Recent Developments in Virtual Acoustic Imaging Systems

Institute of Sound and Vibration Research, University of Southampton, England SO17 1BJ
pan@isvr.soton.ac.uk

Abstract
This paper will present a brief review of progress in the development of loudspeaker-based systems for producing the illusion in a listener of the existence of a virtual source of sound at a pre-determined location relative to the listener. A fundamental consideration in the design of such systems is the conditioning of the matrix of transfer functions relating loudspeaker inputs to the signals reproduced at the listener’s ears. This plays a major part in dictating the bandwidth of effective operation of such systems and a brief description will be given of loudspeaker/listener geometries for ensuring successful operation over a wide band of frequencies. The efficient implementation of inverse filters that ensure the reproduction of the desired signals at the ears of the listener is an important issue and recent developments in filter design procedures will be outlined. The sensitivity of performance to listener position will also be discussed and suggestions given for the rate of adaptation required of the audio signal processing system if the illusion is to be preserved under conditions of a moving listener. Finally, a review of strategies will be presented for tracking the position of the listener by processing optical images of the space in which reproduction is sought.

1. Introduction

Loudspeaker-based virtual acoustic imaging systems rely on the design of a matrix of inverse filters that is the inverse of the matrix of transfer functions relating the loudspeaker input signals to the signals generated at the ears of the listener. This inverse filter matrix is then used to operate on the signals that are desired at the listener’s ears in order to generate the loudspeaker input signals. The design of these filters is a major factor in ensuring the success of such systems and it is important to understand the basic geometrical considerations that determine their characteristics. The efficiency of implementation of the filters is also a significant factor as the implementation of such systems is now being considered for a wide range of communication systems. A second major consideration is the robustness of such systems to movements of the listener’s head. It turns out that loudspeaker/listener geometry also plays a major role in influencing this factor. However, once the geometry has been optimized, it is still necessary to appreciate that the spatial extent of the reproduction of the desired signals is limited. Therefore it is necessary to examine strategies for adapting the inverse filter matrix to compensate for movements of the head of the listener relative to the loudspeakers used for signal transmission. The third important consideration then becomes the development of suitable techniques for tracking the position of the listener’s head relative to the loudspeakers.

2. The influence of geometry on inverse filter design

The basic signal processing block diagram is shown in Fig. 1. $C(k)$ is the matrix of transfer functions in the discrete frequency domain that relates the vector of loudspeaker input signals $v(k)$ to the vector of signals $w(k)$ produced at the listener’s ears. It is therefore assumed that $w(k) = C(k)v(k)$. The matrix $H_x(k)$ of cross-talk cancellation filters operates on the vector of binaural signals $u(k)$ in order to deduce the vector $v(k) = H_x(k)u(k)$. The signals $u(k)$ are typically synthesised by convolving the signal associated with an intended virtual source with a pair of filters representing the transfer functions from the position of the virtual source to the ears of the listener.

![Image of signal processing for virtual acoustic imaging](image)

The basic requirement of the cross-talk cancellation matrix is to ensure that the reproduced signals are simply a delayed version of the binaural signals. Thus we wish to ensure that $w(k)$ is made equal to the desired signals $d(k)$ at the listener’s ears that are in turn given by $u(k)e^{-j\omega\Delta}$ where $\Delta$ represents the delay. It therefore follows that...
\[ C(k)H_x(k) \approx e^{-j\Delta}I \] (1)

where \( I \) denotes the identity matrix. Therefore at a given discrete frequency \( k \) the cross-talk cancellation matrix is given by \( H_x(k) \approx C^{-1}(k)e^{-j\omega} \). It is thus the condition number of the matrix \( C(k) \) that is of central importance in determining the nature of the matrix \( H_x(k) \) of cross-talk cancellation filters. In particular it should be noted that if \( C(k) \) is badly conditioned, then the errors in reproducing the sound at the listener’s ears will be proportional to the condition number (the ratio of the maximum to the minimum singular values) of \( C(k) \). It has been shown [1,2,3] that, at frequencies where the condition number of the matrix \( C(k) \) is large, there is less robustness to errors in the design of the cross-talk cancellation filters and that, in particular, the sensitivity of the performance of the system in response to listener head movement is proportionally reduced.

It is therefore particularly important to note the influence of the geometrical arrangement of loudspeakers on the condition number of \( C(k) \). Fig. 2 shows the dependence of the condition number of \( C \) of the (two by two) matrix \( C(k) \) on the angle subtended at the ears of a listener by a pair of point monopole sources. Fig. 2 shows this dependence when there is no assumed effect of listener head diffraction on the transfer functions between the outputs (volume acceleration) of the two point sources and the pressures produced at two points 18cm apart that are used to model the positions of the listener’s ears. (Full details of the geometry are given in [3]). Fig. 3 shows the effect of including a model of the listener head diffraction that is based on the classical theory of acoustic scattering by a rigid sphere [4]. (Again, full details are given in [3]). Finally, Takeuchi [5] has shown that the condition number of the transfer function matrix when based on measurements of the Head Related Transfer Function (HRTF) of the KEMAR dummy head shows almost identical trends to those shown in Fig.3.

It is evident from these figures that the free field model of the listener HRTF dominates the basic behaviour of the condition number as a function of loudspeaker span, although the inclusion of head diffraction does reduce the extremes of variation of the condition number. Of particular note is the dependence on loudspeaker span of the first maximum (shown dark) that corresponds to a path-length difference of one half of the acoustic wavelength occurring between the distance from

![Figure 2](image2.png)

**Figure 2.** The dependence of the condition number of the matrix \( C(k) \) on loudspeaker span and frequency for a free-field model [3].

![Figure 3](image3.png)

**Figure 3.** The dependence of the condition number of the matrix \( C(k) \) on loudspeaker span and frequency for a rigid sphere model of the listener’s head [3].

one of the sources to the two ears of the listener (in the assumed symmetrical arrangement of loudspeakers relative to the listener). It has been shown [1,2,3] that this maximum for a given loudspeaker span defines the “ringing frequency” first identified by Kirkeby et al [6] as exhibiting an undesirable response in the time domain. Thus for example, at a loudspeaker span of sixty degrees, this frequency lies in the region of about 2kHz. However, if the loudspeaker span is reduced to about ten degrees, then this ringing frequency is pushed up to about 11kHz. This fundamental property of the two-source/two-field point transfer function matrix led to the development of the so-called “Stereo Dipole” (SD) virtual sound imaging system [6,7,8] which shows considerable robustness to listener head movement [9] albeit over a limited but extremely useful operational
bandwidth. The logical extension of this type of system design is to ensure that the inversion problem is always well-conditioned by using a number of pairs of loudspeakers each of which transmits a range of frequencies over which the condition number of the HRTF matrix is always guaranteed to have a relatively low value. This philosophy motivates the development of the “Optimal Source Distribution” (OSD) for virtual acoustic imaging where in principle it is desirable to have a continuous distribution of source strength as a function of angular position relative to the listener [5,10]. In practice however, highly effective systems can be built [11,12] by using, for example, three pairs of loudspeakers to transmit signals in three frequency ranges. Thus a pair of loudspeakers at a subtended angle of about six degrees may be used for frequencies above about 3.5kHz, a pair at an angle of about thirty degrees for frequencies between 450Hz and 3.5kHz and a pair at one hundred and eighty degrees (placed on either side of the listener) for frequencies below 450Hz. Subjective experiments have shown how highly effective such systems can be [11].

3. Inverse filter design

The accurate delivery of the requisite time histories to the ears of the listener is crucially dependent on the design of the matrix of cross-talk cancellation filters. A highly effective technique is that described by Kirkeby et al [12] that relies on the use of regularization to limit the duration of the inverse filters in the time domain. The technique is based on the regularized least squares solution for the cross-talk cancellation matrix that is given by

\[ H_x(z) = \left[ \frac{1}{\text{det}(C(z))} \text{adj}(C(z)) \right] \]

(3)

where the modeling delay has been omitted. However, an appropriate decomposition of the transfer function \( \text{det}(C(z)) \) can be found by means of solving the following equation in which the polynomials \( C_{\text{min}}(z) \) and \( C_{\text{max}}(z) \) are of strictly minimum and strictly maximum phase respectively

\[ \text{det}(C(z)) = C_{\text{min}}(z)C_{\text{max}}(z) \]

(4)

Having computed the polynomials \( C_{\text{min}}(z) \) and \( C_{\text{max}}(z) \), the inverse filter matrix \( H_x(z) \) can be decomposed as

\[ H_x(z) = \frac{B_1(z)}{C_{\text{min}}(z)} \text{adj}(C(z)) + \frac{B_2(z)}{C_{\text{max}}(z)} \text{adj}(C(z)) \]

(5)

where the rational transfer functions comprising \( C_{\text{min}}(z) \) in the denominator are directly realizable and the rational transfer functions comprising \( C_{\text{max}}(z) \) in the denominator are realizable in backward time (see reference [14]). In the typical case where the time-domain models of the measured HRTFs \( C(z) \) comprise a few tens or hundreds of samples, the numerical solution of equations (4) and (5) for the determination of the unknown polynomials \( C_{\text{min}}(z) \), \( C_{\text{max}}(z) \), \( B_1(z) \) and \( B_2(z) \) is not trivial. A technique utilising the cepstrum for the solution of equation (4) and the related Diophantine equation for the solution of the lower part of equation (5) is proposed in [14] and it is shown to achieve virtually perfect modeling of the inverse filter matrix \( H_x(z) \). In the form outlined above the inversion method is restricted to off-line application as the whole input signal has to be known in advance for the implementation of the backward-time filtering. However it is possible to envisage a block-processing algorithm that allows the on-line implementation of the method at the expense of a modest increase in the input-output latency.

4. Sensitivity to listener head movement

Detailed numerical and subjective studies have been successful in identifying the spatial extent of the region in which listener head movements can be tolerated without loss of localization of the image [15]. Since this region is by necessity restricted in spatial extent, then it is natural to assess the degree to which listener head movement can be compensated for by updating the inverse filters. Rose [16] has undertaken a comprehensive study of the rate at which the inverse filters have to be updated in order to preserve a stable image perceived by the listener. It emerges from subjective experiments that the rate at which the filters...
have to be updated will depend upon both the movement of the listener relative to the loudspeakers and the position of the virtual source relative to the loudspeakers. For example, it transpires that the greatest degree of image stability for relative motion in the horizontal plane occurs when the head of the listener moves relative to the loudspeakers in a direction that is towards the direction of the virtual source. A typical value of the increment of head displacement that requires filter up-date is about 3cm, although this value is evidently dependent on a number of factors including the direction of the relative motion and, of course, the type and bandwidth of the reproduced signals. Some elementary image processing algorithms have been implemented [16] that enable the tracking of the head of the listener and the subsequent automatic update of the inverse filters. These have been based on simple concepts of template matching together with straightforward optimization algorithms for ensuring the head position is found that minimizes the error between the image and the template evaluated over a number of pixels. Subjective experiments have demonstrated the practical feasibility of the technique in producing a stable image for lateral listener head movements. The successful integration of both image and audio processing remains a significant challenge in this field and a recent study [17] has clearly illustrated the potential for developing head tracking algorithms that are both potentially more effective and efficient to implement.

5. Conclusions

A brief review has been presented of the development of loudspeaker based virtual acoustic imaging systems. A good understanding has been reached of the main factors affecting system performance and more efficient filter design techniques are being developed. Considerable scope remains for the development of systems that are adaptive to listener position.

References