Audio Engineering Society

Convention Paper

Presented at the 120th Convention
2006 May 20–23 Paris, France

Improved Time Delay Analysis/Synthesis for Parametric Stereo Audio Coding

Christophe Tournery[1] and Christof Faller[1]

[1]*Audiovisual Communications Laboratory, EPFL Lausanne, Switzerland*

Correspondence should be addressed to C. Tournery (christophe.tournery@epfl.ch)

ABSTRACT

For parametric stereo and multi-channel audio coding, it has been proposed to use level difference, time difference, and coherence cues between audio channels to represent the perceptual spatial features of stereo and multi-channel audio signals. In practice, it has turned out that by merely considering level difference and coherence cues a high audio quality can already be achieved. Time difference cue analysis/synthesis did not contribute much to a higher audio quality, or, even decreases audio quality when not done properly. However, for binaural audio signals, e.g. binaural recordings or signals mixed with HRTFs, time differences play an important role. We investigate problems of time difference analysis/synthesis with such critical signals and propose algorithms for improving it. A subjective evaluation indicates significant improvement over our previous time difference analysis/synthesis.

1. INTRODUCTION

Recently, the coding gain for stereo and multi-channel audio coding has been significantly improved by representing the spatial aspects of stereo and multi-channel audio signals with perceptual spatial parameters. Such techniques have been investigated in a number of papers on Binaural Cue Coding (BCC) [1, 2] and Parametric Stereo (PS) [3]. A BCC scheme is illustrated in Figure 1. Such a BCC scheme has previously been denoted "BCC for natural rendering" or "BCC type II". In the following when referring to BCC we always mean this flavor of BCC. While these techniques propose the use of level difference, time difference (or phase difference [3]), and coherence as perceptual spatial cues, in practice mostly only level difference and coherence cues have been used. The de-correlation effect of relatively large time delays is indeed captured with the coherence cues. Thus, for loudspeaker playback it is not clear whether time difference cues improve audio quality compared to not considering time difference cues. Other factors leading to abandoning time difference cues are that often audio content is mixed with only amplitude panning (thus time difference cues are not present) and the fact that time difference analysis/synthesis does not perform as well as an-

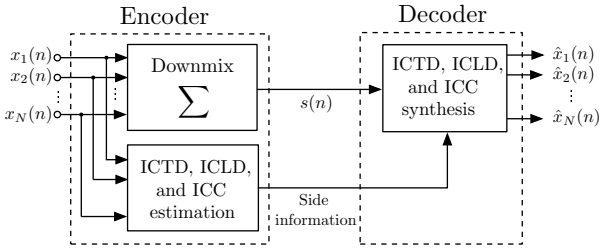

Fig. 1: Binaural Cue Coding (BCC) scheme. A multichannel audio signal is downmixed to a single channel. The single channel and side information is transmitted to the decoder.

ticipated.

If BCC or PS is to be applied to binaural recordings or signals mixed with head related transfer functions (HRTFs) or binaural room impulse responses (BRIRs), time difference cues are essential. Informal listening revealed that without time difference cues the spatial image width is significantly less wide or externalized than when time differences are used. In this paper, we propose improvements for time difference analysis and synthesis processing and apply it to this important class of signals.

We are addressing the following issues related to ICTD analysis and synthesis:

- At higher frequencies time delays can not be simply measured as phase delays, due to the "phase wrapping problem". We are proposing a computationally efficient algorithm which estimates at higher frequencies the group delay as opposed to the phase delay.
- The time delays vary not only as a function of frequency but also as a function of time. The variation of time delays potentially causes artifacts in the overlapping part of two frames with different time delays.

Throughout the paper, the three cues level difference, time difference, and coherence will be denoted *inter-channel level difference* (ICLD), *inter-channel time difference* (ICTD), and *inter-channel coherence* (ICC), respectively.

The paper is organized as follows. Section 2 describes the short-time Fourier transform (STFT) based processing we are using. The ICTD estimation algorithm, using phase delay estimation at low frequencies and group delay estimation at high frequencies is described in Section 3. ICTD synthesis is described in Section 4. A subjective evaluation, comparing the previously published ICTD synthesis to the proposed scheme, is described in Section 5. The conclusions are in Section 6.

2. TIME-FREQUENCY PROCESSING

In this section, we are describing the short-time Fourier transform (STFT) based processing which is used for the BCC analysis and synthesis scheme we are investigating in this paper.

2.1. STFT processing

The use of the STFT is in the following motivated by its suitability for BCC synthesis. Generally speaking, BCC synthesis applies time varying filtering to its mono input signal to generate its output signals. A short-time Fourier transform (STFT) is used for this purpose. Given a signal $s(n)$, its STFT spectra are denoted $S(k,i)$, where k is the spectrum time index and i is the frequency index.

A frame of N samples is multiplied with a window before a N-point DFT is applied. We use a Hann window with zero padding at both sides,

$$w_a(l) = \begin{cases} 0 & \text{for} \quad 0 \le l < Z \quad \text{or} \\ & \quad N-Z \le l < N \\ \sin^2\left(\frac{(l-Z)\pi}{W}\right) & \text{for} \quad Z \le l < Z+W\,, \end{cases} \tag{1}$$

where Z is the width of the zero region before and after the non-zero part of the window. Figure 2 shows the described window schematically.

The non-zero window span is W and the size of the transform is $N = 2Z+W$. Adjacent windows are overlapping and are shifted by $W/2$ samples (hop size). The window was chosen such that the overlapping windows add up to a constant value of 1. Therefore, for the inverse transform there is no need for additional windowing. A plain inverse DFT of size N with time advance of successive frames by $W/2$ samples is used. If the spectrum is not modified, perfect reconstruction is achieved by overlap add.

With appropriate zero padding, filtering with a filter of any length can be implemented by multiplying the STFT spectra with frequency responses of the filter [4]. Thus, the described STFT is suitable for BCC synthesis (application of scale factors for ICLD synthesis, delays for ICTD synthesis, and filtering for ICC synthesis [4]).

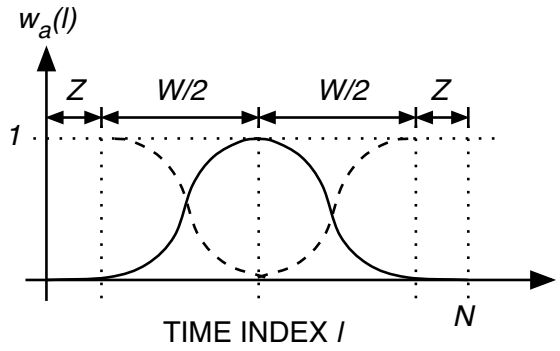

Fig. 2: Analysis window. The time-span of the window W is shorter than the DFT length N such that non-circular time-shifts within the range $[-Z,\ Z]$ are possible. The window is advanced by $W/2$ samples.

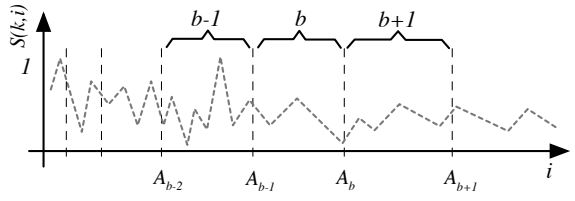

Fig. 3: The spectral coefficients belonging to one partition are $X(k,i)$ with $A_{b-1} \le i < A_b$.

2.2. Perceptually motivated spectral resolution

The uniform spectral resolution of the STFT is not well adapted to human perception. Therefore, the uniformly spaced spectral coefficients $S(k,i)$ $(0 \le i \le N/2)$ are grouped into B non-overlapping partitions with bandwidths better adapted to perception. Only the first $N/2+1$ spectral coefficients of the spectrum are considered because the spectrum is symmetric. The indices of the STFT coefficients $S(k,i)$ which belong to the partition with index b $(1 \le b \le B)$ are $i \in \{A_{b-1}, A_{b-1}+1, \ldots, A_b - 1\}$ with $A_0 = 0$, as is illustrated in Figure 3. The signals represented by the spectral coefficients of the partitions correspond to the perceptually motivated subband decomposition used by BCC. Thus, within each such partition only one set of inter-channel cues (ICLD, ICTD, ICC) is synthesized for each channel pair.

For our experiments we used $W = 640$, $Z = 192$, and $N = 1024$ for a sampling rate of 44.1 kHz. We used $B = 20$ partitions, each having a bandwidth of approximately 2 ERB [5]. Figure 4 illustrates the partitions used for the given parameters. Note that the last partition is smaller than two ERB due to the cutoff at the Nyquist frequency.

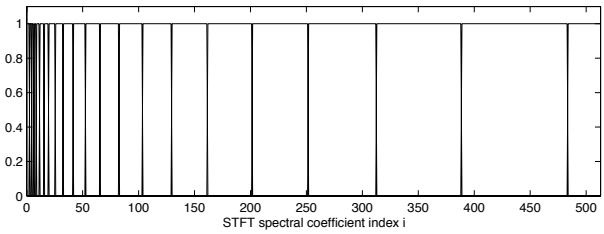

Fig. 4: The spectral coefficients of the uniform STFT spectrum are grouped to mimic the non-uniform frequency resolution of the auditory system.

3. TIME DELAY ESTIMATION

Estimation of ICLD and ICC within partitions is described in [6]. Here we are describing in detail how to estimate ICTD, while avoiding phase wrapping at higher frequencies.

3.1. Phase delay and group delay estimation

For partitions at frequencies below $\frac{1}{2\tau_{\max}}$ Hz, where $[-\tau_{\max},\ \tau_{\max}]$ is the range of delays to be estimated in seconds, no phase wrapping occurs and the time delay for each partition b is computed as the slope of the phase difference between left and right channels,

$$\Phi(k,i) = \arg\left(X_1(k,i)X_2^*(k,i)\right). \tag{2}$$

Linear regression is applied to estimate the slope of $\Phi(k,i)$ which is proportional to the delay. As long as no phase wrapping occurs, the regression line goes through the origin and the fitted equation is

$$\hat{\Phi}(k,i) = a_1 i, \tag{3}$$

where $\hat{\Phi}(k,i)$ is the predicted value of the phase difference. For each partition b, the group delay in expressed as a function of the slope a_1 as

$$\mathrm{ICTD}(k,b) = \frac{a_1 N}{2\pi}. \tag{4}$$

Equation 3 can be written in matrix form as $\mathbf{H}a_1 = \mathbf{y}$, where $\mathbf{H}$ is a $B \times 1$ matrix of the spectral coefficient indices, i.e. $H_{1,1} = A_{b-1}$, $H_{2,1} = A_{b-1} + 1$,

$\ldots$, $H_{A_b-A_{b-1},1} = A_b - 1$, and $y_1 = \Phi(k, A_{b-1})$, $y_2 = \Phi(k, A_{b-1}+1)$, $\ldots$, $y_{A_b-A_{b-1}} = \Phi(k, A_b - 1)$. The linear least squares solution derived from the normal equation [7],

$$\mathbf{H}^{\mathrm{T}}\mathbf{H}a_1 = \mathbf{H}^{\mathrm{T}}\mathbf{y}\,, \tag{5}$$

is given by

$$a_1 = \left(\mathbf{H}^{\mathrm{T}}\mathbf{H}\right)^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{y}\,. \tag{6}$$

We take into account the energy present in the left and right spectrum with a weighted linear least squares regression whose solution is similar to Equation 6,

$$a_1 = \left(\mathbf{H}^{\mathrm{T}}\mathbf{W}\mathbf{H}\right)^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{W}\mathbf{y}\,, \tag{7}$$

where $\mathbf{W}$ is a diagonal matrix with its non-zero diagonal elements determining the weight for each measurement [7]. We choose the weights as a function of the magnitude of the cross-spectrum between left and right,

$$W_{i,i} = |X_1(k,i)X_2^*(k,i)|^{\beta}\,, \tag{8}$$

where $\beta > 1$ gives more emphasis on the cross-spectrum coefficients with high energy. Informal listening and fitting experiments show that $\beta = 3$ gives good results and is the value used in the following.

Above $\frac{1}{2\tau_{\max}}$ Hz phase wrapping occurs and requires to unwrap $\Phi(k,i)$ before applying the linear regression. Unwrapping is done by adding or subtracting 2π to $\Phi(k,j)$, $j \in \{i+1,\ldots,A_b-1\}$, every time $|\Phi(k,i+1)-\Phi(k,i)|$ is larger than π.

The group delay is then estimated as before except that the fitted line does not go through the origin anymore. Therefore the fitted equation is

$$\hat{\Phi}(k,i) = a_1 i + a_0\,, \tag{9}$$

which can be written in matrix form as

$$\mathbf{H}\begin{bmatrix} a_1 & a_0 \end{bmatrix}^{\mathrm{T}} = \mathbf{y}\,, \tag{10}$$

where $\mathbf{y}$ and the first column of $\mathbf{H}$ are defined as previously, and the second column of $\mathbf{H}$ contains only ones.

Given a_1 computed from Equation 10, the ICTD(k,b) is obtained with Equation 4. To illustrate the fitting process, Figure 5 shows for one particular band the phase difference between left and right, the fitted line and the normalized power of the cross-spectrum. Note how part of the data is almost ignored by the regression process due to its low power.

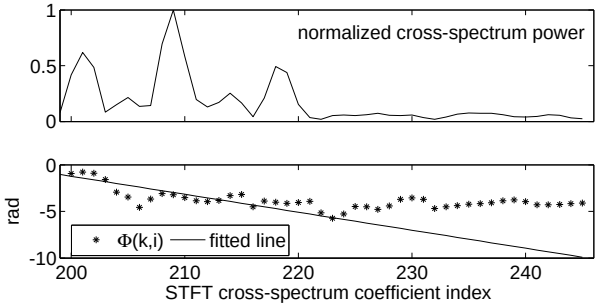

Fig. 5: Phase difference $\Phi(k,i)$ and normalized power of the cross-spectrum, and linear regression on $\Phi(k,i)$.

3.2. Perceptually motivated adjustment of the estimated ICTD

For the type of audio items we concentrate on, stereo signals processed with BRIRs for headphone playback, informal listening revealed that ICTD synthesis introduces so many artifacts that the quality is clearly lower than when no ICTD cues are synthesized (i.e. ICTD$=0$). Optimization in terms of adjusting the time or frequency resolution for ICTD synthesis did not yield enough improvement. We also experimented with an algorithm which compensates for the energy loss in the overlapping region between successive STFTs, without notable success.

Thus, we attacked the problem from a different perspective. The idea was to consider ICTD cues only when they are perceptually relevant. Note that for headphone playback (ideally) the ICTD, ICLD, and ICC cues are identical to the interaural cues at the ear entrances. Motivated by the "cue selection" auditory source localization model [8], we hypothesized that the ICTD cues are only perceptually relevant when the ICC is relatively high. ICC has a range between zero and one, where one indicates that the left and right partition signals are coherent. Thus, we assumed that ICTD are important when ICC is relatively close to one.

Additionally, when the signal is tonal we avoid to change the ICTD in time, since the ear is much more sensitive to artifacts for tonal signals than for non-tonal signals, as is indicated by psychoacoustic masking experiments [9]. For this purpose, we estimate in each partition, as a function of time, a tonality measure denoted TON(k,b). TON$(k,b) = 1$ indicates that the signal is tonal and stationary, while TON$(k,b) = 0$ indicates that the signal is noise-like or a transient.

Given the described considerations, "smoothed" ICTD cues are computed given the estimated ICTD cues:

$$\begin{aligned}\mathrm{ICTD}_{\mathrm{sm}}(k,b) = \alpha\mathrm{ICTD}(k,b) \\ +(1-\alpha)\mathrm{ICTD}(k,b-1)\,,\end{aligned} \tag{11}$$

with the following forgetting factor

$$\alpha = (1-\mathrm{TON}(k,b))\,\mathrm{ICC}(k,b)\,. \tag{12}$$

This results in that when the signal is not tonal and the ICC is high, the ICTD is quickly adjusted to the estimated ICTD. In the other cases, the ICTD is only modified slowly in time, avoiding artifacts due to overlap add of signals with different delays.

Additionally, we assume that when the ICLD and ICTD give contradictory localization information, they are also perceptually irrelevant. Thus, in this case the ICLD cue is assumed to be more reliable and the ICTD is set to zero. This process is applied to the estimated cues prior to applying Equation 11. It is described in the following in detail.

Considering the range of permissible values for ICLD and ICTD cues let

$$\mathrm{ICLD}_{\mathrm{nrm}}(k,b) = \frac{\mathrm{ICLD}(k,b)}{\mathrm{ICLD}_{\mathrm{max}}} \tag{13}$$

$$\mathrm{ICTD}_{\mathrm{nrm}}(k,b) = \frac{\mathrm{ICTD}(k,b)}{\mathrm{ICTD}_{\mathrm{max}}}\,, \tag{14}$$

which can be interpreted as normalized directional information between left and right. $\mathrm{ICLD}_{\mathrm{max}}$ and $\mathrm{ICTD}_{\mathrm{max}}$ define the permissible range of values that the corresponding cues can take. We used $\mathrm{ICLD}_{\mathrm{max}} = 20$ dB and $\mathrm{ICTD}_{\mathrm{max}} = f_s\tau_{\mathrm{max}}$ where $\tau_{\mathrm{max}} = 0.7$ msec. To prevent contradicting level and time difference cues, we set $\mathrm{ICTD}(k,b)$ to zero whenever

$$\left|\mathrm{ICLD}_{\mathrm{nrm}}(k,b) - \mathrm{ICTD}_{\mathrm{nrm}}(k,b)\right| > d \tag{15}$$

where d determines the degree of contradiction between the cues until the ICTD is set to zero. We use $d = 1$ which only sets ICTD to zero for highly contradicting cues.

An example ICTD cues resulting from the described smoothing and setting to zero processing is shown in Figure 6. The figure also shows the originally estimated ICTD cues and the ICC and tonality measures.

The classic duplex theory [10] of sound localization, states that at high frequencies time delay cues are less salient than level difference cues. Motivated by this we set the ICTD cues above a certain frequency to zero. We tested cut-off frequencies of 1, 1.5, 2, and 2.5 kHz. Informal listening indicated that ICTD synthesis was necessary up to 2.5 kHz in order that there was no notable reduction in the width of the auditory spatial image. Thus we synthesize ICTD cues up to 2.5 kHz.

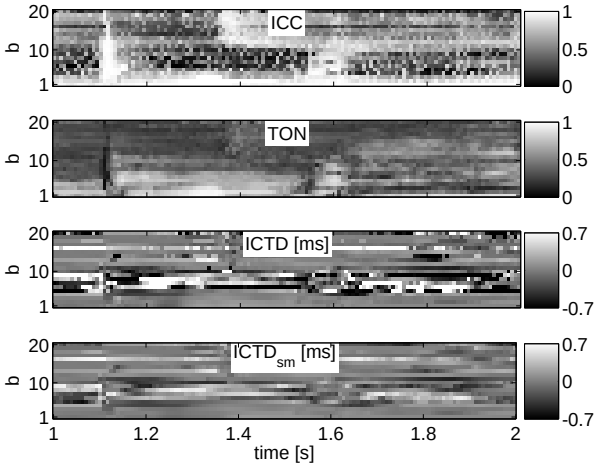

Fig. 6: ICC, tonality, non-smoothed and smoothed ICTD cues [ms] for one second of Excerpt F (Table 1).

4. TIME DELAY SYNTHESIS

The synthesis of ICLD and ICTD within partitions is described in detail in [6]. Additionally, we are using ICC synthesis as is described in [4].

Given the STFT spectrum of the sum signal, $X(k,i)$, the left and right spectra with the desired ICTDs are computed with

$$\begin{aligned}X_1(k,i) &= \exp\left(j\frac{2\pi n\mathrm{ICTD}_{\mathrm{sm}}(k,i)}{2N}\right)X(k,i) \\ X_2(k,i) &= \exp\left(-j\frac{2\pi n\mathrm{ICTD}_{\mathrm{sm}}(k,i)}{2N}\right)X(k,i)\,.\end{aligned} \tag{16}$$

For obtaining an ICTD value for each STFT coefficient i, the partition ICTD are interpolated.

5. SUBJECTIVE EVALUATION

5.1. Subjects and playback setup

One experienced and ten non-expert subjects participated in the subjective test. The results from one non-expert

listener were not taken into account due to inconsistent rating of the test items (hidden reference often wrongly detected). For audio playback an *Apple PowerBook G4* laptop computer was used with an external digital audio-out device (*M-Audio Sonica Theater*) connected directly to *Sennheiser HD600* headphones.

5.2. Stimuli

Ten different stereo audio excerpts, each with a duration of 10 s sampled at 44.1 kHz, were used in the test. The first two were used as training items and excluded from the results. Table 1 summarizes the content of the considered excerpts.

All selected excerpts have a pronounced wide spatial image, i.e. auditory events are localized widely between left and right. Most items are also ambient, except item G. The items include critical signal components such as transients, tonal parts, prominent vocal parts, which are critical for BCC synthesis.

Each excerpt was processed in three different ways:

No ICTD Only ICC and ICLD cues are synthesized (ICTD= 0).

ICTD ICTD cues are estimated as described in Section 3.1 but are not smoothed before synthesis. ICC, ICLD and ICTD cues are synthesized.

ICTD sm Similar to "ICTD" but ICTD cues are processed as described in Section 3.2 prior to synthesis. ICC, ICLD, and ICTD_{sm} cues are synthesized.

Table 1: List of the audio excerpts used.

excerpt	*name*	*category*
A	A Room Of Your Own	pop/rock
B	Blue Eyes	pop/rock
C	Bovio	latin
D	Carnival	classical
E	Ella Y Yo	latin
F	He Perdido Contigo	latin
G	Help	pop/rock
H	Herr Herr	classical

5.3. Test method

The subjects were asked to grade different specific degradations and the overall audio quality of the processed excerpts with respect to the known reference, i.e. the original excerpt. The three different grading tasks of this test are summarized in Table 2. Task 1 assesses the sound stage width disregarding any distortion in the audio quality. The motivation for this task was to assess to what degree the image width is reduced by BCC synthesis. Informal listening indicated that ICTD is crucial for a wide (and externalized) spatial image. Task 2 evaluates degradation introduced by BCC not related to image width. The goal of this task is to capture distortions or spatial image artifacts other than image width. Task 3 assesses the overall preference of the subjects.

Table 2: Tasks and scales of the subjective test.

task		*scale*
1	image width	wide (5) . . . narrow (1)
2	audio quality ignoring image width	ITU-R 5-grade impairment
3	overall audio quality	ITU-R 5-grade impairment

Table 3: ITU-R 5-grade impairment scale.

grade	*scale*
5	no difference
4	slight difference, not annoying
3	slightly annoying
2	annoying
1	very annoying

The test was carried out in a sound insulated room and the listeners were free to adjust the volume to a level comfortable to them. During the test, each subject was able to access the reference and randomly access the item processed in the three different ways and the hidden reference. Each item could be accessed with a corresponding "Play" button of a graphical user interface as shown on Figure 7. Simultaneous switching between the items was possible, i.e. when another play button was pressed, audio instantly faded to the corresponding processed item. The gradings were entered via graphical sliders that were permanently visible for all test items and could be adjusted at any time to reflect the proper grading and ranking.

It is important to note that subjects were specifically asked to pay attention to the rank order of the test items. The feature of being able to play the items according to their rank order greatly facilitates this task as opposed

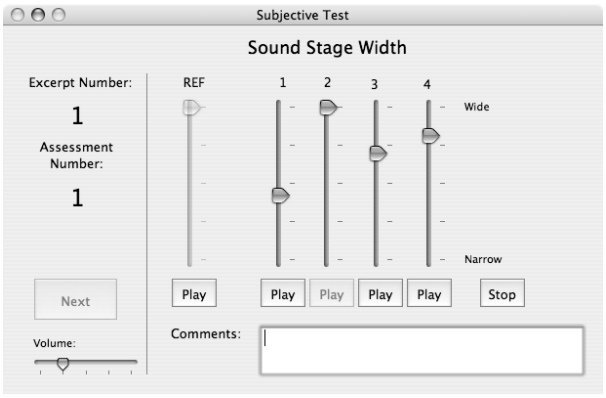

Fig. 7: User interface of the subjective test software.

to other testing schemes that allow to listen only once to each item in a pre-defined order. The ordering of the hidden reference and the processed items was randomly chosen for each subject and each excerpt but not changed during the three different tasks performed for each excerpt. The philosophy of this test method corresponds closely to MUSHRA [11].

5.4. Results

The results of the subjective test are shown in Figures 8, 9, and 10. For each excerpt the mean over all subjects and 95% confidence intervals of the rating are shown. Further, the overall mean over all excerpts and listeners and corresponding 95% confidence interval are shown.

Image width: The gradings for image width are shown in Figure 8. As expected, without ICTD synthesis the spatial image is most narrow. The regular ICTD synthesis has a more wide spatial image than the proposed ICTD synthesis (ICTD sm). However, the barely overlapping confidence intervals of the overall average gradings indicate that the proposed ICTD synthesis clearly improves image width compared to no ICTD synthesis.

During informal listening prior to the subjective test, we were not aware of the image width difference between the regular and proposed ICTD synthesis. Different tuning of the smoothing scheme may further improve image width.

Audio quality, disregarding image width: In task 2 the audio quality is assessed with respect to the reference without considering the spatial image width.

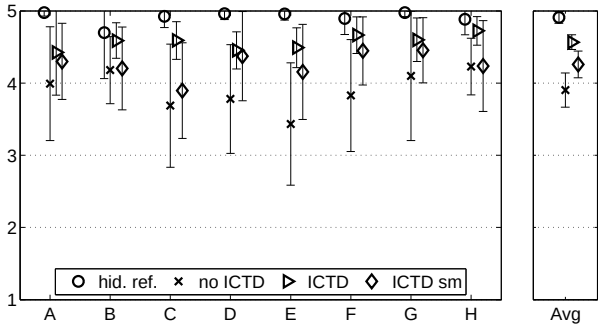

Fig. 8: Subjective test results for task 1: image width. The grading averaged over all listeners are shown for each excerpt and averaged over all excerpts. For each grading its 95% confidence interval is shown. A grading of 5 indicates the same width as the reference and 1 indicates "narrow".

The results in Figure 9 show that both, the proposed ICTD synthesis (ICTD sm) and no ICTD synthesis, have significantly less distortions than the regular ICTD synthesis. The proposed ICTD synthesis has about the same amount of distortions than no ICTD synthesis. This indicates that the proposed scheme is successful in eliminating the problems on audio quality of the regular ICTD synthesis.

Overall audio quality: The overall quality gradings in Figure 10 show the integral impact of all noticeable degradations on audio quality and provide an indication of the overall audio quality, including distortions and spatial image attributes. The results imply that the proposed ICTD synthesis (ICTD sm) provides significantly better overall audio quality than regular ICTD synthesis. The barely overlapping confidence intervals of the overall average gradings indicate that the proposed ICTD synthesis benefits the audio quality compared to no ICTD synthesis. As implied by the image width gradings, the benefit in audio quality is expected to be due to the improved image width.

Discussion of the results: Regular ICTD synthesis hurts the overall audio quality more than it gives benefit compared to no ICTD synthesis. The proposed scheme which synthesizes the estimated ICTD only in cases when they are expected to be perceptually relevant is able to largely remove the artifacts of regular ICTD synthesis, while

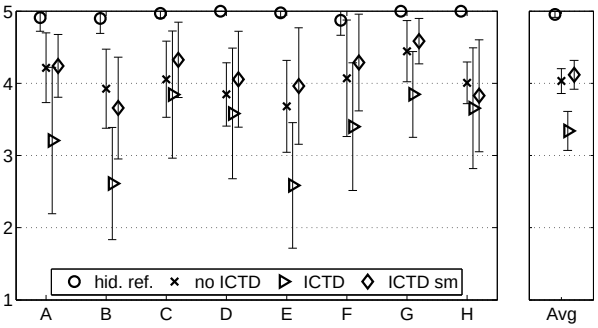

Fig. 9: Subjective test results for task 2: Audio quality disregarding image width. The grading averaged over all listeners are shown for each excerpt and averaged over all excerpts. For each grading its 95% confidence interval is shown.

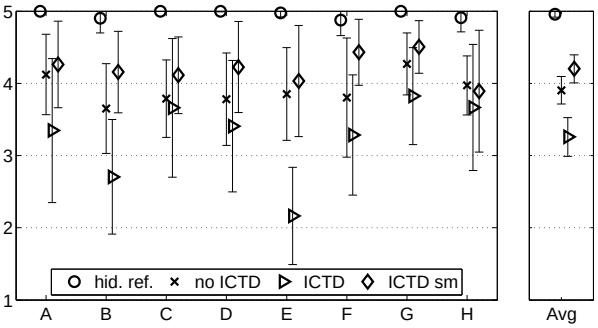

Fig. 10: Subjective test results for task 3: Overall audio quality. The grading averaged over all listeners are shown for each excerpt and averaged over all excerpts. For each grading its 95% confidence interval is shown.

still providing the benefit of an improved spatial image width. For the binaural signals generated from stereo signals with BRIRs, spatial image width is particularly important since it gives an indication about the degree of externalization [12] which "survives" BCC encoding and decoding. The proposed scheme is not perfect in this respect and we were maybe too conservative since regular ICTD synthesis results in a wider spatial image. We were not aware of this prior to the subjective test and plan to further tune the scheme for improved image width without more distortions.

6. CONCLUSIONS

While for average stereo audio content time difference (ICTD) synthesis is in many cases not important, i.e. there is not an obvious benefit in spatial image width due to ICTD synthesis compared to no ICTD synthesis, time difference synthesis is important for binaural audio signals such as binaural recordings or signals generated with HRTF or BRIR filtering. We investigated ICTD analysis and synthesis using stereo signals filtered with BRIRs. Spatial image width in this case is clearly reduced (externalization is largely lost) when no ICTD synthesis is used. On the other hand ICTD synthesis introduces so many distortions than it results in overall worse audio quality. The goal of this work was to improve ICTD analysis and synthesis at least to the point where ICTD synthesis gives a benefit compared to no ICTD synthesis.

We proposed a BCC analysis scheme which circumvents the phase wrapping problem at higher frequencies by estimating group delays using a weighted linear regression based algorithm. To reduce the artifacts of ICTD synthesis, the estimated ICTD are only synthesized when they are believed to be perceptually most important. In the other cases, the ICTD are only varied slowly in time to avoid the artifacts introduced by the overlap add with time varying delays. The estimated ICTD are only synthesized when the coherence between the audio channels is high, implying the perceptual importance of the corresponding ICTD.

The results of a subjective test implies that the proposed scheme largely eliminates the artifacts of the previous ICTD synthesis. Also, there is a benefit of the proposed ICTD synthesis compared to no ICTD synthesis in terms of spatial image width.

7. APPENDIX - CONFIDENCE INTERVALS

In the following the statistics used for confidence interval computation is described. The confidence interval for excerpt e processed with method i is computed as recommended in MUSHRA [11], i.e.

$$[\overline{u}_{e,i} - \delta_{e,i},\ \overline{u}_{e,i} + \delta_{e,i}]\ , \qquad \text{(A-1)}$$

where $\overline{u}_{e,i}$ is the average rating over all listeners, and

$$\delta_{e,i} = t_{0.05}\frac{\sigma_{e,i}}{\sqrt{N}}\,. \qquad \text{(A-2)}$$

The number of listeners is $N = 10$, and $t_{0.05} = 2.101$ is the t value corresponding to 10 samples for each method

i. The standard deviation is given by

$$\sigma_{e,i} = \sqrt{\sum_{l=1}^{N} \frac{(u_{e,i,l} - \overline{u}_{e,i})^2}{N-1}}, \qquad \text{(A-3)}$$

where $u_{e,i,l}$ is the rating of excerpt e processed with method i given by listener l.

When computing the confidence interval over all excerpts and all listeners, we remove the mean rating of each excerpt instead of the overall mean rating over all excerpts and all listeners. The reason is that separate excerpts are expected to have different means depending on their audio content. Therefore the confidence interval is given by

$$[\overline{u}_i - \delta_i,\ \overline{u}_i + \delta_i], \qquad \text{(A-4)}$$

where $\overline{u}_i$ is the average rating of method i over all excerpts and all listeners, and

$$\delta_i = t_{0.05} \frac{\sigma_i}{\sqrt{MN}}, \qquad \text{(A-5)}$$

where $M = 8$ is the number of excerpts and the t value corresponding to $MN = 80$ samples for each method i is $t_{0.05} = 1.98$. The standard deviation is given by

$$\sigma_i = \sqrt{\sum_{e=1}^{M} \sum_{l=1}^{N} \frac{(u_{e,i,l} - \overline{u}_{e,i})^2}{MN-1}}. \qquad \text{(A-6)}$$

8. REFERENCES

[1] C. Faller and F. Baumgarte, "Efficient representation of spatial audio using perceptual parametrization," in *Proc. IEEE Workshop on Appl. of Sig. Proc. to Audio and Acoust.*, Oct. 2001, pp. 199–202.

[2] C. Faller, *Parametric Coding of Spatial Audio*, Ph.D. thesis, Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland, July 2004, Thesis No. 3062, http://library.epfl.ch/theses/?nr=3062.

[3] E. Schuijers, W. Oomen, A. C. den Brinker, and A. J. Gerrits, "Advances parametric coding for high-quality audio," in *Proc. MPCA*, Nov. 2002.

[4] C. Faller, "Parametric multi-channel audio coding: Synthesis of coherence cues," *IEEE Trans. on Speech and Audio Proc.*, vol. 14, no. 1, pp. 299–310, Jan. 2006.

[5] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hear. Res.*, vol. 47, pp. 103–138, 1990.

[6] C. Faller and F. Baumgarte, "Binaural Cue Coding - Part II: Schemes and applications," *IEEE Trans. on Speech and Audio Proc.*, vol. 11, no. 6, pp. 520–531, Nov. 2003.

[7] Norman R. Draper and Harry Smith, *Applied Regression Analysis*, Wiley series in probability and mathematical statistics. Wiley-Interscience Publication, New York, third edition, 1998.

[8] C. Faller and J. Merimaa, "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," *J. Acoust. Soc. Am.*, vol. 116, no. 5, pp. 3075–3089, Nov. 2004.

[9] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*, Springer, New York, 1999.

[10] Rayleigh, "On our perception of sound direction," *Philos. Mag.*, pp. 13:214–232, 1907, (J. W. Strutt).

[11] Rec. ITU-R BS.1534, *Method for the subjective assessment of intermediate quality levels of coding systems*, ITU, 2003, http://www.itu.org.

[12] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*, The MIT Press, Cambridge, Massachusetts, USA, revised edition, 1997.