Recent work has shown that smart speakers can be used for calibrating larger audio systems by placing multiple smart speakers in a room and using them to estimate the relative positions of loudspeakers and the listener. This allows for a more accurate calibration of the audio system, resulting in improved sound quality and localization accuracy.

The self-calibration problem using subarrays is typically referred to as interarray calibration [3] or array configuration calibration [9], and a number of methods have been proposed under various assumptions. A recent approach in [10] (and later improved in [11]) produces a high estimation accuracy, but requires that the raw microphone data (or a sparse spike representation thereof) are exchanged between the subarrays. Moreover, the method only assumes unsynchronised sources and does not take into account that the various subarrays cannot estimate the source positions with the same accuracy. As demonstrated in [12], a better estimation accuracy and robustness to outliers can be obtained if these uncertainties are taken into account. Whereas [11, 12] assume only unsynchronised sources at unknown locations, [13] assumes that each subarray has exactly one synchronised source whose location is known relative to the subarray. This corresponds to a scenario where an audio system consists of only smart speakers.

In this paper, we propose a method for creating a map over synchronised sources (e.g., loudspeakers), using only a central processing unit, and it also takes localisation errors into account when combining the estimated maps of each subarray into a global map. As opposed to existing methods, the proposed method works for a combination of synchronised and unsynchronised sources, and the relative positions of the synchronised sources do not have to be known. Finally, the method does not require at least five sources as in [12], but works even for a simple stereo setup.
Many source localisation algorithms already exist in the scientific literature for various array geometries. In principle, any array geometry can be used as long as at least three\(^1\) sensors (not on the same line) are used. Let an array have \(K \geq 3\) sensors, each with their own direction-dependent and known impulse response vector \(h_k(\theta) \in \mathbb{R}^{N_k}\) where \(\theta\) is the DOA of a source. The source emits a signal which is received by each sensor \(\eta_k\) samples later. In source localisation, the source signal is often (implicitly) assumed to be \(N\)-periodic [14] so that a time-shift of an \(N\)-length signal corresponds to a phase-shift in the frequency domain. Note that this assumption is easily satisfied for the synchronised sources since we can design the calibration signal. For a time-shift \(\eta_k\), any \(N\)-periodic signal can be written as [14]

\[
s(n - \eta_k) = \sum_{l=-L}^L \alpha_l \exp(j\omega_0(n - \eta_k))
\]

with \(\alpha_l = \alpha^*_l\) being a complex amplitude (\(\alpha_0\) is real), \(\omega_0 = 2\pi/N\) is the fundamental frequency, and \(L = \lfloor N/2 \rfloor\) is the maximum number of harmonic components. To facilitate a fast implementation, the shifted source signal for \(n = 0, 1, \ldots, N - 1\) can be rewritten in terms of the DFT matrix \(F = \{\exp(j2\pi nr/N)\}_{n,r=0,\ldots,N-1}\) as

\[
s(\eta_k) = N^{-1}FQ(\eta_k)F^H s(0)
\]

where \(Q(\eta_k) = \text{diag}(q(\eta_k))\). The definition of \(q(\eta_k)\) depends on whether \(N\) is even or not. If \(N\) is even, then

\[
q(\eta_k) = \begin{bmatrix}
1 & \exp(-j\omega_0 \eta_k) & \cdots & \exp(-j(L - 1)\omega_0 \eta_k) \\
cos(L\omega_0 \eta_k) & \exp(j(L - 1)\omega_0 \eta_k) & \cdots & \exp(j\omega_0 \eta_k)
\end{bmatrix}^T.
\]

Conversely, if \(N\) is uneven, then

\[
q(\eta_k) = \begin{bmatrix}
1 & \exp(-j\omega_0 \eta_k) & \cdots & \exp(-jL\omega_0 \eta_k) \\
\exp(jL\omega_0 \eta_k) & \cdots & \exp(j\omega_0 \eta_k)
\end{bmatrix}^T.
\]

Each sensor records \(N\) samples which are a noisy version of the shifted source signal convolved with the corresponding sensor response. This can be written as

\[
y_k = \frac{\beta}{\eta_k}H_k(\theta)s(\eta_k) + e_k
\]

where \(\beta > 0\) is an unknown gain and \(H_k(\theta)\) is a convolution matrix. Since the source signal is \(N\)-periodic, the convolution matrix is circulant and is, therefore, diagonalised by the DFT matrix \(F\). Thus, we have that

\[
y_k = \frac{\beta}{\eta_k} \frac{1}{N} FA_k(\theta)F^H \frac{1}{N} FQ(\eta_k)F^H s(0) + e_k
\]

\[
= G_k(p)s(0) + e_k
\]

where \(A_k(\theta)\) is a diagonal matrix containing the DFT of \(h_k(\theta)\), \(G_k(p) = \frac{1}{N^2} FA_k(\theta)Q(\eta_k)F^H\), and \(p\) is the position of the source. To estimate this source position, we seek the parameters which minimise the squared error \(\sum_{k=1}^K e_k^T e_k\). Equivalently, but more efficiently, the minimisation can be performed by minimising the residual sum of squares w.r.t. the source position \(p\). Thus, we first replace the linear parameters in (7) with their least-squares estimates and then minimise the squared residual. When the source signal \(s(0)\) is known, \(\beta\) is the linear parameter. Conversely, we cannot distinguish between \(s(0)\) and \(\beta\) when both are unknown so the product \(s(0)/\beta\) is the linear parameters in the case of an unknown source signal.

The described signal model and estimation procedure can be used for any array geometry. In the experiments, we have used a uniform circular array (UCA) since the DOA estimation performance is independent of the direction of the source [15, 16] and fast estimation algorithms for it exist [17]. Moreover, a UCA is often used in smart speakers.

\(^1\)In 3D, at last four sensors (not in the same plane) are required.
3.1. Quality matrices

The quality matrices represent how accurately the sources are estimated by the sensor array. This information is very useful when the local coordinate systems are combined into a global coordinate system. It is also absolutely essential when estimates of synchronised and unsynchronised sources are mixed since we can estimate the range of the former much more accurately than for the latter. As we detail below, we compute the quality matrices from the observed Fisher information matrices (FIMs). We focus the attention to the case of synchronised sources, but the same derivation can be followed for the case of unsynchronised sources.

We assume that the noise $e_k$ is white and Gaussian, so that the recorded data are distributed as $y_k \sim N(\mu_k(\vartheta), \sigma^2 I_N)$ where $\mu_k(\vartheta) = G_k(p) s(\theta) \hat{\theta}$ and $\hat{\theta} = [\beta \ p^T]^T$. The FIM is then defined as [18, Sec. 3.9]

$$I(\theta) = \frac{1}{\sigma^2} \sum_{k=1}^{K} \left( \frac{\partial \mu_k(\vartheta)}{\partial \theta} \right)^T \frac{\partial \mu_k(\vartheta)}{\partial \theta} = \frac{1}{\sigma^2} \begin{bmatrix} a & b \ b^T & C \end{bmatrix}.$$  

The inverse FIM is, therefore, given by

$$I^{-1}(\theta) = \sigma^2 \left[ \begin{array}{cc} X & \times \\ \times & (C - ba^{-1}b^T)^{-1} \end{array} \right]$$

from which we can extract the inverse quality matrix to

$$V^{-1} = \sigma^{-2} \left( C - ba^{-1}b^T \right)^{1/2}. \quad (10)$$

The observed FIM is obtained from the FIM by replacing the true parameter values with their estimates. Using the observed FIM as an estimate of the unknown FIM works in our experience well, unless the estimated source location is far from the true one. This is much more likely to happen for unsynchronised sources since the range estimate is very uncertain when the array radius is small relative to the range. A simple heuristic fix for this is to assume a big value for the range estimates so that effectively only the DOA estimates are used in fitting.

4. FITTING

So far, we have described how each sensor array computes estimates of the source positions and how the associated quality matrices are computed. In this section, we combine all this information into one global map of all the sensors and sources.

Assume that the true coordinates of $S$ sources in a reference coordinate system are given as the columns in the matrix $X \in \mathbb{R}^{2 \times S}$. In the coordinate system of the $m$th sensor array, these global coordinates are observed rotated and translated as

$$X_m = Q_m X + t_m 1^T \quad (11)$$

where $Q_m \in \mathbb{R}^{2 \times 2}$ and $t_m \in \mathbb{R}^{2 \times 1}$ are a rotation matrix and a translation vector, respectively. Without loss of generality, we assume that the coordinate system of sensor array 1 is the reference coordinate system so that $Q_1 = I_2$ and $t_1 = 0$. Unfortunately, we do not observe $X_m$ directly, but only the noisy version

$$y_m = \text{vec}(Y_m) = \text{vec}(X_m) + W_m e_m \quad (12)$$

where vec(·) is the vectorisation operator, $W_m \in \mathbb{R}^{2S \times 2S}$ is a block diagonal matrix of the form $W_m = \text{diag}(V_{m1}, \ldots, V_{mS})$, and $e_m = \text{vec}(E_m) \in \mathbb{R}^{2S \times 1}$. The quality matrix $V_m \in \mathbb{R}^{2 \times 2}$ is given by (10). Combining (11) and (12) gives the signal model

$$y_m = \begin{cases} x + W_1 e_1 & m = 1 \\ A(Q_m) B \begin{bmatrix} x \\ t_m \end{bmatrix} + W_m e_m & m = 2, \ldots, M \end{cases}$$

where $A(Q_m) = I_S \otimes Q_m, B = 1 \otimes I_2, x = \text{vec}(X)$, and $\otimes$ denotes the Kronecker product. The task is now to estimate $X$ given the quality matrices in $\{W_m\}_{m=1}^M$ and the observations $\{y_m\}_{m=1}^M$. By stacking all the $y_m$’s on top of each other for $m = 1, \ldots, M$, we obtain the signal model

$$y = H(Q) z + W \epsilon \quad (13)$$

where $Q = [Q_2^T \cdots Q_M^T]^T$ and

$$H(Q) = \begin{bmatrix} I_{2S} & 0 \\ G(Q) & I_{M-1} \otimes B \end{bmatrix}, \quad G(Q) = \begin{bmatrix} A^T(Q_2) & \cdots & A^T(Q_M) \end{bmatrix}^T \quad (14)$$

$$z = [x^T \ t_2^T \cdots \ t_M^T]^T \quad (15)$$

$$W = \text{diag}(W_1, \ldots, W_M) \quad (16)$$

For a known $Q$, the weighted least squares estimates of $X$ and $\{t_m\}_{m=1}^M$ are obtained from

$$\hat{z}(Q) = \left[ H^T(Q) W^{-2} H(Q) \right]^{-1} H^T(Q) W^{-2} y \quad (17)$$

The constrained estimator of $Q$, which minimises the residual sum of squares, is

$$\hat{Q} = \arg\max_{Q \in \mathbb{R}^{2(M-1) \times 2}} y^T W^{-2} H(Q) \hat{z}(Q) \quad (19)$$

s.t. $Q^T Q_m = I_2$ for $m = 2, \ldots, M$ and $\det(Q_m) = 1$ for $m = 2, \ldots, M$. It is well known from generalised Procrustes analysis, that a closed-form solution to the above problem is not available unless $M = 2$ and the same weights are applied to each column of $E_m$. In this case, a 2D eigenvalue decomposition can be used in the computation of $Q = Q_2 \in \mathbb{R}^{2 \times 2}$. If $M > 2$ and the same weights are applied to each column of $E_m$, the estimates of $x$ and $Q$ are computed iteratively as detailed in [19] by solving a series of eigenvalue decompositions. However, since the uncertainty in the $x$- and $y$-coordinates can be far from satisfying the condition that the same weights are applied to each column of $E_m$, we will not use the solution from [19] here. Instead, we seek to find a solution for a general weighted matrix. Such an algorithm was proposed in [20], but it seems to be very sensitive to the starting point. Specifically, the authors suggest that at least 20 random starting points should be tried out and that the weighted solution is not suitable to use as a starting point. This is a major drawback of the algorithm, and we, therefore, suggest that something else is done. For the 2D-case, the rotation matrix can be written as

$$Q_m(\theta_m) = \begin{bmatrix} \cos \theta_m & -\sin \theta_m \\ \sin \theta_m & \cos \theta_m \end{bmatrix} \quad (20)$$

Thus, the complete problem in (19) has $M - 1$ nonlinear parameters. In the case of many sensor arrays, it might be computationally very intensive to optimise such a high-dimensional nonlinear objective, so we instead attack the problem as it is traditionally solved in generalised orthogonal Procrustes analysis. That is, we iterate between estimating $X$ and $Q$. The main advantage of this approach is that the estimation of $Q$ given $X$ decouples into $M - 1$ individual 1D nonlinear optimisation problems instead of the high-dimensional problem in (19). Specifically, we have to solve problems of the form

$$\hat{\theta}_m = \arg\min_{\theta_m \in [-\pi, \pi]} (y_m - A(Q_m(\theta_m)) x) W_m^{-1} x, \quad \text{s.t.} \quad \det(Q_m(\theta_m)) = 1 \quad (21)$$
where \( P_{W^{-1}B} = W^{-1}_m B(B^T W^{-2}_m B)^{-1} B^T W^{-1}_m \). In 3D, we instead get a series of 2D nonlinear optimisation problems which are not too costly to solve. Given an estimate of \( Q \), we can compute an estimate of \( X \) from (18). The algorithm can be initialised by setting the initial value of \( X \) equal to the observation matrix with the best quality matrix.

5. EXPERIMENTS

In this section, we present the results from three experiments. First, we illustrate how the quality matrices allow us to combine estimates having very different estimation errors. Second, we evaluate the estimation accuracy as a function of the noise level. And third, we evaluate the estimation accuracy as a function of the reverberation time. All experiments were run using MATLAB, and the code will be available at [http://tinyurl.com/jknvbn](http://tinyurl.com/jknvbn).

In the first experiment, we used four sensor arrays, three synchronised sources, and one unsynchronised source. Each sensor array was a UCA with three microphones and a radius of 0.06 m. The calibration signal was 500 ms of white Gaussian noise which was bandpass filtered from 500 Hz to 1500 Hz. The filtering is performed since real-world loudspeakers have a large group-delay at low frequencies and are very directional at high frequencies. The sampling frequency was 4 kHz and white Gaussian noise was added so that the microphone recordings had an SNR of 10 dB. Fig. 2 shows the results. The true source and sensor positions are marked with black crosses and dots, respectively, and the source position estimates are marked with red stars. The small coloured circles denote the estimated sensor positions, and the coloured ellipses denote a contour of the quality matrices centred on the individual location estimates. For the synchronised sources, the ellipse contours are so small that they are hardly visible in the figure. For the unsynchronised sources, however, the contours indicate that the range estimates are much more uncertain than the angle estimates.

To the best of our knowledge, no other method exists which can directly solve a problem such as the one in the first experiment. Only special cases have been considered so far in the literature, and in the second experiment, we looked at one such special case treated in [13]. Specifically, we considered the case where four sensor arrays and synchronised sources were used. The sources and sensor arrays were co-located meaning that each sensor array knew the position of its own source with a very high precision. The sensor arrays were placed at the coordinates \((1, 1), (2.5, 1), (3, 3), \) and \((1, 2.5)\) in a room of size \((5, 6, 3)\) m. We computed the estimation accuracy of the sources as a function of the SNR for a reverberation time of 250 ms. The reverberation was added using a RIR-generator [21]. The proposed method was compared to two different reference methods which is a variation of [13] and the multidimensional unfolding method (MDU) in [7, 8]. The former consists in that we use the source localisation method of the proposed method to compute the local maps and classical multidimensional scaling for combining these maps. Using the same source localisation algorithm in the first reference method and the proposed method ensured that we evaluated the effect of using the proposed fitting method. For the second reference method [7, 8], we included all prior knowledge about the local geometry of the sensor arrays. As a performance measure, we used the sum of squared errors which is the dissimilarity measure often used in Procrustes analysis. For each SNR, 100 Monte Carlo runs were conducted. In each run, a new noise vector and a small random perturbation of the sensor array positions were generated. The results are shown in Fig. 3. The proposed method outperformed the reference methods across all SNRs. This demonstrates the importance of using weighting matrices, even when no unsynchronised sources are present.

In the third and final experiment, we had the same experimental setup as in the second experiment, except for that we varied the reverberation time and fixed the SNR to 20 dB. The results are given in Fig. 4, and they again show that the proposed method outperformed the reference methods.

6. CONCLUSION

In this paper, we have proposed a new two-step method for calibrating an audio system including one or several smart speakers. The method consists of a source localisation step in which each smart speaker computes a local map over the synchronised sources (e.g., loudspeakers) and unsynchronised sources (e.g., listeners). These local maps are then transmitted to a central unit which combines them into a global map in a fitting step. The fitting is performed according to the quality matrices pertaining to each local map, and they ensure that the most accurate estimates receive the greatest weight in the fitting. Via simulations, we demonstrated that the proposed method outperformed two reference methods for various noise levels and reverberation times.
7. REFERENCES


