A ROBUST TIME DIFFERENCE OF ARRIVAL ESTIMATOR IN REVERBERANT ENVIRONMENTS

Jae-Mo Yang, Chang-Heon Lee, †Seungil Kim and Hong-Goo Kang

Dept. of Electrical and Electronic Eng., Yonsei Univ., Korea
†Info. Tech. Lab., LG Electronics Institute of Technology, Korea
⟨jaemo2879, leech, hgkang⟩@dsp.yonsei.ac.kr, †goodksi@lge.com

ABSTRACT

This paper proposes a robust generalized cross correlation (GCC)-based time difference of arrival (TDOA) estimator in reverberant and noisy environments. Under the assumption that the phase of cross channel power spectrum should be linear in a normal single source environment, the effect of reverberation to cross power spectrum is directly reduced by a recursive estimation method. An adaptive smoothing process is also designed to make the system more robust.

Simulation results with a two-channel microphone system verify that the proposed algorithm significantly improves the accuracy of TDOA by enhancing phase linearity in various reverberant environments. Also, the reliability of proposed system is confirmed by the experiments performed in additional background noise conditions. Finally, we confirm the superiority of the proposed algorithm from the results of real room experiments.

1. INTRODUCTION

In voice activated human-and-machine interaction systems, accurately estimated voice source location or direction can significantly improve system performance [1, 2]. The sound source localization (SSL) using microphone array has been widely studied for past twenty years [3]. The SSL algorithm is generally classified into three different approaches such as steered-beamformer, high resolution spectral estimation and time difference of arrival (TDOA) based methods [4]. Among them, the TDOA based method has been extensively investigated in practical systems because of their high accuracy and reasonable complexity [5, 6, 7]. Conventional TDOA-based SSL algorithms use generalized cross correlation (GCC) between two channel input signals [8]. With the GCC-based concept, the phase transform (PHAT) method is jointly used to improve the performance in low noise and relatively high reverberation environments [7, 9].

Conventional GCC-based SSL systems required a lot of microphones for robust estimation in background noise and reverberation environments. Especially, the reverberation is the principle obstacle of the performance degradation of the GCC-based method. The interferences caused by the reverberation result in somewhat high correlation values at the false time lags that are different from the actual TDOA of the target speaker [9]. Robust TDOA algorithms to be designed for overcoming the problem of the reverberant environments are introduced [5, 6], but their performance are not good enough when if the number of microphones are limited and it is operated in additional background noise condition.

In this paper, we focus on overcoming the problem by considering phase characteristics. We propose a robust method which directly enhances the cross power spectrum by removing the reliably estimated reverberant components. The linear phase characteristic of the cross channel power spectrum and the availability of the direct signal power are key criteria to recursively estimate the temporal reverberant component of room impulse response (RIR). To further enhance the estimation accuracy, an adaptive smoothing scheme is also proposed, which dynamically controls the estimation parameters depending on the stationarity of each signal frame.

The proposed algorithm is combined with the GCC-PHAT method, and then its performance is evaluated in artificially generated and real room environments. Simulation results show that the accuracy of the direction of arrival (DoA) information is much higher than GCC-PHAT only, $i.e.$ its estimated angle is very close to the true angle even in harsh real room environments.

2. BACKGROUND

2.1. Signal model

Assuming that signals radiated by a single source, $s(t)$, impinge on two channel microphones, each received signal can be represented by the following frequency domain formula [6, 7]:

$$X_i(\omega) = S(\omega)H_i(\omega) + N_i(\omega), \quad i = 1, 2 \ , \qquad (1)$$

where $N_i(\omega)$ is the noise sensed by the i^{th} microphone, and $H_i(\omega)$ is the transfer function of RIR between source and i^{th}

© EURASIP, 2009

microphone. $H_i(\omega)$ can be modeled as [10, 11]

$$H_1(\omega) = \alpha_0 + \sum_{k=1}^{\infty} \alpha_k e^{-j\omega\tau_{\alpha,k}},$$
$$H_2(\omega) = \beta_0 e^{-j\omega\tau_\theta} + \sum_{k=1}^{\infty} \beta_k e^{-j\omega\tau_{\beta,k}}, \quad (2)$$

where α_k and β_k are attenuation factors normally less than one, τ_θ is TDOA between two input signals, and $\tau_{\alpha,k}$, $\tau_{\beta,k}$ are time delays caused by the reverberation. The first term in each of Eq. (2) is a direct component from source to microphone while the second term is a reverberant component related to RIR.

2.2. TDOA estimation with GCC

The GCC function is represented by [8]:

$$\hat{\tau}_{GCC} = \arg\max_{\tau} \int_{-\infty}^{\infty} \frac{1}{\Psi(\omega)} G(\omega) e^{-j\omega\tau} d\omega, \quad (3)$$

where $G(\omega)$ is a cross power spectrum of two input signals and $\Psi(\omega)$ is a weighting function used to enhance estimation performance. In the PHAT method, the magnitude of the cross power spectrum,$\Psi(\omega) = |G(\omega)|$, is used as a weighting function. Though the PHAT had originated by an ad-hoc technique, it has been known to be robust for reverberant environments but with relatively low background noise level. The reason can be found from the following observation. In reverberant environment, the signal to reverberation ratio is approximately equal across for all frequencies because the power of interference caused by the reverberation is directly proportional to the target signal power [2]. Therefore, it is reasonable to give equal importance to all frequencies when reverberation is the primary cause of the problem. Recently, the GCC-PHAT is also shown to be optimal such that it can be considered as a specialized form of the maximum-likelihood (ML) TDOA estimator in a reverberant environment [7, 9].

However, our experimental results show that the performance of the GCC-PHAT degrades severely as background noise level becomes higher. Also, a prior research showed that the performance of the PHAT method is superior when the input SNR is higher than $20dB$, but its performance drops significantly when the SNR becomes low [7]. The problem can be overcome or at least reduced if we utilize the linear phase characteristics of cross-power spectrum. In the following section, a full details of novel approaches of the proposed algorithm is described.

3. THE PROPOSED FRAMEWORK

Figure 1 depicts the framework of whole TDOA estimation processing including two recursively connected sub-blocks proposed in this paper. At first, reverberation components estimated in the previous frame are subtracted from the cross power spectrum, and TDOA is estimated using the standard GCC-PHAT algorithm. The estimated TDOA information is used to re-calibrate reverberation components to be used for the next frame processing. Two sub-blocks depicted in the grey regions represent the proposed algorithm: reduction of reverberant components (Sub-block1) and modeling of reverberant components (Sub-block2).

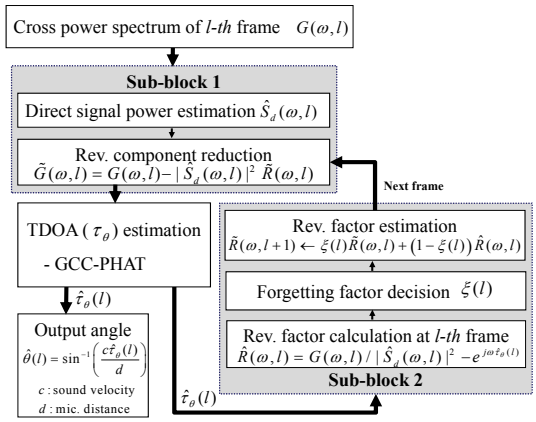

Fig. 1. Framework of reverberant component reduction.

3.1. Sub-block 1 : Reverberant component reduction

3.1.1. Reduction methodology

If background noises given in Eq. (1) are uncorrelated, the cross power spectrum at the l^{th} frame can be represented by

$$\begin{aligned} G(\omega,l) &= X_1(\omega,l)X_2^*(\omega,l) \\ &= |S_{drt}(\omega,l)|^2 \left\{e^{j\omega\tau_\theta} + R(\omega)\right\}, \end{aligned} \quad (4)$$

where the power of the direct signal component is

$$|S_{drt}(\omega,l)|^2 = \alpha_0\beta_0 |S(\omega,l)|^2, \quad (5)$$

and the reverberant component of the RIR is

$$\begin{aligned} R(\omega) = &\sum_{k=1}^{\infty} \frac{\beta_k}{\beta_0} e^{j\omega\tau_{\beta,k}} + \sum_{k=1}^{\infty} \frac{\alpha_k}{\alpha_0} e^{-j\omega(\tau_{\alpha,k}-\tau_\theta)} \\ &+ \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} \frac{\alpha_k\beta_j}{\alpha_0\beta_0} e^{-j\omega(\tau_{\alpha,k}-\tau_{\beta,j})}. \end{aligned} \quad (6)$$

If the power of the direct signal component, $|\hat{S}_d(\omega,l)|^2$, is known, and the reverberant component of the RIR, $\tilde{R}(\omega,l)$, is well estimated, the reverberant component can be simply removed by the following equation.

$$\begin{aligned} \tilde{G}(\omega,l) &= G(\omega,l) - |\hat{S}_d(\omega,l)|^2 \tilde{R}(\omega,l) \\ &\approx |S_{drt}(\omega)|^2 e^{j\omega\tau_\theta}, \end{aligned} \quad (7)$$

where we assume that two conditions, $|S_{drt}(\omega,l)|^2 \approx |\hat{S}_d(\omega,l)|^2$ and $R(\omega) \approx \tilde{R}(\omega,l)$ are satisfied. To obtain the best performance, two terms, $|\hat{S}_d(\omega,l)|^2$ and $\tilde{R}(\omega,l)$, need to be accurately estimated since both terms are highly related to TDOA

estimation. Since it is difficult to find a closed-loop solution, an adaptive method that gradually tracks the actual value is proposed in this paper. To further enhance the reliability of the estimation accuracy, the estimated values of direct signal power and the reverberant component of RIR in stationary regions are used. The reliability factor is computed by the coherence of cross power spectrum in successive frames. Detailed methods to estimate $|\hat{S}_d(\omega,l)|^2$ and $\tilde{R}(\omega,l)$ are given in next two subsections.

3.1.2. Power estimation of the direct signal component

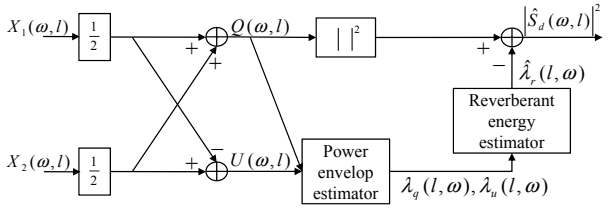

Fig. 2. GSC-based direct signal power estimation.

To estimate the power of the direct signal component, we adopt a two-channel generalized side-lobe canceller (GSC) structure [12]. Figure 2 shows a simplified block diagram to estimate the direct signal power. In this method, the power envelop of the delay-and-sum beamformer (DSB) output, $Q(\omega,l)$, and the delay-and-subtract output used for a reference signal, $U(\omega,l)$, are obtained by using first-order recursive equations:

$$\begin{aligned}\lambda_q(\omega,l) &= \gamma\lambda_q(\omega,l-1) + (1-\gamma)\left|Q(\omega,l)\right|^2,\\ \lambda_u(\omega,l) &= \gamma\lambda_u(\omega,l-1) + (1-\gamma)\left|U(\omega,l)\right|^2,\end{aligned} \tag{8}$$

where γ is a forgetting factor set close to, but less than, one. Then, the energy of reverberant residual components,$\hat{\lambda}_r(\omega,l)$, is obtained as follows :

$$\hat{\lambda}_r(\omega,l) = W(\omega,l)\lambda_u(\omega,l), \tag{9}$$

where $W(\omega,l)$ is a frequency dependent gain that is adaptively updated by using a quadratic cost function, $J_W = \{\lambda_e(\omega,l)\}^2$, where the error, $\lambda_e(\omega,l)$, is equal to $\lambda_q(\omega,l) - \hat{\lambda}_r(\omega,l)$ [12]. Finally, the direct signal power is estimated using a spectral subtraction method :

$$|\hat{S}_d(\omega,l)|^2 = |Q(\omega,l)|^2 - \hat{\lambda}_r(\omega,l). \tag{10}$$

In $Habets$'s de-reverberation method [12], a post filter is applied to the DSB output,$Q(\omega,l)$, however, the spectral subtraction method given in Eq. (10), is good enough in our application because only the power envelop of the direct signal component is needed.

3.2. Sub-block 2 : Modeling of reverberant component

The estimated TDOA, $\hat{\tau}_\theta(l)$, is recursively used to estimate the reverberant component of RIR. Generally, the TDOA is well estimated in the stationary region of speech signal, $e.g.$ vowel interval. The coherence of cross power spectrum is used as a factor of measuring the frame reliability. The smoothed reverberant component is obtained using a recursive equation with a variable forgetting factor.

3.2.1. Reverberant component at the l^{th} frame

From Eq. (4), the reverberant component of RIR can be represented by

$$\hat{R}(\omega,l) = \frac{G(\omega,l)}{\left|\hat{S}_d(\omega,l)\right|^2} - e^{j\omega\hat{\tau}_\theta(l)}, \tag{11}$$

where $\hat{\tau}_\theta(l)$ is the estimated TDOA and l is a frame index. By substituting, $G(\omega,l)$, $\hat{S}_d(\omega,l)$, and $\hat{\tau}_\theta(l)$ into Eq. (11), $\hat{R}(\omega,l)$ approximates to a reverberant component of RIR, $R(\omega)$, if the power of direct signal and TDOA are well defined.

3.2.2. Estimated reverberant component smoothing

If a speech signal is stationary over a long enough interval, the reverberant component can be directly modeled by Eq. (11) and removed by Eq. (7). However, the stationary property of normal speech signal is only guaranteed over $20 \sim 30ms$ [4], it would be better to introduce a smoothing process to the estimated reverberant component of RIR. The following first order recursive equation is adopted in the proposed approach:

$$\tilde{R}(\omega,l+1) \leftarrow \xi(l)\tilde{R}(\omega,l) + (1-\xi(l))\,\hat{R}(\omega,l). \tag{12}$$

The variable forgetting factor, $\xi(l)$, is determined by measuring the coherence of cross power spectrum in successive frames:

$$\xi(l) = \left\{ \begin{array}{cc} 1-\varepsilon\zeta(l) & \zeta(l) > threshold \\ 1 & otherwise \end{array} \right\}, \tag{13}$$

where ε, $0 < \varepsilon \ll 1$, is a positive constant set to the fluctuation range of the forgetting factor.

The normalized coherence of cross power spectrum between current and previous frames, $\zeta(l)$, is calculated as follows:

$$\zeta(l) = \frac{\sum_{m=0}^{M-1} G^*(\omega_m,l-1)G(\omega_m,l)}{\sqrt{\sum_{m=0}^{M-1}|G(\omega_m,l-1)|^2 \sum_{m=0}^{M-1}|G(\omega_m,l)|^2}}, \tag{14}$$

where M is the size of the discrete Fourier transform (DFT).The idea of proposed adaptation process is summarized as following three cases:

case 1. $\zeta(l) < threshold$: **Non-stationary region**
▷ Maintaining $\tilde{R}(\omega, l)$.
case 2. $threshold < \zeta(l) \ll 1$: **Weakly-stationary region**
▷ Use small portion of $\hat{R}(\omega, l)$ to update.
case 3. $threshold \ll \zeta(l) < 1$: **Strongly-stationary region**
▷ Use large portion of $\hat{R}(\omega, l)$ to update.

4. SIMULATION RESULTS

The proposed algorithm is implemented into a conventional GCC-PHAT estimator and its performance is directly compared. The room environment is artificially generated by the modified frequency domain image source model (ISM) with negative reflection coefficients [10]. The reverberation time, T_{60}, is measured by $Lehmann$'s energy decay curve (EDC) [10]. The level of the additive white Gaussian noise (WGN) varies from 0 to $40dB$ as the reverberation time is increased from 0 to $500msec$. The room size is set to $10 \times 7 \times 3m$, the distance between microphone and speaker is $5m$ at the front 0° direction and the distance between two microphones is $0.08m$. The sampling frequency is $8000Hz$ and $64msec$ Hamming window is applied with 50% overlap.

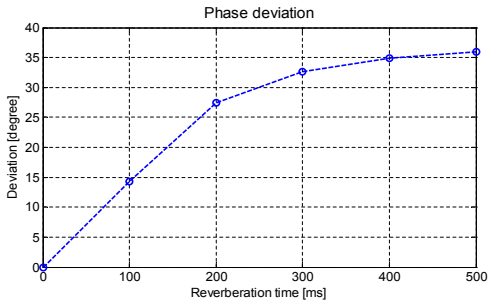

Fig. 3. Accuracy of the cross power spectrum phase in a view of phase deviation (SNR $50dB$)

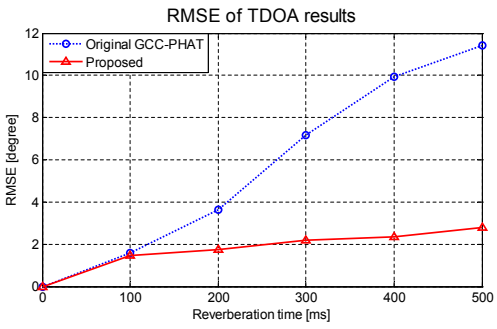

Fig. 4. Performance comparison of TDOA estimation using GCC-PHAT before and after applying the proposed processing in a reverberant environment (SNR $50dB$)

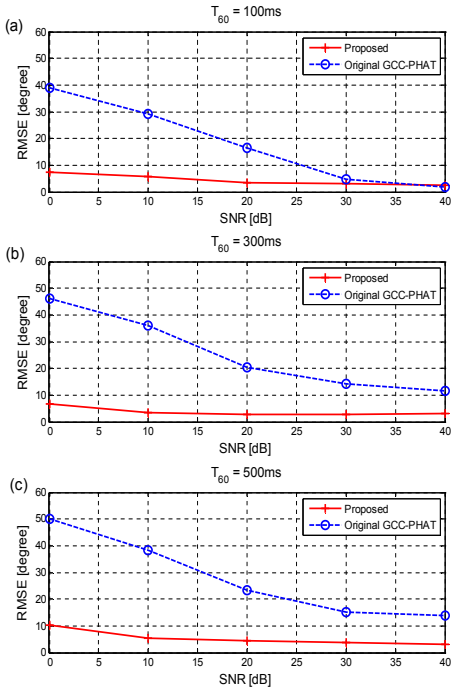

Fig. 5. Performance comparison of TDOA estimation using GCC-PHAT in a noisy and reverberant environment (a) $T_{60} = 100ms$, (b) $T_{60} = 300ms$, (c) $T_{60} = 500ms$.

Figure 3 depicts the phase deviation of the cross power spectrum in reverberation only considered environment. We convert the standard deviation of phase to the degree value (please refer to Fig. 1 : Output angle block). As shown in Fig. 3, the phase deviation increases as the reverberation time becomes longer. Figure 4 depicts the root mean squared error (RMSE) of TDOA estimation when the proposed algorithm is on and off with the GCC-PHAT in the same environment to Fig 3. Since the information of TDOA is strongly related to the phase accuracy of cross power spectrum, the performance of original GCC-PHAT marked by circles is not good in high level reverberation environment. By the proposed method, the performance of TDOA estimation is remarkably improved as the RMSE becomes less than 3°.

Figure 5 depicts the RMSE of TDOA estimation in noisy and reverberant environment. The performance of conventional GCC-PHAT only algorithm seriously degrades in low SNR environments even in short duration of reverberation. And its RMSE always becomes larger than 10 degrees in high level reverberant environments as depicted in Fig. 4(b) and 4(c). On the contrary, the performance of the proposed algorithm remarkably improves in all of the simulation environments. Especially, RMSE becomes below 10 degrees when input SNR is higher than $0dB$ in all of the reverberant environments. By the results, we conclude that the performance of the proposed algorithm is robust to harsh environment.

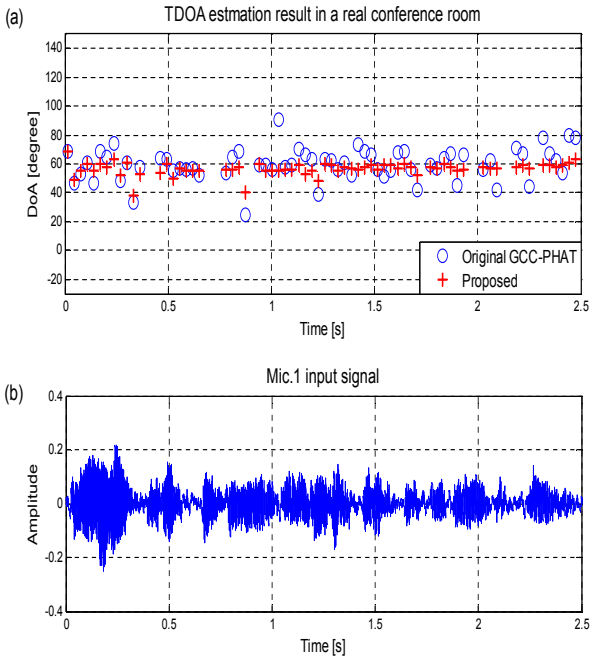

Fig. 6. Performance comparison of TDOA estimation using GCC-PHAT in a real conference room (a) DoA result, (b) Mic. 1 input signal.

5. REAL ROOM EXPERIMENT

We have also tested the performance of the proposed algorithm in a real conference room whose geometry of recoding environment is same to the synthetic room we designed in section 4. The target speaker is located at the 60 degrees from the front direction. There exist a fan noise from the ceiling and an acoustic reverberation. The average SNR including observation noise of microphone is measured as approximately $7dB$ and the reverberation time is about 0.5 second. Conversational speech was recorded in around 2.5 seconds.

Figure 6 shows the result of TDOA estimation using the GCC-PHAT algorithm. The circle marks in Fig. 6(a) show the TDOA result of the original GCC-PHAT with $32ms$ frame shift, and the cross marks depict the result after applying the proposed algorithm. As shown in the figure, the estimated DoA of the proposed algorithm converges to the true value with a very small deviation in a second.

6. CONCLUSION

In this paper, we proposed a reverberant component reduction algorithm for TDOA estimation. Under the assumption that the phase of cross power spectrum should be linear in a normal single source environment, we estimated the reverberant component of cross power spectrum of two-channel input signals. To guarantee reliable estimation, the reverberant components were adaptively smoothed based on the stationarity criterion of speech signal. From the results of the synthetic environment generated by the modified image source model (ISM), we verified that the performance of TDOA estimation was remarkably improved even under severe reverberant and noisy environments. Also, we verified the superiority of the proposed algorithm in real conference room. In the future, we are going to apply the proposed algorithm to a moving speaker. In this case, the smoothing parameters need to be appropriately controlled to follow a time varying room environment.

7. REFERENCES

[1] V.M. Trifa, A. Koene, J. Moren, and G. Cheng, "Real-time acoustic source localization in noisy environments for human-robot multimodal interaction," *Robot and Human interactive Comm.*, pp. 393–398, Aug. 2007.

[2] B. Mungamuru and P. Aarabi, "Enhanced sound localization," *Systems, Man, and Cybernetics, Part B, IEEE Trans.*, vol. 34, pp. 1526 – 1540, June 2004.

[3] M.S. Brandstein and H. Silverman, "A practical methodology for speech localization with microphone arrays," *Computer, Speech, and Language*, vol. 11, pp. 91–126, 1997.

[4] M.S. Brandstein, "A framework for speech source localization using sensor arrays," *Ph.D. thesis, Dept. Elec., Eng., Brown Univ.*, May 1995.

[5] B. Champagne, S. Bedard, and A. Stephenne, "Performance of time-delay estimation in the presence of room reverberation," *Speech and Audio Processing, IEEE Trans.*, vol. 4, pp. 148–152, March 1996.

[6] M.S. Brandstein and H.F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," *ICASSP-97.*, vol. 1, pp. 375 – 378, April 1997.

[7] Cha Zhang, D. Florencio, and Zhengyou Zhang., "Why does *phat* work well in low noise, reverberative environments?," *ICASSP2008.*, pp. 2565–2568, April 2008.

[8] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *Acoustics, Speech and Signal Processing, IEEE Trans.*, vol. 24, pp. 320–327, Aug. 1976.

[9] T. Gustafsson, Bhaskar D. Rao, and Mohan Trivedi, "Sound localization in reverberant environments : Modeling and statistical analysis," *IEEE trans. on speech and audio processing*, vol. 11, pp. 791–803, November 2003.

[10] Lehmann E.A. and Johansson A.M., "Prediction of energy decay in room impulse responses simulated with an image-source model," *J Acoust Soc Am.*, vol. 124, pp. 269–277, July 2008.

[11] Jont B. Allen and David A. Berkley, "Image method for efficiently simulating small room acoustics," *J Acoust Soc Am.*, vol. 65, pp. 943–950, April 1979.

[12] E.A.P. Habets and S. Gannot, "Dual-microphone speech dereverberation using a reference signal," *ICASSP2008.*, vol. 4, pp. 901–904, April 2008.