Audio Engineering Society

Convention Paper 10635

Presented at the 154th Convention
2023 May 13–15, Espoo, Helsinki, Finland

This paper was peer-reviewed as a complete manuscript for presentation at this convention. This paper is available in the AES E-Library (http://www.aes.org/e-lib), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Dynamic Adaptation in Geometrical Acoustic CTC

Alberto Vancheri[1], Tiziano Leidi[1], Thierry Heeb[1], Loris Grossi[1], and Noah Spagnoli[1]

[1] *University of Applied Sciences and Arts of Southern Switzerland*

Correspondence should be addressed to Tiziano Leidi (tiziano.leidi@supsi.ch)

ABSTRACT

By controlling sound delivery at the listener's ears, Crosstalk Cancellation (CTC) allows for 3D audio reproduction from a limited number of speakers by simulation of binaural cues. Originally pioneered in the sixties, CTC is currently an active field of research due to increased interest in Augmented Reality (AR) and Virtual Reality (VR) applications and widespread availability of immersive audio content. In this paper, we present an extension of our multiband, geometrical acoustics inspired, CTC solution able to support a freely moving user. Unlike the static case, support of a moving user requires the ability to update CTC filters in real-time. Being rooted in the time-domain, our solution offers natural support for continuous adaptation to changing conditions. The effectiveness of the proposed solution is verified by laboratory experiments.

1 Introduction

Humans are capable of perceiving 3D audio using only two ears. By analyzing the differences between the ear signals, our brain is capable of reconstructing the full 3D audio scene. Crosstalk Cancellation (CTC) is a signal processing technique allowing for control of sound delivery at the listener's ears from a set of speakers. In particular, it can be configured such that contributions from a given source speaker are received almost unaltered at one of the listener's ears and cancelled out at the other. If the source signal contains binaural encoded content, the binaural cues will be rendered properly at the listener's left and right ears, resulting in an immersive 3D audio listening experience.

CTC is an active field of research since many years and multiple different approaches have been studied. A key milestone in the development of CTC solutions has been the Recursive Ambiophonic Crosstalk Elimination (RACE) presented by [1] in 2007. This is a time-domain processing approach that has inspired many modern solutions, including our own approach to CTC. In our initial work [2], we introduced a variation of the RACE algorithm based on geometrical modelling of the acoustical propagation paths and inspired by the ray tracing technique used in computer graphics rendering. A key advantage of the proposed approach is its high robustness obtained by regularization of the order of the cancellation signals cues. In our subsequent work [3], we extended our approach to support frequency-dependent CTC through the introduction of multibands CTC processing. The resulting CTC solution has been shown to have high cancellation effectiveness and low coloration over a wide frequency band for a user in a central listening position.

Proper rendering of 3D audio is a key element for upcoming Augmented Reality (AR) and Virtual Reality (VR) applications. Unlike traditional entertainment such as watching a movie where the user is mainly in a static position, AR and VR applications allow for free user movements. In such a context, a CTC system

needs to be able to continuously adapt to track the actual dynamics of the user's position. In particular, this requires that the configuration of the CTC system can be updated in (near) real-time without audible artifacts.

Section 2 provides a short overview of CTC principles and reviews some existing solutions. Section 3 reviews the background of our approach to CTC, whilst sections 4 to 6 introduce the theoretical and practical sides of the newly proposed solution supporting dynamic adaptation for a freely moving user. In section 7, we present laboratory experiments and measurements showing the performance of the newly proposed dynamic approach to CTC. Results show that the new CTC system is clearly able to support dynamic adaptation for a freely moving user with high cancellation effectiveness and no audible artifacts.

2 CTC overview

3D audio reproduction over loudspeakers using CTC has been a busy field of research for a long time. Recently, the widespread availability of immersive audio content and upcoming AR and VR applications have further boosted interest in the domain. The principles of CTC were pioneered by Bauer [4] in the early sixties and the first patent in the field was filed by Atal et al. in 1966 [5]. The first commercial applications were brought to market about 20 years later by Cooper Bauck Transaural. Good backgrounders on CTC technologies can be found in the works of Masiero et al. [6] and Gardner [7].

Mathematically, a two loudspeakers CTC system can be described by the following z-domain matrix equation:

$$E(z) = H(z)S(z) \tag{1}$$

where $E(z) = (E_1(z), E_2(z))^T$ represents the left ear and right ear signals, $S(z) = (S_1(z), S_2(z))^T$ the left speaker and right speaker signals, and $H(z)$ is a 2x2 matrix whose coefficients h_{ij} are the transfer functions from speaker i to ear j.

By inserting a pre-processing filter (represented by a matrix $CTC(z)$) at the input of the system, the resulting transfer function becomes:

$$E(z) = H(z)CTC(z)S(z) \tag{2}$$

Ideal crosstalk cancellation is achieved if:

$$E(z) = kz^{-\delta}S(z) \tag{3}$$

where k is a gain factor and $z^{-\delta}$ is pure delay. This results in the ear signals being delayed homothetic copies of the speaker signals. In other words, $CTC(z)$ is an approximation of the inverse of the forward path matrix $H(z)$, up to the gain factor k, and combined with a delay for causality reasons. Computation of $CTC(z)$ is thus similar to a matrix inversion problem. When $H(z)$ is non-minimum phase, the inversion problem is ill-defined and only approximate inversion can be used. The same formalism can obviously be extended to systems with more than two loudspeakers.

In his work [8], Parodi shows that correct sound source localization requires cancellation levels of 20 dB and more. At the same time, very high levels of boost (reaching 30 dB and more) may be required at frequencies where the matrix inversion is problematic according to Choueiri [9]. At such frequencies, matrix inversion approximation errors can result in high deviations between expected and computed values, resulting in reduced crosstalk cancellation effectiveness.

In addition, $CTC(z)$ needs to be adapted to the actual listener position as shown by Lee and Lee [10] to take the effective propagation matrix $H(z)$ from speakers to ears into account. A modified RACE algorithm supporting non-central user positions has been presented by Cecchi et al. [11]. Whilst our own approach to CTC ([2], [3]) is also loosely based on RACE, it differs from Cecchi's work by using truncated impulse responses instead of full recursion, which provides benefits in terms of system stability.

For typical AR or VR applications, the user is further allowed to freely move in a given spatial region. In such a case, $H(z)$ becomes time variant and, consequently, the crosstalk cancellation filter $CTC(z)$ has to be updated in real-time to track the user's instantaneous position and orientation. A significant part of recent research in the field of crosstalk cancellation has been focused on the real-time computation and smooth regularization of the inverse approximation $CTC(z)$ of the time-varying system forward path $H(z)$ resulting from user tracking.

Many different approaches to user-position adaptive CTC have been presented. For instance, Bruschi et al. present a method based on a combination of RACE and beam-forming supporting dynamic user tracking ([12]). A similar concept based on super directive beam-forming has been described by Ma et al. ([13]). Another approach is proposed by Choueiri where optimal multibands CTC filters are derived in real-time

to track user movements ([9]). Finally, Qiao discloses an interesting solution for user-tracking enabled CTC based on an extension of the analytical spectral division method ([14]), a method initially designed for the creation of acoustically bright or dark sound zones.

For a moving user, the effects of transitions between different cancellation filters $CTC(z)$ corresponding to successive user positions must also be considered. In particular, changes between successive filters should be free of audible artifacts. This calls for smooth, continuous changes between cancellation filters. Our approach, being rooted in the time-domain, provides natural support for time-varying cancellation filters $CTC(z)$, as they can be updated and smoothly interpolated on a sample by sample basis.

3 Background of our time-domain approach

Our approach to crosstalk cancellation is based on the notion of cancellation complex, a system composed of three sources S_0, S_1 and S_2 and two receivers E_1 and E_2 (the ears of the user), as depicted in figure 1. The ear E_1 is called the target ear. In this paper, $x_i(t)$ will indicate a signal emitted from source S_i and $y_j(t)$ a signal received at the ear E_j. The speaker S_0 outputs a signal $x_0(t)$. The sources S_1 and S_2 work in such a way that the signal $y_1(t)$ received at the target ear is the propagation of $x_0(t)$ from S_0 to E_1, whereas the signal received at the non target ear E_2 is $y_2(t) = 0$. Sources S_1 and S_2 provide the recursive cancellation signals needed to attain this goal. In case of a two loudspeaker system, S_0 and S_1 will be the same speaker. It is not possible to share the same loudspeaker for S_0 and S_2.

There are six acoustic paths relevant for the computation of the impulses responses of the system, each labelled with a couple of indexes (i, j), where i refers to the source and takes values in $\{0,1,2\}$ and j refers to the ears and takes values in $\{1,2\}$. For instance the path $(0,2)$ refers to the acoustic path between source S_0 and non target ear E_2.

To the acoustic path (i, j), we associate a propagation time τ_{ij} and a gain g_{ij} given by the following formulas:

$$\tau_{ij} = \frac{L_i}{c} + \tau_{ij}^{(H)} \tag{4}$$

$$g_{ij} = \frac{1}{L_i} g_{ij}^{(H)} \tag{5}$$

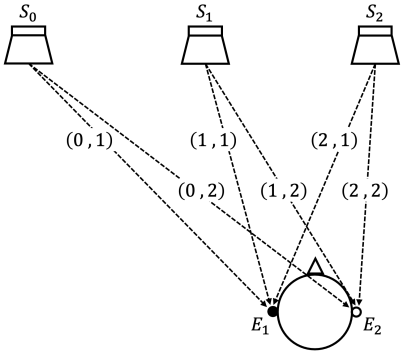


Fig. 1: Structure of a cancellation complex. The acoustic paths between the speaker S_i and the ear E_j is labelled with (i, j). The original sound is emitted from the speaker S_0 and is directed to the target ear E_1 along the path $(0,1)$. The speakers S_1 and S_2 cooperate to cancel the crosstalk generated along the path $(0,2)$.

where L_i is the distance of the center of the head from the source S_i and c is the speed of sound. Both gains and delays are made of two contributions, the first due to the propagation of the sound from the source to the center of the head and the second related to the presence of the head. The values of the head related contributions $\tau_{ij}^{(H)}$ and $g_{ij}^{(H)}$ depend on the incidence angle of the sound and on the frequency, and are usually described with lookup tables. In order to construct suitable lookup tables, we subdivide the frequency range between 0 Hz and the Nyquist frequency into several bands, as described in our previous work [3]. Incidence angles are sampled on a discrete grid. The lookup tables are accessed with an index referring to the frequency band and with the incidence angles of the plane wave emitted from the source, and give back values of coefficients $\tau_{ij}^{(H)}$ and $g_{ij}^{(H)}$ (see equations 4 and 5). The exact values used for these coefficients are then obtained by low-order interpolation of the values read in the tables.

Lookup tables can be produced either by derivation from direct measures acquired for example with in-ear microphones, or generated with analytical HRTF models based e.g. on the Woodworth formula and IIR filters as presented in the work by Brown and Duda [15]. The development of high quality tables is a current research topic of our group, but will not be the focus of this paper.

The propagation law resulting from this approach is described by the impulse response:

$$h_{ij}(t) = g_{ij}\delta(t - \tau_{ij}) \tag{6}$$

In order to lighten the notation in the remaining part of the paper, we will not label formula with indexes referring to the selected frequency band. The dependency on frequency has been addressed in a previous paper [3] where a multiband version of our algorithm has been developed. The procedure described in the next sections can simply be repeated in each cancellation band, an approach that has been applied for the practical measures presented in the results section 7.

The basic operation that has to be repeated iteratively for computing the impulse responses of the cancellation complex is the cancellation from a speaker j of a crosstalk produced at the ear k by a sound emitted from a source i. We will label such an operation with the triple of indexes (ijk). Assuming the propagation law 6, it is easy to show that, if the source i emits a pulse $\delta(t)$, the crosstalk at the ear k will be $g_{ik}\delta(t - \tau_{ik})$. To cancel this crosstalk, the source j must send a signal $-\frac{g_{ik}}{g_{jk}}\delta(t - \tau_{ik} + \tau_{jk})$. This shows that the impulse response $h_{ijk}(t)$ connecting the source i to j via the ear k is given by:

$$h_{ijk}(t) = -\frac{g_{ik}}{g_{jk}}\delta(t - (\tau_{ik} - \tau_{jk})) \tag{7}$$

The cancellation signal will produce a new crosstalk that needs in turn to be cancelled and so on in a recursive way. The recursive computation starts with the triple (022) and goes on alternating the triples (211) and (122).

4 The problem of dynamic adaptation in time-domain CTC

The goal of this paper is to introduce the principles of a time-domain based CTC model that enables smooth transitions between consecutive user positions detected by a user tracking device. The presented approach aims for independence from specific tracking systems, which can be based on stereoscopic or time of flight cameras, or integrated in virtual reality or motion capture systems. At a sufficiently high tracking frame rate, such as when using commodity tracking sensors working at 25 to 50 fps, the introduction of smooth transitions between tracked user positions is an approach that allows for a fluid dynamic adaptation of the CTC process. With "smooth transition" we mean that, during the displacement between two consecutive positions, the user does not perceive artifacts connected with phase or amplitude discontinuities, and at the same time the cancellation effectiveness is maintained sufficiently strong to avoid unnatural losses of CTC spatiality perception. Phase or amplitude discontinuities tend to appear in the CTC signal because of the noisiness of the tracking system, whereas losses of CTC spatiality might be the consequence of an excessive latency of the whole tracking, computation and reproduction system. The objective concerning CTC spatiality can be restated by saying that the system must adjust the parameters τ and g appearing in the impulse responses 7 in such a way that the CTC sweetspot fluidly follows (ideally sample by sample) the user during movement. In static CTC, the sweetspot is the region of the head configuration space where the cancellation effectiveness remains above a sufficient threshold that allows for correct perception of the CTC effect [8]. The dimensions and position of the sweetspot mainly depends on the position of the used reproduction and cancellation loudspeakers and on how CTC algorithms are parameterized. Therefore, if the user performs a movement, the system must reconfigure the CTC signal generation process in such a way that the sweetspot chases the user as depicted in figure 2.

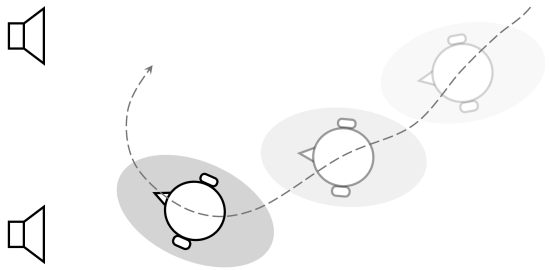


Fig. 2: The sweetspot (represented by a grey ellipse) has to follow the user while moving. The dashed grey line describes an hypothetical path followed by the listener.

An extension of the approach sketched in section 3 to support a moving user faces several difficulties. Good cancellation performances require a robust low-noise model $\alpha(t)$ of the moving user, where $\alpha(t)$ is a time dependent vector containing the head configuration,

i.e. the coordinates of the center of the head and the rotation angles. This model must be derived from the tracking information. Commodity tracking systems provide data streams of these quantities, which in their raw form might however result as excessively noisy in the context of CTC. Indeed, the accuracy of the model $\alpha(t)$ will be influenced by the capabilities of the tracking system, by the type of processing performed to reduce noise affecting the measures, as well as by the delay introduced by the smoothing of data. In addition, the model should ideally be able to forecast the position of the user for a time in the future in the order of the reciprocal of the frame rate and to compensate system latency.

At an audio sample rate of 48 kHz and for a reference frequency of 1000 Hz, an error of an audio sample in the alignment of the cancellation signal with the crosstalk roughly corresponds to an attenuation of the crosstalk of about 18 dB, a performance that has to be considered as borderline for optimal cancellation performances. Considering that, during a single audio sample, the sound wave covers about 0.7 cm at a velocity of 343 m/s, we see that the challenge of good cancellation performances for a moving user is a difficult one.

Real-time implementation is also conditioned by the underlying hardware and overall latency constraints of the system. Whilst a PC-based implementation benefits from abundant computing power, it may fall short from meeting the low latency required for good CTC reactiveness to movement. On the other hand, embedded processors such as DSPs can offer very low latency but may suffer from limited computing power. Understanding trade-offs between model complexity and achievable performance is thus crucial for the realization of efficient, real-world implementations of dynamic CTC systems. Also take into consideration, that with a low-latency system, the mentioned forecast behaviour attempting to precisely chase the exact position of the user might reveal more effective. Excessive latency becomes instead detrimental and might eliminate any potential positive effect provided by the forecast functionality.

5 An accurate time-domain CTC model for a moving user

We will now introduce an accurate model of a time-variant CTC.

A feasible generalisation to a moving user of the partial impulse responses in equation 7 can be based on the following basic assumptions:

- The sound propagates in spherical waves in such a way that the direct propagation time from the sources to the center of the head (hence not inclusive of the head related contributions nor room reflections) is considered to depend only on the distance between the source and the center of the head, as well as on the speed of sound.
- The sound wave is not perturbed by the movement of the head. In other words, the sound field, when the head is in a given position, is not dependent on its state of motion and coincides with the sound field that would be measured for a head in a static position.
- The propagation law of sound signals is given by equation 6 with time dependent coefficients τ and g.

The second assumption enables us to use the lookup tables generated for a user in a static position in the case of a moving one. This assumption is reasonable if the movement of the head is slow compared to the speed of sound c, which is always the case.

We will compute the time-dependent generalisation of the impulse responses in equation 7 with reference to a generic triple (ijk) where a crosstalk generated along the path (i,k) by a pulse emitted at time t_0 must be cancelled along the path (j,k). As a consequence of the assumptions above, if the pulse sent by the source i reaches the center of the head at the time t_H, it will be received at the ear k at the time t_E given by:

$$t_E = t_H + \tau_{ik}^{(H)}(t_H) \tag{8}$$

where $\tau_{ik}^{(H)}(t_H)$ is the head related delay along the path (ik) derived from the lookup tables accessed with the head in the configuration $\alpha(t_H)$, where $\alpha(t)$ is the model of head movement mentioned in section 4.

The time t_H is given by the unique solution of the equation:

$$t_H = t_0 + \frac{L_i(t_H)}{c} \tag{9}$$

where $L_i(t_H)$ is the distance of the center of the head from the source i at time t_H.The existence and uniqueness of the solution of equation 9 is given by the trivial fact that the pulse will hit the head once and only once.

Equation 9 can easily be solved if we assume that the radial velocity v_i of the center of the head with respect to the source i is constant during the propagation. Under this assumption, the solution of the equation is:

$$t_H = t_0 + \frac{L_i(t_0)}{c - v_i} \tag{10}$$

where $L_i(t_0)$ is the distance between the source i and the center of the head at the time t_0. Both the radial velocity v_i and the distance $L_i(t_0)$ can be determined by the model $\alpha(t)$.

In order to compute the timing of the cancellation signal, we need the time t_P when the pulse must be emitted from the source j to reach the ear k at the time t_E, equation 8. This time is given by the equation:

$$t_P = t_E - \left(\frac{L_j(t_0) + v_j \cdot (t_H - t_0)}{c} + \tau_{jk}^{(H)}(t_H) \right) \tag{11}$$

where, as in equation 10, we have introduced the constant radial velocity v_j of the center of the head with respect to the source j and the distance $L_j(t_0)$ between the source j and the center of the head at the emission time t_0.

Finally we need the formula for the gains along the path (l,k), with $l = i, j$, given by:

$$g_{lk} = \frac{1}{L_l(t_0) + v_l \cdot (t_H - t_0)} g_{lk}^{(H)}(t_H) \tag{12}$$

The gain g_P of the pulse to be emitted from the source j is, as in equation 7, the ratio of g_{ik} over g_{jk}:

$$g_P = \frac{L_j(t_0) + v_j \cdot (t_H - t_0)}{L_i(t_0) + v_i \cdot (t_H - t_0)} \cdot \frac{g_{ik}^{(H)}(t_H)}{g_{jk}^{(H)}(t_H)} \tag{13}$$

Now we have all the elements to write an algorithm for the computation of the partial time variant impulse responses 7 for a generic triple (ijk) for a pulse emitted at the time t_0:

- compute the distances $L_i(t_0)$ and $L_j(t_0)$ and the related radial velocities v_i and v_j from the model of the user's movement $\alpha(t)$;
- compute the time t_H using equation 10;
- compute the head configuration using the model $\alpha(t_H)$;
- using $\alpha(t_H)$, compute the incidence angles of the sound propagating along the paths (i,k) and (j,k) over the head and use the lookup tables to compute the delays and gains $\tau_{ik}^{(H)}(t_H)$, $\tau_{jk}^{(H)}(t_H)$, $g_{ik}^{(H)}(t_H)$, and $g_{jk}^{(H)}(t_H)$;
- using equations 8 and 11 compute the releasing time of the cancellation signal from the source j;
- using equations 13 compute the gain of the released cancellation signal from the source j;
- use the formula:

$$h_{ijk}(t) = -g_P \delta(t - t_P) \tag{14}$$

  to compute the dynamical generalisation of the partial impulse responses equation 7.

Like in the static case, the algorithm sketched above is used recursively starting with the triple (022) and going on alternating the triples (211) and (122). The iterations will be truncated at a maximal order N chosen in accordance with some optimisation criteria [2]. We will refer to this solution as the fully dynamic model.

6 Proposed implementation

The fully dynamic model sketched in section 5 is computationally expensive and includes tasks like the computation of time derivatives to evaluate radial velocities v_i and v_j that may lead to large errors due to numerical instabilities which may amplify the noise present in the user tracking information. Robustness, as well as a strong simplification, can be obtained if in the equations 10, 11 and 13, we approximate the ratios $\frac{v_i}{c}$ and $\frac{v_j}{c}$ with zero. In this case $t_H = t_0 + \frac{L_i(\alpha(t_0))}{c}$ and no time derivatives have to be computed. We will call the model obtained with these approximations the quasi-static model.

We will now describe a possible implementation for the model of movement introduced in section 4 and

used in section 5 to construct the partial, time variant impulse responses equation 14, which includes the forecast functionality described in section 4. The configuration of the user is described by a time dependent vector $\alpha(t)$ with 6 components containing the coordinates of the center of the head and three rotation angles. The components of $\alpha(t)$ are least square polynomials $P_j(t)$, with $0 \leq j \leq 5$, of degree d fitting the tracking data provided by the tracking system. The polynomials are computed using the data from the last n_F frames acquired by the tracking system and are updated when a new frame is acquired.

In order to make the model more robust, the sequence of the n_F frames acquired by the tracking system and used to compute the polynomials $P_j(t)$ is low-pass filtered with a first order IIR filter or with a minimum phase Bessel filter. This might introduce an additional delay that has to be considered when tuning the model parameters. Altogether, the model $\alpha(t)$ includes as free parameters the degree d of the polynomials, the number n_F of supporting points for the computation of the polynomials and the parameters of the frame filters. For the choice of these parameters a trade-off is required among four objectives: minimize the potential delay introduced by the filters, minimize the impact of noise on the evaluation of the polynomials, minimize the jump in $\alpha(t)$ occurring when a new frame is acquired and minimize the computational load. An aggressive filtering reduces the impact of noise and jumps in $\alpha(t)$ but might introduce delayed responses, whereas a mild filtering might increase the reactivity of the system to the movement of the user but leaves room for possible artifacts.

An additional improvement in accuracy and computational efficiency is made available by a trick we use in the computation of the impulse responses. The direct application of the algorithm sketched in section 5 would require an iterative computation of the partial responses 14 for each audio sample to be processed in the source audio channel. More precisely, a single audio sample generates a sequence of pulses released at different times that require multiple access to the model of movement $\alpha(t)$ and to the lookup tables. Instead of following this direct approach, we consider, at each time t, all the audio samples inducing, at some order in the recursion, an emission at time t. The advantage of this alternative approach is twofold. On the one hand, the individuation of all these samples requires only the configuration of the head at the current time t, leading to a more efficient computation. On the other hand, by accessing the model of movement α at the current time t, the result is less dependent on the polynomial extrapolation of the head configuration.

The use of first order IIR filters or minimum phase Bessel filters to reduce noise in the tracking data followed by least square polynomials to obtain a smooth forecast of the spatial configuration of the user does not guarantee that audio artifacts are completely avoided. Indeed, there could be other sources of signal irregularities besides recurring noise present in the tracking signal. This could be the case, for instance, for occlusions (e.g. in case of a temporary shadowing of the tracking sensors), or for more or less isolated anomalies due to tracking system confusion (e.g. for not ideal illumination or anomalous sources of light). Furthermore, for example for rapid modification of the movement direction, when the polynomials are updated, the model $\alpha(t)$ might change in a relatively abrupt way because the polynomials $P_j(t)$ are recomputed using new supporting points. We will not address this type of issues in this paper. We only mention that several additional smoothing approaches can be used, from the detection and elimination of outliers in the tracking signal, up to a temporary switch-off of the CTC signal in case of extreme jumps in the tracking data. In general, in order to reduce discontinuities in the audio output, a fade out and fade in procedure can be applied when a new frame is updated.

7 Experimental results

In this section, the effectiveness of our implementation of the newly proposed dynamic approach to CTC is assessed by means of laboratory experiments and measurements, conducted in similar conditions and using the same setup as in our previous works [2][3]. The setup consists of two loudspeakers mounted on two stands in front of the user, placed 60 cm away from each other.

The audio system used for the laboratory experiments operates at a sampling rate of $F_s = 48kHz$ and exploits 5 frequency bands ($B^{(1)}$ from $0Hz$ to $800Hz$, $B^{(2)}$ from $800Hz$ to $2kHz$, $B^{(3)}$ from $2kHz$ to $4kHz$, $B^{(4)}$ from $4kHz$ to $5.5kHz$ and $B^{(5)}$ from $5.5kHz$ to $F_s/2$). CTC is applied on bands $B^{(2)}$, $B^{(3)}$ and $B^{(4)}$ and the emitted signal corresponds to the sum of all bands.

A multisine of period $L = 4096$ samples is used as input excitation signal. A multisine signal has the same

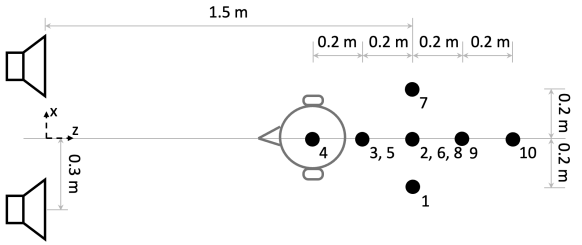


Fig. 3: Path followed by the user during the recordings.

spectral content as a single Dirac pulse in terms of individual components' magnitudes but components' phases are modified such that the signal's energy is spread as uniformly as possible over the L samples of the signal. This allows for fast and accurate steady-state response measurement of a system's transfer function as shown in [16].

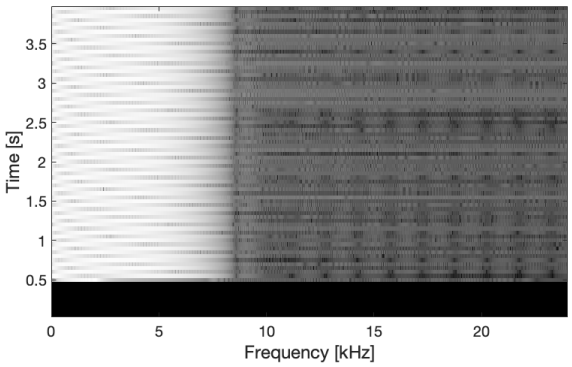
The recordings used to compute the effectiveness of the proposed approach are obtained by equipping a freely moving person with in-ear microphones. The tracking system features high quality sensors operating at 50 fps.

During the recordings, the user moves in the room following the pattern described in figure 3 starting at position 1 up to position 10, changing location (roughly) every 10 seconds.

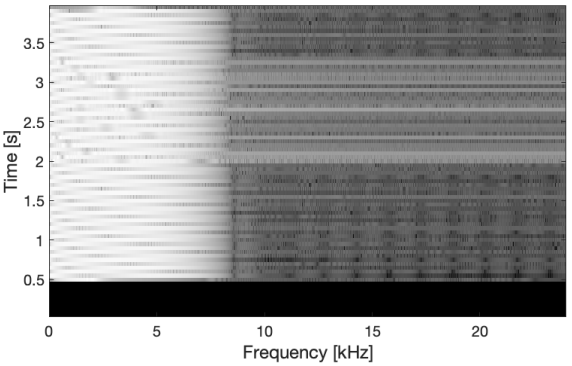
The pattern of movement in figure 3 has been chosen to limit incidence angles of the sound wave on the head. Indeed, currently available lookup tables do not support high level performances for large angles (high quality lookup tables is not the topic of this paper and is part of our current research). Furthermore, the cancellation effectiveness indicator used in the experiments is not reliable when the shadowing effect due to the head is too large, as explained below.

The robustness and performance of the approach is analysed using the following criteria: the presence of artifacts during the movement of the user and the effectiveness of the cancellation before, during and immediately after the movement.

The presence of artifacts in the generated signals is detected with two different strategies: firstly by listening to the generated audio and secondly by performing a spectral analysis of the signals. The results of our analysis provide evidence that there are no audible artifacts introduced by movement support.



(a) Signal without artifacts.



(b) Signal with intentionally added artifacts.

Fig. 4: Detail of the spectrogram.

Figure 4 is an example of the spectral analysis performed on two portions of the signal, one without artifacts and the second one with intentionally added artifacts, consisting in phase discontinuities at the audible limit. Consequently, the spectrogram can be used to provide evidence that no discontinuities or artifacts are present.

The cancellation effectiveness is estimated by computing the level difference between the signals received at the target and non target ears. The difference in level between target and non target ear is biased by the shadowing of the head. However, if the incidence angles are not too large, the shadowing is not too severe and the measures are able to give evidence of effective cancellation. A more appropriate measure would be based on the level difference at the non target ear when the cancellation is on and off. This type of measure, described in papers [2], [3] requires two different measures in the same conditions, one with enabled cancellation and the other with disabled cancellation, and is not easy to

implement in a dynamic situation.

The top part of figure 5 depicts the cancellation effectiveness when a multisine signal is emitted on the left channel while the user is moving following the path represented in figure 3. The bottom part of figure 5 shows the x and z position of the moving user.

It can be observed that the cancellation effectiveness is only marginally affected by the movement. The dynamic approach to CTC provides stability before and immediately after the movement and is still capable to produce, although with reduced efficiency, a spatiality effect while the user is freely moving. Latency is the main cause of the slightly reduced performance during movement.

8 Summary and conclusions

In this paper, we presented a novel time-domain approach to CTC capable of supporting arbitrary user positions and movements. Based on an approximation of the acoustical propagation model from the speakers to the ears and an approximation of the user's trajectory, the proposed solution brings dynamic reconfiguration to the concept of cancellation complexes introduced in our previous works ([2], [3]). Being implemented in the time-domain, our approach is not limited to frame-by-frame update of the CTC filters, but can also do so on a sample-by-sample basis whilst still operating with zero latency.

Laboratory experiments have confirmed the effectiveness of the proposed solution and its capacity to handle a moving listener in arbitrary spatial positions. It has been shown that CTC effectiveness is only marginally affected compared to the case of a static listener in a central position and that no audible side effects are introduced by supporting a moving user. This enables artifacts-free 3D audio perception for a freely moving user, opening the door for high-quality AR and VR applications.

Our near-term research priority is the generation of high-quality lookup tables, especially for high incidence angles, based on physical measurements and/or analytical models.

This work is an extension of an initial research program funded by Innosuisse, the Swiss funding agencies for innovative technologies, under grant 42471.1 IP-ICT INXS-3D.

References

[1] Glasgal, R., "360 Localization via 4.x RACE Processing," *Audio Engineering Society 123rd Convention*, 2007.

[2] Vancheri, A., Leidi, T., Heeb, T., Grossi, L., Spagnoli, N., and Weiss, D., "Geometrical Acoustics Approach to Crosstalk Cancellation," in *Proceedings of the Audio Engineering Society 152nd Convention*, 2022.

[3] Vancheri, A., Leidi, T., Heeb, T., Grossi, L., and Spagnoli, N., "Multiband time-domain crosstalk cancellation," in *Proceedings of the Audio Engineering Society 153rd Convention*, 2022.

[4] Bauer, B. B., "Stereophonic Earphones and Binaural Loudspeakers," *J. Audio Eng. Soc*, 9(2), pp. 148–151, 1961.

[5] Atal, B. S. and Schroeder, M. R., "Apparent sound source translator," 1966, uS Patent 3,236,949.

[6] Masiero, B. S., Fels, J., and Vorländer, M., "Review of the crosstalk cancellation filter technique," 2011.

[7] Gardner, W. G., "3-D Audio Using Loudspeakers," 1998.

[8] Lacouture Parodi, Y., *"A systematic study of binaural reproduction systems through loudspeakers: A multiple stereo-dipole approach"*, Ph.D. thesis, 2010.

[9] Choueiri, E. Y., "Optimal Crosstalk Cancellation for Binaural Audio with Two Loudspeakers," Self-published, 2010.

[10] Lee, K.-S. and Lee, S.-P., "A real-time audio system for adjusting the sweet spot to the listener's position," *IEEE Transactions on Consumer Electronics*, 56, 2010.

[11] Cecchi, S., Primavera, A., Virgulti, M., Bettarelli, F., Li, J., and Piazza, F., "An efficient implementation of acoustic crosstalk cancellation for 3D audio rendering," in *2014 IEEE ChinaSIP*, pp. 212–216, 2014, doi:10.1109/ChinaSIP.2014.6889234.

[12] Bruschi, V., Cecchi, S., Bruschi, V., Ortolani, N., and Piazza, F., "Immersive sound reproduction in real environments using a linear loudspeaker array," 2019.

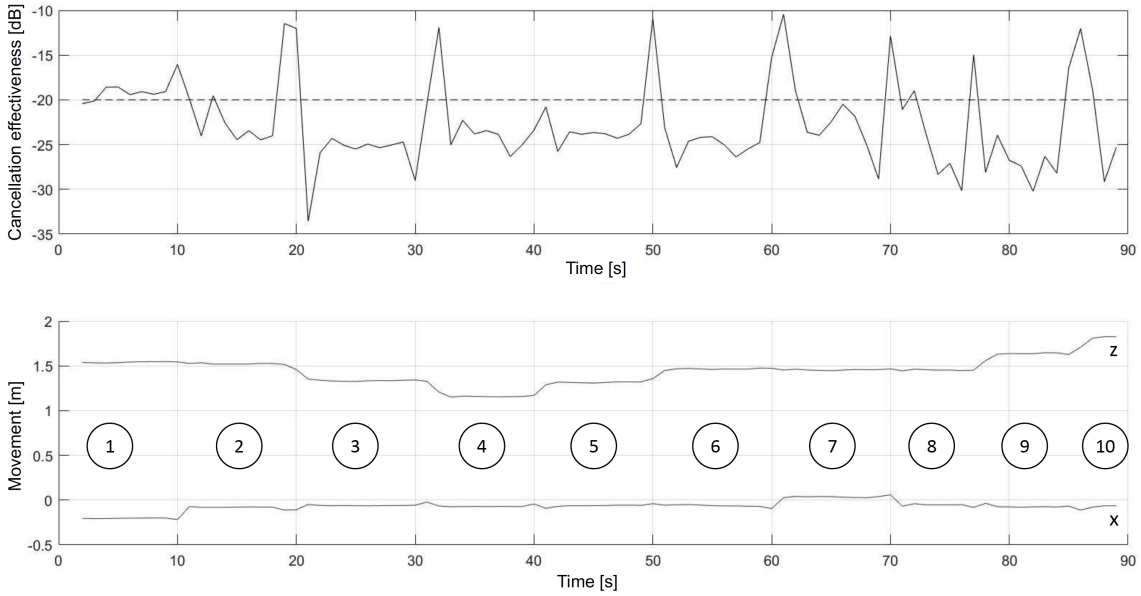


Fig. 5: Cancellation effectiveness with respect to the movement of the user. x and z directions are represented in figure 3. Numbers surrounded by a circle represent the positions illustrated in figure 3.

[13] Ma, X., Hohnerlein, C., and Ahrens, J., "Concept and Perceptual Validation of Listener-Position Adaptive Superdirective Crosstalk Cancellation Using a Linear Loudspeaker Array," *J. Audio Eng. Soc*, 67(11), pp. 871–881, 2019.

[14] Qiao, Y. and Choueiri, E., "Real-time Implementation of the Spectral Division Method for Binaural Personal Audio Delivery with Head Tracking," in *Proceedings of the Audio Engineering Society 151st Convention*, 2021.

[15] Brown, C. and Duda, R., "An efficient HRTF model for 3-D sound," in *Proceedings of 1997 Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 4 pp.–, 1997, doi: 10.1109/ASPAA.1997.625596.

[16] Beis, "Instant Audio Frequency Response Measurements," 2015, www.beis.de/Elektronik/FreqResp/InstFreqRespMeasure.html.