AN ALGEBRAIC THEORY OF 3D SOUND SYNTHESIS WITH LOUDSPEAKERS

PAUL G. FLIKKEMA

Department of Electrical Engineering, P.O. Box 15600, Flagstaff, Arizona 86011 USA
Paul.Flikkema@nau.edu

The problem of reproducing J signals with K loudspeakers is considered. A algebraic, time-domain synthesis approach is developed that extends the multi-input multi-output inverse theorem (MINT) of Miyoshi and Kaneda. The approach is general and can encompass joint surround sound synthesis and loudspeaker/room correction. First, a discrete time-domain matrix description is developed that captures the effects of amplifiers, loudspeakers, and room acoustics. Based on this model, it is shown that exact synthesis is possible with practical reproduction apparatus only if $K > J$. Sufficient conditions are also presented. The results are specialized to the case of zero-crosstalk transaural sound reproduction, and the theoretical importance of loudspeaker time alignment is illustrated. Finally, minimum-power exact synthesis is briefly described.

INTRODUCTION

How many loudspeakers are required to faithfully reproduce an acoustic signal at a particular set of points? Assuming that the desired signal is available, the key problems are amplifier/loudspeaker responses that depart from the ideal, room acoustics, and crosstalk cancellation. Existing approaches are based on inverse filtering and can be divided into two classes: non-adaptive and adaptive. Adaptive techniques [1, 2] are based on random signal models and therefore only approximate the true inverse, as pointed out in [3]. In particular, signal transients will not be tracked exactly due to memory of the adaptive filter, which may cause distortion of the signal. The non-adaptive exact inverse (MINT—multi-input multi-output inverse theorem) approach by Miyoshi and Kaneda [3] is based on the z-transform. Cross-talk cancellation and 3D sound synthesis techniques can also be defined using the z- or Fourier transform [2, 4] (see also [5] and the references therein). These techniques are based on complex exponential signal models defined over at least the positive time axis. For example, frequency-domain approaches consider only the steady-state response after initial transients.

In this paper we generalize the MINT approach by considering finite-duration signals in the time domain. The exact multiple-input, multiple-output (MIMO) response is characterized by a Toeplitz-block matrix, and therefore the approach is based on linear algebra. Rather than inverting or cancelling an undesired response, we consider whether it is possible to synthesize a collection of inputs that can generate the desired collection of outputs. The desired output set is general; in this paper we treat the synthesis of sound at a set of point locations, and the theory therefore applies to transaural, ambisonic, and ambiophonic sound reproduction.

This paper is organized as follows. Section 1 defines the time-domain matrix description of a MIMO loudspeaker sound synthesis system. Section 2 forms the core of the paper, and presents results on the achievability of exact synthesis, using transaural sound reproduction as an example case. The resulting interpretation of relative loudspeaker-receiver delays is given in Section 3. Section 4 touches on the implementation of exact synthesis, with conclusions following in Section 5.

We use the following nomenclature: $\text{len}(\boldsymbol{h})$ is the length of the vector $\boldsymbol{h}$. The transpose of a vector or matrix is given by the operator $(\cdot)^T$. The M-dimensional real vector space is denoted by $\mathbb{R}^M$, and the range space of the matrix $\boldsymbol{H}$ is $\mathcal{R}(\boldsymbol{H})$.

1. MIMO TIME-DOMAIN MODEL

Assume that there are K input signals (each uniquely associated with a loudspeaker) and J receivers (e.g., microphones or ears), and that the loudspeakers and receivers are at fixed locations. Let the finite-length discrete-time impulse response from loudspeaker k to receiver j (including amplifier, loudspeaker, and room effects) be denoted by the column vector $\boldsymbol{h}^{(jk)}$. (The response can be truncated based on some criterion if necessary.) Then for an output $\boldsymbol{f}^{(j)}$ of length L, any corresponding input $\boldsymbol{s}^{(k)}$ is of length $L - \text{len}(\boldsymbol{h}^{(jk)}) + 1$. If such a pair exists, they are related by

$$\boldsymbol{f}^{(j)} = \boldsymbol{H}^{(jk)} \boldsymbol{s}^{(k)}, \tag{1}$$

where $\boldsymbol{H}^{(jk)}$ is Toeplitz with first column the concatenation of $\boldsymbol{h}^{(jk)}$ and $L - \text{len}(\boldsymbol{h}^{(jk)})$ zeros, and whose k-th column is shifted down one position (with zero insertion at the top) from column $k-1$.

The composite output vector can be defined as the stacking of the vectors $\boldsymbol{f}^{(j)}$, $j = 1, \ldots, J$ into one vector $\boldsymbol{f}$.

Similarly, we can construct the composite input vector $\boldsymbol{s}$ from $\boldsymbol{s}^{(k)}, k = 1, \ldots, K$.
Using these definitions, we can define the *time-domain response matrix* (TDRM) $\boldsymbol{H}$ as follows. It is composed of the Toeplitz blocks

$$\boldsymbol{H}^{(jk)}, \quad j \in \{1, 2, \ldots, J\}, \quad k \in \{1, 2, \ldots, K\}, \quad (2)$$

so that $\boldsymbol{H}$ is of dimension $JL \times \sum_{k=1}^{K} \text{len}(\boldsymbol{s}^{(k)})$, and

$$\boldsymbol{f} = \boldsymbol{H}\boldsymbol{s}. \quad (3)$$

This model is correct only if all the loudspeaker-receiver channels are time aligned, i.e., if the first element of $\boldsymbol{h}^{(jk)}$ is non-zero for all j, k. This issue is discussed later; until then we assume time alignment to simplify the presentation.
This model can take into account non-ideal amplifier, loudspeaker and room characteristics at multiple receiver locations. The only requirements are linearity and time invariance, so that an impulse response from each input signal to each receiver can be defined.
The description of $\boldsymbol{H}$ depends on the synthesis requirement. The simplest is the specification of a set of J pressure (scalar) responses at different locations, e.g., $J/2$ binaural signals in the case of transaural synthesis. On the other hand, the model also admits specification of velocity (vector) responses by considering three orthogonal components at each location; for example, provision of these velocity components spatially coincident with a pressure response will yield ambisonic reproduction.
We are now in the position to ask: For a given TDRM, is it possible to synthesize input signals so that any desired set of output signals $\boldsymbol{d}$ will be produced at the receivers? In other words, what are the conditions on J, K, and $\boldsymbol{H}$ such that it is possible to generate a signal $\boldsymbol{s}$ that yields $\boldsymbol{d} = \boldsymbol{H}\boldsymbol{s}$?

2. THE ACHIEVABILITY OF EXACT SYNTHESIS

The above question can be restated in terms of the following

Definition. A TDRM $\boldsymbol{H}$ is *exact synthesis* if for any desired output $\boldsymbol{d}$ there exists an input $\boldsymbol{s}$ such that $\boldsymbol{d} = \boldsymbol{H}\boldsymbol{s}$.

From standard results in linear algebra, we have the following result:

Theorem 1. A TDRM $\boldsymbol{H}$ is *exact synthesis* iff $\text{rank}(\boldsymbol{H}) \geq \text{len}(\boldsymbol{d})$, or equivalently, iff $\mathcal{R}(\boldsymbol{H}) = \mathbb{R}^{JL}$.

Consequently, exact synthesis requires that $\boldsymbol{H}$ be at least as wide as it is tall.
Let $\boldsymbol{I}_L$ be the $L \times L$ identify matrix, and $\boldsymbol{0}_L$ be the $L \times L$ matrix of zeros. If the transmission path from each loudspeaker to each receiver is ideal, then the submatrices of the TDRM are

$$\boldsymbol{H}^{(jk)} = \begin{cases} \boldsymbol{I}_L, & j = k \\ \boldsymbol{0}_L, & j \neq k \end{cases}, \quad (4)$$

where we have ignored scaling factors for attenuation loss without loss of generality. Hence $\boldsymbol{H}$ is the identity matrix, and each intended signal reaches its desired receiver with zero distortion or crosstalk. Therefore, in the case of ideal (unit-pulse) responses and zero crosstalk, we can have exact synthesis when $K = J$.
We can now state a necessary condition for exact synthesis.

Theorem 2. In any practical sound synthesis system, exact synthesis can be achieved only if the number of loudspeakers exceeds the number of receivers, i.e., $K > J$.

This can be seen by noting first that in any practical system, the submatrices will be taller than they are wide. Hence Theorem 1 requires that $K > J$.
In the case of transaural sound, we desire to reproduce two signals at two locations, and the TDRM is

$$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{H}^{(11)} & \boldsymbol{H}^{(12)} \\ \boldsymbol{H}^{(21)} & \boldsymbol{H}^{(22)} \end{bmatrix}, \quad (5)$$

where each submatrix has L rows. Assume that perfect ipsilateral responses can be achieved so that $\boldsymbol{H}^{(kk)} = \boldsymbol{I}_L$. However, due to crosstalk, the contralateral TDRM's $\boldsymbol{H}^{(jk)}$, $j \neq k$, must have fewer than L columns in practice. It follows that $\boldsymbol{H}$, even though full rank, must have a range space that is a proper subspace of $\mathbb{R}^{2L}$. Hence, even under these ideal conditions, exact synthesis is not possible with two loudspeakers, and it can be seen that the situation only worsens if the ipsilateral responses are not ideal.
This observation leads to the question of sufficiency—the question that concludes Section 2. It is clear that $\boldsymbol{H}$ must have at least JL linearly independent columns for exact synthesis.
More useful would be a result that defines the minimum number of loudspeakers needed for exact synthesis, and that can be based on an easily measurable characteristic of the multi-input multi-output electro-acoustic channel. The structure of the TDRM can be exploited to obtain such a sharper result.
Since the output space is of dimension JL, we must have by Theorem 1 that

$$\sum_{k=1}^{K} \text{len}(\boldsymbol{s}^{(k)}) \geq JL. \quad (6)$$

Suppose that the impulse responses are truncated to a common length, so that the input sequences are also of

a common length $\text{len}(\boldsymbol{s}^{(k)}) = M$ for all k. Then Theorem 1 requires that $KM \geq JL$, or

$$K \geq \frac{L}{M} J. \tag{7}$$

This result can be summarized:

Theorem 3. Exact synthesis can be achieved if (i) $\boldsymbol{H}$ is full rank (ii) the number of loudspeakers K satisfies (7). Moreover, the minimum number of loudspeakers needed for exact synthesis is the smallest integer exceeding the right-hand side of (7).

Since L and M are respectively the output sequence and input sequence lengths, we have that $L/M > 1$ (due to the non-ideal impulse response of the environment). More importantly, this result implies that the greater the dispersion (i.e., as L/M grows), the more loudspeakers are required for exact synthesis. It can now be seen that, for finite-duration signals, Theorem 2 is simply a weaker version of Theorem 3, since the latter theorem takes into account the actual dispersion.
Now consider the asymptotic case when the signal duration grows without bound. Assume that the longest impulse response has duration I, and that all responses are zero-padded (if necessary) to this length. Then

$$\lim_{L\to\infty} \frac{L}{M} = \lim_{L\to\infty} \frac{L}{L+I-1} = 1.$$

Thus we have the following sufficiency result:

Corollary. As the input signal length grows arbitrarily large, exact synthesis can be achieved if $\boldsymbol{H}$ is full rank and $K \geq J$.

In other words, as the signal lengths grow, the submatrices become square in the limit.
It is of interest to determine the relationship between this matrix condition and the well-known MINT condition regarding transfer function zeros. Consider a multi-input, single-output (MISO) system. In the z-transform domain, we have

$$F(z) = \sum_{k=1}^{K} S^{(k)}(z) H^{(k)}(z).$$

If all K channels share a set of common zeros, then we have $H^{(k)}(z) = C(z)G^{(k)}(z)$ for all k, and

$$F(z) = C(z) \sum_{k=1}^{K} S^{(k)}(z) G^{(k)}(z).$$

In the time domain, this implies that

$$\boldsymbol{f} = \boldsymbol{C}\boldsymbol{G}\boldsymbol{s}.$$

When $C(z) = 1$ (i.e., no common zeros), $\boldsymbol{C}$ is the identity matrix and $\boldsymbol{G} = \boldsymbol{H}$. Otherwise, $\boldsymbol{C}$ will be taller than it is wide and a loss of rank results. Hence we can state that common zeros in the set of SISO transfer functions $\{H^{(jk)}(z)\}_{k=1}^{K}$ making up any of the J MISO transfer functions prevent exact synthesis.
In the MINT approach, it was found that any common zeros in the z-domain prevent exact inversion when two loudspeakers are employed to reproduce one signal. Three key advantages of the present approach are (i) it covers the case of arbitrary numbers of loudspeakers and receivers, (ii) amplifier/speaker characteristics and room acoustics are treated jointly, and (iii) it addresses exact reproduction of transients. Finally, it shows the effect of transfer function zeros on the reproduction of finite-duration signals.

3. TIME ALIGNMENT

Absolute and relative delays are known to affect the ability to invert room responses [6]. In conventional (i.e., not room-adaptive) surround sound, loudspeaker-listener distances are as close to uniform as possible, creating the "sweet spot" where a more lifelike sound stage is perceived.
Up to this point, we have assumed time alignment in our modeling. From the perspective of exact synthesis, the presence of relative delays implies that in each of the K submatrices indexed by a fixed j, only the submatrices associated with the minimum propagation time to receiver j are Toeplitz: all others will have a block of zeros above the Toeplitz block. This clearly implies a loss of rank and hence a degraded ability to synthesize the desired signal vector.
Note that non-zero relative delays do not imply loss of exact reproduction; it simply means that a particular loudspeaker (or loudspeakers) may be underutilized. Because of this, Theorem 3 can be extended to include the time alignment requirement. Fortunately, relative time delays can be easily removed from early channels using simple digitally-implemented room-adaptive delays; the correct delays can be determined by sounding.

4. MINIMUM-POWER SYNTHESIS

Conventional approaches to 3D sound attempt to undo the effects of the loudspeakers and room acoustics using pre-equalization of the desired signals. In contrast, we propose the computation of a new set of signals that will reproduce the desired signals (at the desired locations). In this approach, all loudspeakers will in general contribute (in varying degrees) to all received signals: the signals are synthesized so that the loudspeakers optimally (in the sense of minimum overall power) cooperate to generate the desired responses.
Assuming that Theorem 3 is satisfied, the next step is the implementation of exact synthesis. By Theorem 3, there exists at least one signal vector $\boldsymbol{s}$ that can reproduce any

desired vector $\boldsymbol{d}$ via

$$\boldsymbol{d} = \boldsymbol{H}\boldsymbol{s}. \tag{8}$$

Due to the fact that each loudspeaker contributes a large number of linearly independent vectors to $\boldsymbol{H}$, it is unlikely that $\boldsymbol{H}$ will be square. Thus (8) is an underdetermined system with an infinite number of solutions. Probably the most desirable solution is the one requiring minimum power. This corresponds to the minimum Euclidean-norm solution

$$\hat{\boldsymbol{s}} = \boldsymbol{H}^{\dagger}\boldsymbol{d}, \tag{9}$$

where $\boldsymbol{H}^{\dagger}$ is the pseudoinverse, or Moore-Penrose inverse, of $\boldsymbol{H}$ (see, e.g., [7]).

A well-known issue in loudspeaker/room correction is whether all speaker-receiver channels are minimum phase. If strong reflections are present and the receiver is off-axis from a loudspeaker, a channel may not be. While it is a good idea to locate loudspeakers to ensure minimum-phase channels, this may not always be possible. In this context it is worth noting that this synthesis approach applies to non-minimum phase channels. Work is on-going to quantify the additional cost (in number of speakers or power) in this case.

5. CONCLUSION

This paper introduced an algebraic time-domain theory of 3D sound synthesis based on finite-duration signals and a discrete-time linear model of reproduction apparatus and room response. The condition of exact synthesis was defined, and necessary and sufficient conditions for exact synthesis were found. The approach handles arbitrary numbers of loudspeakers and receivers, and the finite-duration model accurately captures signal transient behavior. Finally, synthesis (e.g., transaural or ambisonic approaches) and correction of loudspeaker-room response are treated jointly.

The focus of this paper was sound reproduction, including pressure and velocity responses, at a collection of fixed points. It is well-known that exact synthesis of a sound *field* requires an infinite number of loudspeakers. Therefore the fidelity of sound field approximation with a given number of loudspeakers is of great interest, and is currently under study.

REFERENCES

[1] P. A. Nelson, H. Hamada, and S. J. Elliott, "Adaptive inverse filters for stereophonic sound reproduction," *IEEE Transactions on Signal Processing*, vol. 40, pp. 1621–1632, July 1992. 1

[2] J. Garas, *Adaptive 3D Sound Systems*. Kluwer Academic, 2000. 1, 1

[3] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, pp. 145–152, Feb. 1988. 1, 1

[4] U. Horbach and A. Karamustafaoglu, "Numerical simulation of wave fileds created by loudspeaker arrays," in *107th AES Convention*, 1996. AES Preprint 5021. 1

[5] C. Kyriakakis, P. Tsakalides, and T. Holman, "Surrounded by sound: acquisition and rendering methods for immersive audio," *IEEE Signal Processing Magazine*, pp. 55–66, Jan. 1999. 1

[6] P. A. Nelson, F. Orduña-Bustamante, and H. Hamada, "Inverse filter design and equalization zones for multichannel sound reproduction," *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 185–192, May 1995. 3

[7] Å. Björck, *Numerical Methods for Least Squares Problems*. SIAM Press, 1996. 4