published in: *International Workshop on Acoustic Signal Enhacement (IWAENC) 2014*

GEOMETRY CALIBRATION OF MULTIPLE MICROPHONE ARRAYS IN HIGHLY REVERBERANT ENVIRONMENTS

Axel Plinge and Gernot A. Fink

Department of Computer Science, TU Dortmund University, Dortmund, Germany

ABSTRACT

Microphone arrays can be used for a number of applications such as speaker diarization and tracking. For these, it is necessary to calibrate their geometry with good precision. Manual measurement is cumbersome and impractical for ad hoc configurations as distributed sensor nodes. So an fast automated calibration method that provides sufficient accuracy is required. It is even more convenient if data from the target application itself can be used so that the system can be calibrated online during its use. In this paper, we propose an automated geometry calibration method that outperforms existing state-of-the-art approaches. It does not require speakers at the nodes and works well in high reverberation. It was evaluated with real recordings in a smart room. By simply playing a white noise signal from a mobile phone at a few positions around the arrays, a calibration error of below 2 cm and 2° was achieved. By identification of speech events at different positions, the same method can be applied online; Here an error of 10 cm and 3° was achieved.

Index Terms— microphone array, ad hoc, distributed sensor network, geometry calibration, speaker tracking

1. INTRODUCTION

For applications such as speaker diarization and tracking [1]. multiple distributed microphone arrays are utilized, whose geometry has to be known. Manual measurements are cumbersome and hardly practical in applications with many arrays or ad hoc configurations of acoustic nodes in distributed sensor networks. It is desirable that the calibration requires little effort or can be achieved online during the actual application. High reverberation is found in typical target scenarios such as smart rooms. Here, existing methods achieve only limited accuracy. Pure acoustic methods provide only means of relative geometry estimation. The rotation, translation and mirroring of the geometry has to be aligned for integration with video data and multi modal tracking.

Most existing methods require a dedicated calibration step and impose some special constraints that are not required for the target application. If the sensor nodes are equipped with an additional speaker, absolute time of arrival (ToA) measurements can be produced by playing calibration sounds at each sensor node. With given distances of all microphones to a number of base points, the geometry can be calculated using multidimensional scaling [2]. This was used with consumer devices like laptops [3], smartphones [4] and experimental sensor nodes [5] that all have at least one speaker and at least one microphone in known relative distance.

When the acoustic sensor nodes are not equipped with speakers, the geometry has to be inferred from unknown source positions that are not aligned with the sensors. To measure time difference of arrival (TDoA) between pairs of sensors, there has to be strict time synchronization because otherwise the measurement contains an unknown drift or jitter based offset. If the arrays are not physically connected, synchronization can be achieved wirelessly by dedicated algorithms [6, 7] otherwise only the time offsets between the sensors can be estimated [8]. From the TDoA estimates, the geometry can be inferred. Using a dedicated calibration step in which signals such as white noise or sweep chirps are played with a speaker allows reasonably good estimation [9]. Around 10 cm accuracy was achieved with noise in an reverberant smart room using data set matching (DSM) [10] or affine estimation [11].

In passive estimation, only speech events may be used. Using speech of a single moving person on a random trajectory, the relative geometry of the nodes was computed using data set matching (DSM) [10] and the random sampling consensus (RANSAC) method [12] with an accuracy of around 25 cm. When using small microphone arrays, the direction of arrival (DoA) at each node can be computed and a relative geometry may be inferred. The scaling can be estimated separately using TDoA information [13]. Recently a multimodal approach using visual speaker localization for absolute position and orientation calibration was proposed [14].

In this paper, a method for relative geometry calibration of distributed microphone arrays in highly reverberant environments is presented. Salient sound events and their DoA at each acoustic sensor node are estimated by a robust method [15] that was chosen because it provides reliable localization and detection of speakers in the targeted indoor environments with high reverberation by means of neurobiologically in-

This work was supported by the German Research Foundation (DFG) under contract number Fi 799/5-1.

The authors thank Florian Jacob and the anonymous reviewers for their helpful suggestions.

spired strategies [16]. The inter-array TDoA is computed by correlation. The geometry is estimated by hierarchical error minimization with respect to both measurements. For highest accuracy, white noise, e.g., played from a mobile phone can be used in a dedicated calibration session. The method also works with speech events which allows it to be applied online, e.g., during conference sessions. For online application, a classification step should be added [17] to exclude sounds that do not provide reliable acoustic localization like footfall noise, chair movement, doors, windows etc.

2. METHOD

In the following a set of $M > 2$ distributed sensor nodes with small microphone arrays is considered whose geometry is to be calibrated. The individual array geometries are assumed known and the sampling is synchronized. The M arrays are at unknown 2D positions $\boldsymbol{r}_m \in \mathbb{R}^2$ with unknown orientations $o_m \in [-\pi, \pi]$. Only the relative geometry can be estimated, so $\boldsymbol{r}_0$ and o_0 are fixed to an arbitrary value. First, the geometry of individual arrays $m > 0$ relative to the array with index 0 is estimated. Then the geometry of all arrays including the pairs (n, m) with $m, n > 0$ is estimated jointly.

2.1. Measurement and Target Function

At a set of fixed unknown source positions s_i, sound is played or spoken and received by all microphone arrays. A DoA localization method [15] is used that allows to isolate the events and compute an angle $\Theta_{i,m}$ for each of the sound events. The events themselves are identified automatically as time segments with low DoA variance. An inter-array TDoA estimate $d_{i,(m,n)}$ is computed for each sound event by correlation.

For each pair m, n of arrays, the position and orientation have to fulfill the geometric relations with respect to the measured TDoA and DoA as illustrated in Fig. 1. This is used to calculate an error for the estimate in the following way: Given an estimate of the orientations $\hat{o}_{m,n}$ and positions $\hat{\boldsymbol{r}}_{m,n}$, the measured DoAs $\Theta_{i,m}$ and $\Theta_{i,n}$ can be used to compute the source position $\hat{\boldsymbol{s}}_{i,(m,n)}$ by triangulation

$$\begin{aligned}\hat{\boldsymbol{s}}_{i,(m,n)} &= \hat{\boldsymbol{r}}_m + \hat{k}_{i,m} \begin{pmatrix} \cos\left(\hat{o}_m + \Theta_{i,m}\right) \\ \sin\left(\hat{o}_m + \Theta_{i,m}\right) \end{pmatrix} \\ &= \hat{\boldsymbol{r}}_n + \hat{k}_{i,n} \begin{pmatrix} \cos\left(\hat{o}_n + \Theta_{i,n}\right) \\ \sin\left(\hat{o}_n + \Theta_{i,n}\right) \end{pmatrix} . \end{aligned} \tag{1}$$

The line intersection provides estimates $\hat{k}_{i,m}$ and $\hat{k}_{i,n}$ of the distances to the source, cp. Fig. 1, so the relative distance

$$||\hat{\boldsymbol{s}}_{i,(m,n)} - \hat{\boldsymbol{r}}_m|| - ||\hat{\boldsymbol{s}}_{i,(m,n)} - \hat{\boldsymbol{r}}_n|| = \hat{k}_{i,m} - \hat{k}_{i,n} \tag{2}$$

allows to compute the error with respect to the measured TDoA $d_{i,(n,m)}$ multiplied by the speed of sound c as

$$\begin{aligned}\epsilon_{i,(n,m)} = \\ ||\hat{\boldsymbol{s}}_{i,(m,n)} - \hat{\boldsymbol{r}}_m|| - ||\hat{\boldsymbol{s}}_{i,(m,n)} - \hat{\boldsymbol{r}}_n|| - c \cdot d_{i,(m,n)} .\end{aligned} \tag{3}$$

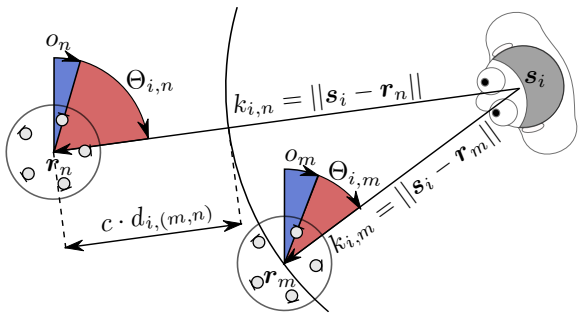

Fig. 1. Geometry relations of two arrays at $\boldsymbol{r}_m$ and $\boldsymbol{r}_n$ and a source at $\boldsymbol{s}_i$ in relation to the measured DoAs $\Theta_{i,m}$, $\Theta_{i,n}$ and the TDoA $d_{i,(n,m)}$ where $||$ is the Euclidean norm.

By solving (1) and evaluating (3) consecutively, an error for any estimate for $\boldsymbol{r}$ and o with respect to both DoA and TDoA measurements is computed.

2.2. Estimation

This is used to estimate the geometry, first for pairs $(0, m)$ and then jointly for all microphone arrays. The problem is underdetermined for a single speaker position, so random index subsets $\Omega \subset \{3 \dots T\}$ with size $I = |\Omega|$ of all T sound events are used.

For each array with index $m > 0$ the fitting of the relative geometry estimates $\hat{o}_{0,m}$, $\hat{\boldsymbol{r}}_{0,m}$ is evaluated by computing the squared sum of the errors over all sound events. By minimizing this, the parameters are estimated:

$$o_m^*(\Omega), \boldsymbol{r}_m^*(\Omega) = \underset{o_m, \boldsymbol{r}_m}{\operatorname{argmin}} \sum_{i \in \Omega} \epsilon_{i,(0,m)}^2 . \tag{4}$$

This is done in two steps. First an estimate is computed by exhaustive search on a coarse grid of, e.g., 10 cm and 2°. Then the estimate is refined by gradient descent using the Broyden-Fletcher-Goldfarb-Shanno algorithm [18]. It estimates the Hessian based on evaluating the function in a local neighborhood, so no explicit derivation is needed.

The estimate based on individual pairs $(0, m)$ is used as a starting point to search for a joint estimate. First a joint estimate of the source positions is computed using (1)

$$\hat{\boldsymbol{s}}_i = \frac{2}{M(M-1)} \sum_{m<n} \hat{\boldsymbol{s}}_{i,(n,m)} , \tag{5}$$

then the over all error is evaluated using (3)

$$\epsilon_i^2 = \frac{2}{M(M-1)} \sum_{m<n} \epsilon_{i,(m,n)}^2 . \tag{6}$$

The joint geometry estimate is formulated as stacked vectors of the positions $\boldsymbol{R} = (r_1 \dots r_M)^T$ and orientations

1. Measurement
   - Record sound at fixed positions around the arrays.
   - Extract I sound events.
   - Compute DoAs $\Theta_{i,\cdot}$ and inter-array TDoA $d_{i,(\cdot,\cdot)}$.

2. Estimation using random sets Ω of I sound events
   - Compute geometry estimates $o^*_m(\Omega), \boldsymbol{R}^*_m(\Omega)$ for each pair $(0, m)$ of arrays by exhaustive search followed by gradient descent.
   - Compute joint estimates $\boldsymbol{o}^*(\Omega), \boldsymbol{R}^*(\Omega)$ for all arrays by gradient descent initialized with the individual estimates.
   - Keep the estimate if $\epsilon_\Omega < 20\,\text{cm}$. Chose the next Ω until N such estimates are found.
   - Compute weighted mean $\boldsymbol{o}^*, \boldsymbol{R}^*$ of all estimates.

Fig. 2. Proposed geometry calibration procedure.

$\boldsymbol{o} = (o_1 \dots o_M)^T$. It is computed by gradient descent for a set Ω of source positions

$$\boldsymbol{o}^*(\Omega), \boldsymbol{R}^*(\Omega) = \underset{\boldsymbol{o}, \boldsymbol{R}}{\operatorname{argmin}} \left({\epsilon_\Omega}^2 = \frac{1}{|\Omega|} \sum_{i \in \Omega} \epsilon_i^2 \right) . \quad (7)$$

The procedure is repeated to improve the estimation and remove bias that can be the result of an individual error in measurement or the choice of source positions. To remove outliers, only estimates with an error $\epsilon_\Omega < 20\,\text{cm}$ are kept. New Ω are chosen until we have several such estimates, e.g. $N = 40$. The positions $\boldsymbol{R}^*$ and orientations $\boldsymbol{o}^*$ are computed as average weighted by $1/\epsilon_\Omega$:

$$\boldsymbol{R}^* = 1/ \left(\sum_\Omega 1/\epsilon_\Omega \right) \left(\sum_\Omega \frac{\boldsymbol{R}^*(\Omega)}{\epsilon_\Omega} \right) \quad (8)$$

$$\boldsymbol{o}^* = 1/ \left(\sum_\Omega 1/\epsilon_\Omega \right) \left(\sum_\Omega \frac{\boldsymbol{o}^*(\Omega)}{\epsilon_\Omega} \right) . \quad (9)$$

3. RESULTS

In order to test the real-world performance, recordings were made in a highly reverberant $3.7 \times 6.8 \times 2.6\,\text{m}^3$ conference room of a smart house installation at our university. Three circular microphone arrays with 5 microphones in a 5 cm radius were embedded in a table. Each array was captured by a separate sound card at 48 kHz. The sound cards were synchronized, recordings of coherent white noise showed a remaining jitter of 22 μs between them. A reverberation time of $670 \pm 89\,\text{ms}$ over the microphone signals was calculated using a blind estimation algorithm [19]. Experiments where done with human speech and white noise. In both experiments, the identical 10 positions around the table shown in Fig. 3 were used.

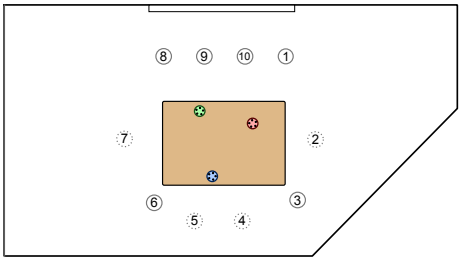

Fig. 3. Positions of the microphone arrays (colored cicles) and the calibration sounds (numbered circles, the dotted circles mark sitting, the solid a standing human speaker).

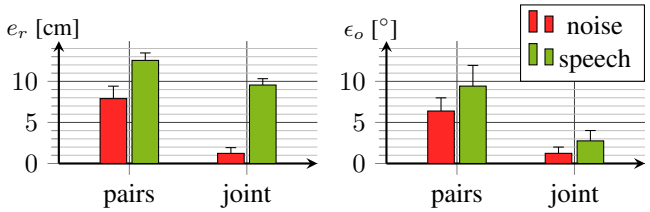

Fig. 4. Calibration error and its standard deviation of pairwise and joint estimation using $N = 60$ sets of $I = 6$ positions over ten experiments.

3.1. White noise

In the first experiment, a 20 s white noise sample was played from a mobile phone at all 10 positions slightly above table height. The DoAs were computed using [1] and had an RMS around 5° compared to the ground truth. The TDoA estimates were computed using unweighted crosscorrelation with windows of 2^{17} samples. The corresponding distance estimates had an RMS of 3 cm compared to the ground truth. Both estimates showed very little variance over the 20 s period, so much shorter times may be used. If distortions by ambient sounds are expected, outliers should be removed by median filtering.

3.2. Human speaker

In the second experiment, a human speaker spoke one sentence at all 10 positions. The speech was localized using [1]. Time segments with little angular variance were grouped to detect the utterances. The DoA estimates had an RMS around 5° compared to the ground truth. The TDoA estimates were computed using SRP PHAT with windows of 2^{16} samples; The corresponding distance estimates had an RMS of 8 cm compared to the ground truth.

3.3. Calibration

The method was applied 10 times using $N = 60$ and $I = 6$. In Fig. 4, the translation and rotation calibration errors ϵ_r, ϵ_o for different steps are plotted for both speech and noise. Using noise, the mean of the joint estimates shows an impressively

method	microphones	signal			T_{60} [s]	e_r [cm]	e_o[°]
ToA [4]	5 x 1 + speaker	chirps	node positions		0.60 (real)	6.8	?
ToA [5]	4 x 4 + speaker	MLS	node positions		0.26 (real)	1.5	?
TDoA [10]	4 x 4 circ.	noise	random walk	60 s	0.60 (real)	9.3	?
TDoA [10]	4 x 4 circ.	speech	random walk	30 s	0.60 (real)	11.0	?
TDoA [12]	3 x 5 circ.	speech	random walk	360 s	0.50 (sim.)	25.0	4.0
TDoA [∗]	3 x 5 circ.	noise	10 positions	300 s	0.67 (real)	1.2	1.3
TDoA [∗]	3 x 5 circ.	speech	10 positions	80 s	0.67 (real)	9.5	2.8
DoA+V [14]	3 x 5 circ. + 5 cameras	speech	10 positions	80 s	0.67 (real)	6.6	1.9

Table 1. Comparison of different geometry calibration approaches with the proposed method [∗]

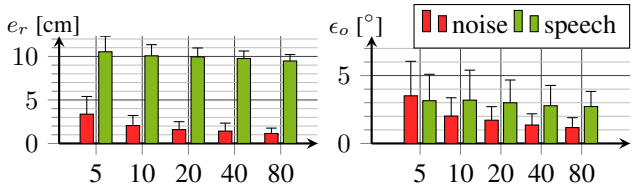

Fig. 5. Calibration error and its standard deviation for increasing number N of sets of $I = 6$ positions.

#	calibration	ϵ_a [°]	ϵ_l [cm]	P [%]	R [%]
1	measurement	3.50	17.3	100.0	94.1
	auto. noise [∗]	3.40	17.0	100.0	94.1
	auto. speech [∗]	3.86	29.9	98.7	93.6
2	measurement	4.87	19.8	100.0	94.3
	auto. noise [∗]	4.71	16.6	100.0	94.6
	auto. speech [∗]	4.99	26.8	100.0	94.6

Table 2. Acoustic speaker tracking results for manual measurement and calibration using the proposed method [∗] for the calibration sequence (#1) and a subsequent recording (#2).

low 1.2±0.6 cm and 1.3±0.7°. For speech, the joint mean has an error of 9.5±0.7 cm and 2.8±1.3°. In Fig. 5, the results for different numbers N of sets are shown. When using twenty or more sets, the accuracy is already close to its minimum error.

3.4. Application to speaker tracking

To test the applicability of the proposed method, the calibration results were used as the basis for tracking a speaker using [1]. Both the calibration recording and a separate recording of a speaker taking up 1 positions were tested. The tracking performance using the measurement and calibration using white noise and speech is shown in table 2. The angular and metric tracking errors $\epsilon_{a,l}$ are given. The angular tracking error is similar for measurement and calibration with noise, (even slightly less for the latter) and slightly increased for the speech based calibration. The Euclidean errors behave similar. For the speech based calibration, the tracking RMS increases by about 10 cm corresponding to the estimation error. A distance of 0.5 m is used as margin Euclidean error for precision and recall to reflect what error may be tolerable for practical applications. The results for noise based calibration and measurement are similar, using the speech based calibration the precision (P) and recall (R) decrease only slightly.

3.5. Comparison to other methods

Table 1 compares the proposed method to others. Notably only a ToA method using a speaker mounted at the four microphone arrays and maximum length sequence signals (MLS) [5] achieves a translation error close to the proposed method using noise. Using speech, [10] comes close to the accuracy achieved by the proposed method. Experiments with noise are better with about 9.3 cm but still not close to the 1.2 cm of the proposed approach. In their experiments the microphone arrays were mounted on the ceiling resulting in a larger distance from the sound source. The accuracy of our recent approach using visual speaker localization [14] lies between speech and noise results.

4. CONCLUSION

A fully automated geometry calibration procedure for sensor nodes with small microphone arrays distributed in an ad hoc fashion was developed. It uses a set of sound events from positions around the arrays; These can be natural speech or artificial noise played form a handheld device. No prior knowledge of the positions or timing of the sound events is required. The method was tested with real world indoor recordings, showing its robustness against reverberation. The speech based calibration is slightly worse, most likely due to less accurate TDoA estimates. The calibration is very precise and outperforms existing state-of-the-art TDoA approaches. It is also more precise that ToA approaches using an additional speaker mounted at each of the arrays. When applying the results in speaker tracking, the noise based calibration leads to identical results as calibration by measurement. The speech based calibration produces slightly inferior results that are well within boundaries for practical use. The method is a true alternative to cumbersome manual measurement. In a conference scenario, speech events at fixed positions occur naturally. This allows the proposed method to be used online.

5. REFERENCES

[1] Axel Plinge and Gernot A. Fink, "Multi-Speaker Tracking using Multiple Distributed Microphone Arrays," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Florence, Italy, 2014.

[2] Stanley T. Birchfield, "Geometric Microphone Array Calibration by Multidimensional Scaling," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 2003.

[3] Vikas C. Raykar, Igor V. Kozintsev, and Rainer Lienhart, "Position Calibration of Microphones and Loudspeakers in Distributed Computing Platforms," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 70–83, 2005.

[4] Marius H. Hennecke and Gernot A. Fink, "Towards Acoustic Self-Localization of Ad Hoc Smartphone Arrays," in *Workshop on Hands-Free Speech Communication and Microphone Arrays*, Edinburgh, UK, 2011, pp. 127–132.

[5] Pasi Pertila, Mikael Mieskolainen, and Matti S. Hamalainen, "Closed-form Self-localization of Asynchronous Microphone Arrays," in *Joint Workshop on Hands-Free Speech Communication and Microphone Arrays*, 2011, pp. 139–144.

[6] Shmulik Markovich-Golan, Sharon Gannot, and Israel Cohen, "Blind Sampling Rate Offset Estimation and Compensation in Wireless Acoustic Sensor Networks with Application to Beamforming," in *Int. Workshop on Acoustic Signal Enhancement*, Aachen, Germany, 2012.

[7] Joerg Schmalenstroeer, Patrick Jebramcik, and Reinhold Haeb-Umbach, "A Gossiping Approach to Sampling Clock Synchronization in Wireless Acoustic Sensor Networks," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Florence, Italy, 2014.

[8] Pasi Pertila, Matti S. Hamalainen, and Mikael Mieskolainen, "Passive Temporal Offset Estimation of Multichannel Recordings of an Ad-Hoc Microphone Array," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 21, no. 11, pp. 2393–2402, Nov. 2013.

[9] S. Daniele Valente, Marco Tagliasacchi, E. Antonacci, Paolo Bestagini, Augusto Sarti, and Stefano Tubaro, "Geometric Calibration of Distributed Microphone Arrays from Acoustic Source Correspondences," in *IEEE Workshop on Multimedia Signal Processing*, 2010.

[10] Marius H. Hennecke, Thomas Plötz, Gernot A. Fink, and Reinhold Haeb-Umbach, "A Hierarchical Approach to Unsupervised Shape Calibration of Microphone Array Networks," in *IEEE Workshop on Statistical Signal Processing*, Cardiff, Wales, UK, 2009, pp. 257–260.

[11] Sebastian Thrun, "Affine Structure From Sound," in *Conference on Neural Information Processing Systems*, 2005, vol. 18, pp. 1353–1360.

[12] Florian Jacob, Joerg Schmalenstroeer, and Reinhold Haeb-Umbach, "DoA-based Microphone Array Postion Self-Calibration using Circular Statistisc," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vancover, Canada, 2013.

[13] Joerg Schmalenstroeer, Florian Jacob, Reinhold Haeb-Umbach, Marius H. Hennecke, and Gernot A. Fink, "Unsupervised Geometry Calibration of Acoustic Sensor Networks using Source Correspondences," in *Interspeech*, Florence, Italy, 2011.

[14] Axel Plinge and Gernot A. Fink, "Geometry Calibration of Distributed Microphone Arrays Exploiting Audio-Visual Correspondences," in *European Signal Processing Conference*, Lisbon, Portugal, 2014.

[15] Axel Plinge and Gernot A. Fink, "Online Multi-Speaker Tracking Using Multiple Microphone Arrays Informed by Auditory Scene Analysis," in *European Signal Processing Conference*, Marrakesh, Morocco, 2013.

[16] Axel Plinge, Marius H. Hennecke, and Gernot A. Fink, "Robust Neuro-Fuzzy Speaker Localization Using a Circular Microphone Array," in *Int. Workshop on Acoustic Echo and Noise Control*, Tel Aviv, Israel, 2010.

[17] Axel Plinge, Rene Grzeszick, and Gernot A. Fink, "A Bag-of-Features Approach to Acoustic Event Detection," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Florence, Italy, 2014.

[18] Richard H. Byrd, Peihuang Lu, and Jorge Nocedal, "A Limited Memory Algorithm for Bound Constrained Optimization," *SIAM Journal on Scientific and Statistical Computing*, vol. 16, pp. 1190–1208, 1995.

[19] Heinrich W. Löllmann, Emre Yilmaz, Marco Jeub, and Peter Vary, "An Improved Algorithm for Blind Reverberation Time Estimation," in *Int. Workshop on Acoustic Echo and Noise Control*, Tel Aviv, Israel, 2010.