Review of the crosstalk cancellation filter technique
Bruno Masiero, Janina Fels, Michael Vorländer
Institut für Technische Akustik, RWTH Aachen University, Germany. Email: bma@akustik.rwth-aachen.de

Abstract
It is long known that to correctly reproduce a binaural signal through a pair of loudspeakers, this signal has to be pre-filtered to compensate for the crosstalk effect that will otherwise ruin the spatial clues contained on the binaural signal. In this work we make a brief review of the design technique for crosstalk cancellation (CTC) filters specially focused on its application in virtual reality installations and discuss still open dilemmas. CTC filters are designed based on the transfer paths between loudspeakers and listener ears and usually delivers channel separations of over 20 dB in a small region around the listeners head, the so-called sweet spot. If the listener moves away from the sweet spot, channel separation will deteriorate and the spatial cues from the binaural signal will be lost. A great deal of researchers tried to answer the question on how to increase the size of the sweep spot through a more skilled system setup, specially important for virtual reality applications where the users should be allowed to freely move and rotate their heads. For such applications dynamic CTC systems can be implemented with the aid of a head-tracking device to determine distance and orientation of the listener’s head to the loudspeakers, allowing a constant update of the cancellation filters.

Introduction
Reproduction of binaural signals through loudspeakers was first proposed by Bauer in 1961 [1] and the first patent was obtained by Atal et. al. in 1966 [2]. Only about 20 years later came the first commercial application by the company Cooper Bauck Transaural. The crosstalk cancellation technique also had a big appeal in the academic world with research groups all over the world dedicated to that matter.

The fundamental idea on how to perform crosstalk cancellation has not changed since its original proposition. However the practical implementation of CTC filter design and adequate transducers did improve considerably. Bauck and Cooper were the first to notice that closely placed loudspeaker provide a larger sweet spot and to verify that ideal loudspeaker distribution was frequency dependent [3]. Nelson and his group went further and analyzed the condition number of the plant transfer matrix to explain this behavior and proposed the optimal source distribution (OSD) [4]. This system provides a broader sweet spot with a frequency dependent positioning of the loudspeakers and simplified CTC filtering. Parodi and Rubak recently verified that the use of elevated loudspeakers provide a wider sweet spot [5].

An interesting by-product of the OSD research was to show that low frequencies (≤400Hz) will always require a very high energy compensation with a narrow sweet spot. Fortunately localization cues do not occur in this frequency range, so no crosstalk cancellation is actually need for low frequencies.

The OSD approach works well for static situation, e.g., for a listener seated in front of a television. The other way around in virtual reality applications the user must be allowed to freely move inside the virtual environment and being so, a system like OSD would not be adequate, as it would require the use of a very large number of high frequency transducers distributed all around the reproduction space. For such applications a reduced number of full-range loudspeakers are still used and CTC filter design receives increased important [6]. Several methods have been proposed on how to calculate CTC filters and will be discussed later in this paper. Usually the research of new filter designs only analyses the resulting channel separation delivered by the filter in question, ignoring the very important aspect of sweet spot size, probably because the sweet spot size mainly depends on the loudspeaker and listener position, i.e., on the plant transfer matrix. Different choices of filter design will not give a sweet spot size larger than the one obtained with the optimal filter design.

Crosstalk and its Cancellation
Binaural signals can be reproduced through loudspeakers. Nevertheless, since the signal reproduced by a loudspeaker will be heard by the listener’s both ears (see crosstalk effect on Fig. 1), a set of filters has to be used to achieve the required channel separation between ears, usually called Crosstalk Cancellation (CTC) filters.

From Fig. 1 it is easy to see that this system can be described in frequency domain as follows ($z = e^{-j\omega}$):

\begin{align}
  e_L(z) &= H_{LL}(z)s_L(z) + H_{RL}(z)s_R(z), \\
  e_R(z) &= H_{LR}(z)s_L(z) + H_{RR}(z)s_R(z),
\end{align}

where $e_L(z)$ and $e_R(z)$ are respectively the signals arriving at left and right ears and $s_L(z)$ and $s_R(z)$ are the left and right channels of a binaural signal. Assuming that

\begin{align*}
  e(z) &= \begin{bmatrix} e_L(z) & e_R(z) \end{bmatrix}^T, \\
  s(z) &= \begin{bmatrix} s_L(z) & s_R(z) \end{bmatrix}^T
\end{align*}

and

\begin{align}
  \mathbf{H}(z) &= \begin{bmatrix} H_{LL}(z) & H_{RL}(z) \\
  H_{LR}(z) & H_{RR}(z) \end{bmatrix},
\end{align}

the optimal CTC filter can be found by solving the following minimization problem:

\begin{align}
  \min_{\mathbf{H}(z)} \| e(z) \|_2 \\
  \text{subject to} \quad \| \mathbf{H}(z) \|_{2,1} \leq 1,
\end{align}

where \( \| \cdot \|_{2,1} \) denotes the nuclear norm of a matrix

\begin{align}
  \| \mathbf{H}(z) \|_{2,1} &= \sum_{i=1}^m \| H_{ii}(z) \|_2,
\end{align}

for some matrix partitioning of \( \mathbf{H}(z) \). The nuclear norm of a matrix is defined as the sum of the singular values of the matrix.
Figure 1: Diagram of binaural reproduction through loudspeakers. The CTC filters are displayed in the upper part while the direct (solid line) and crosstalk acoustic paths (dashed line) are displayed in the lower part of the diagram.

Eq. 1 can be described in matrix form as

$$e(z) = H(z)s(z). \quad (2)$$

The crosstalk path can be canceled out (or at least be considerably attenuated) with an adequate filter structure, as depicted in the upper part of Fig. 1. This filter structure should always be placed between the input signal and the loudspeakers and can be represented as matrix $C$, resulting in the complete transmission path

$$e(z) = H(z)C(z)s(z). \quad (3)$$

Correct binaural reproduction is achieved if $e(z) = s(z) \cdot e^{j\omega \Delta}$, where $\Delta$ is a delay to account for the acoustic lag between loudspeakers and listener position. Thus, the solution to the crosstalk cancellation matrix is given by

$$C(z) = H^{-1}(z)e^{j\omega \Delta}. \quad (4)$$

The reproduction of binaural signals does not have to be made necessarily through two loudspeakers [7]. If more than two loudspeakers are used, matrix $H$ will no longer be a square matrix and therefore will not be invertible. In this case, the pseudo-inverse can be used, yielding

$$C(z) = \left( H^H(z)H(z) \right)^{-1}H^H(z)e^{j\omega \Delta} \quad (5)$$

where $(\cdot)^H$ represents the conjugate transpose operation.

This solution in frequency domain will deliver the best possible channel separation for a given listener-loudspeaker setup. It has, however, some drawbacks. Channel separation is achieved by means of constructive and destructive wave interference. For some frequencies, a very high amplitude is required only to be later canceled at the listener ears. The overall gain of the CTC filter has to be reduced to avoid clipping at these frequencies and with that the dynamic range of the reproduced binaural signal can be drastically reduced. In spatial-time domain, these frequencies will present a long ringing behavior, what could also be understood as a very narrow sweet spot. These effects can be avoided or at least ameliorated by skillfully positioning the loudspeakers, as discussed in the next chapter.

Furthermore, when calculated using Discrete Fourier Transform, the obtained CTC filters will suffer from cyclic aliasing and non-causality. Possible workarounds are to extend the frequency resolution of the original transfer matrix $H$ and to apply time windowing, to apply a regularization constrain or to calculate the inverse filters in the time domain. These and other CTC filter design techniques will be discussed later in this paper.

Figure 2: Tradeoff aspects of a crosstalk cancellation system.

Source Position

Sweet spot size, dynamic range and channel separation depend not only on frequency but also on the position of the sound sources. Thus, sources optimally positioned can deliver a better listening experience with fewer spectral coloration and more freedom of movement for the listener.

If the sound source pair is placed very near to each other, e.g. spanning an angle of 12° with the center of the head, then a relatively broad sweet spot will be achieved, but only for higher frequencies, as can be seen in top left plot of Fig. 3. This setup was called "Stereo Dipole" by Kirkeby et. al. on their earlier publications on crosstalk cancellation systems [8]. On the other hand, if the sources are further apart, e.g. spanning an angle of 60° with the center of the head, then improved channel separation will be verified in the lower frequencies while the sweet spot size for higher frequencies will diminish. Parodi and Rubak also noted that elevated sources provide a broader sweet spot as the transfer path variance decreases for elevated sources [5].

To optimize the sound source position for a static CTC system, Takeuchi and Nelson analyzed the condition number of the plant matrix [4]. The condition number of a matrix delivers an estimate of how much an error
argued that for an overall system, with only two sound sources, it is then better to have an intermediary source span, e.g. 60° apart [9].

So far we have only considered a pair of sound sources, that considerably restrict the area where crosstalk cancellation is viable. To overcome this limitation, system with three (or more) sound sources should be employed. The sound sources can be driven either pair wise, switching between the pairs that works best for each position, or driven all simultaneously. Lenzl describes a system with four loudspeakers displaced in the vertices of a square, which choses, according to the orientation of the listener, between four pairs of neighboring loudspeakers (90° configuration) or two pair of opposite loudspeakers (180° configuration). This method has the disadvantage that different configuration of loudspeaker pairs present different channel separation (and eventually coloration) to the reproduced binaural signal and cross-fading between loudspeaker pairs might generate audible artifacts [10].

Using more than two loudspeakers simultaneously was already suggested early in 1990’s [7]. Bauck and Cooper analyzed many loudspeaker configuration for improved binaural reproduction, most of them, though, not applicable for virtual reality applications. Recently, Parodi and Rubak compared a two and four loudspeaker CTC setup regarding sweet spot size and verified that the four channel setup will result in a reduction in the sweet spot size [5]. On the other hand, the dynamic range of the resulting filters is improved without any considerable variation in channel separation and such a system can deliver a smooth variation as the user changes orientation. As this setup have twice more loudspeakers to be modeled than before, chances of plant modeling error increases, so the importance of a precise tracking increases as well.

Filter Design
Calculation of CTC filters either in time or frequency domain is equivalent. Equation (4) can be rewritten in time domain as

\[ C(t) = \hat{H}^{-1}(t)I(\Delta). \]  

where \( \hat{H} \) is the concatenation of the convolution matrices of the elements of \( H \) and \( I(\Delta) \) is a diagonal matrix with two delayed impulses in the diagonal [11, 12].

As matrix inversion has computational complexity of approximately \( O(n^3) \), inversion in frequency domain has the advantage that its computational requirements are considerably smaller than computation in time domain. Nevertheless, calculation in frequency domain suffers from some drawbacks.

Kirkeby et. al. showed that an ideal CTC filter can be written as an infinite summation of delayed impulses [8], meaning that CTC filters are infinitely long. When the discrete fourier transform (DFT) is used to calculate the filter in frequency domain, care must be taken with cyclic aliasing.

The observations of Takeuchi and Nelson were applied to a static system. They could even show that several listener could be seated alongside and also enjoy the same binaural signal. But in a virtual reality (VR) environment the listener must be able to move freely inside the reproduction space. Ensuring that for every position in the VR space pairs of adequately spaced sources with direct sight to the listener exists would require a huge amount of loudspeakers distributed around the space, increasing cost and complexity of the system. So, in this case, a compromise between number of sources and crosstalk cancellation quality must be achieved. Bai and Lee did analyze different source configurations and

---

1Note that the condition number of an invertible matrix is identical to the condition number of its inverse matrix.
Regularization

As mentioned earlier for some given frequencies the CTC system might require the transaural signal to be played with very high energy. This causes not only loss of dynamic range, but also introduce the so-called “ringing frequencies”. Regularization, proposed by Kirkeby et. al [13], add an extra requirement on the maximum energy of the transaural signal, leading to the following solution to the inversion problem

\[ C(z) = \left( H^H(z)H(z) + \beta B(z)H(z) \right)^{-1} H^H(z)e^{j\omega \Delta}, \]

where \( B(z) \) is the shape-factor of the regularization that define which frequencies are to be regularized and \( \beta \) is the gain factor that defines how much those frequencies are to be regularized, actually acting as a trade-off factor between channel separation and dynamic loss. As a byproduct, regularization reduces the size of the CTC filters while increasing the non-causal behavior of the filters\(^2\).

Causality

The computationally expensive calculation of the CTC filters in time domain has the advantage of producing causal filters. Causality constrain can be implemented in frequency domain (thus, with reduced computational costs) by using the spectral factorization method [14, 15]

\[ C(z) = \frac{1}{\det(S(z))} \left[ \frac{\text{adj}(S(z)) H^H(z)e^{j\omega \Delta}}{\det(S(z))} \right]_+, \]

where \( S(z) = \left( H^H(z)H(z) \right) \), \( \text{adj}(\cdot) \) is the adjugate matrix, \( \det(\cdot) \) the determinant, \( (\cdot)^+ \) and \( (\cdot)^- \) denote respectively the minimum causal stable and minimum anti-causal stable parts and \([\cdot]_+\) denotes the causal part. As \( S(z) \) is a hermitian matrix, \( \det(S(z)) \) is real. The spectral factorization can be then easily implemented in cepstral domain, allocating the first half of the cepstrum for the causal stable part and the second half for the anti-causal stable part.

Note that for an improved channel separation, the value of \( \Delta \) should be a bit larger than the acoustic delay between loudspeakers and listener, allowing for some pre-ringing.

Other methods

Equations (4) and (6) can be interpreted as the solution to a \( \ell_2 \) minimization problem. The crosstalk cancellation problem, can, however, be casted in other formats, e.g. as a mini-max optimization that is equivalent to a \( \ell_{\infty} \) minimization problem [16]. This method searches for a solution which maximizes the crosstalk cancellation evenly in the whole spectrum, providing clear improvement for low frequencies. The major drawback of mini-max optimization is its very elevated computation cost, being currently still inappropriate for real time applications.

An approach prone for real time applications is the use of adaptive filtering [17]. This technique has the advantage that no plant model is needed for the filter inversion, constantly delivering good channel separation. On the other hand this technique suffers from the usual limitations regarding plant estimation common in active noise control applications and requires the listeners to place microphones in the entrance of each ear, again undesirable for virtual reality applications that try to minimize the use of wearable hardware.

Plant Modeling

Ideally, the transfer path from each loudspeaker to each of the listener ears would have to be know to achieve a perfect channel separation. For an individual virtual reality application this could be achieved by having a dataset of individual HRTFs plus a dataset of the loudspeakers’ directivity, but for commercial applications, such effort would not be practicable and a generic plant model would be needed.

Akeroyd et. al. evaluated the channel separation for unmatched plant model [18]. Their numerical analysis showed that unmatched filters will result in a considerably smaller channel separation. Nevertheless, no perceptual tests were made to verify if the provided channel separation was enough for a realistic binaural impression. A similar test was conducted by Parodi and Rubak indicating that in order to avoid lateralization the channel separation should be in the range between 15 dB and 20 dB depending on the reproduced binaural stimuli [12].

On the other hand, Cooper and Bauck report that the use of simplified HRTF, with a frequency dependent smoothing applied, or even a spherical-model head produced satisfying results for a wide range of listeners [3]. They conclude, nevertheless, saying that further investigation in this topic is needed.

Multiuser

More than one user crosstalk cancellation system, already proven by Bauck and Cooper to be mathematically feasible, is possible under severe practical limitations, since achievable channel separation considerably decreases and the signal levels at the receivers have to be reduced due to the unstable crosstalk cancellation filter that result from an ill-conditioned transfer matrix [19]. Simulation to optimize source arrangement through the minimization of transfer matrix’ condition number shows that ideal position varies significantly for neighboring frequencies, frustrating a practical implementation of this setup, since very narrow frequency dividing filters would be needed. A band optimized system reduces the overall conditioning of the transfer matrix and reduces the number of sources needed and is possibly the optimal compromise to guarantee stability to this system. Nonetheless, the number of required loudspeakers would still be considerably high and practical implementation still questionable.

\(^2\)Note that regularization can also be applied to calculation in time domain [11].
Summary
This paper discussed the design difficulties for a crosstalk cancellation for virtual reality applications, mainly discussing how sound sources should be optimally placed and the available techniques for CTC filter design, both aspects already well described within a vast literature. There are, however, some aspects that are still not completely understood, as for example the level of detail required for describing the transmission path or how to expand the system to multiple listener. In this paper the influence of the reproduction room in the system, a very important aspect of the crosstalk cancellation in virtual reality installations, was intentionally left untouched, as this topic has still a very incipient literature [20].

References