Audio Engineering Society
Convention Paper
Presented at the 113th Convention
2002 October 5–8 Los Angeles, CA, USA

*This convention paper has been reproduced from the author's advance manuscript, without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. Additional papers may be obtained by sending request and remittance to Audio Engineering Society, 60 East 42nd Street, New York, New York 10165-2520, USA; also see* `www.aes.org`. *All rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the* Journal of the Audio Engineering Society.

# Gaussian Mixture Model Based Methods for Virtual Microphone Signal Synthesis

Athanasios Mouchtaris[1], Shrikanth S. Narayanan[1], and Chris Kyriakakis[1]

[1] *Integrated Media Systems Center (IMSC), University of Southern California, Los Angeles, CA, 90089-2564, USA*

Correspondence should be addressed to Athanasios Mouchtaris (`mouchtar@sipi.usc.edu`)

**ABSTRACT**

Multichannel audio can immerse a group of listeners in a seamless aural environment. However, several issues must be addressed, such as the excessive transmission requirements of multichannel audio, as well as the fact that to-date only a handful of music recordings have been made with multiple channels. Previously, we proposed a system capable of synthesizing the multiple channels of a virtual multichannel recording from a smaller set of reference recordings. In this paper these methods are extended to provide a more general coverage of the problem. The emphasis here is on time-varying filtering techniques that can be used to enhance particular instruments in the recording, which is desired in order to simulate virtual microphones in several locations close and around the sound source.

## 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

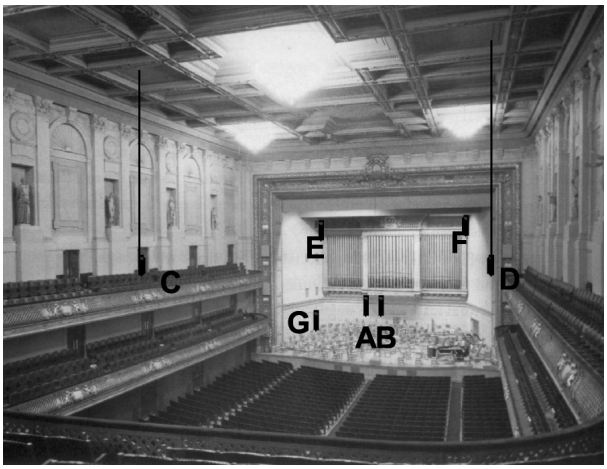


Fig. 1: An example of how microphones may be arranged in a recording venue for a multichannel recording. In the virtual microphone synthesis algorithm, microphones A and B are the main reference pair from which the remaining microphone signals can be derived. Virtual microphones C and D capture the hall reverberation, while virtual microphones E and F capture the reflections from the orchestra stage. Virtual microphone G can be used to capture individual instruments such as the tympani. These signals can then be mixed and played back through a multichannel audio system that recreates the spatial realism of a large hall.

be addressed. Multichannel audio imposes excessive requirements to the transmission medium. A system we previously proposed [7, 8], attempted to address this issue by offering the alternative to resynthesize 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*, was 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), 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 into two different categories, namely reverberant and spot microphone recordings. For simulating recordings of microphones placed far from the orchestra (reverberant microphones), 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), 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 next section outlines the spectral conversion method that is employed for the resynthesis problem and is followed by the section on the adaptation method that allows for using these conversion parameters to an arbitrary recording (synthesis problem). Finally, the algorithms described are validated by simulation results and possible directions for future research are given.

## SPECTRAL CONVERSION

The approach followed for spot microphone resynthesis is based on spectral conversion methods that have been successfully employed to speech synthesis applications [1, 12, 5]. A training data set is created from the existing reference and target recordings by applying a short sliding window and extracting the parameters that model the short-term spectral envelope (in this paper we use the cepstral coefficients [9]). 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, $[\boldsymbol{x}_1\boldsymbol{x}_2\ldots\boldsymbol{x}_n]$ 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 [1, 12, 5, 2] and the references therein). In [8] the algorithms based on Gaussian mixture models (GMM, [12, 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) that can be modeled 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 normal multivariate distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ and $p(\omega_i)$ is the prior probability of class $\omega_i$. The parameters of the GMM, *i.e.* the mean vectors, covariance matrices and priors, can be estimated using the expectation maximization (EM) algorithm [10].

The analysis that follows focuses on the conversion of [12]. A GMM pdf is assumed for the reference spectral vectors and the function $\mathcal{F}$ is designed such that the error

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

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

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

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

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

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

## 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 to address 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 training target 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. Such a transformation can be found based on maximum likelihood constrained adaptation that is described in [4, 3] and was developed for the task of speaker adaptation for speech recognition.

We start by applying a GMM as in (1) for the reference random vector $\boldsymbol{x}$ of an existing multichannel recording for which the resynthesis method of the previous section has been applied. The random vector $\boldsymbol{x}'$ corresponds to the reference recording of the stereophonic recording to which the synthesis methods are to be applied (for which no target recording is available). We assume that target random vector $\boldsymbol{x}'$ is related to reference random vector $\boldsymbol{x}$ by a probabilistic linear transformation

$$\boldsymbol{x}' = \begin{cases} \mathbf{A}_1\boldsymbol{x} + \boldsymbol{b}_1 & \text{with probability } p(\lambda_1|\omega_i) \\ \mathbf{A}_2\boldsymbol{x} + \boldsymbol{b}_2 & \text{with probability } p(\lambda_2|\omega_i) \\ \vdots & \vdots \\ \mathbf{A}_N\boldsymbol{x} + \boldsymbol{b}_N & \text{with probability } p(\lambda_N|\omega_i) \end{cases} \tag{5}$$

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 the constraint

$$\sum_{j=1}^{N} p(\lambda_j|\omega_i) = 1, \quad i = 1, \dots, M \tag{6}$$

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

$$\mathrm{g}(\boldsymbol{x}'|\omega_i, \lambda_j) = \mathcal{N}(\boldsymbol{x}'; \mathbf{A}_j\boldsymbol{\mu}_i^x + \boldsymbol{b}_j, \mathbf{A}_j\boldsymbol{\Sigma}_i^{xx}\mathbf{A}_j^T) \tag{7}$$

| Band Nr. | Frequency Range | | LPC Order | Mixtures | |
|---|---|---|---|---|---|
| | Low (kHz) | High (kHz) | | 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.

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

$$\mathrm{g}(\boldsymbol{x}^{'}) = \sum_{i=1}^{M}\sum_{j=1}^{N} p(\omega_i)p(\lambda_j|\omega_i)\mathcal{N}(\boldsymbol{x}^{'}; \mathbf{A}_j\boldsymbol{\mu}_i^{x}+\boldsymbol{b}_j, \mathbf{A}_j\boldsymbol{\Sigma}_i^{xx}\mathbf{A}_j^{T}) \tag{8}$$

Thus $\boldsymbol{x}^{'}$ is modeled also as a GMM, with $M \times N$ Gaussian mixtures. The matrices $\mathbf{A}_j$, the vectors $\boldsymbol{b}_j$ and the conditional probabilities $p(\lambda_j|\omega_i)$ can be estimated using maximum likelihood estimation techniques. As explained in [4, 3], the EM algorithm can be applied to this case as well, in a similar manner to estimating the parameters of a GMM from observed data. In essence, it is a linearly constrained estimation of the GMM parameters.

The purpose of adopting the transformation (5) is to use it in order to obtain a target training sequence for the synthesis problem. The assumption, as previously mentioned, is that this function represents the statistical correspondence between the two available recordings. It is then justifiable (especially in the absence of further information) 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

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

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. Since it is not clear what parameters $\boldsymbol{v}_i$ and $\boldsymbol{\Gamma}_i$ represent, we follow the analysis of [12], where the form of the conversion function proposed is explained by examining the limit-case of a single class GMM for $\boldsymbol{x}$ (*i.e.* a Gaussian distribution). In that case, and assuming the source and target vectors are jointly Gaussian, the optimal conversion function in mean-squared sense will be

$$\begin{aligned} \mathcal{F}(\boldsymbol{x}_k) &= \mathrm{E}(\boldsymbol{y}|\boldsymbol{x}_k) \\ &= \boldsymbol{\mu}^{y} + \boldsymbol{\Sigma}^{yx}\boldsymbol{\Sigma}^{xx^{-1}}(\boldsymbol{x}_k - \boldsymbol{\mu}^{x}) \\ &= \boldsymbol{v} + \boldsymbol{\Gamma}\boldsymbol{\Sigma}^{xx^{-1}}(\boldsymbol{x}_k - \boldsymbol{\mu}^{x}) \end{aligned} \tag{10}$$

where $\mathrm{E}(\cdot)$ denotes the expectation operator. So, in the limit-case, it holds that

$$\boldsymbol{v} = \boldsymbol{\mu}^{y},\ \boldsymbol{\Gamma} = \boldsymbol{\Sigma}^{yx} \tag{11}$$

We also examine the simple case where (5) and (9) become

$$\boldsymbol{x}^{'} = \mathbf{A}\boldsymbol{x}+\boldsymbol{b},\ \boldsymbol{y}^{'} = \mathbf{A}\boldsymbol{y}+\boldsymbol{b} \tag{12}$$

Since under these conditions

$$\boldsymbol{\mu}^{x^{'}} = \mathbf{A}\boldsymbol{\mu}^{x}+\boldsymbol{b},\ \boldsymbol{\mu}^{y^{'}} = \mathbf{A}\boldsymbol{\mu}^{y}+\boldsymbol{b} \tag{13}$$

and

$$\boldsymbol{\Sigma}^{x^{'}x^{'}} = \mathbf{A}\boldsymbol{\Sigma}^{xx}\mathbf{A}^{T},\ \boldsymbol{\Sigma}^{y^{'}x^{'}} = \mathbf{A}\boldsymbol{\Sigma}^{yx}\mathbf{A}^{T} \tag{14}$$

it is then apparent that the parameters $\boldsymbol{v}^{'}$ and $\boldsymbol{\Gamma}^{'}$ for the conversion function for the synthesis case will be

$$\boldsymbol{v}^{'} = \mathbf{A}\boldsymbol{v}+\boldsymbol{b},\ \boldsymbol{\Gamma}^{'} = \mathbf{A}\boldsymbol{\Gamma}\mathbf{A}^{T} \tag{15}$$

The conversion function for the limit-case becomes

$$\begin{aligned} \mathcal{F}(\boldsymbol{x}_k^{'}) &= \mathrm{E}(\boldsymbol{y}^{'}|\boldsymbol{x}_k^{'}) \\ &= \boldsymbol{\mu}^{y^{'}} + \boldsymbol{\Sigma}^{y^{'}x^{'}}\boldsymbol{\Sigma}^{x^{'}x^{'}-1}\left(\boldsymbol{x}_k^{'} - \boldsymbol{\mu}^{x^{'}}\right) \\ &= \mathbf{A}\boldsymbol{v}+\boldsymbol{b}+\mathbf{A}\boldsymbol{\Gamma}\boldsymbol{\Sigma}^{xx^{-1}}\mathbf{A}^{-1}\left(\boldsymbol{x}_k^{'} - \mathbf{A}\boldsymbol{\mu}^{x} - \boldsymbol{b}\right) \end{aligned} \tag{16}$$

| SC Method | Cepstral Distance | | Centroids per Band |
|---|---|---|---|
| | Train | Test | |
| Full | 0.6451 | 0.7144 | Table 1 |
| Diag | 0.5918 | 0.7460 | Table 1 |

Table 2: Normalized distances for LSE method for full and diagonal conversion.

By analogy then, it is justifiable to conclude that the conversion function for synthesis will be

$$\mathcal{F}(\boldsymbol{x}_k^{'}) = \sum_{i=1}^{M}\sum_{j=1}^{N} p(\omega_i|\boldsymbol{x}_k^{'}) p(\lambda_j|\boldsymbol{x}_k^{'},\omega_i)\Big[\mathbf{A}_j\boldsymbol{v}_i + \boldsymbol{b}_j + \mathbf{A}_j\boldsymbol{\Gamma}_i\boldsymbol{\Sigma}_i^{xx^{-1}}\mathbf{A}_j^{-1}\left(\boldsymbol{x}_k^{'} - \mathbf{A}_j\boldsymbol{\mu}_i^{x} - \boldsymbol{b}_j\right)\Big] \quad (17)$$

where

$$p(\omega_i|\boldsymbol{x}_k^{'}) = \frac{p(\omega_i)\sum_{j=1}^{N} p(\lambda_j|\omega_i)\mathrm{g}(\boldsymbol{x}_k^{'}|\omega_i,\lambda_j)}{\sum_{i=1}^{M}\sum_{j=1}^{N} p(\omega_i)p(\lambda_j|\omega_i)\mathrm{g}(\boldsymbol{x}_k^{'}|\omega_i,\lambda_j)} \quad (18)$$

and

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

and $\mathrm{g}(\boldsymbol{x}^{'}|\omega_i,\lambda_j)$ is given from (7). Thus, all the parameters of the conversion function (17) are known from the resynthesis stage of the algorithm.

## 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 of this paper. The objective was to recreate 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 [11], 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, 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 LSE spectral conversion algorithm with full and diagonal covariance matrices [12], 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 con-

| Band Nr. | LPC Order | GMM Classes | Number of 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.

version were also given in [8]. 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. As explained in [4, 3], the adaptation methods described are less computationally demanding when applied to GMM's with diagonal covariance matrices. Thus, it was apparent that it would be more efficient to combine these methods with the diagonal conversion algorithm of [12].

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 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, validate the methods derived here and support 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 components, 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.

## CONCLUSIONS

We termed as multichannel audio resynthesis the task of recreating the multiple microphone recordings of an

| Adaptation Method | Cepstral Distance | | Components per Band |
|---|---|---|---|
| | Same | Other | |
| None | 0.9454 | 1.3777 | Table 3 |
| M-1 | 1.1227 | 1.1482 | Table 3 |
| M-2 | 1.0034 | 1.1348 | Table 3 |
| M-3 | 0.8794 | 1.0995 | Table 3 |
| M-4 | 0.8589 | 1.0728 | Table 3 |

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

existing multichannel audio recording, with the purpose of efficient transmission and as a first step to multichannel audio synthesis. The synthesis problem 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. It should be noted the methods described in this paper will not yield acceptable results for all types of sounds. Transient sounds in general cannot be adequately processed by simply modifying their short-term spectral envelope. The special case of percussive drum-like sounds was examined in [8] because of their acoustical significance and because models for these sounds are available (see for example [6]). More work is also needed in this area for identifying other types of sounds which these methods cannot adequately address and possible alternative solutions for these cases.

## ACKNOWLEDGMENTS

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

## REFERENCES

[1] 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)*, pages 655–658, New York, NY, April 1988.

[2] G. Baudoin and Y. Stylianou. On the transformation of the speech spectrum for voice conversion. In *IEEE Proc. Int. Conf. Spoken Language Processing (ICSLP)*, pages 1405–1408, Philadephia, PA, October 1996.

[3] V. D. Diakoloukas and V. V. Digalakis. Maximum-likelihood stochastic-transformation adaptation of Hidden Markov Models. *IEEE Trans. Speech and Audio Processing*, 7(2):177–187, March 1999.

[4] V. V. Digalakis, D. Rtischev, and L. G. Neumeyer. Speaker adaptation using constrained estimation of Gaussian mixtures. *IEEE Trans. Speech and Audio Processing*, 3(5):357–366, September 1995.

[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)*, pages 285–289, Seattle, WA, May 1998.

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

[7] 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.

[8] A. Mouchtaris, S. S. Narayanan, and C. Kyriakakis. Multiresolution spectral conversion for multichannel audio resynthesis. To appear *IEEE Proc. Int. Conf. Multimedia and Expo (ICME 2002).*

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

[10] D. A. Reynolds and R. C. Rose. Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Trans. Speech and Audio Processing*, 3(1):72–83, January 1995.

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

[12] Y. Stylianou, O. Cappe, and E. Moulines. Continuous probabilistic transform for voice conversion. *IEEE Trans. Speech and Audio Processing*, 6(2):131–142, March 1998.