Multichannel Audio Synthesis by Subband-Based Spectral Conversion and Parameter Adaptation

Athanasios Mouchtaris, *Member, IEEE*, Shrikanth S. Narayanan, *Senior Member, IEEE*, and Chris Kyriakakis, *Member, IEEE*

***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, which would significantly enhance the recording's acoustic impression. We approach this problem by extending the model employed for the resynthesis problem. This is accomplished by adapting the resynthesis conversion parameters to the statistical properties of the recording that we wish to enhance. This parameter adaptation is similar to the task adaptation employed in speech recognition, when a specific model is applied to a different environment (speaker, language or channel). One particular approach to this problem is shown here to be quite advantageous toward solving the multichannel audio synthesis problem as well.**

***Index Terms*—Audio recording, audio resynthesis, audio systems, Gaussian mixture model, multichannel audio, virtual microphones.**

I. 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 (denoted as *reference* channels or recordings in this work, *e.g.*, the left and right channels in a traditional stereophonic recording). The solution, termed multichannel audio *resynthesis*, focused 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 in a 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. For this reason, stem recordings are also referred to as *target* recordings in this work. We have obtained such stem microphone recordings from two orchestra halls in the U.S. by placing microphones at various locations throughout the hall. By recording a performance with a total of sixteen microphones, we then designed a system that *recreates* these recordings (thus named *virtual microphone* recordings) from the main microphone pair. These resynthesized stem recordings are then mixed in order to produce the final multichannel audio recording. The distinction of the recordings is made depending on the location of the microphone in the venue, thus resulting in two different categories, namely *reverberant* and *spot* microphone recordings.

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). For simulating recordings of such microphones, infinite impulse response (IIR) filters were designed from existing multichannel recordings made in a particular concert hall [1]. Our objective was to estimate the appropriate filters that capture the concert hall acoustical properties from a given set of stem microphone recordings. These IIR filters designed were shown to be capable of recreating the acoustical properties of the venue at specific locations.

Spot microphones are microphones that are placed close to the sound source (*e.g.*, G in Fig. 1). These microphones introduce a very 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 in [2] focuses on this problem and is based on spectral conversion.

Manuscript received August 14, 2002; revised December 16, 2003. This work was supported by the Integrated Media Systems Center, a National Science Foundation Engineering Research Center, under Cooperative Agreement EEC-9529152. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the National Science Foundation. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Gerald Schuller.

A. Mouchtaris was with the Integrated Media Systems Center and Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089 USA. He is now with the Electrical and Systems Engineering Department, University of Pennsylvania, Philadelphia, PA 19104 USA (e-mail: mouchtar@ieee.org).

S. S. Narayanan and C. Kyriakakis are with the Integrated Media Systems Center and Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089 USA (e-mail: shri@sipi.usc.edu; ckyriak@imsc.usc.edu).

Digital Object Identifier 10.1109/TSA.2004.841061

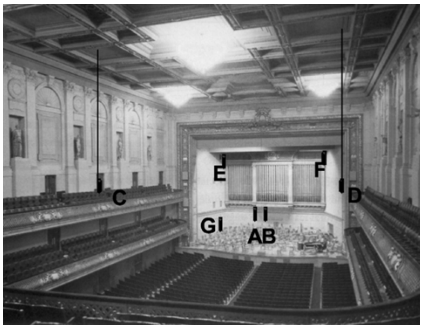

Fig. 1. Example of how microphones may be arranged in a recording venue for a multichannel recording. In the virtual microphone synthesis algorithm, microphones A and B are the main reference pair from which the remaining microphone signals can be derived. Virtual microphones C and D capture the hall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

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. Note 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 for the resynthesis problem can be readily applied to arbitrary recordings. Their time-invariant nature offers the advantage that these filters can be applied to any recording although having been designed based on a specific 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 address in this paper, thus in Sections II–IV only the spot microphone synthesis case is examined.

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 objective of such a system is to decrease the transmission requirements of multichannel audio, by allowing some of the channels to be recreated at the receiving end. 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. The objective in this case is to significantly increase the realism of an existing stereo recording. Our approach is to take advantage of the resynthesis parameters that have been derived based on an existing target channel of a different recording. In other words, given a multimicrophone recording, we attempt to extract the time-varying filters that relate two different channels and apply them to a different stereo recording. Given the time-varying nature of these filters, some method must be derived to adapt them to the characteristics of the stereo recording. In order to achieve that, the stereo and multimicrophone recordings are related with the Gaussian mixture model (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. Note that, since the synthesis problem is addressed based on the estimated filters from an existing multimicrophone recording, it is essential that we first describe the procedure and give some results for the resynthesis algorithm as well. The last step, then, is to attack the synthesis problem based on the resynthesis parameters and their adaptation to a specific stereo recording.

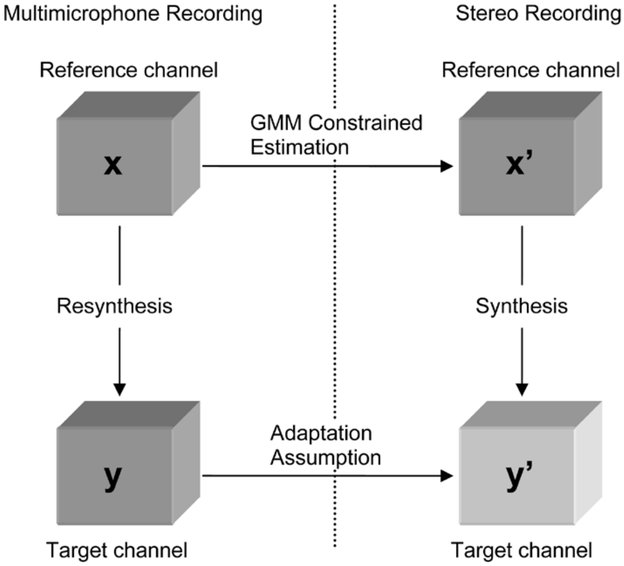

Fig. 2. Block diagram outlining multichannel audio resynthesis and synthesis. Resynthesis corresponds to existing multichannel audio recordings while synthesis corresponds to stereo recordings. The objective of resynthesis is to recreate the multiple channels of the recording (target channels) from a smaller set of reference channels. The objective of synthesis is to completely synthesize target channels from one or two reference channels, thus converting the stereo recording for multichannel rendering. Resynthesis parameters can be used for the synthesis task, by adapting them through GMM-constrained estimation and the adaptation assumption explained in the text.

The remainder of this paper is organized as follows. In Section II, spot microphone resynthesis is addressed, since it is an inherent part of the synthesis problem. The next step toward deriving the synthesis algorithm is the adaptation of the resynthesis parameters to a different recording. Model adaptation schemes are discussed in Section III. In Section IV, we test the effectiveness of spectral conversion methods when applied to resynthesis, especially focusing on computationally efficient model structures of conversion. The performance of the adaptation methods toward achieving a solution for the synthesis problem is also examined in Section IV. Finally, a brief discussion of the derived methods is given in Section V and possible directions for future research are proposed.

II. Spectral Conversion for Resynthesis

The methods for spot microphones are geared toward enhancing certain instruments in the reference recording. A microphone placed near a particular instrument mainly captures a “dry” (nonreverberant) version of the instrument and some leakage from nearby instruments. The instruments close to the target microphone are far more prominent in the target recording than in the reference recording. Our objective is to retain the perceptual advantages of the multichannel recording, as a first step toward addressing the problem. This, in effect, means that our objective is to enhance the desired voices/instruments in the reference recording, even if the resynthesized signal is not identical with the desired. As mentioned in the previous section, we were able to produce identical responses for the reverberant microphones case, however the spot microphone case proved to be far more demanding.

For the spot microphones case, nonstationarity of the audio signals is the focus of this paper; the spectral conversion methods attempt to address this problem. The problem arises from the fact that the objective of our method is to enhance a particular instrument in the reference recording. The instrument to be enhanced has a frequency response that significantly varies in time, and as a result a time-invariant filter would not produce meaningful results. Our methods are based on the fact that the reference and target responses are highly related (same performance recorded simultaneously with different microphones). Based on this observation, the desired transfer function, although constantly varying in time, can be estimated based on the reference recording with the use of the spectral conversion methods. For the spot microphones case, each target microphone captures mainly a specific type of instruments while the reference microphone "weighs" all instruments approximately equally. This corresponds to the dependence of the spot microphones on their location with respect to the orchestra. Although the response of these microphones depends on the acoustics of the hall as well, this dependence is not considered perceptually significant (for reasons explained in Section II-A), and this greatly simplifies the solution. The methods proposed here result in one conversion function for each pair of spot and reference microphones (with the reference microphone remaining the same in all cases), so that all target waveforms can be resynthesized from only one recording.

A. Spectral Conversion

Our initial experiments for the spot microphones case, detailed in the next paragraph, motivated us to focus on modifying 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 (LP) analysis [3], and the residual signal is the result of inverse filtering the audio signal of the current frame 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 [4].

It is interesting to describe one of our initial experiments that led us to focus on the short-term spectral envelope and, as a consequence, on the spectral conversion methods that are described next. In this simple experiment, we attempted to synthesize the desired response (in this case the response captured by the microphone placed close to the chorus of the orchestra), by using the reference residual and the cepstral coefficients obtained from the desired response. In other words, we were interested to test the result of our resynthesis methods in the ideal case where the desired sequence of cepstral coefficients was correctly "predicted". The result was an audio signal which sounded more reverberant than the desired signal (for reasons explained later in this section), but otherwise extremely similar in all respects. Thus, deriving an algorithm that correctly predicts the desired sequence of cepstral coefficients from the reference cepstral coefficients of the respective frame, would result in a resynthesized signal very close to the desired. The problem as stated, is exactly the problem statement of spectral conversion, which aims to design a mapping function from the reference to the target space, whose parameters remain constant for a particular pair of reference and target sources. 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 [5]–[7]. The objective of voice conversion is to modify a speech waveform so that the context remains as is but appears to be spoken by a specific (target) speaker. Although the application is completely different, the approach followed is very suitable for our problem. In voice conversion pitch and time-scaling need to be considered, while in the application examined here this is not necessary. This is true 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.

At this point, it is of interest to mention that the spectral conversion methods are useful for modifying the spectral coloration of the signal, and the target response is resynthesized using the modified spectral envelope along with the residual derived from the reference recording. Note that short-term analysis indicates using windows in the order of 50 ms, which means that the residual (in effect the modeling error) contains the reverberation which cannot be modeled with the short-term spectral envelope. As a result, the resynthesized response might sound more reverberant than the target response, depending on how reverberant the reference response originally is. Our concern, though, is mostly to enhance a specific instrument within the reference recording, without focusing on dereverberating the signal. In most cases this will not be an issue, given that usually the reference recordings are not highly reverberant.

Assuming that a sequence $[\boldsymbol{x}_1 \boldsymbol{x}_2 \dots \boldsymbol{x}_n]$ 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 $[\boldsymbol{y}_1 \boldsymbol{y}_2 \dots \boldsymbol{y}_n]$ (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 [5]–[8] and the references therein). In this paper, the algorithms based on vector quantization (VQ) ([5]) and GMMs ([6], [7]) were implemented and compared.

1) Spectral Conversion Based on VQ: Under this approach, the spectral vectors of the reference and target signals (training data) are vector quantized using the well-known modified K-means clustering algorithm (see for example [9] 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) Spectral Conversion Based on GMMs: 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 as

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

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 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 [10].

As already mentioned, the function $\mathcal{F}$ is designed so that the spectral vectors $\boldsymbol{y}_k$ and $\mathcal{F}(\boldsymbol{x}_k)$ are close in some sense. In [6], the function $\mathcal{F}$ is designed such that the error

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

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, *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 ω_i, $p(\omega_i|\boldsymbol{x}_k)$ can be computed by applying Bayes' theorem

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

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

A different solution for function $\mathcal{F}$ results when a different function than (2) is minimized [7]. 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

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

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 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

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

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 joint density estimation (JDE) method.

B. Diagonal Implementation

The GMM-based LSE spectral conversion algorithm can be implemented with the covariance matrix having no structural restrictions or restricted to be diagonal [6], 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 second part of this paper. As explained in [11] and [12], 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 [6] for LSE and the diagonal conversion for JDE implemented in this paper, as explained next.

It is important to note that the covariance matrix 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 (6). This will be zero if the covariance of $\boldsymbol{z}$ is diagonal. In order to obtain the same structure as in the diagonal LSE conversion, we must restrict the matrices $\boldsymbol{\Sigma}_i^{xx}$, $\boldsymbol{\Sigma}_i^{yy}$, $\boldsymbol{\Sigma}_i^{xy}$, and $\boldsymbol{\Sigma}_i^{yx}$ in (6) 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 [13], that the inverse of $\boldsymbol{\Sigma}_i^{zz}$ will be

$$\boldsymbol{\Sigma}_i^{zz^{-1}} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{bmatrix} \tag{7}$$

where

$$\begin{aligned}\mathbf{A} &= \left(\boldsymbol{\Sigma}_i^{xx} - \boldsymbol{\Sigma}_i^{xy}\boldsymbol{\Sigma}_i^{yy^{-1}}\boldsymbol{\Sigma}_i^{yx}\right)^{-1}\\ &= \boldsymbol{\Sigma}_i^{xx^{-1}} + \boldsymbol{\Sigma}_i^{xx^{-1}}\boldsymbol{\Sigma}_i^{xy}\mathbf{C}\boldsymbol{\Sigma}_i^{yx}\boldsymbol{\Sigma}_i^{xx^{-1}}\\ \mathbf{B} &= -\mathbf{A}\boldsymbol{\Sigma}_i^{xy}\boldsymbol{\Sigma}_i^{yy^{-1}} = -\boldsymbol{\Sigma}_i^{xx^{-1}}\boldsymbol{\Sigma}_i^{xy}\mathbf{C}\\ \mathbf{C} &= \left(\boldsymbol{\Sigma}_i^{yy} - \boldsymbol{\Sigma}_i^{yx}\boldsymbol{\Sigma}_i^{xx^{-1}}\boldsymbol{\Sigma}_i^{xy}\right)^{-1}\\ &= \boldsymbol{\Sigma}_i^{yy^{-1}} + \boldsymbol{\Sigma}_i^{yy^{-1}}\boldsymbol{\Sigma}_i^{yx}\mathbf{A}\boldsymbol{\Sigma}_i^{xy}\boldsymbol{\Sigma}_i^{yy^{-1}}.\end{aligned} \tag{8}$$

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 $\boldsymbol{\Sigma}_i^{zz}$, the joint pdf of $\boldsymbol{x}$ and $\boldsymbol{y}$ can be written as

$$\mathrm{g}(\boldsymbol{x}, \mathbf{y}) = \frac{\exp\left(-\frac{1}{2}(\boldsymbol{x}^T\mathbf{A}\boldsymbol{x} + \mathbf{y}^T\mathbf{C}\mathbf{y} + 2\boldsymbol{x}^T\mathbf{B}\mathbf{y})\right)}{(2\pi)^K\sqrt{|\boldsymbol{\Sigma}_i^{zz}|}} \tag{9}$$

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

$$|\boldsymbol{\Sigma}_i^{zz}| = |\boldsymbol{\Sigma}_i^{yy}| \left|\boldsymbol{\Sigma}_i^{xx} - \boldsymbol{\Sigma}_i^{xy}{\boldsymbol{\Sigma}_i^{yy}}^{-1}\boldsymbol{\Sigma}_i^{yx}\right|. \qquad (10)$$

C. Subband Processing

Audio signals contain information over a larger bandwidth than speech signals. The sampling rate for audio signals is usually 44.1 or 48 kHz compared to 8 or 16 kHz for speech. Moreover, since high acoustical quality for audio is essential, it is important to consider the entire spectrum in detail. 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 presented in the previous section (the signals were demodulated and decimated after they were passed through the filterbanks and before the linear predictive analysis). Perfect reconstruction filter banks, based on wavelets [14], 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 from passband to stopband is desirable. 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.

D. Transient Sounds Consideration

The spectral conversion methods described earlier will not produce the desired result in all cases. Transient sounds, such as percussive sounds in music, cannot be adequately processed by altering their spectral envelope and must be examined separately. An example of an analysis/synthesis model that treats transient sounds separately and is very suitable as an alternative to the subband-based residual/LP model that we employed, is described in [15]. It is suitable since it also models the audio signal in different bands, in each one as a sinusoidal/residual model [16], [17]. The sinusoidal parameters can be treated in the same manner as the LP coefficients during spectral conversion [18]. We are currently considering this model for improving the produced sound quality of our system. However, no structured model is proposed in [15] for transient sounds. In the remainder of this section, the special case of percussive sounds is addressed.

The case of percussive drum-like sounds is considered of particular importance in music analysis. 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 [19] (see also [20]–[22]), but for the enhancement purposes investigated in this paper, the emphasis is given to the residual instead of the LP parameters. The idea is 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 [19], 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 proved to be quite successful and further details are given in Section IV-D. The drawback of this approach is that a robust algorithm is required for identifying the particular instrument instances in the reference recording. A possible improvement of the proposed method would be to extract all instances of the instrument from the target response and use some clustering technique for choosing the residual that is more appropriate in the resynthesis stage. The reason is that the residual/LP model introduces modeling error which is larger in the spectral valleys of the AR spectrum; thus, better results would be obtained by using a residual which corresponds to an AR filter as close as possible to the resynthesis AR filter. However, this approach would again require robustly identifying all the instances of the instrument.

The methods described in this section can be used for resynthesizing recordings of microphones that are placed close to the orchestra. For this case, the desired responses (stem recordings) are available and are required in order to derive the conversion functions. In the synthesis problem, the desired responses are not available. In Section III, we attempt to address this lack of training data by adapting the parameters derived for the resynthesis problem, based on the derived statistics of the available reference recording of the synthesis problem. As we demonstrate, the desired waveforms can be synthesized by taking advantage of techniques developed for speech recognition parameter adaptation.

III. ML Constrained Adaptation for Synthesis

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 enhanced using 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 [11], [12], 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

$$\boldsymbol{x}' = \begin{cases} \mathbf{A}_1\boldsymbol{x} + \boldsymbol{b}_1, & \text{with probability } p(\lambda_1|\omega_i) \\ \mathbf{A}_2\boldsymbol{x} + \boldsymbol{b}_2, & \text{with probability } p(\lambda_2|\omega_i) \\ \vdots & \vdots \\ \mathbf{A}_N\boldsymbol{x} + \boldsymbol{b}_N, & \text{with probability } p(\lambda_N|\omega_i). \end{cases} \tag{11}$$

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 the constraint

$$\sum_{j=1}^{N} p\left(\lambda_j|\omega_i\right) = 1, \quad i = 1, \ldots, M \tag{12}$$

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

$$\mathrm{g}(\boldsymbol{x}'|\omega_i, \lambda_j) = \mathcal{N}\left(\boldsymbol{x}'; \mathbf{A}_j\boldsymbol{\mu}_i^x + \boldsymbol{b}_j, \mathbf{A}_j\boldsymbol{\Sigma}_i^{xx}\mathbf{A}_j^T\right) \tag{13}$$

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

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

The matrices $\mathbf{A}_j$, the vectors $\boldsymbol{b}_j$, and the 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 (11) 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 response 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

$$\mathbf{y}' = \begin{cases} \mathbf{A}_1\mathbf{y} + \boldsymbol{b}_1, & \text{with probability } p(\lambda_1|\omega_i) \\ \mathbf{A}_2\mathbf{y} + \boldsymbol{b}_2, & \text{with probability } p(\lambda_2|\omega_i) \\ \vdots & \vdots \\ \mathbf{A}_N\mathbf{y} + \boldsymbol{b}_N, & \text{with probability } p(\lambda_N|\omega_i). \end{cases} \tag{15}$$

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

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. A solution is provided for adapting the parameters of both the JDE and LSE resynthesis methods. This derived conversion function for synthesis will allow the synthesis of the target response from the reference channel of the stereo recording as depicted in Fig. 2

A. LSE Parameter Adaptation

Since it is not clear what parameters $\boldsymbol{v}_i$ and $\boldsymbol{\Gamma}_i$ represent, we follow the analysis of [6], 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

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

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

$$\boldsymbol{v} = \boldsymbol{\mu}^y, \quad \boldsymbol{\Gamma} = \boldsymbol{\Sigma}^{yx}. \tag{17}$$

We also examine the simple case where (11) and (15) become

$$\boldsymbol{x}' = \mathbf{A}\boldsymbol{x} + \boldsymbol{b}, \quad \mathbf{y}' = \mathbf{A}\mathbf{y} + \boldsymbol{b}. \tag{18}$$

Since under these conditions

$$\boldsymbol{\mu}^{x'} = \mathbf{A}\boldsymbol{\mu}^x + \boldsymbol{b}, \quad \boldsymbol{\mu}^{y'} = \mathbf{A}\boldsymbol{\mu}^y + \boldsymbol{b} \tag{19}$$

and

$$\boldsymbol{\Sigma}^{x'x'} = \mathbf{A}\boldsymbol{\Sigma}^{xx}\mathbf{A}^T, \quad \boldsymbol{\Sigma}^{y'x'} = \mathbf{A}\boldsymbol{\Sigma}^{yx}\mathbf{A}^T \tag{20}$$

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

$$\boldsymbol{v}' = \mathbf{A}\boldsymbol{v} + \boldsymbol{b}, \quad \boldsymbol{\Gamma}' = \mathbf{A}\boldsymbol{\Gamma}\mathbf{A}^T. \tag{21}$$

The conversion function for the limit-case becomes

$$\begin{aligned} \mathcal{F}(\boldsymbol{x}'_k) &= \mathrm{E}(\mathbf{y}'|\boldsymbol{x}'_k) \\ &= \boldsymbol{\mu}^{y'} + \boldsymbol{\Sigma}^{y'x'}{\boldsymbol{\Sigma}^{x'x'}}^{-1}\left(\boldsymbol{x}'_k - \boldsymbol{\mu}^{x'}\right) \\ &= \mathbf{A}\boldsymbol{v} + \boldsymbol{b} + \mathbf{A}\boldsymbol{\Gamma}{\boldsymbol{\Sigma}^{xx}}^{-1}\mathbf{A}^{-1}\left(\boldsymbol{x}'_k - \mathbf{A}\boldsymbol{\mu}^x - \boldsymbol{b}\right). \end{aligned} \tag{22}$$

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

$$\mathcal{F}(\boldsymbol{x}'_k) = \sum_{i=1}^{M}\sum_{j=1}^{N} p(\omega_i|\boldsymbol{x}'_k)p(\lambda_j|\boldsymbol{x}'_k,\omega_i) \left[\mathbf{A}_j\boldsymbol{v}_i + \boldsymbol{b}_j + \mathbf{A}_j\boldsymbol{\Gamma}_i\boldsymbol{\Sigma}_i^{xx^{-1}}\mathbf{A}_j^{-1}\left(\boldsymbol{x}'_k - \mathbf{A}_j\boldsymbol{\mu}_i^x - \boldsymbol{b}_j\right)\right] \quad (23)$$

where

$$p(\omega_i|\boldsymbol{x}'_k) = \frac{p(\omega_i)\sum_{j=1}^{N} p(\lambda_j|\omega_i)\mathrm{g}(\boldsymbol{x}'_k|\omega_i,\lambda_j)}{\sum_{i=1}^{M}\sum_{j=1}^{N} p(\omega_i)p(\lambda_j|\omega_i)\mathrm{g}(\boldsymbol{x}'_k|\omega_i,\lambda_j)} \quad (24)$$

and

$$p(\lambda_j|\boldsymbol{x}'_k,\omega_i) = \frac{p(\lambda_j|\omega_i)\mathrm{g}(\boldsymbol{x}'_k|\omega_i,\lambda_j)}{\sum_{j=1}^{N} p(\lambda_j|\omega_i)\mathrm{g}(\boldsymbol{x}'_k|\omega_i,\lambda_j)} \quad (25)$$

and $\mathrm{g}(\boldsymbol{x}'|\omega_i,\lambda_j)$ is given from (13). Thus, all the parameters of the conversion function (23) are known from the resynthesis stage of the algorithm.

B. JDE Parameter Adaptation

Given the linearity of the transformations (11) and (15) and the fact that for a particular class ω_i, $\boldsymbol{x}$ and $\boldsymbol{y}$ will be jointly Gaussian, $\boldsymbol{x}'$ and $\boldsymbol{y}'$ will also be jointly Gaussian for a particular class ω_i and λ_j. Thus

$$\begin{aligned}\mathrm{E}(\mathbf{y}'|\boldsymbol{x}'_k,\omega_i,\lambda_j) &= \boldsymbol{\mu}_i^{y'} + \boldsymbol{\Sigma}_i^{y'x'}\boldsymbol{\Sigma}_i^{x'x'^{-1}}\left(\boldsymbol{x}'_k - \boldsymbol{\mu}_i^{x'}\right)\\ &= \mathbf{A}_j\boldsymbol{\mu}_i^y + \boldsymbol{b}_j + \mathbf{A}_j\boldsymbol{\Sigma}_i^{yx}\boldsymbol{\Sigma}_i^{xx^{-1}}\mathbf{A}_j^{-1}\\ &\quad\times\left(\boldsymbol{x}'_k - \mathbf{A}_j\boldsymbol{\mu}_i^x - \boldsymbol{b}_j\right)\end{aligned} \quad (26)$$

since

$$\boldsymbol{\Sigma}_i^{x'x'} = \mathbf{A}_j\boldsymbol{\Sigma}_i^{xx}\mathbf{A}_j^T, \quad \boldsymbol{\Sigma}_i^{y'x'} = \mathbf{A}_j\boldsymbol{\Sigma}_i^{yx}\mathbf{A}_j^T \quad (27)$$

and

$$\boldsymbol{\mu}_i^{x'} = \mathbf{A}_j\boldsymbol{\mu}_i^x + \boldsymbol{b}_j, \quad \boldsymbol{\mu}_i^{y'} = \mathbf{A}_j\boldsymbol{\mu}_i^y + \boldsymbol{b}_j. \quad (28)$$

It is also true that under the above analysis, the pdf of $\boldsymbol{y}'$ will be

$$\mathrm{g}(\mathbf{y}') = \sum_{i=1}^{M}\sum_{j=1}^{N} p(\omega_i)p(\lambda_j|\omega_i) \times\mathcal{N}\left(\mathbf{y}';\mathbf{A}_j\boldsymbol{\mu}_i^y + \boldsymbol{b}_j,\mathbf{A}_j\boldsymbol{\Sigma}_i^{yy}\mathbf{A}_j^T\right). \quad (29)$$

Finally, the conversion function for synthesis will be

$$\begin{aligned}\mathcal{F}(\boldsymbol{x}'_k) &= \mathrm{E}(\mathbf{y}'|\boldsymbol{x}'_k)\\ &= \sum_{i=1}^{M}\sum_{j=1}^{N} p(\omega_i|\boldsymbol{x}'_k)p(\lambda_j|\boldsymbol{x}'_k,\omega_i)\\ &\quad\times\left[\mathbf{A}_j\boldsymbol{\mu}_i^y + \boldsymbol{b}_j + \mathbf{A}_j\boldsymbol{\Sigma}_i^{yx}\boldsymbol{\Sigma}_i^{xx^{-1}}\mathbf{A}_j^{-1}\right.\\ &\quad\left.\left(\boldsymbol{x}'_k - \mathbf{A}_j\boldsymbol{\mu}_i^x - \boldsymbol{b}_j\right)\right]\end{aligned} \quad (30)$$

where $p(\omega_i|\boldsymbol{x}'_k)$ and $p(\lambda_j|\boldsymbol{x}'_k,\omega_i)$ are given from (24) and (25) respectively, and $\mathrm{g}(\boldsymbol{x}'|\omega_i,\lambda_j)$ is given from (13). Again, all the parameters of the conversion function (30) are known from the resynthesis stage of the algorithm. It is of interest to note that the conversion function derived for the JDE synthesis problem is optimal in mean-squared sense while the conversion function for LSE synthesis is not optimal in any sense. Also, in the case of diagonal conversion, the correction of the covariance matrices cancels in (23) and (30), thus the adaptation essentially occurs only for the means.

IV. Results and Discussion

In this section, we apply the algorithms derived for synthesis to two multimicrophone recordings, one of classical music and one of modern music, that we obtained from two US concert halls, using a large number of microphones in various locations. As explained earlier, 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. We first test the resynthesis performance, since the success of the synthesis algorithm is highly dependent on the resynthesis results. Objective results are given that measure how close the derived cepstral coefficients are to the desired ones. The same test is next taken for the adaptation performance for multichannel synthesis, but now the desired responses are available only for testing, since there is no training phase in this case (instead, there is now an adaptation phase). Subjective results from listening tests are also given, as an additional means of validating our methods. Finally, some issues related with percussive sound resynthesis and synthesis are briefly discussed.

A. Spectral Conversion Performance

The three spectral conversion methods outlined in Section II-A were implemented and tested using a multichannel recording, obtained as described in Section I of this paper. The objective was to recreate the channel 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. More generally, it might hold that a spot microphone might enhance more than one types of musical sources. Usually, such microphones are placed with a particular type of instruments in mind which is easy to discern by acoustical examination, but, in general, careful selection of the training data should result in the desirable result even in complex cases.

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. Given that our focus was on modifying the short-term spectral properties of the reference signal, the analysis window we used was a 2048 sample window for a 44.1-kHz sampling rate. This is a typical value often used when the objective is to alter the short-term

TABLE I
PARAMETERS FOR THE CHORUS MICROPHONE **RESYNTHESIS** EXAMPLE

Band	Frequency Range		LP	Mixtures	
Nr.	Low (kHz)	High (kHz)	Order	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 II
NORMALIZED DISTANCES FOR LSE-, JDE- AND VQ-BASED METHODS, FOR FULL AND DIAGONAL CONVERSION

SC	Covari-	Ceps. Distance		Centroids
Method	ance	Train	Test	per Band
LSE	Full	0.6451	0.7144	Table I
	Diag	0.5918	0.7460	Table I
JDE	Full	0.6629	0.7445	Table I
	Diag	0.6524	0.7508	Table I
VQ	N/A	1.2903	1.3338	1024

TABLE III
PARAMETERS FOR THE CHORUS MICROPHONE **SYNTHESIS** EXAMPLE

Band	LP	GMM	Components			
Nr.	Order	Classes	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 IV
NORMALIZED DISTANCES FOR LSE METHOD WITHOUT ADAPTATION ("NONE") AND WITH SEVERAL COMPONENTS ADAPTATION (M-1 TO M-4) FOR DIAGONAL CONVERSION

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

TABLE V
NORMALIZED DISTANCES FOR JDE METHOD WITHOUT ADAPTATION ("NONE") AND SEVERAL COMPONENTS ADAPTATION (M-1 TO M-4) FOR DIAGONAL CONVERSION

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

spectral properties of audio signals, and was found to produce good sound quality results in our case as well. Results were evaluated through both objective and subjective performance criteria (results from the listening tests are given in Section IV-C). The spectral conversion methods were found to provide promising enhancement results. The experimental conditions for the objective tests are given in Table I. 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 was considered. The order of the LP 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 II, the average quadratic cepstral distance (averaged over all vectors and all eight bands) is given for each method, for the training data as well as for the data used for testing (9 s 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 LP parameters). The two cases tested are the JDE and LSE spectral conversion algorithms with full and diagonal covariance matrices, as explained in Section II-B. 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. Results for full conversion were also given in [2]. Here, we test the efficiency of both diagonal and full conversion when applied to the resynthesis problem.

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 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 II compared to the number of centroids used for the GMM-based algorithms. However, the results for the VQ-based algorithm were still unacceptable both from the objective and subjective perspectives (a higher number of centroids was tested, up to 8192, without any significant improvement).

B. Adaptation Performance

The experimental conditions for the synthesis example (spectral conversion followed by parameter adaptation) are given in Table III. 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 in Section IV-A).

In Tables IV and V, the average quadratic cepstral distance for the synthesis example is given, for the LSE and JDE methods respectively. The objective is to test the performance of the adaptation method for two different cases. The first case is 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 derived as explained in the resynthesis algorithm, 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 Tables IV and V (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 Tables IV and V 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 (more evident for LSE synthesis 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 is when the GMM parameters correspond 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 which we obtained from a live modern music performance as explained in Section I. The GMM parameters were derived 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 Tables IV and V (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 degradation of performance stems from the fact that the conversion parameters were derived based on a completely different recording than the one they were applied to. It is expected that in order to take full advantage of the adaptation methods, the training database should contain a large number of recordings of nature as diverse as possible.

C. Subjective Tests

The effectiveness of the proposed methods was measured by employing subjective tests as well. Two different tests were designed, an ABX test and a perceptual preference test, based on previously conducted listening tests for tasks such as voice conversion (*e.g.*, [6]). Note that the resynthesis method used for obtaining the synthesized waveforms for the tests was the one of (3), while the adaptation method was the one of (23). This decision was made since the two resynthesis methods as well as the two adaptation methods that were proposed in this work result in acoustically identical recordings, for the same design conditions (training dataset, number of conversion parameters, number of adaptation parameters). For this particular test, we used the number of parameters of Table I for synthesizing the waveforms. The ABX test was designed for evaluating the performance of the resynthesis methods. Since the synthesis algorithms are based on the resynthesis parameters, it is essential to test how effective acoustically the resynthesis algorithms perform. The performance of the synthesis methods will always be bounded by the performance on the resynthesis parameters on which it is based on. On the other hand, since the resynthesis parameters are derived for a particular recording, their application to a different recording results in significant quality degradation. The adaptation methods for synthesis attempt to address this issue, by adapting the resynthesis parameters to a different recording. Thus, we decided to test the synthesis performance by employing a perceptual preference test, that evaluates the acoustical quality of the synthesized waveform (*i.e.*, after adapting the resynthesis parameters to the statistics of the new recording) compared to the waveform that is synthesized by directly applying the resynthesis parameters (*i.e.*, without adaptation of the parameters). The number of parameters used for synthesizing the test waveforms is the same as in case M-4 of Table III, and the case examined is the one termed as "Same" in Tables IV and V, *i.e.*, when a small percentage of the available training dataset is used in order to simulate a stereo recording similar to the multimicrophone recording available.

One important issue that distinguishes the ABX test designed here is the selection of the waveforms that are labeled as A, B, or X. In a conventional ABX test, A and B would represent the target and source waveforms, and X would represent the synthesized waveform. Then, the subjects participating in the listening test would have to choose whether X is perceptually closer to A or to B. In our test, X was always the target response, while A and B were the source waveform and synthesized waveform, in random order. As already mentioned, the objective of the methods designed in this work for resynthesis and synthesis is not necessarily to synthesize identical responses to the target waveform. The objective is to synthesize a waveform that retains most of the key properties of the target response, mainly emphasizing particular instruments or voices and diminishing others. The ABX test was designed in this way so that the listener will compare the source and target waveforms and decide which one is closer perceptually to the desired response, although the two responses might not necessarily sound identical.

In the particular ABX test we designed, the training database was derived in such manner as to emphasize the chorus of the orchestra in the reference recording while diminishing the other instruments present, mainly woodwind and stringed instruments. The listeners then were asked to decide whether X was perceptually closer to A or B, based on whether they per-

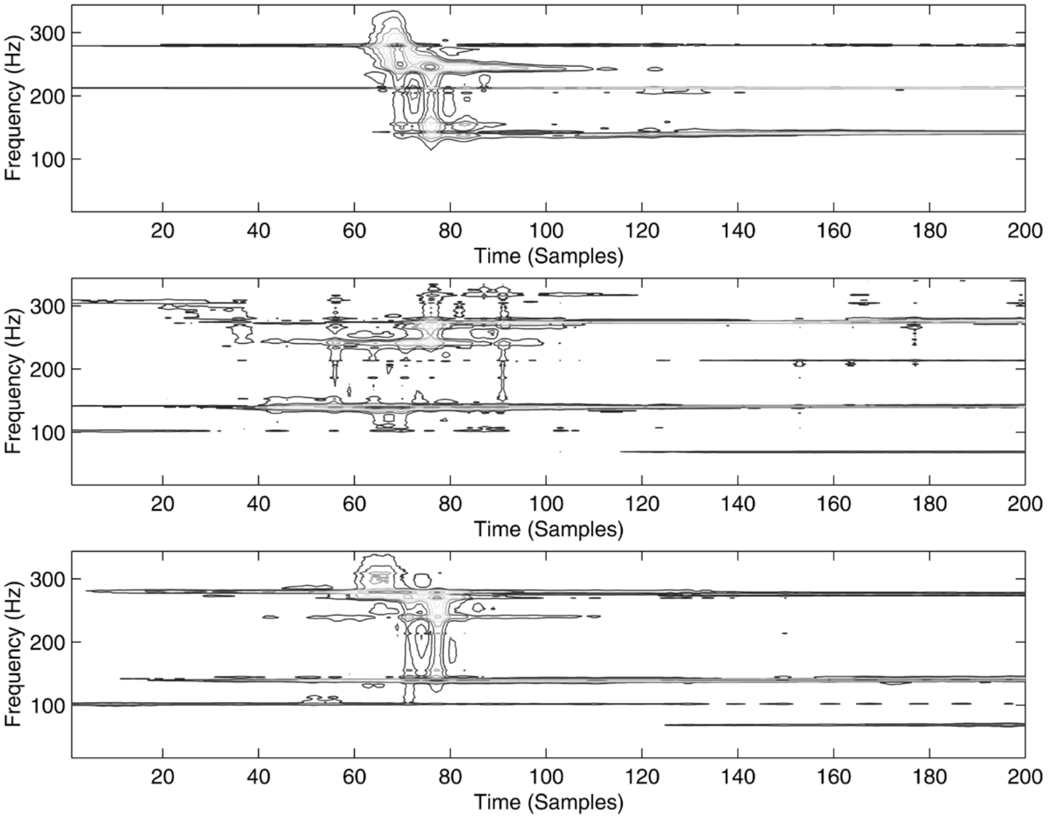

Fig. 3. Choi–Williams distribution of the (top) desired, (middle) reference, and (bottom) resynthesized waveforms at the time points during a tympani strike (samples 60–80).

TABLE VI
RESULTS FROM THE ABX AND PREFERENCE LISTENING TESTS

	ABX Test	Pref. Test
Results correct	72%	89%

ceived (for each one of the three waveforms) the voices being prominent, and the woodwind and stringed instruments in the background, or the opposite. We synthesized three waveforms, about 9 s each, that were not included in the training data. These waveforms, along with the corresponding source and target recordings, were presented to all twelve participants of the listening test in random order, with the choice of A, B, and X, as explained in the previous paragraph (three ABX tests per listener). The results of this test are shown in Table VI, and it is evident that the resynthesis performance is quite satisfactory, considering the complexity of the task. The results labeled as “Results correct” in the table for this test correspond to the case when the synthesized waveform was identified as closer to the target response. It is of interest to note that most listeners noticed a quality degradation for the synthesized waveform; however, in most cases, this did not seem to influence their decision. Quality improvement of the synthesized signals is a very important issue that must be considered for further improving the proposed methods. Among the parameters that are related with the resulting signal quality, the most important are the design and size of the training database, the number of GMM parameters, as well as the underlying model used for representing audio signals.

For the perceptual preference test, the same number of listeners had to decide whether the waveform with adaptation or without adaptation was of better acoustical quality. The number of synthesized waveforms was three pairs (each pair corresponding to the same source waveform, synthesized with and without adaptation), and were presented in random order to the twelve participants of the test (three perceptual preference tests per listener). The results are shown in Table VI, and it is clear that the adaptation methods performance is satisfactory. The results labeled as “Results correct” in the table for this test, correspond to the case when the synthesized waveform after adaptation was identified as the one of superior quality. The quality of the synthesized waveforms is worse when compared to the original waveforms, but adaptation produces much improved results when compared to the spectral conversion without adaptation, as the results of the test demonstrate.

We believe other types of tests will be very useful to evaluate the methods proposed here as well. One issue is localization, *i.e.*, whether there is perceptual improvement when all the synthesized waveforms along with the original recordings are mixed and rendered through a multichannel audio system. We are currently experimenting on such issues and methods on quantifying the results.

D. Percussive Sound Synthesis

The algorithm described in Section II-A considering the special case of percussive sound resynthesis was tested. Fig. 3 shows the time-frequency evolution of a tympani instance using the Choi-Williams distribution [23], a distribution that achieves the high resolution needed in such cases of impulsive signal nature. Fig. 3 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 the sense of proximity to the audio signal, as informal listening tests clearly demonstrated.

It is of interest to consider possible methods to manipulate such sounds for the synthesis case as well. The method described for percussive sound resynthesis must be modified for the case of synthesis, where the exact excitation signal is not available. One possible solution to this problem is to use the excitation signal of the same type of instrument obtained from another recording and modify it so as to provide an acoustically closer result to the desired. The idea is that this second recording contains isolated or enhanced instances of the instrument that can be used to extract a residual signal similar to what we would have in the resynthesis case. Application of this approach gave satisfying results although the sound produced differs when compared to the actual target recording (available only for testing in this case). This is true since the excitation signal, even with no modeling error, characterizes the way the instrument is excited, which might be different for different recordings. However, the synthesized signal usually contains no significant artifacts and the main difficulty is that of locating the exact instances of the instrument, as noted in Section II-D as well.

V. Conclusions

We termed multichannel audio resynthesis as 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 toward 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 by deriving more subjective evaluation tests in addition to the objective and subjective measures that were utilized in this paper. For example, it would be interesting to derive listening tests that measure the amount of localization that the synthesized multichannel recordings achieve, since this is one of the essential goals of the derived algorithms. It is also of interest to verify the validity of our assumption, that a more diverse collection of recordings would result in easily generalizable GMM parameters. A large number of multimicrophone recordings of various types of music performances is required in this case, which is far from trivial to acquire.

It should be noted that 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 because of their acoustical significance and because models for these sounds are available. 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.

Finally, the listening tests we performed made clear that the quality of the synthesized signals must be improved. A different model for audio signals, such as a sinusoidal model, might possibly result in perceptual improvement of the proposed algorithms. We are also investigating the effect of the design of the training database on the resulting signal quality, for various synthesis objectives.

Acknowledgment

The authors would like to thank Prof. J. Van der Spiegel for helping to organize the listening tests, as well as all the persons who participated in these tests.

References

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

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

[3] S. Haykin, *Adaptive Filter Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1996.

[4] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 2, pp. 236–242, Apr. 1984.

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

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

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

[8] G. Baudoin and Y. Stylianou, "On the transformation of the speech spectrum for voice conversion," in *Proc. IEEE Int. Conf. Spoken Language Processing (ICSLP)*, Philadephia, PA, Oct. 1996, pp. 1405–1408.

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

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

[11] V. V. Digalakis, D. Rtischev, and L. G. Neumeyer, "Speaker adaptation using constrained estimation of Gaussian mixtures," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 5, pp. 357–366, Sep. 1995.

[12] V. D. Diakoloukas and V. V. Digalakis, "Maximum-likelihood stochastic-transformation adaptation of Hidden Markov models," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 2, pp. 177–187, Mar. 1999.

[13] J. M. Mendel, *Lessons in Estimation Theory for Signal Processing, Communications and Control*. Englewood Cliffs, NJ: Prentice-Hall, 1995.

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

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

[16] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-34, no. 4, pp. 744–754, Aug. 1986.

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

[18] O. Cappe and E. Moulines, "Regularization techniques for discrete cepstrum estimation," *IEEE Signal Process. Lett.*, vol. 3, pp. 100–102, Apr. 1996.

[19] J. Laroche and J.-L. Meillier, "Multichannel excitation/filter modeling of percussive sounds with application to the piano," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 329–344, Apr. 1994.

[20] R. B. Sussman and M. Kahrs, "Analysis and resynthesis of musical instrument sounds using energy separation," in *IEEE Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Atlanta, GA, May 1996, pp. 997–1000.

[21] M. W. Macon, A. McCree, L. Wai-Ming, and V. Viswanathan, "Efficient analysis/synthesis of percussion musical instrument sounds using an all-pole model," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Seattle, WA, May 1998, pp. 3589–3592.

[22] J. Laroche, "A new analysis/synthesis system of musical signals using Prony's method-application to heavily damped percussive sounds," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Glasgow, U.K., May 1989, pp. 2053–2056.

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

Athanasios Mouchtaris (S'02–M'04) received the Diploma degree in electrical engineering from Aristotle University of Thessaloniki, Thessaloniki, Greece, in 1997 and the M.S. and Ph.D. degrees in electrical engineering from the University of Southern California, Los Angeles, in 1999 and 2003, respectively.

He is currently a Postdoctoral Researcher in the Electrical and Systems Engineering Department, University of Pennsylvania, Philadelphia. His research interests include signal processing for immersive audio environments, spatial audio, multichannel audio synthesis, speech synthesis with emphasis on voice conversion, and speech enhancement.

Dr. Mouchtaris is a member of Eta Kappa Nu.

Shrikanth S. Narayanan (M'87–SM'02) received the M.S. degree in 1990, the Engineer degree in 1992, and the Ph.D degree in 1995, all from the University of California, Los Angeles (UCLA).

He was with AT&T Research (originally AT&T Bell Labs), first as a Senior Member and later as a Principal Member of its Technical Staff from 1995 to 2000. Currently, he is an Associate Professor of Electrical Engineering, Linguistics, and Computer Science at the University of Southern California (USC), Los Angeles. He is a Research Area Director of the Integrated Media Systems Center and a National Science Foundation (NSF) Engineering Research Center at USC. His research interests include signals and systems with applications to speech, language, multimodal, and biomedical problems. He has published over 125 papers and holds three U.S. patents.

Dr. Narayanan was an Associate Editor of the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING (2000–2003) and serves on the Speech Communication Technical Committee of the Acoustical Society of America, and the Speech Processing and Multimedia Signal Processing Technical Committees of the IEEE Signal Processing society. He is a member of Tau-Beta-Pi, Phi Kappa Phi, and Eta-Kappa-Nu. He received the NSF CAREER Award, the USC Engineering Junior Research Award, and a Center for Interdisciplinary Research Faculty Fellowship.

Chris Kyriakakis (M'97) received the B.S. degree from the California Institute of Technology, Pasadena, in 1985 and the M.S. and Ph.D. degrees from the University of Southern California (USC), Los Angeles, in 1987 and 1993, respectively.

He is the Director of the Immersive Audio Laboratory within the School of Engineering, USC. He became a Faculty Member in the Department of Electrical Engineering, USC, in 1996 and an Associate Professor in 2002. His research interests include microphone array signal processing, virtual microphones, multichannel audio rendering, head-related transfer function modeling, room acoustics, psychoacoustics, and multichannel audio streaming over high-bandwidth networks.

Dr. Kyriakakis is a member of the Audio Engineering Society and the Acoustical Society of America.