EURASIP J. Applied Signal Processing
Special Issue: Digital Audio for Multimedia
Submitted

# Virtual Microphones for Multichannel Audio Resynthesis

**Athanasios Mouchtaris**
*Integrated Media Systems Center (IMSC), Electrical Engineering-Systems Department,*
*University of Southern California, 3740 McClintock Ave. EEB 428, Los Angeles, CA 90089-2564, USA*
*Email: mouchtar@sipi.usc.edu*

**Shrikanth S. Narayanan**
*Integrated Media Systems Center (IMSC), Electrical Engineering-Systems Department,*
*University of Southern California, 3740 McClintock Ave. EEB 430, Los Angeles, CA 90089-2564, USA*
*Email: shri@sipi.usc.edu*

**Chris Kyriakakis** *
*Integrated Media Systems Center (IMSC), Electrical Engineering-Systems Department,*
*University of Southern California, 3740 McClintock Ave. EEB 432, Los Angeles, CA 90089-2564, USA*
*Email: ckyriak@imsc.usc.edu*

Multichannel audio offers significant advantages for music reproduction that include the ability to provide better localization and envelopment, as well as reduced imaging distortion. On the other hand, multichannel audio is one of the most demanding media types in terms of transmission requirements. A novel architecture was previously proposed, allowing delivery of uncompressed multichannel audio over high-bandwidth communications networks. In most cases, however, bandwidth limitations prohibit transmission of multiple audio channels. In such cases, an alternative would be to transmit only one or two reference channels and recreate the rest of the channels at the receiving end. In this paper, we propose a system that is capable of synthesizing the required signals from a smaller set of signals recorded in a particular venue. These synthesized "virtual" microphone signals can be used to produce multichannel recordings that accurately capture the acoustics of the particular venue. Applications of the proposed system include transmission of multichannel audio over the current Internet infrastructure and, as an extension of the methods proposed here, remastering of existing monophonic and stereophonic recordings for multichannel rendering.

**Keywords and phrases:** Multichannel audio, Gaussian Mixture Model, distortion measures, virtual microphones, audio resynthesis, multiresolution analysis.

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

# 1 Introduction

Multichannel audio can enhance the sense of immersion for a group of listeners by reproducing the sounds that would originate from several directions around the listeners, thus simulating the way we perceive sound in a real acoustical space. On the other hand, multichannel audio is one of the most demanding media types in terms of transmission requirements. A novel architecture allowing delivery of uncompressed multichannel audio over high-bandwidth communications networks was presented in [1]. As suggested there, for applications in which bandwidth limitations prohibit transmission of multiple audio channels, an alternative would be to transmit only one or two channels (denoted as *reference* channels or recordings in this work, *e.g.* the left and right signals in a traditional stereo recording) and reconstruct the remaining channels at the receiving end. The system proposed in this paper provides a solution for reconstructing the channels of a specific recording from the reference channels and is particularly suitable for live concert hall performances. The proposed method is based on information of the acoustics of a specific concert hall and the microphone locations with respect to the orchestra, information that can be extracted from the specific multichannel recording.

Before proceeding to the description of the method proposed, a brief outline of the basis of our approach is given. A number of microphones are used to capture several characteristics of the venue, resulting in an equal number of *stem recordings* (or *elements*). Fig. 1, provides an example of how microphones may be arranged in a recording venue in a multichannel recording. These recordings are then mixed and played back through a multichannel audio system that attempts to recreate the spatial realism of the recording venue. Our objective is to design a system based on available stem recordings that is able to recreate all of these recordings from the reference channels at the receiving end (thus, stem recordings are also referred to as *target* recordings here). The result would be a significant reduction in transmission requirements, while enabling mixing at the receiving end. Consequently, such a system would be suitable for completely resynthesizing any number of channels in the initial recording (*i.e.* no information needs to be transmitted about the target recordings other than the conversion parameters). This is different than what commercial systems accomplish today. In addition, the system proposed in this paper is a structured representation of multichannel audio that lends itself to other possible applications such as multichannel audio synthesis which is briefly described later in this section. By examining the acoustical characteristics of the various stem recordings, the distinction of microphones is made into reverberant and spot microphones.

Spot microphones are microphones that are placed close to the sound source (*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 here that focuses on this problem is based on spectral conversion (SC). The special case of percussive drum-like sounds is separately examined since these sounds are of impulsive nature and cannot be addressed by spectral conversion methods. These sounds are of particular interest however, since they greatly affect our perception of proximity to the orchestra.

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

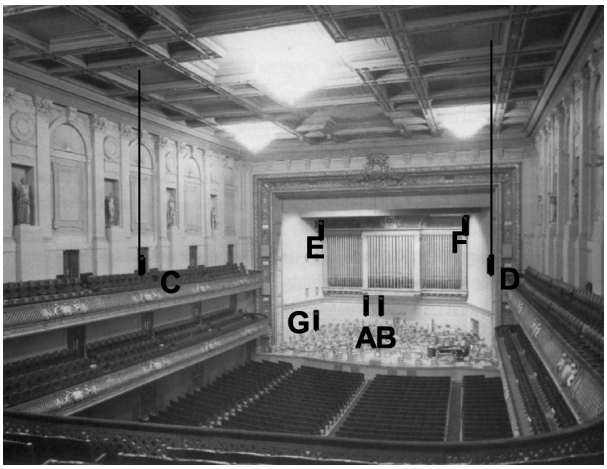


Figure 1: An example of how microphones may be arranged in a recording venue for a multichannel recording. In the virtual microphone synthesis algorithm, microphones A and B are the main reference pair from which the remaining microphone signals can be derived. Virtual microphones C and D capture the 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.

because they mainly capture reverberant information (that can be reproduced by the surround channels in a multichannel playback system). The recordings captured by these microphones can be synthesized by filtering the reference recordings through linear time-invariant (LTI) filters, designed using the methods that will be described in later sections of this paper. Existing reverberation methods use a combination of comb and all-pass filters to effectively add reverberation to the existing monophonic or stereophonic signal. Our objective is to estimate the appropriate filters that capture the concert hall acoustical properties from a given set of stem microphone recordings. We describe an algorithm that is based on a spectral estimation approach and is particularly suitable for generating such filters for large venues with long reverberation times. Ideally, the resulting filter implements the spectral modification induced by the hall acoustics.

We have obtained such stem microphone recordings from two orchestra halls in the US 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. It should be noted that the methods proposed here intend to provide a solution for the problem of resynthesizing *existing* multichannel recordings from a smaller subset of these recordings. The problem of completely synthesizing multichannel recordings from stereophonic (or monophonic) recordings, thus greatly augmenting the listening experience, is not addressed here. The synthesis problem is a topic of related research to appear in a future publication. However, it is important to distinguish the cases where these two problems (synthesis and resynthesis) differ. For reverberant microphones, since the result of our method is a group of LTI filters, both problems are addressed at the same time. The filters designed are capable of recreating the acoustic properties of the venue where the specific recordings took place. If these filters are applied to an arbitrary (non-reverberant) recording, the resulting signal will contain the venue characteristics at the

particular microphone location. In such manner, it is possible to completely synthesize reverberant stem recordings and synthesize a multichannel recording. In contrary, this will not be possible for the stem microphone methods. As it will be clear later, the algorithms described here are based on the specific recordings that are available. The result is a group of spectral conversion functions that are designed by estimating the unknown parameters based on training data that are available from the target recordings. These functions cannot be applied to an arbitrary signal and produce meaningful results. This is an important issue when addressing the synthesis problem and will not be the topic of this paper.

The remainder of this paper is organized as follows. In Section 2 the spot microphone resynthesis problem is addressed. Spectral conversion methods are described and applied to the problem in different subbands of the audio signal. The special case of percussive sounds is also examined. In Section 3 the reverberant microphone resynthesis problem is examined. The issue of defining an objective measure of the method's performance arises which is addressed by defining a normalized mutual information measure. Finally, a brief discussion of the results is given in Section 4 and possible directions for future research on the subject are proposed.

# 2 Spot Microphone Resynthesis

## 2.1 Spectral Conversion

The goal is to modify the short-term spectral properties of the reference audio signal in order to recreate the desired one. The short-term spectral properties are extracted by using a short sliding window with overlapping (resulting in a sequence of signal segments or frames). Each frame is modeled as an autoregressive (AR) filter excited by a residual signal. The AR filter coefficients are found by means of linear predictive analysis (LPC, [2]) and the residual signal is the result of inverse filtering the audio signal of the current frame 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 [3].

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

Such a mapping function can be designed by following the approach of voice conversion algorithms [4–6]. 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. The method to do that is briefly described next.

Assuming that a sequence $[\boldsymbol{x}_1\boldsymbol{x}_2\ldots\boldsymbol{x}_n]$ of reference spectral vectors (*e.g.* line spectral frequencies (LSF's), cepstral coefficients, *etc.* ) is given, as well as the corresponding sequence of target spectral vectors $[\boldsymbol{y}_1\boldsymbol{y}_2\ldots\boldsymbol{y}_n]$ (training data from the reference and target recordings respectively), a function $\mathcal{F}(\cdot)$ can be designed which, when applied to vector $\boldsymbol{x}_k$, produces a vector close in some sense to vector $\boldsymbol{y}_k$. Many algorithms have been described for designing this function (see [4–7] and the references therein). Here the algorithms based on vector quantization (VQ, [4]) and Gaussian mixture models (GMM, [5, 6]) were implemented and compared.

### 2.1.1 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 [8] 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.

### 2.1.2 Spectral Conversion based on GMM

In this case, the assumption made is that the sequence of spectral vectors $\boldsymbol{x}_k$ is a realization of a random vector $\boldsymbol{x}$ with a probability density function (pdf) that can be modeled as a mixture of $M$ multivariate Gaussian pdf's. Thus, the pdf of $\boldsymbol{x}$, $\mathrm{g}(\boldsymbol{x})$, can be written 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 $\omega_i$. The parameters of the GMM, *i.e.* the mean vectors, covariance matrices and priors, can be estimated using the expectation maximization (EM) algorithm [9].

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 [5], 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 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 $\omega_i$, $p(\omega_i|\boldsymbol{x}_k)$ 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 $\boldsymbol{\Gamma}_i$, $i = 1, \ldots, M$) 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 [6]. Assuming that $\boldsymbol{x}$ and $\boldsymbol{y}$ are jointly Gaussian for each class $\omega_i$, 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,

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

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.

### 2.2 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 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. Perfect reconstruction filter banks, based on wavelets [10], 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.

### 2.3 Residual Processing for Percussive Sounds

The SC methods described earlier will not produce the desired result in all cases. Transient sounds 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 [11]. It is suitable since it also models the audio signal in different bands, in each one as a sinusoidal/residual model [12,13]. The sinusoidal parameters can be treated in the same manner as the LP coefficients during spectral conversion [14]. We are currently considering this model for improving the produced sound quality of our system. However, no structured model is proposed in [11] for transient sounds. In the remainder of this section, the special case of percussive sounds is addressed.

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

Table 1: Parameters for the chorus microphone example.

The case of percussive drum-like sounds is considered of particular importance. 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 [15] (see also [16–18]), 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 [15], 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 the next section. The drawback of this approach is that a robust algorithm is required for identifying the particular instrument instances in the reference recording. 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.

### 2.4 Implementation Details

The three spectral conversion methods outlined in Section 2.1 were implemented and tested using a multichannel recording, obtained as described in Section 1 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. A database of about 10,000 spectral vectors for each band was created so that only parts

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

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

of the recording where the chorus is present are used, with the choice of spectral vectors being the cepstral coefficients. Parts of the chorus recording were selected so that there were no segments of silence included. Results were evaluated through informal listening tests and through objective performance criteria. The SC methods were found to provide promising enhancement results. The experimental conditions are given in Table 1. The number of octave bands used was 8, a choice that gives particular emphasis on the frequency band 0-5 kHz and at the same time does not impose excessive computational demands. The frequency range 0-5 kHz is particularly important for the specific case of chorus recording resynthesis since this is the frequency range where the human voice is mostly concentrated. For producing 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.

In Table 2, the average quadratic cepstral distance (averaged over all vectors and all 8 bands) is given for each method, for the training data as well as for the data used for testing (9 sec. of music from the same recording). The cepstral distance is normalized with the average quadratic distance between the reference and the target waveforms (*i.e.* without any conversion of the LPC parameters). The improvement is large for both the GMM-based algorithms, with the LSE algorithm being slightly better, for both the training and testing data. The VQ-based algorithm, in contrast, produced a deterioration in performance which was audible as well. This can be explained based on the fact that the GMM-based methods result in a conversion function which is continuous with respect to the spectral vectors. The VQ-based method, on the other hand, produces audible artifacts introduced by spectral discontinuities because the conversion is based on a limited number of existing spectral vectors. This is the reason why a large number of centroids was used for the VQ-based algorithm as seen in Table 2 compared to the number of centroids used for the GMM-based algorithms. However, the results were still unacceptable both from the objective and subjective perspectives.

The algorithm described in Section 2.1 considering the special case of percussive sound resynthesis was tested as well. Fig. 2 shows the time-frequency evolution of a tympani instance using the Choi-Williams distribution [19], a distribution that achieves the high resolution needed in such cases of impulsive signal nature. Fig. 2 clearly demonstrates the improvement in drum-like sound resynthesis. The impulsiveness of the signal at around samples 60-80 is observed in the desired response and verified in the synthesized waveform. The attack part is clearly enhanced, significantly adding naturalness in the audio signal, as our informal listening tests clearly demonstrated.

The methods described in this section can be used for synthesizing recordings of microphones that are placed close to the orchestra. Of importance in this case were the short-term spectral properties of the audio signals. Thus, linear time-invariant filters were not suitable and the time-frequency properties of the waveforms had to be exploited in order to obtain a solution. In the next section, we focus on microphones placed far

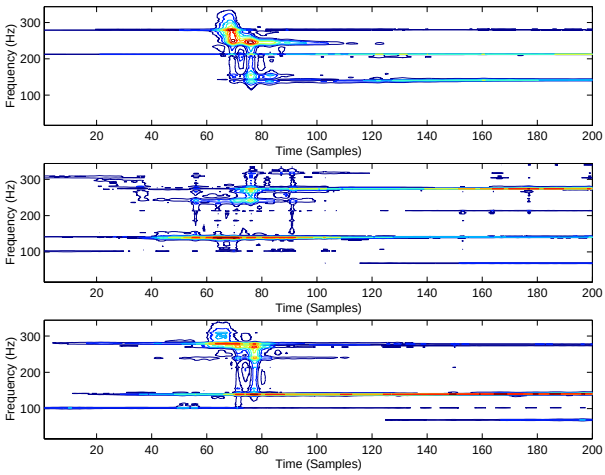


Figure 2: Choi-Williams distribution of the desired (top), reference (middle) and synthesized (bottom) waveforms at the time points during a tympani strike (samples 60-80).

from the orchestra and thus contain mainly reverberant signals. As we demonstrate, the desired waveforms can be synthesized by taking advantage of the long-term spectral properties of the reference and the desired signals.

# 3 Reverberant Microphone Signal Synthesis

The problem of synthesizing a virtual microphone signal from a signal recorded at a different position in the room can be described as follows. Given two processes $s_1$ and $s_2$, determine the optimal filter $H$ that can be applied to $s_1$ (the reference microphone signal) so that the resulting process $s_2'$ (the virtual microphone signal) is as close as possible to $s_2$. The optimality of the resulting filter $H$ is based on how "close" $s_2'$ is to $s_2$. For the case of microphone signals, the distance between these two processes must be measured in a way that is psychoacoustically valid. We can treat this as a typical system identification problem. However, there are several unique aspects that need to be considered, the most important being that the physical system is characterized by a long impulse response. For a typical large symphony hall the reverberation time is approximately 2 sec., which would require a filter of more than 96000 taps to describe the reverberation process (for a typical sampling rate of 48 kHz).

## 3.1 IIR Filter Design

There are several possible approaches to the problem. One is to use classical estimation theoretic techniques such as least-squares or Wiener filtering based algorithms to estimate the hall environment with a long finite-duration impulse response (FIR) or infinite-duration impulse response (IIR) filter. Adaptive algorithms such as LMS [2] can provide an acceptable solution in such system identification problems while least-squares methods suffer prohibitive computational demands. For LMS the limitation lies in the fact that the input and the output are non-stationary signals making its convergence quite slow. In addition, the required length of the filter is very large so such algorithms would prove to be inefficient for this problem. Although it is possible to prewhiten the input of the adaptive algorithm (see for example [2, 20] and references therein), so that convergence is improved, these algorithms still did not prove to be efficient for this problem.

An alternative to the aforementioned methods for treating system identification problems, is to use spectral estimation techniques based on the cross-spectrum [21]. These methods are divided into parametric and non-parametric. Non-parametric methods, based on averaging techniques such as the averaged periodogram (Welch spectral estimate) [22–24] are considered more appropriate for the case of long observations and for non-stationary conditions since no model is assumed for the observed data (a different approach based on the cross-spectrum which, instead of averaging, solves an overdetermined system of equations can be found in [25]). After the frequency response of the filter is estimated, an IIR filter can be designed based on that response. The advantage of this approach is that IIR filters are a more natural choice of modeling the physical system under consideration and can be expected to be very efficient in approximating the spectral properties of the recording venue. In addition an IIR filter would implement the desired frequency response with a significantly lower order compared to an FIR filter. Caution must, of course, be taken in order to ensure the stability of the filters.

To summarize, if we could define a power spectral density $S_{s_1}(\omega)$ for signal $s_1$ and $S_{s_2}(\omega)$ for signal $s_2$, then it would be possible to design filter $H(\omega)$ that can be applied to process $s_1$ resulting in process $s_2'$, which is intended to be an estimate of $s_2$. The filter $H(\omega)$ can be estimated by means of spectral estimation techniques. Furthermore, if $S_{s_1}(\omega)$ is modeled by an all-pole approximation $|1/A_{p1}|^2$ and $S_{s_2}(\omega)$ similarly as $|1/A_{p2}|^2$ then $H = A_{p1}/A_{p2}$, if $H$ is restricted to be the minimum phase spectral factor of $|H(\omega)|^2$. This results in a stable IIR filter that can be designed efficiently but is minimum phase. The analysis that follows provides the details for designing $H$.

The estimation of $H(\omega)$ is based on computing the cross-spectrum $S_{s_2 s_1}$ of signals $s_2$ and $s_1$ and the auto spectrum $S_{s_1}$ of signal $s_1$. It is true that if these signals were stationary then

$$S_{s_2 s_1}(\omega) = H(\omega) S_{s_1}(\omega) \tag{7}$$

The difficulties arising in the design of filter $H$ are due to the non-stationary nature of audio signals. This issue can be partly addressed if the signals are divided into segments short enough that can be considered of approximately stationary nature. It must be noted, however, that these segments must be large enough so that they can be considered long compared to the length of the impulse response that must be estimated, in order to avoid edge effects (as explained in [26], where a similar procedure is followed for the case of blind deconvolution for audio signal restoration).

For interval $i$, composed from $M$ (real) samples $s_1^{(i)}(0), \ldots, s_1^{(i)}(M-1)$, the empirical transfer function estimate (ETFE, [21]) is computed as

$$\hat{H}^{(i)}(\omega) = \frac{S_2^{(i)}(\omega)}{S_1^{(i)}(\omega)} \tag{8}$$

where

$$S_1^{(i)}(\omega) = \sum_{n=0}^{M-1} s_1^{(i)}(n) e^{-j\omega n} \tag{9}$$

is the Fourier transform of the segment samples. This cannot be considered an accurate estimate of $H(\omega)$ though, since the filter $H^{(i)}(\omega)$ will be valid only for frequencies corresponding to the harmonics of segment $i$ (under the valid assumption of quasi-periodic nature of the audio signal for each segment). An intuitive procedure would be to obtain the estimate of the spectral properties of the recording venue $\hat{H}(\omega)$ by averaging all the estimates available. Since the ETFE is the result of frequency division, it is apparent that in frequencies where $S_{s_1}(\omega)$ is close to zero, the ETFE would become unstable, so

a more robust procedure would be to estimate $H$ using a weighted average of the $K$ segments available [21], *i.e.*

$$\hat{H}(\omega) = \frac{\sum_{i=0}^{K-1} \beta^{(i)}(\omega) H^{(i)}(\omega)}{\sum_{i=0}^{K-1} \beta^{(i)}(\omega)} \tag{10}$$

A sensible choice of weights would be

$$\beta^{(i)}(\omega) = |S_1^{(i)}(\omega)|^2 \tag{11}$$

It can be easily shown that estimating $H$ under this approach is equivalent to estimating the auto-spectrum of $s_1$ and the cross-spectrum of $s_2$ and $s_1$ using the Cooley-Tukey spectral estimate [23] (in essence Welch spectral estimation with rectangular windowing of the data and no overlapping). In other words, defining the power spectrum estimate under the Cooley-Tukey procedure as

$$S_{s_1}^{CT}(\omega) = \frac{1}{K} \sum_{i=0}^{K-1} |S_1^{(i)}(\omega)|^2 \tag{12}$$

where $S(\omega)$ is defined as previously, and a similar expression for the cross-spectrum

$$S_{s_2 s_1}^{CT}(\omega) = \frac{1}{K} \sum_{i=0}^{K-1} S_2^{(i)}(\omega) S_1^{(i)*}(\omega) \tag{13}$$

then, it holds that

$$\hat{H}(\omega) = \frac{S_{s_2 s_1}^{CT}(\omega)}{S_{s_1}^{CT}(\omega)} \tag{14}$$

which is analogous to (7). Thus, for a stationary signal, the averaging of the estimated filters is justifiable. A window can additionally be used to further smooth the spectra.

The method described is meaningful for the special case of audio signals, despite their non-stationarity. It is well known that the averaged periodogram provides a smoothed version of the periodogram. Considering that it is true even for non-stationary (but of finite length) signals that

$$S_2(\omega) S_1^*(\omega) = H(\omega) |S_1(\omega)|^2 \tag{15}$$

then averaging in essence smoothes the frequency response of $H$. This is justifiable since it is true that a non-smoothed $H$ will contain details that are of no acoustical significance. Further smoothing can yield a lower order IIR filter, by taking advantage of AR modeling. Considering signal $s_1$, the inverse Fourier transform of its power spectrum $S_{s_1}(\omega)$ derived as described earlier will yield the sequence $r_{s_1}(m)$. If this sequence is viewed as the autocorrelation of $s_1$ and samples $r_{s_1}(0), \cdots, r_{s_1}(p+1)$ are inserted in the Wiener-Hopf equations for linear prediction (with the AR order $p$ being significantly smaller than the number of samples of each block $M$, for smoothing the spectra)

$$\begin{bmatrix} r_{s_1}(0) & r_{s_1}(1) & \cdots & r_{s_1}(p-1) \\ r_{s_1}(1) & r_{s_1}(0) & \cdots & r_{s_1}(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ r_{s_1}(p-1) & r_{s_1}(p-2) & \cdots & r_{s_1}(0) \end{bmatrix} \begin{bmatrix} a_{p1}(1) \\ a_{p1}(2) \\ \vdots \\ a_{p1}(p) \end{bmatrix} = \begin{bmatrix} r_{s_1}(1) \\ r_{s_1}(2) \\ \vdots \\ r_{s_1}(p) \end{bmatrix} \tag{16}$$

then, the coefficients $a_{p1}(i)$ result in an approximation of $S_{s_1}(\omega)$ (omitting the constant gain term which is not of importance in this case)

$$S_{s_1}(\omega) = \left| \frac{1}{A_{p1}(\omega)} \right|^2 \tag{17}$$

where

$$A_{p1}(\omega) = 1 + \sum_{l=1}^{p} a_{p1}(l) e^{-j\omega l} \tag{18}$$

A similar expression holds for $S_{s_2}(\omega)$. $S_{s_1}$ and $S_{s_2}$ can be computed as in (12). Using the fact that

$$S_{s_2}(\omega) = |H(\omega)|^2 S_{s_1}(\omega) \tag{19}$$

and restricting $H$ to be minimum phase, we find from the spectral factorization of (19) a solution for $H$ is

$$H(\omega) = \frac{A_{p1}(\omega)}{A_{p2}(\omega)} \tag{20}$$

Filter $H$ can be designed very efficiently even for very large filter orders following this method since equation (16) can be solved using the Levinson-Durbin recursion. This filter will be IIR and stable.

A problem with the aforementioned design method is that the filter $H$ is restricted to be of minimum phase. It is of interest to mention that in our experiments the minimum phase assumption proved to be perceptually acceptable. This can be possibly attributed to the fact that if the minimum phase filter $H$ captures a significant part of the hall reverberation, then the listener's ear will be less sensitive to the phase distortion [27]. It is not possible, however, to generalize this observation and the performance of this last step in the filter design will possibly vary depending on the particular characteristics of the venue captured in the multichannel recording.

### 3.2 Mutual Information as a Spectral Distortion Measure

As previously mentioned, we need to apply the above procedure in blocks of data of the two processes $s_1$ and $s_2$. In our experiments, we chose signal block lengths of 100,000 samples (long blocks of data are required due to the long the reverberation time of the hall as explained earlier). We then experimented with various orders of filters $A_{p1}$ and $A_{p2}$. As expected, relatively high orders were required to reproduce $s_2$ from $s_1$ with an acceptable error between $s_2'$ (the resynthesized process) and $s_2$ (the target recording). The performance was assessed through blind A/B/X listening evaluation. An order of 10,000 coefficients for both the numerator and denominator of $H$ resulted in an error between the original and synthesized signals that was not detectable by listeners. We also evaluated the performance of the filter by synthesizing blocks from a part of the signal other than the one that was used for designing the filter. Again, the A/B/X evaluation showed that for orders higher than 10,000 the synthesized signal was indistinguishable from the original. Although such high order filters are impractical for real-time applications, the performance of our method is an indication that the model is valid and therefore motivating us to further investigate filter optimization. This method can be used for off-line applications such as remastering of old recordings. A real-time version was also implemented using the Lake DSP Huron digital audio convolution workstation. With this system we are able to synthesize 12 virtual microphone stem recordings from a monophonic or stereophonic compact disc (CD) in real time.

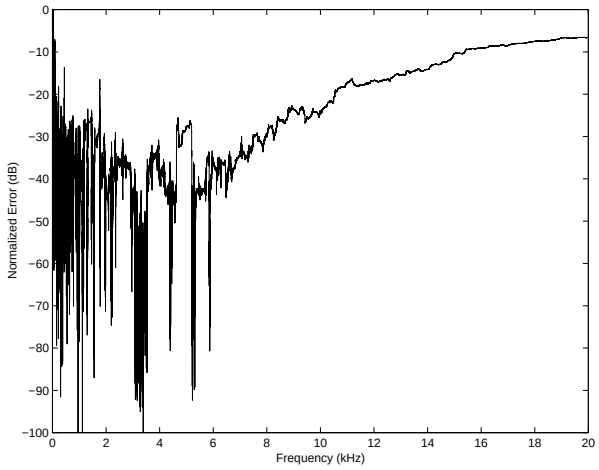


Figure 3: Normalized error between original and synthesized microphone signals as a function of frequency.

To obtain an objective measure of the performance it is necessary to derive a mathematical measure of the distance between the synthesized and the original processes. The difficulty in defining such a measure is that it must also be psychoacoustically valid. This problem has been addressed in speech processing where measures such as the log spectral distance and the Itakura-Saito distance are used [28]. In our case, we need to compare the spectral characteristics of long sequences with spectra that contain a large number of peaks and dips that are narrow enough to be imperceptible to the human ear. In other words, the focus is on the long-term spectral properties of the audio signals, while spectral distortion measures have been developed for comparing the short-term spectral properties of signals. To overcome comparison inaccuracies that would be mathematical rather than psychoacoustical in nature, we chose to perform 1/3 octave smoothing [29] and compare the resulting smoothed spectral cues. The results are shown in Fig. 3 in which we compare the spectra of the original (measured) microphone signal and the synthesized signal. The two spectra are practically indistinguishable below 10 kHz. Although the error increases at higher frequencies, the listening evaluations show that this is not perceptually significant. One problem that was encountered while comparing the 1/3 octave smoothed spectra was the fact that the average error was not reduced with increasing filter order as rapidly as the results of the listening tests suggested. To address this inconsistency we experimented with various distortion measures.

These measures included the RMS log spectral distance, the truncated cepstral distance, and the Itakura distance (for a description of all these measures see for example [8]). The results, however, were still not in line with what the listening evaluations indicated. This led us to a measure that is commonly used in pattern comparison and is known as the mutual information (see for example [30]). By definition, the mutual information of two random variables $X$ and $Y$ with joint probability density function (pdf) $p(x, y)$ and marginal pdf's $p(x)$ and $p(y)$ is the relative entropy between the joint distribution and the product distribution, *i.e.*

$$I(X;Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{21}$$

It is easy to prove that

$$I(X;Y) = H(X) - H(X|Y) \quad (22)$$
$$= H(Y) - H(Y|X) \quad (23)$$

and also

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \quad (24)$$

where $H(X)$ is the entropy of $X$

$$H(X) = -\sum_{x\in\mathcal{X}} p(x)\log p(x) \quad (25)$$

similarly, $H(Y)$ is the entropy of $Y$. $H(X|Y)$ is the conditional entropy defined as

$$H(X|Y) = \sum_{y\in\mathcal{Y}} p(y)H(X|Y=y) \quad (26)$$
$$= -\sum_{y\in\mathcal{Y}} p(y)\sum_{x\in\mathcal{X}} p(x|y)\log p(x|y) \quad (27)$$

while $H(X,Y)$ is the joint entropy defined as

$$H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y) \quad (28)$$

The mutual information is always positive. Since our interest is in comparing two vectors $X$ and $Y$ with $Y$ being the desired response, it is useful to use a modified definition for the mutual information, the Normalized Mutual Information (NMI) $I_N(X;Y)$ which can be defined as

$$I_N(X;Y) = \frac{H(Y) - H(Y|X)}{H(Y)} \quad (29)$$
$$= \frac{I(X;Y)}{H(Y)} \quad (30)$$

This version of the mutual information is mentioned in [30, p. 47] and has been applied in many applications as an optimization measure (*e.g.* radar remote sensing applications [31]). Obviously,

$$0 \le I_N(X;Y) \le 1$$

The NMI obtains its minimum value when $X$ and $Y$ are statistically independent and its maximum values when $X = Y$. The NMI does not constitute a metric since it lacks symmetry. On the other hand, the NMI is invariant to amplitude differences [32], which is a very important property especially for comparing audio waveforms.

The spectra of the original and the synthesized responses were compared using the NMI for various filter orders and the results are depicted in Fig. 4. The NMI increases with filter order both when considering the raw spectra, as well as when we used the spectra that were smoothed using AR modeling (spectral envelope by all-pole modeling with Linear Predictive coefficients). We believe that the NMI calculated using the smoothed spectra is the measure that closely approximates the results we achieved from the listening tests. As can be seen from the figure, the NMI for a filter order of 20,000 is 0.9386 (*i.e.*, close to unity which corresponds to indistinguishable similarity) for the LPC spectra while the NMI for the same order but for the raw spectra is 0.5124. Furthermore, the fact that both the raw and smoothed NMI measures increase monotonically in the same fashion indicates that the smoothing is valid since it only reduces the "distance" between the two waveforms in a proportionate way for all the synthesized waveforms (order 0 in the diagram corresponds to no filtering – it is the distance between the original and the reference waveforms).

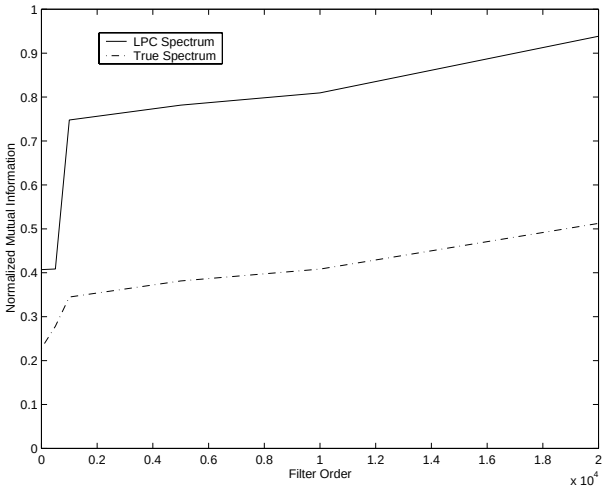


Figure 4: Normalized Mutual Information between original and synthesized microphone signals as a function of filter order.

# 4 Conclusions and Future Research

Multichannel audio resynthesis is a new and important application that allows transmission of only one or two channels of multichannel audio and resynthesis of the remaining channels at the receiving end. It offers the advantage that the stem microphone recordings can be resynthesized at the receiving end, which makes this system suitable for many professional applications and, at the same time, poses no restrictions on the number of channels of the initial multichannel recording. The distinction was made of the methods employed, depending on the location of the "virtual" microphones, namely spot and reverberant microphones. Reverberant microphones are those that are placed at some distance from the sound source (*e.g.* the orchestra) and therefore, contain more reverberation. On the other hand, spot microphones are located close to individual sources (*e.g.*, near a particular musical instrument). This is a completely different problem because placing such microphones near individual sources with varying spectral characteristics results in signals whose frequency content will depend highly on the microphone positions.

Spot microphones were treated separately by applying spectral conversion techniques for altering the short-term spectral properties of the reference audio signals. Spectral conversion algorithms that have been used successfully for voice conversion can be adopted for the task of multichannel audio resynthesis quite favorably. Three of the most common spectral conversion methods have been compared and our objective results, in accordance with our informal listening tests, have indicated that GMM-based spectral conversion can produce extremely successful results. Residual signal enhancement was also found to be essential for the special case of percussive sound resynthesis. Our current research is focused on audio quality improvement for the proposed methods, conducting formal listening tests as well as extensions of this research for the purpose of remastering existing monophonic and stereophonic recordings for multichannel rendering.

For the reverberant microphone recordings, we have described a method for synthesizing the desired audio signals, based on spectral estimation techniques. The emphasis in this case is on the long-term spectral properties of the signals since the reverberation process is considered to be long in duration (*e.g.* 2 seconds for large concert halls). An

IIR filtering solution was proposed for addressing the long reverberation-time problem, with associated long impulse responses for the filters to be designed. The issue of objectively estimating the performance of our methods arose, which was treated by proposing the normalized mutual information as a measure of spectral distance that was found to be very suitable for comparing the long-term spectral properties of audio signals. The IIR filters designed are currently not suitable for real-time applications. We are investigating other possible alternatives for the filter design that will result in more practical solutions.

## References


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

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

[3] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. ASSP-32, pp. 236–242, April 1984.

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

[5] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech and Audio Processing*, vol. 6, pp. 131–142, March 1998.

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

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

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

[9] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech and Audio Processing*, vol. 3, pp. 72–83, January 1995.

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

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

[12] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. ASSP-34, pp. 744–754, August 1986.

[13] X. Serra and J. O. Smith III, "Spectral modeling sythesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Computer Music Journal*, vol. 14, pp. 12–24, Winter 1990.

[14] O. Cappe and E. Moulines, "Regularization techniques for discrete cepstrum estimation," *IEEE Signal Processing Letters*, vol. 3, pp. 100–102, April 1996.

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

[16] 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), pp. 997–1000, May 1996.

[17] 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 *IEEE Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, (Seattle, WA), pp. 3589–3592, May 1998.

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

[19] H.-I. Choi and J. Williams, "Improved time-frequency representation of multicomponent signals using exponential kernels," *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. 37, pp. 862–871, June 1989.

[20] M. Mboup, M. Bonnet, and N. Bershad, "LMS coupled adaptive prediction and system identification: A statistical model and transient mean analysis," *IEEE Trans. Signal Processing*, vol. 42, pp. 2607–2615, October 1994.

[21] L. Ljung, *System Identification: Theory for the User*. Englewood Cliffs, N.J.: Prentice Hall, 1987.

[22] R. B. Blackman and J. W. Tukey, *The Measurement of Power Spectra*. New York, N.Y.: Dover Publications, 1958.

[23] J. W. Cooley and J. W. Tukey, "An algorithm for the machine calculation of complex Fourier series," *Math. Comp.*, vol. 19, pp. 297–301, April 1965.

[24] P. W. Welch, "The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE Trans. Audio and Electroacoustics*, vol. AU-15, pp. 70–73, June 1967.

[25] O. Shalvi and E. Weinstein, "System identification using nonstationary signals," *IEEE Trans. Signal Processing*, vol. 44, pp. 2055–2063, August 1996.

[26] J. T. G. Stockham, T. M. Cannon, and R. B. Ingebretsen, "Blind deconvolution through digital signal processing," *Proc. IEEE*, vol. 63, pp. 678–692, April 1975.

[27] B. D. Radlovic and R. A. Kennedy, "Nonminimum-phase equalization and its subjective importance in room acoustics," *IEEE Trans. Speech and Audio Processing*, vol. 8, pp. 728–737, November 2000.

[28] F. Itakura and S. Saito, "A statistical method for estimation of speech spectral density and formant frequencies," *Electronics and Communications in Japan*, vol. 53A, pp. 36–43, 1970.

[29] B. C. J. Moore, *An Introduction in the Psychology of Hearing.* San Diego, CA: Academic Press, 1989.

[30] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* Wiley, 1991.

[31] D. B. Trizna, C. Bachmann, M. Sletten, N. Allan, J. Topokovand, and R. Harris, "Projection pursuit classification methods applied to multiband polarimentric SAR imagery," in *Proc. IEEE International Geoscience and Remote Sensing Symposium (IGARSS) 2000*, vol. 1, (Honolulu, HI), pp. 105–107, July 2000.

[32] C. Shekhar and R. Chellappa, "Experimental evaluation of two criteria for pattern comparison and alignment," in *Proc. Fourteenth International Conference on Pattern Recognition*, vol. 1, (Brisbane, Australia), pp. 146–153, August 1998.