymbol{\mu}_i^x,\boldsymbol{\Sigma}_i^{xx})}{\sum_{j=1}^{M} p(\omega_j)\mathcal{N}(\boldsymbol{x}_k;\boldsymbol{\mu}_j^x,\boldsymbol{\Sigma}_j^{xx})} \tag{3}$$

If the source and 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,

$$\boldsymbol{\Sigma}_i^{zz} = \begin{bmatrix} \boldsymbol{\Sigma}_i^{xx} & \boldsymbol{\Sigma}_i^{xy} \\ \boldsymbol{\Sigma}_i^{yx} & \boldsymbol{\Sigma}_i^{yy} \end{bmatrix},\ \boldsymbol{\mu}_i^z = \begin{bmatrix} \boldsymbol{\mu}_i^x \\ \boldsymbol{\mu}_i^y \end{bmatrix} \tag{4}$$

The EM algorithm is applied to $\boldsymbol{z}$. Since this method estimates the desired function based on the joint density of $\boldsymbol{x}$ and $\boldsymbol{y}$, it is denoted as the JDE method.

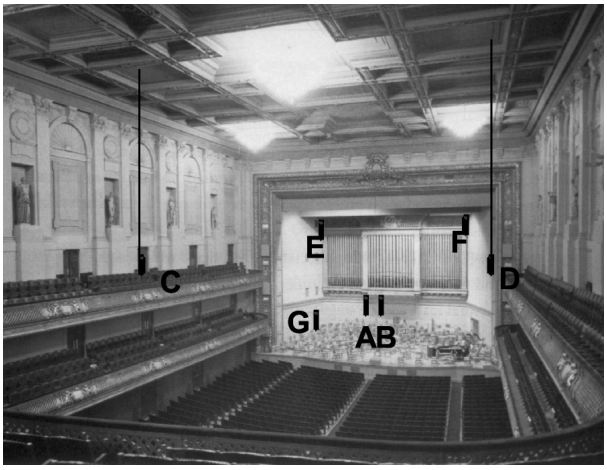

Figure 1: An example of how microphones may be arranged in a recording venue for a multichannel recording.

2.1 Diagonal Implementation

The JDE spectral conversion algorithm can be implemented with the covariance matrices having no structural restrictions or restricted to be diagonal, denoted as full and diagonal conversion respectively. Full conversion is of prohibiting computational complexity when combined with the adaptation algorithm for the synthesis problem examined in the next section. Note that the covariance matrix of $\boldsymbol{z}$ for the JDE method cannot be diagonal because this method is based on the cross-covariance of $\boldsymbol{x}$ and $\boldsymbol{y}$ which is found from (4). This will be zero if the covariance of $\boldsymbol{z}$ is diagonal. Thus, in order to obtain an efficient structure, we must restrict *each* of the matrices $\boldsymbol{\Sigma}_i^{xx}$, $\boldsymbol{\Sigma}_i^{yy}$, $\boldsymbol{\Sigma}_i^{xy}$, and $\boldsymbol{\Sigma}_i^{yx}$ in (4) to be diagonal. For achieving this restriction, we slightly modified the EM algorithm, with the most noteworthy modification being that of obtaining the inverse of $\boldsymbol{\Sigma}_i^{zz}$ by taking advantage of its structure. It is very easy to show [7], that the inverse of $\boldsymbol{\Sigma}_i^{zz}$ will be

$${\boldsymbol{\Sigma}_i^{zz}}^{-1} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{bmatrix} \tag{5}$$

where

$$\begin{aligned}\mathbf{A} &= \left(\boldsymbol{\Sigma}_i^{xx} - \boldsymbol{\Sigma}_i^{xy}{\boldsymbol{\Sigma}_i^{yy}}^{-1}\boldsymbol{\Sigma}_i^{yx}\right)^{-1}\\ \mathbf{B} &= -\mathbf{A}\boldsymbol{\Sigma}_i^{xy}{\boldsymbol{\Sigma}_i^{yy}}^{-1} = -{\boldsymbol{\Sigma}_i^{xx}}^{-1}\boldsymbol{\Sigma}_i^{xy}\mathbf{C}\\ \mathbf{C} &= \left(\boldsymbol{\Sigma}_i^{yy} - \boldsymbol{\Sigma}_i^{yx}{\boldsymbol{\Sigma}_i^{xx}}^{-1}\boldsymbol{\Sigma}_i^{xy}\right)^{-1} \qquad (6)\end{aligned}$$

In the above equations, all matrices, thus their products, sums, and differences are diagonal, so the inversions will be of very low computational demands. Based on this structure for the inverse of $\mathbf{\Sigma}_i^{zz}$, the joint pdf of $\boldsymbol{x}$ and $\boldsymbol{y}$ can be written as

$$\mathrm{g}(\boldsymbol{x},\boldsymbol{y}) = \frac{\exp\left(-\frac{1}{2}\left(\boldsymbol{x}^T\mathbf{A}\boldsymbol{x} + \boldsymbol{y}^T\mathbf{C}\boldsymbol{y} + 2\boldsymbol{x}^T\mathbf{B}\boldsymbol{y}\right)\right)}{(2\pi)^K\sqrt{|\mathbf{\Sigma}_i^{zz}|}} \quad (7)$$

K being the dimensionality of $\boldsymbol{x}$, and the determinant of $\mathbf{\Sigma}_i^{zz}$ equals

$$|\mathbf{\Sigma}_i^{zz}| = |\mathbf{\Sigma}_i^{yy}||\mathbf{\Sigma}_i^{xx} - \mathbf{\Sigma}_i^{xy}\mathbf{\Sigma}_i^{yy^{-1}}\mathbf{\Sigma}_i^{yx}| \quad (8)$$

3 ML Constrained Adaptation

The above approach offers a possible solution to the issue of multichannel audio transmission by allowing transmission of only one or two reference channels along with the filters that can subsequently be used to recreate the remaining channels at the receiving end (virtual microphone resynthesis). Here, we are interested in addressing the issue of virtual microphone synthesis, *i.e.*, applying these filters to arbitrary monophonic or stereophonic recordings in order to enhance particular instrument types and completely synthesize a multichannel recording. This step requires an algorithm that generalizes these filters. In the synthesis case, no target training data will be available so some assumptions must be explicitly made about the target recording. Our approach is to derive a transformation between the reference recording used in the training step of the resynthesis algorithm and the reference recording to be used for the synthesis algorithm, that in some way represents the statistical correspondence between these two recordings. We then assume that the same transformation holds for the two corresponding target recordings and practically test this hypothesis. Techniques for deriving such transformations have been successfully applied in the task of speaker adaptation for speech recognition. In this paper we applied the maximum-likelihood constrained adaptation method [8, 9], which offers the advantage of a simple probabilistic linear transformation leading to a mathematically tractable solution for the synthesis problem.

As in the resynthesis case, we obtain a sequence of spectral vectors from the reference channel of an available multimicrophone recording. These vectors are considered as realizations of a random vector $\boldsymbol{x}$, which is modeled with a GMM as in (1). From the reference channel of the *stereo* recording we also obtain a sequence of spectral vectors, considered as realizations of random vector $\boldsymbol{x}'$. In this manner, we also obtain random vector $\boldsymbol{y}$ from the desired response of the multimicrophone recording, and we denote as $\boldsymbol{y}'$ the random vector that corresponds to the (not available) desired response of the stereo recording. Instead of applying a GMM for $\boldsymbol{x}'$, we attempt to relate the random variables $\boldsymbol{x}'$ and $\boldsymbol{x}$, the motivation being to derive a transformation that relates $\boldsymbol{y}'$ with $\boldsymbol{y}$. We assume that the target random vector $\boldsymbol{x}'$ is related to reference random vector $\boldsymbol{x}$ by a probabilistic linear transformation

$$\boldsymbol{x}' = \begin{cases} \mathbf{A}_1\boldsymbol{x} + \boldsymbol{b}_1 & \text{with probability } p(\lambda_1|\omega_i) \\ \mathbf{A}_2\boldsymbol{x} + \boldsymbol{b}_2 & \text{with probability } p(\lambda_2|\omega_i) \\ \vdots & \vdots \\ \mathbf{A}_N\boldsymbol{x} + \boldsymbol{b}_N & \text{with probability } p(\lambda_N|\omega_i) \end{cases} \quad (9)$$

This equation corresponds to the GMM constrained estimation that relates $\boldsymbol{x}'$ with $\boldsymbol{x}$ in the block diagram of Fig. 2. In the above equation, $\mathbf{A}_j$ denotes a $K \times K$ dimensional matrix (K is the number of components of vector $\boldsymbol{x}$), and $\boldsymbol{b}_j$ is a vector of the same dimension with $\boldsymbol{x}$. Each of the component transformations j is related with a specific Gaussian i of $\boldsymbol{x}$ with probability $p(\lambda_j|\omega_i)$ which satisfy

$$\sum_{j=1}^{N} p(\lambda_j|\omega_i) = 1, \quad i = 1, \ldots, M \quad (10)$$

where M is the number of Gaussians of the GMM that corresponds to the reference vector sequence. Clearly,

$$\mathrm{g}(\boldsymbol{x}'|\omega_i, \lambda_j) = \mathcal{N}(\boldsymbol{x}'; \mathbf{A}_j\boldsymbol{\mu}_i^x + \boldsymbol{b}_j, \mathbf{A}_j\mathbf{\Sigma}_i^{xx}\mathbf{A}_j^T) \quad (11)$$

resulting in the pdf of $\boldsymbol{x}'$

$$\mathrm{g}(\boldsymbol{x}') = \sum_{i=1}^{M}\sum_{j=1}^{N} p(\omega_i)p(\lambda_j|\omega_i)\mathcal{N}(\boldsymbol{x}'; \mathbf{A}_j\boldsymbol{\mu}_i^x + \boldsymbol{b}_j, \mathbf{A}_j\mathbf{\Sigma}_i^{xx}\mathbf{A}_j^T) \quad (12)$$

The matrices $\mathbf{A}_j$, the vectors $\boldsymbol{b}_j$ and the conditional probabilities $p(\omega_i)$ and $p(\lambda_j|\omega_i)$ can be estimated using maximum likelihood estimation techniques. The EM algorithm can be applied to this case in a similar manner to estimating the parameters of a GMM from observed data. In essence, it is a linearly constrained maximum-likelihood estimation of the GMM parameters.

The purpose of adopting the transformation (9) is to use it in order to obtain a target training sequence for the synthesis problem. The assumption is that this function represents the statistical correspondence between the two available recordings. It is then justifiable to apply the same function to the target recording of the multichannel recording to obtain a reference recording for the synthesis problem. The synthesis problem then can be simply solved if the conversion methods mentioned in the previous section are employed. In other words, the assumption made is that the target vector $\boldsymbol{y}'$ for the synthesis problem can be obtained from the available target vector $\boldsymbol{y}$ by

$$\boldsymbol{y}' = \begin{cases} \mathbf{A}_1\boldsymbol{y} + \boldsymbol{b}_1 & \text{with probability } p(\lambda_1|\omega_i) \\ \mathbf{A}_2\boldsymbol{y} + \boldsymbol{b}_2 & \text{with probability } p(\lambda_2|\omega_i) \\ \vdots & \vdots \\ \mathbf{A}_N\boldsymbol{y} + \boldsymbol{b}_N & \text{with probability } p(\lambda_N|\omega_i) \end{cases} \quad (13)$$

This equation corresponds to the adaptation assumption that relates $\boldsymbol{y}'$ with $\boldsymbol{y}$ in the block diagram of Fig. 2.

Note that classes ω_i are the same for $\boldsymbol{x}$ and $\boldsymbol{y}$ since they were assumed to be jointly Gaussian in Section 2. Under this assumption and given the linearity of the transformations (9) and (13), $\boldsymbol{x}^{'}$ and $\boldsymbol{y}^{'}$ will also be jointly Gaussian for a particular class ω_i and λ_j. The pdf of $\boldsymbol{y}^{'}$ will be

$$\mathrm{g}(\boldsymbol{y}^{'}) = \sum_{i=1}^{M}\sum_{j=1}^{N} p(\omega_i)p(\lambda_j|\omega_i)\mathcal{N}(\boldsymbol{y}^{'}; \mathbf{A}_j\boldsymbol{\mu}_i^{y}+\boldsymbol{b}_j, \mathbf{A}_j\boldsymbol{\Sigma}_i^{yy}\mathbf{A}_j^{T}) \tag{14}$$

It is now possible to derive the conversion function for the synthesis problem, based entirely on the parameters derived during the resynthesis stage that correspond to a completely different recording. For a particular class ω_i and λ_j, $\boldsymbol{x}^{'}$ and $\boldsymbol{y}^{'}$ will be jointly Gaussian, thus

$$\begin{aligned}\mathrm{E}(\boldsymbol{y}^{'}|\boldsymbol{x}_k^{'},\omega_i,\lambda_j) &= \boldsymbol{\mu}_i^{y^{'}} + \boldsymbol{\Sigma}_i^{y^{'}x^{'}}\boldsymbol{\Sigma}_i^{x^{'}x^{'}{}^{-1}}\left(\boldsymbol{x}_k^{'}-\boldsymbol{\mu}_i^{x^{'}}\right)\\ &= \mathbf{A}_j\boldsymbol{\mu}_i^{y} + \boldsymbol{b}_j + \mathbf{A}_j\boldsymbol{\Sigma}_i^{yx}\boldsymbol{\Sigma}_i^{xx^{-1}}\mathbf{A}_j^{-1}\\ &\quad \left(\boldsymbol{x}_k^{'}-\mathbf{A}_j\boldsymbol{\mu}_i^{x}-\boldsymbol{b}_j\right)\end{aligned} \tag{15}$$

since

$$\boldsymbol{\Sigma}_i^{y^{'}x^{'}} = \mathbf{A}_j\boldsymbol{\Sigma}_i^{yx}\mathbf{A}_j^{T},\ \boldsymbol{\Sigma}_i^{x^{'}x^{'}} = \mathbf{A}_j\boldsymbol{\Sigma}_i^{xx}\mathbf{A}_j^{T} \tag{16}$$

and

$$\boldsymbol{\mu}_i^{y^{'}} = \mathbf{A}_j\boldsymbol{\mu}_i^{y}+\boldsymbol{b}_j,\ \boldsymbol{\mu}_i^{x^{'}} = \mathbf{A}_j\boldsymbol{\mu}_i^{x}+\boldsymbol{b}_j \tag{17}$$

Finally, the conversion function for synthesis will be

$$\begin{aligned}\mathcal{F}(\boldsymbol{x}_k^{'}) &= \mathrm{E}(\boldsymbol{y}^{'}|\boldsymbol{x}_k^{'}) \qquad (18)\\ &= \sum_{i=1}^{M}\sum_{j=1}^{N} p(\omega_i|\boldsymbol{x}_k^{'})p(\lambda_j|\boldsymbol{x}_k^{'},\omega_i)\Big[\mathbf{A}_j\boldsymbol{\mu}_i^{y}+\boldsymbol{b}_j+\\ &\quad \mathbf{A}_j\boldsymbol{\Sigma}_i^{yx}\boldsymbol{\Sigma}_i^{xx^{-1}}\mathbf{A}_j^{-1}\left(\boldsymbol{x}_k^{'}-\mathbf{A}_j\boldsymbol{\mu}_i^{x}-\boldsymbol{b}_j\right)\Big]\end{aligned}$$

where

$$p(\omega_i|\boldsymbol{x}_k^{'}) = \frac{p(\omega_i)\sum_{j=1}^{N}p(\lambda_j|\omega_i)\mathrm{g}(\boldsymbol{x}_k^{'}|\omega_i,\lambda_j)}{\sum_{i=1}^{M}\sum_{j=1}^{N}p(\omega_i)p(\lambda_j|\omega_i)\mathrm{g}(\boldsymbol{x}_k^{'}|\omega_i,\lambda_j)} \tag{19}$$

and

$$p(\lambda_j|\boldsymbol{x}_k^{'},\omega_i) = \frac{p(\lambda_j|\omega_i)\mathrm{g}(\boldsymbol{x}_k^{'}|\omega_i,\lambda_j)}{\sum_{j=1}^{N}p(\lambda_j|\omega_i)\mathrm{g}(\boldsymbol{x}_k^{'}|\omega_i,\lambda_j)} \tag{20}$$

and $\mathrm{g}(\boldsymbol{x}^{'}|\omega_i,\lambda_j)$ is given from (11). Thus, all the parameters of the conversion function (18) are known from the resynthesis stage of the algorithm.

4 Results and Discussion

The spectral conversion methods outlined in the two previous sections for resynthesis and synthesis were implemented and tested using a multichannel recording of classical music, obtained as described in the first section

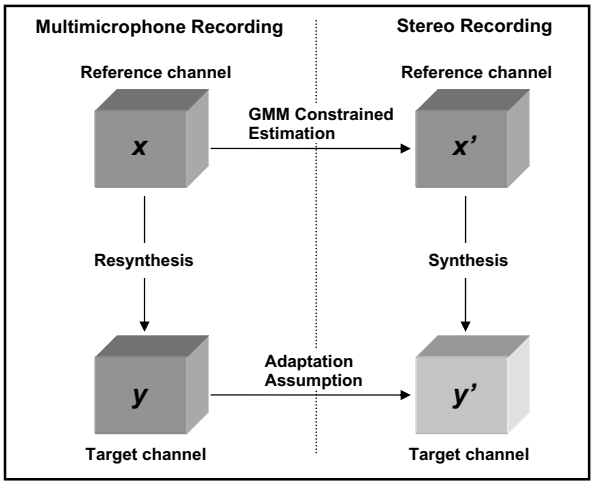

Figure 2: Block diagram outlining multichannel audio resynthesis and synthesis. In synthesis, the target channel is completely synthesized from the reference channel.

of this paper. The objective was to recreate (for resynthesis) or completely synthesize (for synthesis) the channel that mainly captured the chorus of the orchestra. 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 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 methods proposed were found to provide promising enhancement results.

The experimental conditions for the resynthesis example (spectral conversion) and the synthesis example (spectral conversion followed by parameter adaptation) are given in Table 1 and Table 3 respectively. Given that the methods for spectral conversion as well as for model adaptation were originally developed for speech signals, the decision to follow an analysis in subbands seemed natural. 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, 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. The reason is that the short-term spectral envelope is modified separately for each band thus frequency overlapping between adjacent subbands would result in a distorted synthesized signal. The number of octave bands used was 8,

Band Nr.	Frequency (kHz)		LPC Order	Mixtures	
	Low	High		Full	Diag
1	0.0000	0.1723	4	4	8
2	0.1723	0.3446	4	4	8
3	0.3446	0.6891	8	8	16
4	0.6891	1.3782	16	16	32
5	1.3782	2.7563	32	16	64
6	2.7563	5.5125	32	16	64
7	5.5125	11.0250	32	16	64
8	11.0250	22.0500	32	16	64

Table 1: Parameters for the chorus microphone resynthesis example.

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 better 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. The number of GMM components for the synthesis problem is smaller than those of the resynthesis problem due to the increased computational requirements of the described algorithm for adaptation (diagonal conversion is applied for the synthesis problem as explained later in this section).

In Table 2, the average quadratic cepstral distance (averaged over all vectors and all 8 bands) is given for the resynthesis example, 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 two cases tested were the JDE spectral conversion algorithm with full and diagonal covariance matrices [4], denoted as full and diagonal conversion respectively. The difference lies in the fact that in the second case, the covariance matrix for all Gaussians is restricted to be diagonal. This restriction provides a more efficient conversion algorithm in terms of computational requirements, but at the same time requires more GMM components for producing comparable results with full conversion. The improvement is large for both the GMM-based algorithms. Results for full conversion were also given in [2]. Here, we test the efficiency of diagonal conversion to the resynthesis problem since full conversion is of prohibiting computational complexity when combined with the adaptation algorithm for the synthesis problem.

In Table 4, the average quadratic cepstral distance for the synthesis example is given. The objective was to test the performance of the adaptation method for two different cases. The first case was when the GMM parameters correspond to a database obtained from a recording of

SC Method	Ceps. Distance		Centroids per Band
	Train	Test	
Full	0.6629	0.7445	Table 1
Diag	0.6524	0.7508	Table 1

Table 2: Normalized distances for the JDE method, for full and diagonal conversion.

similar nature with the recording that is attempted to be synthesized. Referring to the chorus example, the GMM parameters are obtained as explained in the previous paragraph, by applying the conversion method to a multichannel recording for which the chorus microphone (desired response) is available. If these parameters are applied to another recording of similar nature (*e.g.* both of classical music) the error is quite large as it appears in the second column of Table 4 (denoted as "Same"), in the row denoted as "None" (*i.e.* no adaptation). It should be noted that the error is measured exactly as in the resynthesis case. In other words, the desired response is available for the synthesis case as well but only for measuring the error and not for estimating the conversion parameters. Because of limited availability of such multimicrophone orchestra recordings, the similarity of recordings was simulated by using only a small portion of the available training database (about 5%) for obtaining the GMM parameters. For testing we used the same recordings that were used for testing in the resynthesis example. The results in the second column of Table 4 show a significant improvement in performance by increasing the number of component transformations. It is interesting to note, however, the performance degradation for small numbers of component transformations (cases M-1 and M-2). This can be possibly attributed to the fact that the GMM parameters were obtained from the same recording thus, even with such a small database, they can be expected to capture some of the variability of the cepstral coefficients. On the other hand, adaptation is based on the assumption of the same transformation for the reference and target recordings, which becomes very restricting for such a small number of transformations. The fact that larger numbers of transformation components yield significant reduction of the error, validates the methods derived here and supports the assumptions that were made in the previous section.

The second case examined was when the GMM parameters corresponded to a database obtained from a recording completely different from the recording that is attempted to be synthesized. For this case, we utilized a multimicrophone recording obtained from a live modern music performance. The GMM parameters were obtained from a database constructed from this recording, again the focus being on the vocals of the music. These GMM parameters were applied to the chorus testing recording of the previous examples and the results are given in the third column of Table 4 (denoted as "Other"). An improvement in performance is apparent by increasing the number of transformation com-

Band Nr.	LPC Order	GMM Classes	Components			
			M-1	M-2	M-3	M-4
1	4	4	1	2	2	4
2	4	4	1	2	2	4
3	8	8	1	2	4	8
4	16	16	1	2	8	16
5	32	16	1	2	8	16
6	32	16	1	2	8	16
7	32	16	1	2	8	16
8	32	16	1	2	8	16

Table 3: Parameters for the chorus microphone synthesis example.

ponents, however this case proved to be, as expected, more demanding. The results show that adaptation is very promising for the synthesis problem, but must be applied to a database that corresponds to recordings of nature as diverse as possible.

5 Conclusions

We termed as multichannel audio resynthesis the task of recreating the multiple microphone recordings of an existing multichannel audio recording, from a smaller set of reference signals. Our motivation was to provide a scheme that allows for efficient transmission of multichannel audio through low-bandwidth networks. At the same time, the resynthesis problem arises as a first step towards solving the multichannel audio synthesis problem. Multichannel audio synthesis is the more complex task of completely synthesizing these multiple microphone recordings from an existing monophonic or stereophonic recording, thus making it available for multichannel rendering.

In this paper we applied spectral conversion and adaptation techniques, originally developed for speech synthesis and recognition, to the multichannel audio synthesis problem. The approach was to adapt the GMM parameters developed for the resynthesis problem (where the desired response is available for training the model) to the synthesis problem (no available desired response) by assuming that the reference and target recordings are related with a number of probabilistic linear transformations. The results we obtained were quite promising. Further research is needed in order to validate our methods using a more diverse database of multimicrophone recordings as well as experimenting with other approaches of model adaptation. More work is also needed for identifying types of sounds which these methods cannot adequately address (such as transient sounds) and possible alternative solutions for these cases.

Acknowledgment

This research was funded by the Integrated Media Systems Center, a National Science Foundation Engineering Research Center, Cooperative Agreement EEC-9529152 and Dept. of Army (DAAD-19-99-D-0046).

Adaptation Method	Ceps. Distance		Components per Band
	Same	Other	
None	0.9900	1.2792	Table 3
M-1	0.9938	1.2341	Table 3
M-2	0.9303	1.1865	Table 3
M-3	0.9011	1.1615	Table 3
M-4	0.8786	1.1019	Table 3

Table 4: Normalized distances for the JDE method without adaptation ("None") and several components adaptation (M-1 to M-4) for diagonal conversion.

References

[1] A. Mouchtaris and C. Kyriakakis, "Time-frequency methods for virtual microphone signal synthesis," in *Proc.* 111^{th} *Convention of the Audio Engineering Society (AES), preprint No. 5416*, (New York, NY), November 2001.

[2] A. Mouchtaris, S. S. Narayanan, and C. Kyriakakis, "Multiresolution spectral conversion for multichannel audio resynthesis," in *IEEE Proc. Int. Conf. Multimedia and Expo (ICME)*, vol. 2, (Lausanne, Switzerland), pp. 273–276, August 2002.

[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), pp. 655–658, April 1988.

[4] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech and Audio Processing*, vol. 6, 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), pp. 285–289, May 1998.

[6] 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, pp. 72–83, January 1995.

[7] J. M. Mendel, *Lessons in Estimation Theory for Signal Processing, Communications and Control*. Prentice Hall, 1995.

[8] V. V. Digalakis, D. Rtischev, and L. G. Neumeyer, "Speaker adaptation using constrained estimation of Gaussian mixtures," *IEEE Trans. Speech and Audio Processing*, vol. 3, pp. 357–366, September 1995.

[9] V. D. Diakoloukas and V. V. Digalakis, "Maximum-likelihood stochastic-transformation adaptation of Hidden Markov Models," *IEEE Trans. Speech and Audio Processing*, vol. 7, pp. 177–187, March 1999.