Audio Engineering Society

Convention Paper

Presented at the 138th Convention
2015 May 7–10 Warsaw, Poland

Reproduction of realistic background noise for testing telecommunications devices

Juan David Gil Corrales[1], Wookeun Song[2], and Ewen MacDonald[1]

[1] *Technical University of Denmark, Department of Electrical Engineering, Lyngby, Denmark*

[2] *Brüel & Kjær Sound and Vibration Measurement A/S, Nærum, Denmark*

Correspondence should be addressed to Juan David Gil Corrales (juandagilc@gmail.com)

ABSTRACT

A method for reproduction of sound, based on crosstalk cancellation using inverse filters, was implemented in the context of testing telecommunications devices. The effect of the regularization parameter, number of loudspeakers, type of background noise, and a technique to attenuate audible artefacts, were investigated. The quality of the reproduced sound was evaluated both objectively and subjectively with respect to the reference sounds, at points where telecommunications devices would be potentially placed around the head. The highest regularization value gave the best results, the performance was equally good when using eight or four loudspeakers, and the reproduction method was shown to be robust for different program materials. The proposed technique to reduce audible artefacts increased the perceived similarity.

1. INTRODUCTION

Algorithms for speech enhancement in the presence of background noise are integral to modern telecommunications devices. The development and testing of these algorithms require realistic reproduction of the background noise under controlled conditions, and the influence of the testing facilities has to be minimized to ensure reliable results. In this study, a technique based on the calculation of inverse filters, that compensate for the response of the room and the reproduction system, to accurately reproduce the sound with a limited amount of loudspeakers at particular test positions, is presented.

Different methods for reproduction of sound using multiple loudspeakers exist. One standardized method, ETSI EG 202 396-1 [1], is based on reproduction using four loudspeakers of binaurally recorded signals. The loudspeakers' signals are cal-

culated following an equalization procedure, that compensates for the magnitude difference of the sound pressure level of the reproduced sound at the ears, with respect to the reference one; this equalization procedure does not compensate for the crosstalk. As there are no phase corrections either, the coherence between reproduced and reference sounds can be very poor for mid- and high-frequencies, especially at positions different to the ears, where artefacts in the form of comb filters can be perceived.

Another method used for reproducing sound under laboratory conditions is Higher-Order Ambisonics (HOA). The reference sound can be recorded using a spherical microphone array, which allows measurement of the sound field's spherical harmonic components. A regular loudspeaker setup, usually composed of a large number of loudspeakers, is used to reproduce the spherical harmonic components. The quality of the reproduction at the center of the loudspeaker array depends on the resolution of the microphone and loudspeaker arrays. Even though HOA is highly documented, and has received a lot of attention from the academic community, the required number of loudspeakers constrains implementation of standardize procedures in the telecommunications industry [2].

The recently standardized method for sound field reproduction, used in telecommunications devices testing [2], proposes a technique based on fast deconvolution using regularization as proposed by Kirkeby et al. [3]. The idea behind the method is to calculate a matrix of inverse filters that compensate for the loudspeakers' and room's response at particular positions in space defined by a custom microphone array. The positions of the microphones of this array are, in principle, not important for the calculation because the problem is modelled with electrical signals only [4]. The error, in terms of magnitude spectrum difference between reference and reproduced sounds, is negligible at target positions, i.e. positions where the sound is optimized, and starts increasing as the distance from them increases. High coherence between reference and reproduced sound can also be achieved up to a frequency limit determined by the distance between microphones.

For this investigation, the matrix inversion method was implemented. The effect of the regularization parameter, the number of loudspeakers used for the reproduction, and the robustness with respect to different program materials, on the reproduction quality was investigated in the current study. A technique to attenuate the audible effects of the non-causal part of the impulse response of the inverse filter is proposed to improve the subjective performance of the method.

2. THEORY

The matrix inversion method is based on recordings of the reference sound using a custom microphone array. Impulse response measurements are then performed in the reproduction room from every loudspeaker to every microphone of the microphone array. This matrix of impulse response functions can be inverted using regularization to obtain a set of inverse filters that compensate for the response of the room and the loudspeakers. The problem is illustrated in figure 1, as formulated by Kirkeby et. al [5].

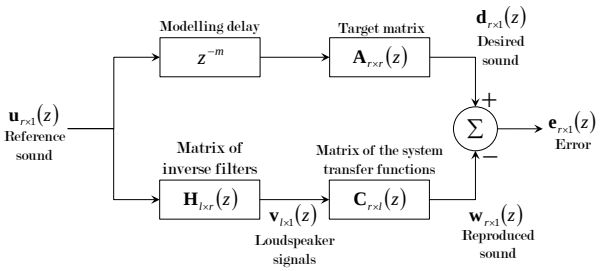

Figure 1: Block diagram form of the multichannel sound reproduction problem using matrix inversion.

The reproduced sound $\mathbf{w}$, is recorded by r microphones, and it is reproduced using l loudspeakers which playback the loudspeaker signals $\mathbf{v}$. The sound is naturally affected by the loudspeakers' and room's impulse responses to each microphone, $\mathbf{C}$. Written as a linear system in equation 1, the problem is as follows:

$$\begin{pmatrix} w_{1,1}(z) \\ w_{2,1}(z) \\ \vdots \\ w_{r,1}(z) \end{pmatrix} = \begin{pmatrix} C_{1,1}(z) & \cdots & C_{1,l}(z) \\ C_{2,1}(z) & \cdots & C_{2,l}(z) \\ \vdots & \ddots & \vdots \\ C_{r,1}(z) & \cdots & C_{r,l}(z) \end{pmatrix} \begin{pmatrix} v_{1,1}(z) \\ v_{2,1}(z) \\ \vdots \\ v_{l,1}(z) \end{pmatrix} \tag{1}$$

The target matrix **A** in the block diagram, is an identity matrix used to define that the desired reproduced sound **d** is recorded using the same microphone positions that are used to record the reference sound **u**. The modelling delay z^{-m} has to be introduced in the processing, to ensure that the inverse filters can be implemented given that they are non-causal. The value of m is half the length of the inverse filter.

From the block diagram, it can be seen that the loudspeakers' signals can be obtained from a set of recordings of the reference sound using the matrix of inverse filters. The final goal of the matrix inversion method is to derive the matrix **H**:

$$\begin{pmatrix} v_{1,1}(z) \\ v_{2,1}(z) \\ \vdots \\ v_{l,1}(z) \end{pmatrix} = \begin{pmatrix} H_{1,1}(z) & \cdots & H_{1,r}(z) \\ H_{2,1}(z) & \cdots & H_{2,r}(z) \\ \vdots & \ddots & \vdots \\ H_{l,1}(z) & \cdots & H_{l,r}(z) \end{pmatrix} \begin{pmatrix} u_{1,1}(z) \\ u_{2,1}(z) \\ \vdots \\ u_{r,1}(z) \end{pmatrix} \tag{2}$$

By requiring that the impulse responses of the inverse filters must be stable, but not necessarily causal (i.e., $|z| = |e^{j\omega\Delta}| = 1$, where ω is the angular frequency, and Δ is the sampling interval), a cost function J can be defined as the sum of the total squared error and the total effort [5]. The total square error represents how well the reference sound is reproduced at the microphone positions. The total effort, represents the energy of the loudspeaker signals, which needs to be enough to achieve low performance error, but limited to protect the loudspeakers from saturation.

$$J = \mathbf{e}(e^{j\omega\Delta})^H \mathbf{e}(e^{j\omega\Delta}) + \beta \mathbf{v}(e^{j\omega\Delta})^H \mathbf{v}(e^{j\omega\Delta}) \tag{3}$$

where $\mathbf{e} = \mathbf{d} - \mathbf{w}$, is the performance error; H is the Hermitian, or conjugate transpose; and β is the regularization parameter.

The solution can be controlled from minimizing only the error, to minimizing only the effort. This is done by changing the regularization parameter β from zero to infinity. For low regularization values, the reproduction is more precise, but the loudspeakers' signals will have higher energy, risking the performance of the loudspeakers and creation of distortion. For high values, the loudspeakers' signals will have less energy, but there can be more error due to the lower energy radiated. One effect of the regularization is that it makes the filter non-causal. A consequence of this is the is the generation of audible artefacts in the form of pre-echoes [6, 7].

For a regularization parameter larger than zero, the cost function is minimized in the least-squares sense by the following vector:

$$\mathbf{v}(e^{j\omega\Delta}) = \frac{\mathbf{C}(e^{j\omega\Delta})^H \mathbf{A}(e^{j\omega\Delta}) \mathbf{u}(e^{j\omega\Delta})}{\mathbf{C}^H(e^{j\omega\Delta})\mathbf{C}(e^{j\omega\Delta}) + \beta\mathbf{I}} \tag{4}$$

where **I** is the identity matrix. By comparing this to equation 2, the matrix of inverse filters, in the frequency domain, is the following:

$$\mathbf{H}(e^{j\omega\Delta}) = \frac{\mathbf{C}(e^{j\omega\Delta})^H \mathbf{A}(e^{j\omega\Delta})}{\mathbf{C}^H(e^{j\omega\Delta})\mathbf{C}(e^{j\omega\Delta}) + \beta\mathbf{I}} \tag{5}$$

3. ELECTRO-ACOUSTIC SYSTEM

3.1. Recording system

As described above, the matrix inversion method makes use of a custom microphone array, which is designed to mark positions at which the sound field is optimized. In figure 2, a sketch of the microphone array used in this study, is shown. It is composed of eight target (1-8) and four validation microphones (v1-v4). The positions of the target microphones have been selected to cover a region around the head where telecommunications devices are usually placed. The validation microphones are set in between some of the target microphones to evaluate the sound reproduction at positions where the reproduction of the sound field is not optimized. The validation microphone positions also represent the possible points in space where telecommunications devices' microphones can be located when mounted on the HATS.

3.2. Reproduction system

A photograph of the reproduction room is shown in figure 3. The loudspeaker array is composed of eight loudspeakers placed in a squared shape around the HATS. The maximum distance is 2m from the center of the square. The loudspeakers are active loudspeakers with concentric units, so the low- and mid-frequency drivers are located in the same axis, presenting a more symmetric radiation with respect to

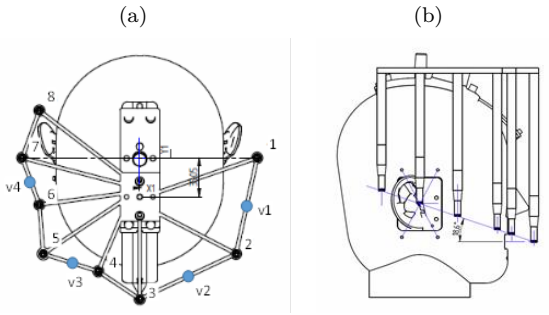

Figure 2: Custom microphone array. (a) Top view, (b) Right side view.

the center of the loudspeaker, which was adjusted to the same height as the Ear Reference Point (ERP).

Two additional sources (9,10) were placed in the room to generate the reference background noise. By having these sources, the reference background noise was always the same and comparison across experiments were possible.

The output gain of the loudspeakers of the array were adjusted by verifying that 90dB SPL were obtained with each loudspeaker at the middle of the loudspeaker array without the presence of the HATS. The reference sound sources, were adjusted to reproduce 80dB SPL, in this way the reproduction system had more dynamic range than the total sound pressure level to be reproduced.

4. OPTIMIZATION PROCEDURE

4.1. Reference sound scene recording

The first step for reproduction of sound using the matrix inversion method is to capture the reference sound to be reproduced. In our study, the reference sound was generated in the same room as the reproduction. Four program materials were selected to verify how robust the method is across different characteristics like the temporal evolution and spectral content of the signals. The program materials used were pink noise, pop music, classical music, and speech. The last three were used in both the objective and subjective evaluation.

4.2. System identification and inverse filters

The impulse response function, from every loudspeaker to every target microphone of the microphone array was measured using sine sweeps with

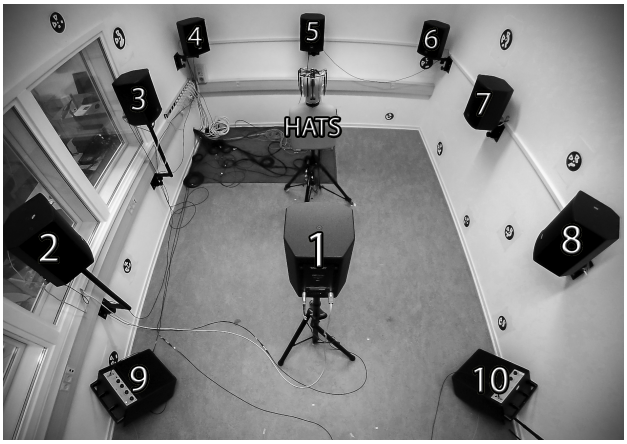

Figure 3: Photograph of the reproduction room, the loudspeaker array (1-8), the reference sound sources (9,10), and the HATS with the custom microphone array.

duration of approximately 3s. Compared to other signals used to measure impulse responses, sine sweeps have the advantage that they separate the non-linear distortion created by the system from the linear part of the impulse response [8]. By truncating at 1.5s, the non-linear components and noise in the impulse response can be reduced. It can be seen in figure 4(a), that a signal-to-noise ratio, i.e. peak-to-noise-floor ratio, of approximately 90dB was achieved.

Using equation 5, it is possible to calculate the matrix of inverse filters. Five different values for the regularization parameter were chosen after preliminary tests: $\beta = 10,\ 1,\ 0.32,\ 0.1,$ and 0.01. These values correspond to regularization thresholds $(20\log_{10}(\beta^{-1}))$ of -20dB, 0dB, 10dB, 20dB, and 40dB, respectively.

The effect of regularization on the inversion of the matrix of impulse response functions is illustrated in figures 4(b) and 5(b). The two inverse filters calculated correspond to regularization thresholds of 20dB and -20dB, respectively of a system with two loudspeakers and two microphones. First, it can be seen that regularization has the effect of making the inverse filter non-causal. Non-causal filters have a characteristic pre-response, which generates strong audible artefacts in the form of pre-echoes [7]. As

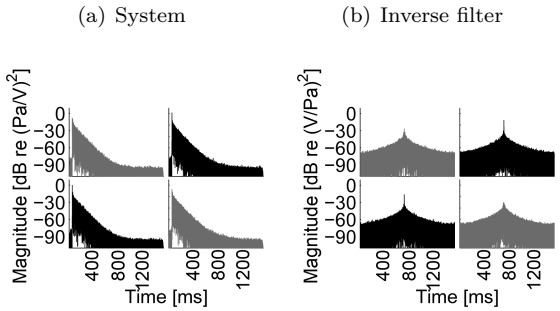

Figure 4: Example of a 2-by-2 system. (a) impulse response from two loudspeakers (columns) to two microphones (rows), (b) impulse response of the inverse filter with regularization threshold of 20dB.

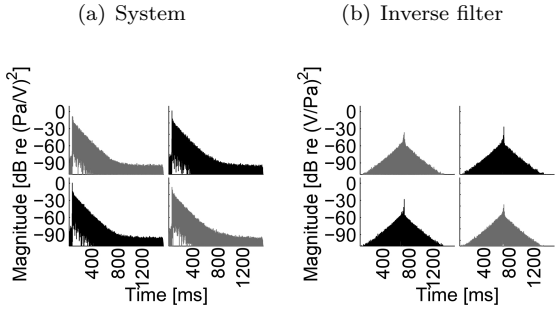

Figure 5: Example of a 2-by-2 system. (a) impulse response from two loudspeakers (columns) to two microphones (rows), (b) impulse response of the inverse filter with regularization threshold of −20dB.

mentioned before, a modelling delay has been introduced in order to be able to implement the filter. However, the audible artefacts caused by the non-causality will still be present in the reproduced sound. Second, the audibility of the artefacts caused by regularization depend on the slope of the rising part of the inverse impulse response function before the peak [9, 10]. It can be seen that this slope depends on the values selected.

In this study, a technique to attenuate the audibility of the artefacts created by the non-causal part of the inverse impulse response function is presented, inspired by the technique that Mukai et al. [11] used to eliminate pre-echo noise of separating filters in blind source separation. The technique consists of post-processing the impulse response of the inverse filter by applying a window to artificially attenuate the response before and after the peak. While the audible artefacts are created by the pre-response only, attenuation before and after the peak showed better results than attenuation only before.

Figure 6 shows the original and post-processed inverse impulse response functions for different values of the regularization threshold, together with a Gaussian window, which, after various preliminary tests, was selected to produce the best results in terms of sound quality and objective error.

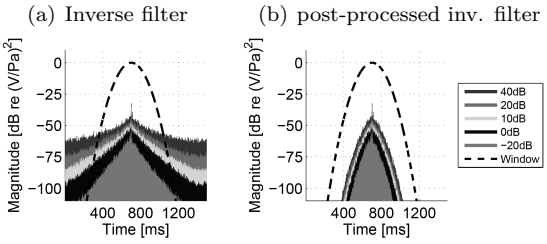

Figure 6: Post-processing Gaussian window applied to one channel of the inverse impulse response function calculated with different regularization thresholds.

5. OBJECTIVE EVALUATION

5.1. Regularization parameter

In the matrix inversion method, the regularization parameter is the most important variable for the quality of the reproduction. This parameter controls how precise the inversion of the matrix of impulse responses should be.

According to Kirkeby et. al [12], an appropriate procedure to select the regularization parameter is to choose a (low) number and lower it more while checking that the loudspeakers signals do not saturate the reproduction system. As it was said above, five regularization thresholds (−20dB, 0dB, 10dB, 20dB, and 40dB) were selected after preliminary tests. None of them led to loudspeaker signals that risked saturation of the system.

Figure 7, shows the maximum error (magnitude spectrum difference) and average coherence between reproduced and reference sounds, calculated across

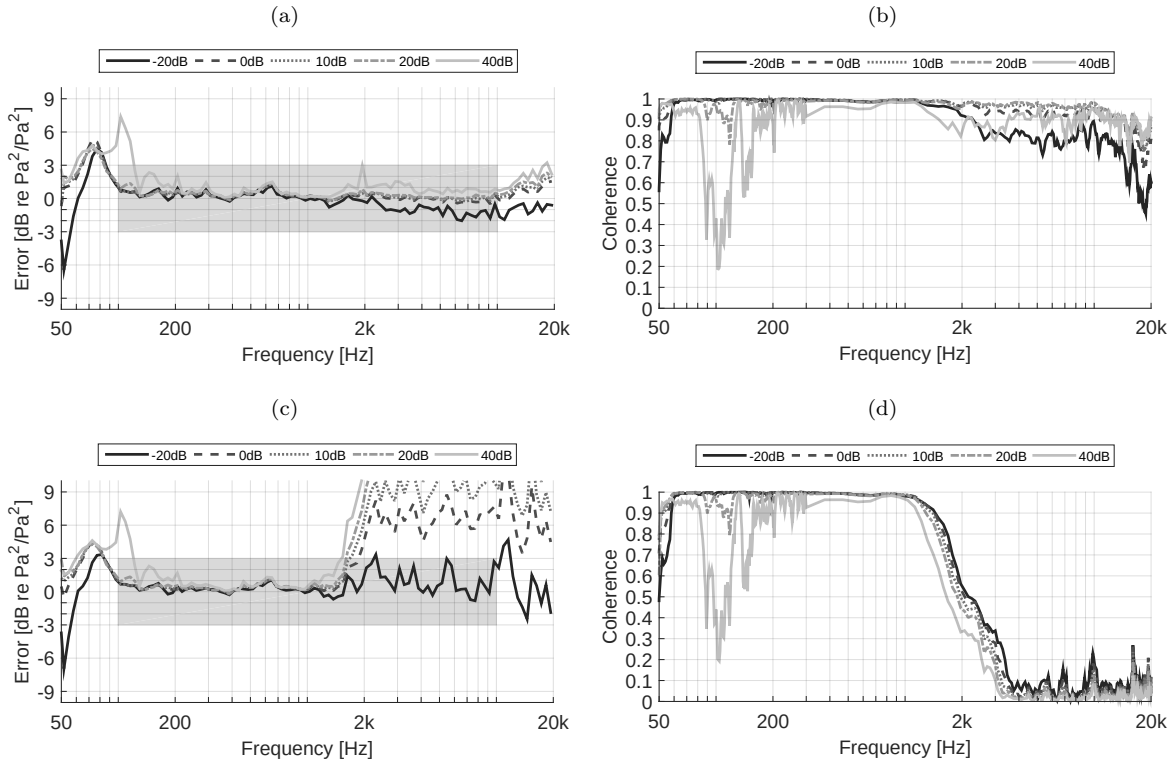

Figure 7: Maximum error and average coherence across target ((a) and (b)) and validation microphones ((c) and (d)), for different regularization thresholds when reproducing pink noise using 8 loudspeakers.

target and validation microphones, for the different regularization thresholds selected. The reproduction was done using 8 loudspeakers and pink noise was selected as the test signal. In figures 7(a) and 7(b), the maximum error and average coherence across the target microphones are shown. Figures 7(c) and 7(d) show the same results for the validation positions.

Comparison between the performance at target and validation microphones, suggests that the regularization threshold of -20dB ($\beta = 10$) is the one that ensures lower errors and better coherence in all microphone positions, even though, for some frequencies, the error at the validation positions exceeds the $\pm$3dB range, considered as the evaluation criteria. This result contradicts the previously mentioned procedure to select the regularization parameter suggested by Kirkeby et. al.

The matrix inversion method does the optimization of the sound field at particular positions defined by the target microphones of the array. The validation microphones are not part of the optimization procedure, therefore higher errors are obtained at these positions as the regularization changes. These errors suggest strong audible artefacts at these positions, which need to be controlled, because these positions represent places where telecommunications devices would be placed in actual terminal testing.

5.2. Number of channels

As a comparison with another standardized method for reproduction of sound [1], which uses four loudspeakers for the reproduction, optimization of the sound field using the matrix inversion method was performed using different number of loudspeakers.

Figure 8 shows the error and coherence at the validation positions for optimization done with three systems composed of 8, 4, and 2 channels. The number of channels make reference to the number of loudspeakers and target microphones used in the

Table 1: Loudspeaker and microphone configuration for different number of channels.

	8CHS	4CHS	2CHS
Loudspeakers	1 - 8	2, 4, 6, 8	2, 8
Target mics.	1 - 8	1, 2, 5, 7	1, 7
Validation mics.	v1 - v4	v1 - v4	v1, v4

optimization. Table 1 summarizes the specific loudspeakers and microphones used in each case.

As can be seen, the reproduction using 8 channels, with respect to 4, does not produce any improvements in terms of magnitude spectrum difference, but the coherence improves at higher frequencies. When two channels were used, due to the reduction of spatial resolution, more errors were obtained and the coherence was significantly reduced at positions not taken into account during the matrix inversion. Even though it is not shown here, at target positions, the performance was the same with the three channels configurations.

5.3. Program materials

Different program materials were selected to cover different temporal and frequency characteristics, as well as different styles of music. Pink noise was tested as it is a traditional broadband test signal in acoustics. Pop music, was tested because it is more dynamic and percussive than classical music, which is softer and continuous. And speech was tested because the final application of this study is the telecommunications industry.

Figure 9, shows that the maximum error and the average coherence across microphones do not change for the selected program materials.

6. SUBJECTIVE EVALUATION

One of the challenges faced in sound reproduction problems are the audible artefacts created by the reproduction method. Depending on the method, these artefacts have different characteristics. The matrix inversion method is prone to artefacts in form of pre-echoes, which appear due to the non-causality created by the matrix inversion of a system that is not well-conditioned.

A listening test was designed in this study to evaluate the quality of the reproduction using the independent variables previously analysed objectively, the regularization parameter, the number of channels, and the program materials. It was based on the ITU-R BS.1534-1 recommendarion [13], which describes a "MUlti Stimulus test with Hidden Reference and Anchor (MUSHRA)". The procedure used in this study cannot be considered a complete MUSHRA because the anchor was not used. However, it allows comparison of different settings simultaneously, in this case the optimization of the sound field with different regularization thresholds, for a given pair of number of channels and program materials. The user interface of the test is shown in figure 10.

The listening test was presented to five normal hearing subjects, with ages between 25 to 37, who were instructed to rate the similarity between a set of sound samples (the reproduced sounds) with respect to a reference sample (the reference sound), according to a scale from 0 to 100 divided and marked with the tags "very different", "different", "similar" and "very similar". The sound samples were presented binaurally using headphones and they were the recordings from the microphones v1 and v4, which were the validation microphones closest to the ears.

In addition to the five regularization thresholds, a subset of reproduced sounds optimized with the post-processing technique to attenuate audible artefacts using the Gaussian window were included. The

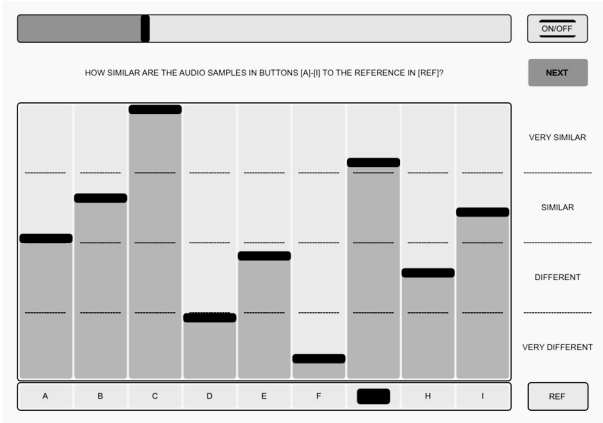

Figure 10: MUSHRA-like test interface.

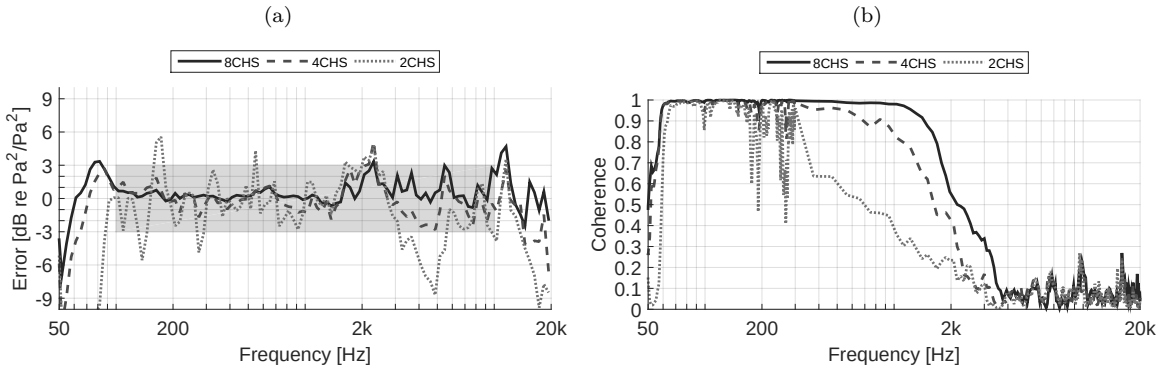

Figure 8: Maximum error and average coherence across validation microphones for optimization of the sound field using different number of loudspeakers. Regularization threshold of −20dB, reproduction of pink noise.

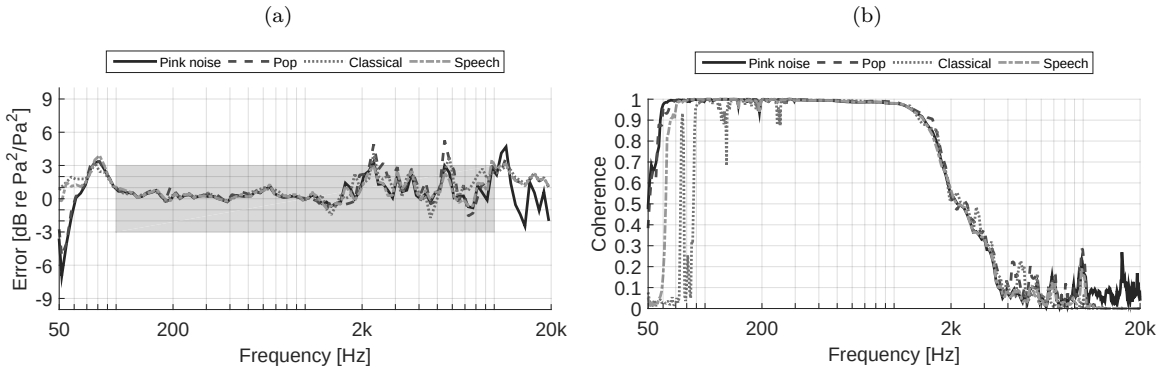

Figure 9: Maximum error and average coherence across validation microphones for optimization of different program materials with 8 loudspeakers. Regularization threshold of −20dB.

highest three regularization thresholds were selected for this evaluation. A total of eight audio samples were presented per condition together with the hidden reference, which was a copy of the sound the subjects were comparing with.

Figure 11 shows the average ratings across subjects, together with error bars representing the 95% confidence interval. Each of the subplots is divided in three groups. The first group shows that all the subjects were able to easily identify the hidden reference among the reproduced sound samples. The second one shows the ratings for the reproduced sound samples processed with the five regularization thresholds without post-processing. The third region shows the ratings for sounds using post-processing of the impulse response of the inverse filters.

The ratings show that the regularization threshold that gave rise to a more similar reproduced sound with respect to the reference is −20dB. For all program materials and all channel configurations, the similarity decreases as the regularization threshold increases (regularization parameter β decreases). This agrees with the results presented in figure 7(c).

In all the conditions, the subjects perceived the samples with the Gaussian window more similar to the reference than without it.

The z-scores of the ratings presented in figure 11 were calculated (not shown here) and they indicated

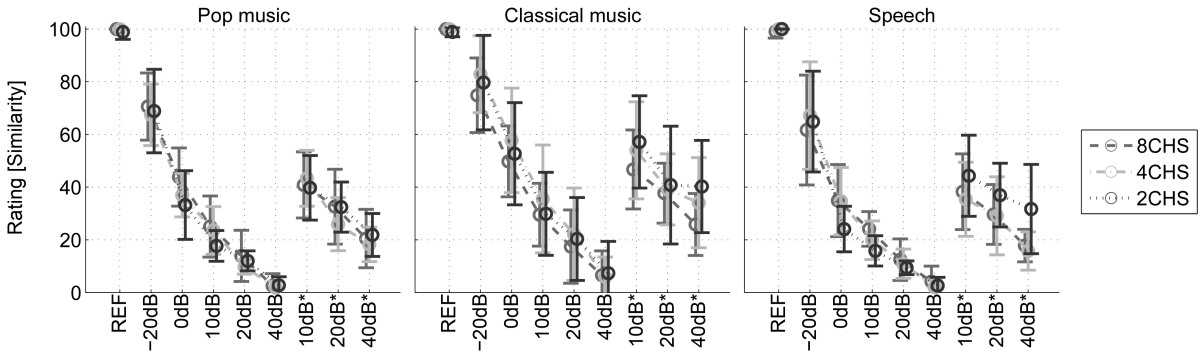

Figure 11: Average ratings across subjects, the error bars represent 95% confidence interval.

that the difference in ratings of the similarity across different program materials and channel configurations were negligible.

7. CONCLUSIONS

The matrix inversion method for reproduction of sound was implemented within the application of testing telecommunications devices. The impact of the regularization parameter, different number of channels used for the optimization and the robustness across different program materials were studied. A technique to improve the subjective performance of the method was proposed and evaluated.

From the selected regularization thresholds, the one that produced the sound the most similar to the reference was the lowest (highest regularization parameter). This is in opposition to previous recommendations found in literature. The objective errors increased and the subjective similarity decreased as the regularization threshold increased.

In terms of the objective measures used in this study, the differences between optimization using 4 or 8 loudspeakers were small in terms of coherence. In comparison, the reproduction using two channels presented more errors and lower coherence due to the reduction of spatial resolution, or increase of the distance between microphones. The subjective evaluation showed no significant effect on the number of channels.

For different program materials, there were no differences in terms of error and coherence, and the similarity ratings of the subjects agreed with this observation.

The main difficulty with the matrix inversion method, the creation of audible artefacts in the form of pre-echoes, was addressed with a technique to reduce the audible artefacts of the non-causal part of the impulse response of the inverse filter. This technique showed an increase in the similarity ratings by approximately 10% compared to unprocessed samples.

One advantage of the matrix inversion method with respect to the procedure described by the ETSI EG 202 396-1, and Higher-Order Ambisonics is that it takes into account the crosstalk cancellation. However, the reproduction of sound is only at local positions defined by the microphone array. The errors obtained at these positions was found to be within ±3dB in the frequency range from 100Hz to 20kHz, and the coherence higher than 0.7 up to 16kHz. The same performance was maintained at validation positions up to 1.5kHz.

8. REFERENCES

[1] European Telecommunication Standards Institute. Speech processing, transmission and quality aspects - Part 1: Background noise simulation technique and background noise database. *ETSI EG 202 396-1*, 4:1–58, 2011.

[2] European Telecommunication Standards Institute. A sound field reproduction method for terminal testing including background noise database. *ETSI TS 103 224*, 1:1–36, 2014.

[3] Ole Kirkeby, Philip A. Nelson, Felipe Orduna-Bustamante, and Hareo Hamada. Fast deconvolution of multichannel systems using regularization. *IEEE Transactions on Speech and Audio Processing*, 6(2):189–194, March 1998.

[4] Philip A. Nelson and Stephen J. Elliott. *Active control of sound*. Academic Press, 1993.

[5] Ole Kirkeby and Philip A. Nelson. Local sound field reproduction using digital signal processing. *Journal of the Acoustical Society of America*, 100(February):1584–1593, 1996.

[6] Hironori Tokuno, Ole Kirkeby, Philip A. Nelson, and Hareo Hamada. Inverse filter of sound reproduction systems using regularization. *IEICE Trans. Fundamentals*, E80-A(5):809–820, 1997.

[7] Scott G. Norcross and Martin Bouchard. Multichannel Inverse Filtering with Minimal-Phase Regularization. *Audio Engineering Society Convention 123*, pages 1–8, 2007.

[8] Swen Müller and Paulo Massarani. Transfer function measurement with sweeps. *Journal of the Audio Engineering Society*, pages 443–471, 2001.

[9] Scott G. Norcross, Gilbert A. Soulodre, and Michel C. Lavoie. Subjective effects of regularization on inverse filtering. *Audio Engineering Society Convention 114*, 2003.

[10] Scott G. Norcross, Gilbert A. Soulodre, and Michel C. Lavoie. Subjective investigations of inverse filtering. *Journal of the Audio Engineering Society*, 52(10):1003–1028, 2004.

[11] Ryo Mukai, Shoko Araki, Hiroshi Sawada, and Shoji Makino. Evaluation of separation and dereverberation performance in frequency domain blind source separation. *Acoustical Science and Technology*, 25(2):119–126, 2004.

[12] Ole Kirkeby, Philip A. Nelson, Per Rubak, and Angelo Farina. Design of cross-talk cancellation networks by using fast deconvolution. *Audio Engineering Society Convention 106*, 1999.

[13] International Telecomunication Union. Recommendation ITU-R BS.1534-1: Method for the subjective assessment of intermediate quality level of coding systems. 2003.