Electroacoustic method for the calibration of a heterogeneous distributed speaker system

Thomas Joubaud
Orange Labs
Cesson-Sévigné, France
thomas.joubaud@orange.com

Grégory Pallone
Orange Labs
Cesson-Sévigné, France
gregory.pallone@orange.com

***Abstract*—We present an electroacoustic method to calibrate a heterogeneous distributed speaker system such as e.g. various Bluetooth speakers interconnected in a client-server architecture. This method allows to extract parameters necessary for an appropriate *clock coordination* (synchronization/syntonization), *equalization* and *spatial configuration* in a single calibration operation. This approach enables immersive audio or multiroom experiences without having to buy a closed dedicated system.**

***Index Terms*—Calibration, distributed speakers, synchronization, equalization, cartography**

I. Introduction

We define a distributed speaker system as a centralized network of speakers able to create a coherent listening system. This system, composed by a single server and several clients, allows use-cases such as multichannel (spatial sound scene rendering), multiroom (listening to music in the whole house), or new ways to experience audio such as notions of "Media Device Orchestration" [1], or "Media Things" [2]. Such systems have already been proposed successfully by manufacturers, as closed ecosystems. The aim of this work is to focus on heterogeneous distributed speaker systems, composed by wired/wireless speakers of different brands and models, and possibly unknown characteristics. Typically, we want to transform different Bluetooth speakers into a homogeneous multi-device system.

A. Distributed Audio problematic

Three factors make the system heterogeneous, that should be corrected into a homogenous one: *clock coordination*, *equalization* and *spatial configuration*.

A clock coordination process must be applied for the speakers to be both synchronized (they start playing at the same time) and syntonized (they play at the same rate). A synchronization fault, identified as an absolute difference between clocks, may lead to audible artifacts such as comb filtering, spatial image shifts or echoes. A syntonization fault, identified as a non-unitary ratio between frequency clocks, will lead at a given moment to synchronization fault, inducing same artifacts but also dynamic variation of them. Moreover, after a given time proportional to the frequency ratio of the clocks, syntonization fault will inevitably lead to buffer over- or under-runs causing audible glitches. A network synchronization protocol can be used to achieve clock coordination between server and clients, such as the Network Time Protocol [3] or the Precision Time Protocol [4]. Other protocols are better suited for wireless devices according to [5]. However, interconnection between these protocols is not always possible, and speakers that do not use such protocols have to be coordinated differently.

The fact that each speaker is different in terms of sound rendering constitutes the second factor of heterogeneity in a distributed audio system. Because the speakers are set up independently, they do not play at the same volume. Moreover, frequency responses of speakers are different so they have to be compensated through equalization.

The third factor of heterogeneity lies in the spatial configuration of the speakers. In the case of multichannel rendering, we can reasonably assume that the speakers are not ideally positioned according to the ITU standards for multichannel setups [6] [7]. Applying a spatial correction can adapt the listening sweet spot to the listener's position in order to improve the rendering [8]. However, such a correction requires at least the distance between the listener and each speaker, if not their exact position in a coordinate system.

B. Contribution and outline

We propose in this paper a calibration method to homogenize a distributed audio system in the sense of clock coordination, equalization and spatial configuration thanks to a single procedure of electroacoustic measurement performed with a calibration device (e.g. a smartphone). The typical setup is based on existing devices considered as black boxes, such as deployed Bluetooth speakers using the mandatory SBC codec [9]. The calibration process aims at estimating the appropriate parameters to synchronize/syntonize, adjust volume/equalization, and adapt the sweet spot. The following section describes these parameters and underlying models, next section discloses the electroacoustic calibration method, and the final section presents results of an experiment before concluding.

II. Models and parameters

The electroacoustic calibration method has to estimate the clock coordination parameters, needed to synchronize and syntonize all the speakers. These parameters are introduced within a clock model in the following subsection. The calibration

method provides the speakers' impulse response needed for the equalization: it strongly depends on the calibration signal chosen in the second subsection describing the measurement. This method also produces the data required by a cartography algorithm described in the third subsection and necessary for the spatial correction.

A. Clock model

Two independently working speakers possess their own clock. We model the computer's clock as a monotonic time function increasing at the rate defined by the crystal oscillator. The clocks of two speakers are usually different and we introduce the following two parameters to quantify these differences [10]:

- the clock *offset*: time difference at the origin between two clocks (i.e. latency);
- the clock *skew*: frequency ratio between two clocks, or first derivative of the clock with respect to time;

Therefore, in a server/client architecture, the client's clock T_c can be expressed as a function of the server's clock T_s following:

$$T_c = \alpha \left(T_s + \theta \right) , \tag{1}$$

with θ the client's offset and α the client's skew with respect to the server. The server imposes both the packet sending cadency and the instants at which the packets should be rendered. The offset is a time in seconds. The skew is a dimensionless quantity equal to the ratio between server and client's clock frequencies $\frac{f_s}{f_c}$. It is generally given in parts-per-million (ppm) by computing $10^6 \left(\frac{f_s}{f_c} - 1 \right)$. By applying on the audio signal an adequate buffering related to θ and a sampling rate conversion related to α, it is possible to compensate for the appropriate delays and skews [11].

B. Calibration signal

Usual network synchronization protocols are based on timestamps exchange between devices [3][4]. For some devices such as proprietary closed systems or Bluetooth speakers, it is not possible to directly use such protocols to coordinate clocks. We investigated another solution based on accessible information such as the one carried by the audio signal: an acoustic equivalent of a network timestamp is an impulsive event. Our solution is to play and record a signal similar to a Dirac distribution to mimic the timestamps approach. The time at which the measured signal reaches its maximum can therefore be assimilated to a timestamp. Using this method, the speaker's impulse response is also evaluated, allowing the estimation of volume and frequency response differences between the speakers. Instead of using a signal approximating a Dirac distribution which provides a poor signal-to-noise ratio (and probably leads to more degradation by the SBC codec), methods based on cross-correlations are preferred, using pseudo-random maximum-length sequences or exponential sine sweeps (ESS) [12].

C. Acoustic cartography and spatial correction

In the context of multichannel audio rendering, even if the speakers are syntonized and synchronized, spatial correction is often necessary to compensate for speaker misplacement. A standard method for spatial correction consists in applying a delay and a gain to the speakers in order to virtually place them on a circle with the center being the listening sweet spot. This technique is usually based on an electroacoustic measurement between every speaker and a calibration microphone placed at the sweet spot of a surround setup. Repositioning the speakers on a circle may not be perceptively sufficient if their angular position is significantly different from the ITU standard [6]. Angular correction methods exist, that try to compensate for the misplacement of the speakers: The MPEG-H 3D Audio codec [13] implements this kind of functionality which has been evaluated in [14].

In order to adapt the gain/delay or angular correction to the particular position of a listener, the positions of the speakers and the listener must be known. The listener's localization could be achieved thanks to different means such as video cues, beacons, audio from microphones or more simply but less precisely by making an assumption on his position. This electroacoustic calibration method aims at computing the positions of the speakers. [15] describes several approaches to perform this cartography, among which the proposed method focuses on pairwise distances. The electroacoustic measurement therefore provides the distances between each pair of speakers. The classical multidimensional scaling algorithm (MDS) is then used to estimate the speakers' position [16]. However, this approach needs the estimated distance matrix to be a Euclidean distance matrix. It appears that this condition is not always fulfilled in the experiments due to errors in distance estimation. In this case, a gradient descent approach called Alternate Coordinate Descent is employed to estimate the speakers' position [17]. This approach is possible in our calibration method since we assume that sensors and sources are co-located during the calibration process, which is not the case in [18].

III. Electroacoustic calibration method

A. Apparatus

We rely on a client-server architecture. The electroacoustic calibration method is based on a recording calibration device equipped with a microphone. We make the assumption that the calibration device's skew is known, e.g. computed using the soundcard's buffer fill level, such as the method described in [11].

B. Calibration protocol

At the beginning of the electroacoustic calibration protocol, the calibration device starts recording. The following steps are then repeated after moving the calibration device in front of each of the N speakers to be calibrated:

1) The calibration device is placed in front of speaker n.
2) The speaker n plays the calibration signal.

3) Every other speaker in the system plays the calibration signal successively.
4) The speaker n plays the signal for the second time.

The recording is finally stopped, leading to a single channel recording of the calibration microphone containing all information needed for parameter analysis.

C. Analysis

1) Acoustic timestamps and impulse response retrieval: The calibration protocol produces a signal containing a single channel with $N\,(N+1)$ measured ESS, ranging from 20 Hz to 20 kHz in 0.2 second. The latter are transformed into impulse responses using windowed analysis and cross-correlation with the original calibration signal [19]. This operation leads to a series of pulses, which can be oversampled for an increased time precision. The acoustic timestamps correspond to prominent peaks in the cross-correlation signal. Their estimation is therefore a multiple peak detection problem [20]. Every local maximum is firstly found as the transition from a positive to a negative slope. The $N\,(N+1)$ greater peaks are kept, with a constraint of minimal duration between two peaks to discard secondary peaks caused by distortions or reflections. The obtained acoustic timestamps are then used for the estimation of the clock coordination parameters and of the distances between the speakers. Any timestamp T' estimated by this method is equal to:

$$T' = \alpha_c\left(\alpha\left(T+\theta\right)+\theta_c\right), \tag{2}$$

with θ and α the clock coordination parameters of the speaker (offset and skew), T the time at which the signal is sent to the speaker, and θ_c and α_c the offset and skew of the calibration device, respectively. If the calibration device is syntonized with the server (possible on a smartphone thanks to an adequate network synchronization protocol), α_c is known, but synchronization is not necessary since θ_c will not be used in the following.

2) Parameters estimation: The parameters related to the speakers are derived from the signal measured by the calibration microphone. The skew α_i of the ith speaker is estimated from the two timestamps T'_0 and T'_1 produced by the steps 2 and 4 of the calibration protocol:

$$\begin{cases} T'_0 &= \alpha_c\left(\alpha_i\left(T_0+\theta_i\right)+\theta_c\right) \\ T'_1 &= \alpha_c\left(\alpha_i\left(T_1+\theta_i\right)+\theta_c\right) \end{cases}. \tag{3}$$

Therefore, the skew is estimated as:

$$\widetilde{\alpha_i} = \frac{T'_1 - T'_0}{\alpha_c\left(T_1 - T_0\right)}. \tag{4}$$

The symbol ˜ denotes estimated values. The difference $T_1 - T_0$ is a configurable duration between two emitted ESS. Increasing this difference leads to a more reliable estimation of $\widetilde{\alpha_i}$ if α is approximately constant over time. The absolute time at which the calibration signal is sent to a speaker is not necessarily known, but the duration between two signals is. For this reason, the absolute latencies of the speakers are not evaluated, but the relative latencies can be estimated quite precisely. The step 2 of the protocol performed for the ith and jth speakers produces two timestamps T'_i and T'_j:

$$\begin{cases} T'_i &= \alpha_c\left(\alpha_i\left(T_i+\theta_i\right)+\theta_c\right) \\ T'_j &= \alpha_c\left(\alpha_j\left(T_j+\theta_j\right)+\theta_c\right) \end{cases}. \tag{5}$$

The offset $\Theta_{ij} = \theta_i - \theta_j$ between the speakers is finally estimated by:

$$\widetilde{\Theta_{ij}} = \frac{T'_i}{\alpha_c\widetilde{\alpha_i}} - \frac{T'_j}{\alpha_c\widetilde{\alpha_j}} - \left(T_i - T_j\right). \tag{6}$$

The estimation bias is equal to $\theta_c\left(\frac{1}{\alpha_i} - \frac{1}{\alpha_j}\right)$ and neglected since the skew difference is usually less than 100 ppm, i.e. 0.01 %. Since sweep measurements are realized twice (step 4 of the protocol) for skew estimation, offset values can be computed twice and used to check the robustness of the estimation.

The relative gains between speakers are estimated from the energy of the impulse responses, but a filtered approach could lead to more robust results depending on the speakers' bandwidth. The relative spectra between speakers can be estimated from the associated frequency responses.

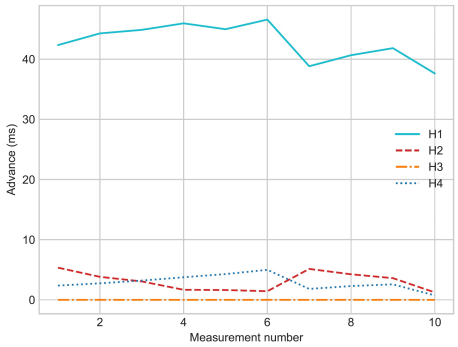
The calibration method also computes the distance between each couple of speakers from the acoustic propagation duration, knowing the speed of sound. In step 3 of the protocol, when the calibration microphone is in front of the jth speaker, the produced timestamps contain the acoustic propagation duration between this speaker and the other ones. For example, the timestamp T'_i is obtained from the ith speaker such that $T'_i = \alpha_c\left(\alpha_i\left(T_i+\theta_i\right)+\lambda_{ij}+\theta_c\right)$, with λ_{ij} the propagation delay between speaker i and j. Exploiting the timestamp T'_j from step 2, λ_{ij} is estimated following the equation:

$$\widetilde{\lambda_{ij}} = \widetilde{\alpha_i}\left(\frac{T'_i}{\alpha_c\widetilde{\alpha_i}} - \frac{T'_j}{\alpha_c\widetilde{\alpha_j}} - \left(T_i - T_j\right) - \widetilde{\Theta_{ij}}\right). \tag{7}$$

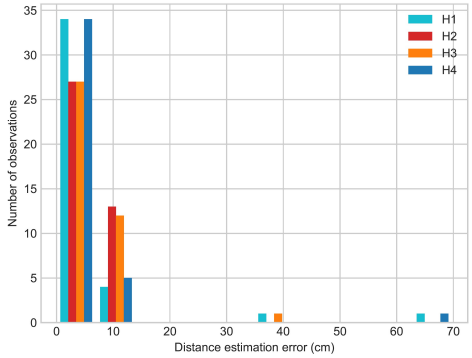
Since propagation delays are measured twice (cf. step 3), pairwise distances can be computed twice and used to check the robustness of the estimation.

IV. Experiment and results

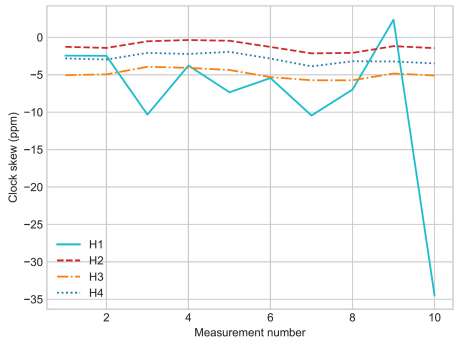
The electroacoustic calibration method described in the previous section is implemented and evaluated in this section. This experiment is conducted in an acoustically treated room (studio). We use our own software called *Soundcast*, based on a server running on a Linux computer, and four clients running on two different Raspberry Pi on the same wired local network. Each client drives a Bluetooth speaker denoted from $H1$ to $H4$. Except for $H3$ and $H4$, the speakers are from different brands and models. The calibration client is launched on the same computer as the server and drives a low-cost USB soundcard connected to a preamplifier linked to a microphone. The estimation of the speakers' offsets, skews, and pairwise distances through the calibration method is performed ten times in order to assess the repeatability and the precision of the method.



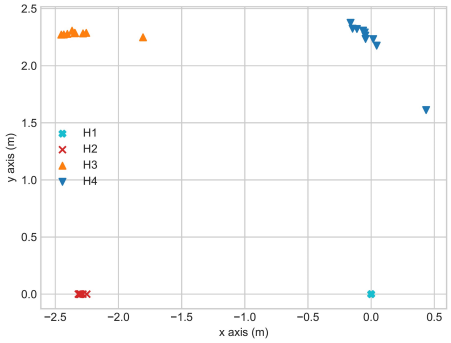
(a) Estimated offset advance in ms



(b) Histogram of the pairwise distances estimation error in cm



(c) Estimated skew in ppm



(d) Estimated position of speakers on a map in m

Fig. 1: Estimation of parameters (offset, skew, position, distances) of the four Bluetooth speakers for the ten measurements.

A. *Clock coordination estimation*

The relative latencies between the speakers are estimated using the most delayed one as reference. The values illustrated in Figure 1a therefore represent the advance (delays to apply in milliseconds) of the speakers compared to the reference ($H3$). $H1$ has a mean advance of 42.8 ms ($\pm$2.9 ms) which is significantly greater than the mean advance of $H2$ and $H4$, 3.1 ms ($\pm$1.5 ms) and 2.9 ms ($\pm$1.2 ms) respectively. The greatest variability is observed with $H1$. The evolution of the values across the ten measurements is similar between $H1$ and $H4$, while it seems to be reversed for $H2$ (we cannot explain this observation yet). Although these offset deviations values are satisfactory in multiroom applications, they may have to be lowered below 1 ms for worst-case multichannel scenarii [21]. Moreover, it should be checked if those values are still valid after a reboot of the speakers.

Figure 1c shows the evolution across the ten measurements of the skew of the four Bluetooth speakers relative to the *Soundcast* server's clock. We observe that the estimation is very stable across measurements for $H2$, $H3$ and $H4$, with a mean value of -1.2, -4.9 and -2.9 ppm, respectively. This stability is confirmed by calculated standard deviations of approximately 0.6 ppm. Furthermore, it can be noticed that the shapes of the curve for those three speakers are very similar. We assume these slight changes are due to slight variations in the skew estimation of the calibration device, which influence is given in equation (4). Finally, the skew estimation for speaker $H1$ appears to be much less stable, with a mean value of -8.1 ppm and a standard deviation of 9.5 ppm. Speaker $H1$ is a device integrating a dynamic phase-locked loop (PLL) mechanism which adapts its sample rate to the data packets arrivals. PLL aims at removing audible clicks caused by buffer over- or underruns. In other words, $H1$ tries to rectify its skew by itself. The last estimated value of nearly -35 ppm, which seems to be an outlier, could be the result of our skew estimation during a strong adaptation of the dynamic PLL.

B. *Equalization estimation*

The relative gains between speakers are extracted from the energy of the impulse responses. No formal evaluation of the estimation has been performed in this work, but this is envisioned in a future work, along with frequency equalization.

C. *Position configuration estimation*

Reference distances between each pair of speakers is measured using a measuring tape and compared with the distances estimated through calibration. Figure 1b shows a histogram of the distance error in centimeters for each speaker, e.g. the

values for $H2$ represent the estimation errors for the distance between $H2$ and all of the other speakers. We observe that all errors are less than or equal to approximately 10 cm, except for one pairwise distance error of almost 40 cm between $H1$ and $H3$ and one of almost 70 cm between $H1$ and $H4$. These high pairwise distance errors belong to the same last measurement for which we also identified an outlier in the estimation of the skew for $H1$. We can conclude that the distance estimation using electroacoustic calibration seems accurate in most cases (the displacement of a speaker by 10 cm often lead to no audible difference).

Figure 1d illustrates the estimated positions of the four speakers on the horizontal plane. The vertical axis is not taken into account since the speakers were approximately placed on a plane and the estimated z-component of the cartography is consequently negligible. As the cartography algorithm uses the pairwise distance matrix, the coordinates are estimated in an arbitrary coordinate system. In the figure, the coordinate system is therefore transformed using Procrustes alignment [22] for each measurement such that $H1$ is placed at the origin, $H2$ is on the x-axis and $H3$ on the horizontal plane. A distinct pair of outliers appears in the point sets of $H3$ and $H4$, and it corresponds to the last measurement. These two outliers are responsible for the noticeable pairwise distance errors in Figure 1b. Discarding these outliers, a metric reflecting the maximum standard deviation of the estimated positions of each speaker is computed as the square root of the maximum eigenvalue of the covariance matrix. This metric is equal to 2 cm for $H2$, 6.5 cm for $H3$ and 9 cm for $H4$. As suggested by the figure, the cartography stands accurate across the ten calibrations, as the deviation values are smaller than the dimensions of the Bluetooth speakers (respectively 18, 21, 11 and 11cm for $H1$ to $H4$).

Preliminary work with simulated data showed that the cartography algorithm works very well as long as the pairwise distance matrix is correct. Cartography errors therefore mainly come from errors in the estimation of the distance between the speakers.

V. Conclusion and future work

We created a heterogeneous distributed speaker system with Bluetooth speakers from different brands and models. Thanks to a novel electroacoustic method, we calibrated the system through a procedure allowing to estimate the offset, skew, gain and position of each speaker in a single operation. We provide in this work objective results as a proof of concept, but subjective evaluation and comparison with other methods are envisioned in the future. Our experiment shows a limit with a speaker which has been identified as implementing a dynamic PLL. We plan to investigate and conclude if this behavior comes intrinsically from a dynamic PLL algorithm, or from a faulty implementation of the PLL leading to erratic skew adaptation. Apart from this device, our approach seems promising since offset and skew estimations are stable across several measurements. The results show an accuracy compatible with multiroom applications but we plan to check if it is enough for multichannel applications where timing precision is crucial. We extracted the relative gains between speakers, but we plan to work on frequency response inversion to give the system a better timbral homogeneity. Concerning spatial configuration, obtained precision is of the order of the speakers' dimensions, which is often satisfactory, especially for speakers with several drivers (e.g. stereo) for which it is difficult to associate a single precise position. Potential applications not only cover the domain of audio-visual content reproduction, but also vocal assistants and interpersonal telecommunication.

References

[1] Francombe, Woodcock, Hughes, Mason, Franck, Pike, Brookes, Davies, Jackson, Cox, Fazi, and Hilton, "Qualitative evaluation of media device orchestration for immersive spatial audio reproduction," *J. Audio Eng. Soc*, vol. 66, no. 6, pp. 414–429, 2018.

[2] MPEG, "IoMT White Paper," 2019, N18879, ISO/IEC JTC 1/SC 29 Coding of audio, picture, multimedia and hypermedia information.

[3] David L. Mills, *Computer Network Time Synchronization - The Network Time Protocol*, CRC Press, 2006.

[4] IEEE, "IEEE standard for a precision clock synchronization protocol for networked measurement and control systems," 2008, IEEE 1588-2008, IEEE Instrumentation and Measurement Society.

[5] Bharath Sundararaman, Ugo Buy, and A. D. Kshemkalyani, "Clock synchronization for wireless sensor networks: a survey," *Ad Hoc Networks*, vol. 3, no. 3, pp. 281–323, may 2005.

[6] ITU, "Multichannel stereophonic sound system with and without accompanying picture," 2012, ITU-R BS.775-3.

[7] ITU, "Multichannel sound technology in home and broadcasting applications," 2019, ITU-R BS.2159-8.

[8] S. Merchel and S. Groth, "Analysis and implementation of a stereophonic play back system for adjusting the "sweet spot" to the listener's position," in *Proceedings of 126th AES Convention*, 2009, p. 809–817.

[9] Bluetooth SIG, "Specification of the bluetooth system, profiles, advanced audio distribution profile version 1.3," 2012.

[10] Y. Wu, Q. Chaudhari, and E. Serpedin, "Clock synchronization of wireless sensor networks," *IEEE Signal Processing Magazine*, vol. 28, no. 1, pp. 124–138, Jan 2011.

[11] Stefan Werner, "An algorithm for audio skew compensation in low latency environments," in *ICMC*, 2005.

[12] Guy-Bart Stan, Jean Jacques Embrechts, and Dominique Archambeau, "Comparison of different impulse response measurement techniques," *Journal of the Audio Engineering Society*, vol. 50, pp. 249–262, 2002.

[13] ISO/IEC 23008-3, "Information technology – High efficiency coding and media delivery in heterogeneous environments – Part 3: 3D Audio," 2015, ISO/IEC JTC 1/SC 29.

[14] Moulin et al., "Perceptual evaluation of loudspeaker misplacement compensation in a multichannel setup using MPEG-H 3D Audio renderer. Application to Channel-Based, Scene-Based, and Object-Based audio materials," in *International Conference on Immersive and Interactive Audio, York, UK*, 2019.

[15] Plinge, Jacob, Haeb-Umbach, and Fink, "Acoustic microphone geometry calibration: An overview and experimental evaluation of state-of-the-art algorithms," *IEEE Signal Processing Magazine*, vol. 33, no. 4, jul 2016.

[16] Ivan Dokmanić, Reza Parhizkar, Juri Ranieri, and Martin Vetterli, "Euclidean distance matrices: Essential theory, algorithms, and applications," *IEEE Signal Processing Magazine*, vol. 32, no. 6, nov 2015.

[17] Reza Parhizkar, *Euclidean Distance Matrices: Properties, Algorithms and Applications*, Ph.D. thesis, EPFL, Suisse, 2013.

[18] R. Heusdens and N. Gaubitch, "Time-delay estimation for TOA-based localization of multiple sensors," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2014, pp. 609–613.

[19] Angelo Farina, "Simultaneous measurement of impulse response and distorsion with a swept-sine technique," in *AES 108th Convention, Paris, France*, nov 2000.

[20] Tom O'Haver, *A Pragmatic Introduction to Signal Processing*, 2014.

[21] Martin, Woszczyk, Corey, and Quesnel, "Sound source localization in a five-channel surround sound reproduction system," 09 1999.

[22] Peter H. Schönemann, "A generalized solution of the orthogonal procrustes problem," *Psychometrika*, vol. 31, no. 1, pp. 1–10, 1966.