

ROBUST TDE-BASED DOA ESTIMATION FOR COMPACT AUDIO ARRAYS

Krishnaraj Varma, Takeshi Ikuma, and A. A. (Louis) Beex

Systems Group — DSP Research Laboratory
The Department of Electrical and Computer Engineering
Virginia Tech, Blacksburg, VA 24061-0111

ABSTRACT

Cross-correlation based time delay estimates (TDE) can be used for direction-of-arrival (DOA) estimation with an acoustic array in not-too-reverberant environments. In order to benefit from the computational efficiency of TDE-based DOA estimation, and concentrating on applications that use a compact microphone array and low sampling frequency, we use a combination of approaches to make TDE-based DOA estimation more robust under reverberant conditions. Cross-correlation interpolation based on the Goertzel algorithm is used for generalized cross-correlation TDE as well as in the steered-response-power algorithm used for comparison. The generalized cross-correlation TDE algorithm precedes a robust TDE-to-DOA process. The latter includes DOA and cross-correlation dependent weighting of the various TDE and removal of an outlier TDE. When the use of TDE from consecutive frames is possible, further performance improvement results from unit-norm DOA adaptation. Performance is evaluated for simulated and recorded data.

1. INTRODUCTION

A robust algorithm is proposed for estimating the DOA of a single acoustic source impinging on a compact 3-D microphone array of arbitrary configuration. The algorithm uses correlation-based pair-wise TDE and is geared towards applications using compact microphone arrays, requiring low sampling frequency and computational effort. Areas of application include microphone arrays mounted on handheld instruments or robots. Additionally, such applications often require inexpensive hardware and minimal power consumption while maintaining acceptable DOA estimation performance. The accuracy and robustness of DOA estimates is largely dependent on robust, high-resolution TDE and a robust TDE to DOA conversion process.

Our algorithm — an improved version of TDE-based

DOA estimation, which utilizes phase transform (PHAT) of cross-correlation estimates — is compared to the steered response power algorithm with PHAT weighting (SRP-PHAT algorithm) [1]. The SRP-PHAT algorithm computes a cumulative generalized cross-correlation value with PHAT weighting (GCC-PHAT) across all microphone pairs at corresponding theoretical delays for all azimuths and elevations of arrival. The DOA that gives the maximum cumulative GCC-PHAT value is assumed to be the true DOA. While it has been argued [1] that the SRP-PHAT algorithm has superior performance over TDE-based algorithms, the TDE-based algorithm requires less computational effort. While improving DOA estimation accuracy toward that of the SRP-PHAT algorithm — by weighted LS combination of TDE to DOA estimates, our method, named WTDE-PHAT, incorporates the Goertzel algorithm to reduce the required computational effort.

2. PROPOSED DOA-ESTIMATION ALGORITHM

The proposed DOA-estimation process consists of three steps: estimation of cross-power spectra, peak detection of corresponding cross-correlation estimates, and weighted LS DOA estimation based on information gathered from the cross-correlation estimates and knowledge of the array configuration. Each step is now described in detail.

2.1. GCC-PHAT based XPSD estimation

In acoustic array processing—especially for indoor applications — a robust TDE method is essential to combat reverberation effects [1]. One of the correlation-based TDE algorithms that is claimed to be more resilient to room reverberation utilizes GCC-PHAT [2]. GCC-PHAT is an ad-hoc technique to pre-whiten the received signals before computing the cross-correlations. If $X_i(k)$ and $X_j(k)$ are the DFT samples of the signals from the i^{th} and j^{th} microphones respectively, then the generalized cross power spectral density (XPSD) between the signals from these two microphones is given by

$$\Phi_{ij}(k) = H_{ij}(k) X_i(k) X_j^*(k) \quad (1)$$

The term $H_{ij}(k)$ in (1) is a generalized pre-filter used to impart some desirable spectral properties on the signals. In the case of PHAT we use

$$H_{ij}(k) = \frac{1}{|X_i(k)||X_j(k)|} \quad (2)$$

By pre-whitening the signals the peaks of the cross-correlation function are made sharper, so that the peaks at the correct delays are better localized. The price paid for this improved localization is decreased robustness. PHAT gives equal weighting to all frequencies — even noise frequencies. Therefore, for signals that are not strictly wideband, the obtained results are based mainly on noise frequencies, and thus tend to be fairly non-robust. Further downstream processing of the time-delay estimates is needed to make the DOA estimate more robust.

2.2. Cross-correlation peak detection

The GCC-PHAT between signals from any pair of microphones can be computed using an inverse-discrete-Fourier-transform operation on the PHAT weighted XPSD of the two signals. The TDE for signals from a pair of microphones corresponds to the cross-correlation lag value where the correlation function peaks (maximizes). The (discrete) correlation-based TDE method has limited resolution for (continuous) time-delay estimates due to the discrete lag values (multiples of the sampling interval) at which cross-correlations can be computed. This becomes detrimental to the DOA estimation algorithm performance when a slow sampling rate is coupled with a physically compact array.

To increase the lag resolution of time-delay estimates, interpolation can be applied to discrete cross-correlation estimates. For example, the cross-correlation functions can be interpolated in the time domain [3], by padding the XPSD in the middle with zeros and using an FFT algorithm that utilizes both input and output pruning. Our proposed interpolation method also uses frequency domain zero-padding — in the form of the computationally efficient Goertzel algorithm [4] with a correction rotation factor — to evaluate the arbitrary (sub-sample) lag at which the cross-correlation maximizes. The Goertzel algorithm allows high-resolution TDE and limits computations to the search space of physically possible delays.

Under the assumption of a far-field source and with knowledge of the exact microphone array geometry, the maximum delay, $\tau_{max,ij}$, for which the peaks of the cross-correlations are expected can be calculated. The maximum peak is searched for over $[-\tau_{max,ij} - \Delta, \tau_{max,ij} + \Delta]$. The search range is widened by Δ , to accommodate some expected noise on the data. The search procedure is initiated by selecting equally spaced lags over the search range and computing their corresponding cross-correlation values using the Goertzel algorithm. Among these cross-correlation values, three-point-patterns (TPP) that exhibit

peaking behavior (*e.g.*, middle point greater than edge points) are extracted. For each TPP, a line search is performed to seek the maximum cross-correlation value. The golden section search method [5] — with a limited number of iterations — was implemented to perform this task. The cross-correlation values of the maximum and second highest peak are stored to compute the weighting factors for the subsequent TDE-to-DOA-estimation.

2.3. Weighted LS TDE-to-DOA-Estimate Conversion

In a typical crude TDE-to-DOA-estimate conversion process, a simple least-squares (LS) solution to a system of equations is used. Consider an arbitrary 3D N -element microphone array that is characterized by the $3 \times N$ matrix, \mathbf{M} , each row of which represents the position of a microphone in Euclidean space. Let the vector $\mathbf{d}(\theta, \phi) = -[\cos\phi\sin\theta \cos\phi\cos\theta \sin\phi]^T$ represent a unit-norm vector in the direction of arrival, with azimuth θ and elevation ϕ . Our aim is to estimate $\mathbf{d}(\theta, \phi)$ and from it θ and ϕ . The relationship between $\mathbf{d}(\theta, \phi)$ and the time delays between microphone pairs (defined by column vector $\boldsymbol{\tau}$) is given by

$$\mathbf{A}\mathbf{d}(\theta, \phi) = \mathbf{v}\boldsymbol{\tau}(\theta, \phi) \quad (3)$$

where \mathbf{A} is a matrix constructed from \mathbf{M} — each row of which represents a vector joining a pair of microphones. Let $\hat{\boldsymbol{\tau}}$ be a vector of time delay estimates obtained from the GCC-PHAT method. The least squares solution for this system of equations is given by

$$\tilde{\mathbf{d}}_{LS}(\theta, \phi) = \mathbf{v}(\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \hat{\boldsymbol{\tau}} \quad (4)$$

To improve the accuracy of this solution, two separate weighting matrices are introduced. The first, \mathbf{Q}_1 , is derived from the nonlinear relationship between TDE and DOA, and the second, \mathbf{Q}_2 , is derived from the peaking characteristics of the cross-correlation estimate. Both are $(N \times N)$ diagonal matrices, so let $\mathbf{Q} = \mathbf{Q}_1 \mathbf{Q}_2$. The resulting weighted system of equations is

$$\mathbf{Q}\mathbf{A}\mathbf{d}(\theta, \phi) = \mathbf{v}\mathbf{Q}\boldsymbol{\tau}(\theta, \phi) \quad (5)$$

and the resulting weighted LS solution is expressed by

$$\tilde{\mathbf{d}}_{wLS}(\theta, \phi) = \mathbf{v}(\mathbf{A}^T \mathbf{Q}\mathbf{A})^{-1} \mathbf{A}^T \mathbf{Q}\hat{\boldsymbol{\tau}}. \quad (6)$$

For a single pair of microphones, separated by D , and a far-field source, the arrival angle θ relative to broadside and time delay τ are related by $\theta = \sin^{-1}(\mathbf{v}\tau/D)$. This relationship indicates that as θ approaches $\pm 90^\circ$, the angle estimate becomes more sensitive to changes in τ . So, small TDE ought to be weighted more than large TDE. Thus, \mathbf{Q}_1 weighting factor chosen for the i^{th} TDE is

$$q_{1i} = 1 - \frac{\max(\tau_{max,i}^2, \tau_i^2)}{\tau_{max,i}^2} \quad (7)$$

The \mathbf{Q}_2 weighting is associated with the TDE confidence level. One good indication of reliable TDE is the number of correlation peaks over the search range. Under non-reverberant conditions, a white source signal would result in a single strong cross-correlation peak. Thus, if the cross-correlation window only contains a single peak, the corresponding TDE weight is the maximum possible, which is 1. If multiple peaks are detected the weight for the i^{th} TDE is computed from

$$q_{2i} = 1 - \max\left(\frac{c_{i2}}{c_{i1}}, 0\right) \quad (8)$$

where c_{i1} and c_{i2} are the values at the largest and second largest correlation peaks of the i^{th} pair of microphones .

3. SINGLE-STEP DOA ESTIMATION RESULTS

The algorithm has been tested with both simulated and recorded data for the same configuration. The room dimension is $L_x = 2.74$ m by $L_y = 3.81$ m by $L_z = 2.44$ m. The array is placed with Microphone 1 (reference microphone) at the coordinate point $[0.75 \ 0.75 \ 0.75]$ m, with one corner of the room defined as the origin. The array consists of four microphones and the other microphones are placed as follows (in m, relative to Microphone 1):

$$\begin{aligned} \mathbf{m}_2 &= [0.095 \ 0 \ 0] \\ \mathbf{m}_3 &= [0.095 \ 0.95 \ 0] \\ \mathbf{m}_4 &= [0 \ 0 \ 0.095] \end{aligned}$$

The loudspeaker source location is $[2.40 \ 2.61 \ 1.45]$ m, equivalent to an azimuth of 41.5° (clockwise from y -axis) and an elevation of 15.8° . This source-microphone configuration exhibits a near-field effect that introduces TDE errors of up to $3 \mu\text{s}$ for each microphone pair. A sampling frequency of 8 kHz is used, with a frame size of 256 (32 ms). The velocity of sound is assumed to be $v = 345$ m/s. Both the simulation and the experiments use 10 speech sentences as source signals, randomly selected from the TIMIT Speech Corpus [6].

The accuracy of the DOA estimates is measured in terms of the RMS DOA error [1]:

$$E_{RMS} = \sqrt{(\hat{\theta} - \theta)^2 + (\hat{\phi} - \phi)^2} \quad (9)$$

3.1. Simulation Procedure and Results

The Monte-Carlo simulations are designed to illustrate the performance of our algorithm in comparison with the SRP-PHAT algorithm. The room is assumed empty with reflective boundaries (*i.e.*, walls, floor, and ceiling). All boundaries are considered to have the same uniform reflection coefficient $\beta \in [0, 1]$. The room acoustics are modeled as a linear time-invariant system, and Peterson's modification [7] of the image model technique of Allen

and Berkley [8] is implemented to compute the acoustic impulse response of the room. The simulated microphone outputs are obtained by convolution of the source signal and the room impulse responses.

The simulation is executed for room reverberation times of 0 ms ($\beta = 0$), 100 ms ($\beta = 0.68$), and 200 ($\beta = 0.82$) ms, as computed from Eyring's formula [9].

$$T_R = -13.82 / \left[v(L_x^{-1} + L_y^{-1} + L_z^{-1}) \ln \beta \right] \quad (10)$$

For each reverberation configuration, four DOA algorithms are tested: TDE-based, TDE-PHAT based, WTDE-PHAT, and SRP-PHAT configured for 1° resolution.

For the simulation, spoken sentences were used as input and 100 half-overlapping frames (~ 1.5 s) were considered. Figure 1 shows the azimuth and elevation arrival angle estimated by the proposed TDE-based algorithm for the moderately reverberant (100 ms) room.

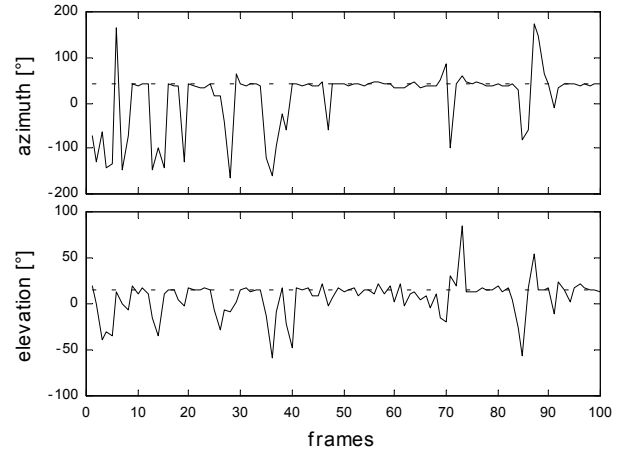


Figure 1: DOA estimates (solid) using proposed algorithm for $T_R = 100$ ms simulation (dashed – actual).

We observe that, when the estimator behaves properly, the DOA estimates are good ($\sim 5^\circ$ error), while when the algorithm breaks down (roughly 50% of the time), its DOA estimates impulsively jump (sometimes to the extremes, *i.e.*, $\pm 180^\circ$ for azimuth and $\pm 90^\circ$ for elevation).

The DOA error rate as a function of RMS error threshold, is shown in Figure 2. The DOA error rate is defined as the fraction of DOA estimates that exceeds the corresponding RMS error threshold. For $T_R = 0$, the TDE-based method outperforms the SRP-PHAT algorithm. Under the 200 ms reverberant condition all algorithms break down (*i.e.*, error rates exhibit affine behavior), including SRP-PHAT. For the moderately reverberant room ($T_R = 100$ ms), the WTDE-PHAT method provides some improvement over the non-weighted TDE-PHAT method, while SRP-PHAT now outperforms all of the TDE-based algorithms. The TDE method without use of PHAT weighting is in break-down already.

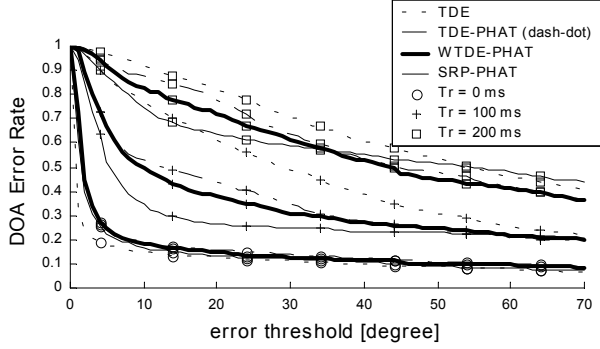


Figure 2: Simulation DOA error rates with speech as test source (256-sample frame size).

3.2. Experimental Results

Experiments were conducted in our lab in a small conference room of the same dimensions as the room in the simulations. The source and array locations and dimensions were also fixed to be the same as in the simulations. The only difference was that the conference room was not an empty room. The measured reverberation time of the room was 90 ms. Figure 3 shows the experimental results for the same speech signals as used in the simulation.

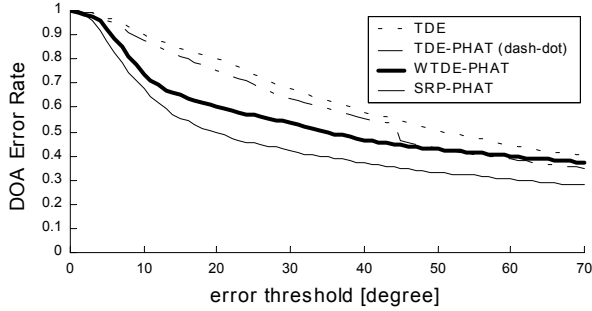


Figure 3: Experimental results with speech as test source.

While there is degradation in performance relative to the simulation results, the same relative performance of the different methods is observed.

4. DOA ESTIMATE IMPROVEMENT BY ROBUST UNIT-NORM ADAPTION

To mitigate the impulsive DOA estimation errors observed in Figure 1, a robust adaptation of DOA estimates is proposed. The numerically efficient unit-norm constrained adaptation [10] is adopted for updating the previous frame DOA estimate using the TDE obtained from the current frame. Furthermore, to reduce anomalous TDE (which cause impulsive DOA errors), a robust TDE outlier detection procedure is used.

4.1. Procedure

To maximize the probability of valid TDE data, two conditions are introduced. First, the signal power must be greater than some preset threshold value. Second, at least four of the diagonal elements of \mathbf{Q}_2 must be non-zero. If

both conditions are met, the adaptation process proceeds.

To reduce the effects of erroneous TDE, each TDE is median filtered with its previous estimates. For length N_{med} median filter, the median filtered TDE at the k^{th} frame, $\hat{\mathbf{t}}_{med,k}$, is

$$\hat{\mathbf{t}}_{med,k} = \text{median}(\hat{\mathbf{t}}_{k-N_{med}+1}, \hat{\mathbf{t}}_{k-N_{med}+2}, \dots, \hat{\mathbf{t}}_k) \quad (11)$$

where $\hat{\mathbf{t}}_k$ is the TDE vector obtained from the k^{th} frame, and the median operation is performed for each element of $\hat{\mathbf{t}}_k$. This follows the exclusion of the (median-filtered) TDE that correspond to excessive discrepancy with the previous DOA estimates. The error between the current TDE and its corresponding value based on the previous DOA is determined by

$$\mathbf{e}_k = \hat{\mathbf{t}}_{med,k} - \mathbf{v}^{-1} \mathbf{A} \hat{\mathbf{d}}_{k-1} \quad (12)$$

where $\hat{\mathbf{d}}_{k-1}$ is the estimated unit-norm DOA vector for the $(k-1)^{\text{th}}$ frame. The TDE error \mathbf{e}_k is the basis of the following outlier detection step. The i^{th} TDE is accepted if

$$|e_{k,i} - \text{median}(\mathbf{e}_k)| < 2 \text{MAD}(\mathbf{e}_k) \quad (13)$$

where $\text{MAD}(\bullet)$ is the median absolute deviation operator, and both median and MAD operations are applied to all elements of \mathbf{e}_k . Let $\bar{\mathbf{t}}_k$, $\bar{\mathbf{Q}}_k$, and $\bar{\mathbf{A}}$ be the respective $\hat{\mathbf{t}}_{med,k}$, \mathbf{Q}_k , and \mathbf{A} from which rows (and columns for \mathbf{Q}_k) have been eliminated that correspond to rejected TDE. The resulting weighted LS criterion is

$$J(\mathbf{d}) = \frac{1}{2} \|\bar{\mathbf{Q}}_k \bar{\mathbf{A}} \mathbf{d} - \mathbf{v} \bar{\mathbf{Q}}_k \bar{\mathbf{t}}_k\|^2 \quad (14)$$

with a unit-norm constraint: $\|\mathbf{d}\|=1$. The DOA estimate $\hat{\mathbf{d}}_{k-1}$ is updated using [10]

$$\hat{\mathbf{d}}_k = \begin{cases} (1 - \mathbf{g}^t \hat{\mathbf{d}}_{k-1}) \hat{\mathbf{d}}_{k-1} + \mathbf{g}, & \text{if } \mathbf{g}^t \hat{\mathbf{d}}_{k-1} > 0 \\ (1 - \mathbf{g}^t \hat{\mathbf{d}}_{k-1}) \hat{\mathbf{d}}_{k-1} + \|\hat{\mathbf{d}}_{k-1}\|^4 \mathbf{g}, & \text{if } \mathbf{g}^t \hat{\mathbf{d}}_{k-1} < 0 \end{cases} \quad (15)$$

where

$$\mathbf{g} = \mu \nabla J(\hat{\mathbf{d}}_{k-1}) \quad (16)$$

with step-size parameter μ . This algorithm only requires multiply-adds, is suitable for fast DSP implementation, and maintains the DOA vector norm close to one.

4.2. Simulation Results

The microphone array and source setup, as well as the 10 signal source files, are the same as in Section 3. Only moderate and high reverberation conditions are considered. The algorithm parameters are defined as $\mu = 50$, $N_{med} = 3$, and the signal power threshold $P_{th} = 10^{-0.5}$. The DOA vector is initialized as $\mathbf{d}_0 = [1, 0, 0]^T$. The number of frames used in the simulation is 200, while

maintaining half overlapping frames (total simulation duration of 3.2 s), to better study the adaptation. Note that the simulation starts after the frame signal level exceeds P_{th} . Lastly, to demonstrate the improvement achieved by the robust operations, adaptation without the robustified system of equations, *i.e.*, (14) with $\hat{\tau}_{med,k}$, \mathbf{Q}_k , and \mathbf{A} is evaluated for comparison. Figure 4 shows the simulation results in terms of DOA RMS error versus frames. The RMS error is averaged over 10 simulation runs.

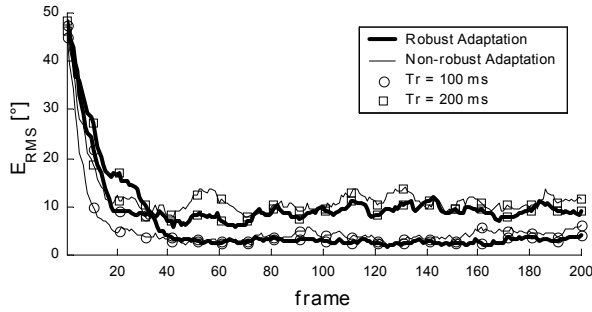


Figure 4: Average DOA RMS error of adapted estimates.

Note that the adaptation transient lasts roughly 30 frames or ~ 500 ms. For fair comparison, DOA estimation with the single-frame procedures is therefore simulated for the extended frame size of 4,096 samples or 512 ms (with 3,968 sample overlap between frames). For this large frame example, the DOA error rates of various algorithms are shown in Figure 5.

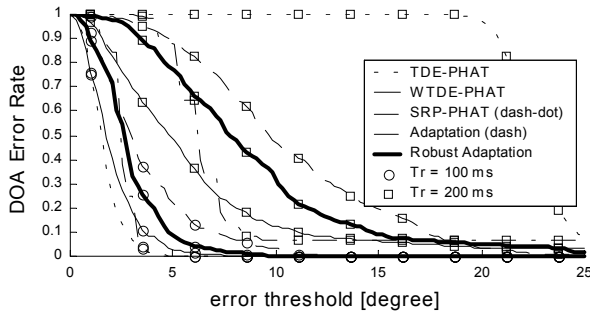


Figure 5: Simulation DOA error rates with speech as test source (4,096-sample frame size).

For the moderately reverberant environment, the single-frame TDE-PHAT method actually performs best. However, in the highly reverberant case it breaks down. We observe that WTDE-PHAT does well for moderate as well as strong reverberation environments. Up to some DOA error rate (0.15 and 0.25 for 100 and 200 ms, respectively) WTDE-PHAT outperforms SRP-PHAT with a larger percentage of its estimates localized near the true DOA. The proposed adaptation methods have a wider pdf than WTDE-PHAT, while its tails are much lighter. Moreover, the robust adaptation method clearly improves the performance of the non-robust adaptation method in both reverberation conditions (strongly for $T_R = 200$ ms).

5. CONCLUSION

We proposed improvements, named WTDE-PHAT, to the GCC-PHAT based DOA estimation algorithm for moderate reverberation environments. The GCC-PHAT based TDE are computed using the Goertzel algorithm, which allows performing peak searches without excessive computational expense. Our weighted-LS TDE-to-DOA-estimate conversion uses both DOA and cross-correlation dependent weighting of the various TDE. For the short-frame size setting, the simulation and experimental results show that the WTDE-PHAT algorithm performs marginally better than the regular LS TDE-PHAT algorithm, though not as well as SRP-PHAT. By incorporating unit-norm adaptation together with robust removal of outlier TDE, the steady-state performance of the robust adaptation method is far better than any single-frame method using the same frame size. Furthermore, the robust adaptive method, when compared to the single-frame method with frame size equivalent to the adaptation transient, performs slightly worse but eliminates impulsive errors (*i.e.*, shorter pdf tails). Also, the WTDE-PHAT algorithm outperforms SRP-PHAT for the larger frame size in both moderate and high reverberation conditions. The robust adaptation method is more advantageous for online estimation due to more uniformly distributed computational requirements.

REFERENCES

- [1] M. Brandstein and D. Ward, ed. Microphone Arrays: Signal Processing Techniques and Applications, Springer-Verlag, Berlin, 2001.
- [2] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. ASSP.*, vol. ASSP-24, pp. 320–327, Aug. 1976.
- [3] S. Holm, "FFT pruning applied to time domain interpolation and peak localization", *IEEE Trans. ASSP.*, vol. ASSP-35, Dec. 1987, pp. 1776-1777.
- [4] J. G. Proakis and D. G. Manolakis, Digital Signal Processing, Prentice Hall, Upper Saddle River, 1996.
- [5] M. S. Bazarara, H. D. Sherali, and C. M. Shetty, Nonlinear Programming: Theory and Algorithm, John Wiley & Sons, New York: 1993.
- [6] DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus, NIST Speech Disc 1-1.1, Oct. 1990.
- [7] P. M. Peterson, "Simulating the response of multiple microphones to a single acoustic source in a reverberant room," *J. Acoust. Soc. Amer.*, vol. 80, pp. 1527–1529, Nov. 1986.
- [8] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, pp. 943–950, Apr. 1979.
- [9] H. Kuttruff, Room Acoustics, 3rd ed., London, U. K.: Elsevier, 1991.
- [10] S. C. Douglas, S. Amari, and S.-Y. Kung, "Gradient adaptation under unit-norm constraints," in *Proc. IEEE Workshop Statistical Signal Array Processing*, Portland, OR, September 1998, pp. 144–14.