OPTIMAL DISTRIBUTED MICROPHONE PHASE ESTIMATION

Marek B. Trawicki & Michael T. Johnson

Marquette University
Speech and Signal Processing Laboratory
P.O. Box 1881
Milwaukee, WI 53201-1881
{marek.trawicki, mike.johnson}@marquette.edu

ABSTRACT

This paper presents a minimum mean-square error spectral phase estimator for speech enhancement in the distributed multiple microphone scenario. The estimator uses Gaussian models for both the speech and noise priors under the assumption of a diffuse incoherent noise field representing ambient noise in a widely dispersed microphone configuration. Experiments demonstrate significant benefits of using the optimal multichannel phase estimator as compared to the noisy phase of a reference channel.

Index Terms— *Acoustic arrays, speech enhancement, amplitude estimation, phase estimation, parameter estimation*

I. INTRODUCTION

For tasks such as speech enhancement and speech recognition, multiple microphone channels can give substantial improvements in SNR/SSNR and recognition accuracy. Most prior research in this area has focused on microphone array configurations, where microphone elements have small and tightly-controlled aperture spacings. This type of configuration leads to solutions such as standard beamforming approaches or other signal combination methods, assuming noise coherence across channels [1-6].

Distributed microphones scenarios, where microphone elements are widely dispersed to give broad acoustic coverage over a region, have not yet received nearly the same level of attention. Many practical task domains fall into this category, including environments such as large offices and conference rooms, broadcast stations, control rooms, airports, etc. In distributed configurations, microphone array assumptions are no longer valid and ambient noise is incoherent across the channels. By using the magnitude-squared coherence function $C_{ij}(f)$ [7] to approximate correlation as a function of frequency and space, the diffuse noise field assumption [8] representing incoherent noise ($C_{ij} < 0.1$) is appropriate for speech frequencies and microphone spacings above about 14 cm.

This work presents an optimal estimator for the source signal spectral phase using a minimum mean-square error criterion. Fundamentally, the work can be viewed as a multichannel extension of the Ephraim Malah single channel estimator [9, 10]. Spectral amplitude estimation is also given in this work, which is similar to the work of Lotter et. al. [11] but reformulated to provide an estimate of the true source signal amplitude rather than the separate estimates of the spectral amplitude at each individual microphone. The phase estimation component introduced here has not been derived previously and leads to a substantially improved estimate of the source phase in multiple channel configurations as well as to a substantially improved overall signal enhancement.

The remainder of this paper is organized into the following sections: system and statistical models (Section II), spectral amplitude estimation (Section III), spectral phase estimation (Section IV), experiments and implementation (Section V), experimental results (Section VI), and conclusion (Section VII).

II. SYSTEM AND MODELS

The time domain additive noise model in the multichannel domain is

$$y_i(t) = c_i s(t - \tau_i) + n_i(t), \quad (1)$$

where $s(t)$ is the true, spatially stationary source signal, τ_i represent signal delay at each channel $i \in [1...M]$, $n_i(t)$ is the incoherent per channel noise, and $c_i \in [0,1]$ are physical attenuation factors. With incoherent noise, signals can be easily aligned through cross-correlation methods without affecting the model so the delay terms τ_i can be dropped. Therefore, the frequency domain model is given as

$$\begin{aligned} Y_i(\lambda,k) &= c_i S(\lambda,k) + N_i(\lambda,k) \\ R_i(\lambda,k) e^{j\vartheta_i(\lambda,k)} &= c_i A(\lambda,k) e^{j\alpha(\lambda,k)} + N_i(\lambda,k), \quad (2) \\ R_i e^{j\vartheta_i} &= c_i A e^{j\alpha} + N_i \end{aligned}$$

where λ and k represent the frame and frequency bin for each microphone i.

Gaussian models are assumed for both the speech prior likelihood of the form

$$p\left(A,\alpha\right)=\frac{A}{\pi\sigma_s^2}\exp\left(-\frac{A^2}{\sigma_s^2}\right) \tag{3}$$

and

$$p\left(Y_i\left|A,\alpha\right.\right)=\frac{1}{\pi\sigma_{N_i}^2}\exp\left(-\frac{\left|Y_i-c_iAe^{j\alpha}\right|^2}{\sigma_{N_i}^2}\right), \tag{4}$$

where σ_S^2 and $\sigma_{N_i}^2$ are the speech and noise spectral variances. Under the diffuse noise field assumption, the noises are independent at each channel so the conditional joint distribution of the noisy spectral coefficients is a product of the independent spectral components

$$\begin{aligned}p\left(Y_1,...,Y_M\left|A,\alpha\right.\right)&=\prod_{i=1}^{M}p\left(Y_i\left|A,\alpha\right.\right)\\&=\prod_{i=1}^{M}\frac{1}{\pi\sigma_{N_i}^2}\exp\left(-\sum_{i=1}^{M}\frac{\left|Y_i-c_iAe^{j\alpha}\right|^2}{\sigma_{N_i}^2}\right).\end{aligned} \tag{5}$$

III. SPECTRAL AMPLITUDE

From the above statistical models and following a similar approach as in Lotter et. al. [11], the minimum mean-square error estimate of the true source spectral amplitude is given as

$$\begin{aligned}\hat{A}_{\mathrm{STSA}}&=\Gamma\left(1.5\right)\left(\frac{\sigma_S^2}{1+\sum_{i=1}^{M}\xi_i}\right)^{\frac{1}{2}}\exp\left(-\frac{v}{2}\right)\\&\times\left[\left(1+v\right)I_0\left(\frac{v}{2}\right)+vI_1\left(\frac{v}{2}\right)\right]\end{aligned} \tag{6}$$

with

$$v=\frac{\left|\sum_{i=1}^{M}\sqrt{\xi_i\gamma_i}\,e^{j\vartheta_i}\right|^2}{1+\sum_{i=1}^{M}\xi_i}. \tag{7}$$

The above solution estimates the true source spectral amplitude given known attenuation factors. By rescaling the attenuation factors to make $c_m=1$ at a specific reference channel, (6) reduces to the multichannel estimator [11] for estimating the spectral amplitude $\hat{A}_m$ at each microphone m. The estimator in (6) also simplifies to the single channel Ephraim Malah estimator [9] for the case of $M=1$.

IV. SPECTRAL PHASE

As discussed in Ephraim and Malah [9] regarding single channel phase estimation, the minimum mean-square error estimation of the complex exponential estimator $e^{j\hat{\alpha}}$ results in a non-unity modulus, which produces an altered and a non-optimal estimate of the spectral amplitude. To prevent the optimal phase estimator from impacting the optimal amplitude estimate, the constrained Lagrange Multiplier optimization approach is taken here to estimate the multichannel phase, where

$$\begin{aligned}&\min_{g,\rho}\ E\left[\left|e^{j\alpha}-g\right|^2\left|Y_1,...,Y_M\right.\right]+\rho\left(\left|g\right|-1\right)\\&\text{subject to }\left|g\right|=1\end{aligned} \tag{8}$$

with

$$g=e^{j\hat{\alpha}}=g_R+jg_I \tag{9}$$

and ρ serving as the Lagrange multiplier. After solving this optimization, the minimum mean-square error phase estimate is

$$\hat{\alpha}=\tan^{-1}\left(\frac{g_I}{g_R}\right) \tag{10}$$

with the ratio between the real and imaginary components given by

$$\frac{g_I}{g_R}=\frac{E\left[\sin\alpha\left|Y_1,...,Y_M\right.\right]}{E\left[\cos\alpha\left|Y_1,...,Y_M\right.\right]}. \tag{11}$$

Specifically, the expectations in (11) are computed as

$$E\left[\cos\alpha\left|Y_1,...,Y_M\right.\right]\propto\cos\psi \tag{12}$$

and

$$E\left[\sin\alpha\left|Y_1,...,Y_M\right.\right]\propto\cos\theta, \tag{13}$$

where

$$\psi=\tan^{-1}\left(b/a\right) \tag{14}$$

and

$$\theta=\sin^{-1}\left(a\Big/\sqrt{a^2+b^2}\right) \tag{15}$$

with

$$a=\sum_{i=1}^{M}\frac{2c_iA}{\sigma_{N_i}^2}\mathrm{Re}\left(Y_i\right) \tag{16}$$

and

$$b=\sum_{i=1}^{M}\frac{2c_iA}{\sigma_{N_i}^2}\mathrm{Im}\left(Y_i\right). \tag{17}$$

By simplifying (11) via (12)-(13) with (14)-(17) and $A_i=c_iA$ and $\sigma_{S_i}^2=c_i^2\sigma_S^2$ per the original additive model, the optimal phase estimator is given as

$$\hat{\alpha}=\tan^{-1}\left(\frac{\sum_{i=1}^{M}\frac{\sqrt{\xi_i}}{\sigma_{N_i}}\mathrm{Im}\left(Y_i\right)}{\sum_{i=1}^{M}\frac{\sqrt{\xi_i}}{\sigma_{N_i}}\mathrm{Re}\left(Y_i\right)}\right), \tag{18}$$

which is an *a priori* SNR weighted sum of the noisy microphone observations. For a single channel case with $M = 1$, this estimator simplifies to the noisy phase.

V. EXPERIMENTS AND IMPLEMENTATION

A. Experimental Setup

Enhancement experiments were conducted using clean speech from the TIMIT [12] corpus corrupted by additive white Gaussian noise uncorrelated across the channels. For the baseline experiments shown here, unity attenuation coefficients were used to generate all data with $c_i = 1$ across all channels. Results were computed using SNR as well as SSNR, but trends in both measures were similar to each other. Thus, only SSNR results are given here.

For analysis, Hanning windowed frames of 256 samples (25.6 ms) were used with 50% overlap between the corresponding frames. Noise estimation was performed on an initial silence region consisting of 5 frames. For each channel, the decision-directed [9] smoothing approach was utilized to recursively-estimate the *a priori* SNR as

$$\begin{aligned}\hat{\xi}_i &= \frac{\sigma_{S_i}^2}{\sigma_{N_i}^2} = \frac{c_i^2 \sigma_S^2}{\sigma_{N_i}^2} \\ &= \alpha_{SNR} \hat{c}_i^2 \frac{\hat{A}^2(\lambda - 1)}{\sigma_{N_i}^2} + (1 - \alpha_{SNR}) P\left[\gamma_i(\lambda) - 1\right]\end{aligned} \tag{19}$$

with the *a posteriori* SNR calculated as

$$\gamma_i = \frac{R_i^2}{\sigma_{N_i}^2}. \tag{20}$$

The smoothing factor was chosen as $\alpha_{SNR} = 0.98$ with thresholds of $\xi_{\min} = 10^{-25/10}$ dB and $\gamma_{\max} = 40$ dB.

B. Attenuation Factor Estimation

For estimating attenuation factors, an arbitrary reference microphone is selected as $c_1 = 1$. Given this assumption, the remaining attenuation factors are directly estimated using the signal powers of the noisy observations across the entire utterance as

$$\hat{c}_i = \frac{\sqrt{\sigma_{y_i}^2 - \sigma_{n_i}^2}}{\sigma_s} = \frac{\sqrt{\sigma_{y_i}^2 - \sigma_{n_i}^2}}{\sqrt{\sigma_{y_1}^2 - \sigma_{n_1}^2}}. \tag{21}$$

VI. EXPERIMENTAL RESULTS

To evaluate the importance of phase estimation, SSNR improvements using the multichannel STSA (6) and phase (18) estimators were compared to SSNR improvements obtained using the multichannel STSA estimator with the noisy phase of the reference channel. Enhancement results were averaged over 10 trial runs for the unity attenuation factor configuration as a function of increasing number of microphone channels.

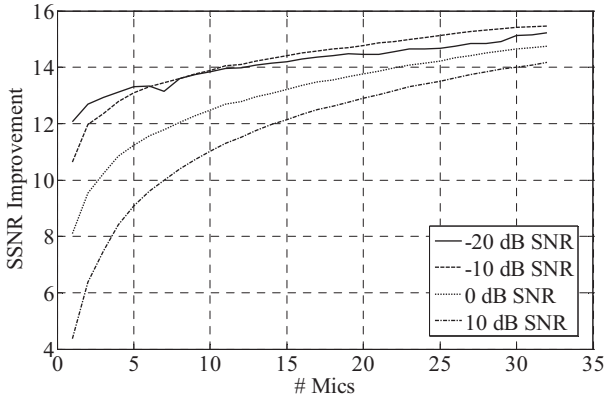

Figure 1 SSNR Improvement

Figure 1 illustrates the overall enhancement from the multichannel STSA and phase estimators. Since the result for $M = 1$ is equivalent to the standard Ephraim Malah STSA filter, improvement versus the single channel case can be easily seen by comparison to the leftmost value in each curve. As can be seen in the figure, there is substantial improvement for all input SNR levels, increasing approximately logarithmically with the number of microphones. In this configuration with unity attenuation factors, all microphones contribute equal information to the enhancement process and the improvement does not asymptote but rather continues to increase with addition of more microphones. Depending on the attenuation factor decay across microphones, other configurations have similar trends but with more slowly increasing or asymptotic performance gains.

Although overall SSNR improvement is highest for the noisiest cases, the net improvement as compared to the single channel case is greatest for the less noisy conditions with the overall improvement slowly converging for an increase in number of microphones.

Figure 2 shows the specific benefit resulting from the new multichannel phase estimator, plotting the net differential between enhancement using multichannel STSA and phase estimators and enhancement using multichannel STSA estimator but with the noisy reference channel phase.

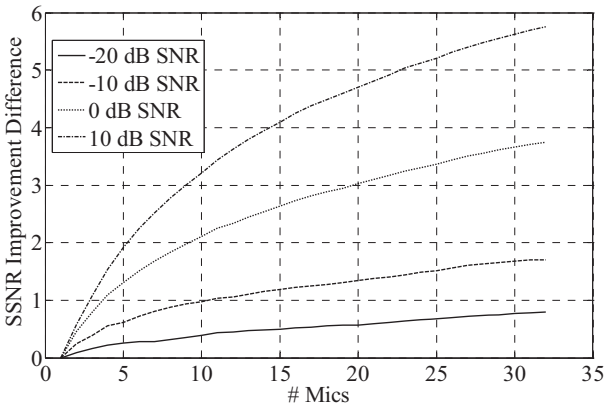

Figure 2 SSNR Improvement of Phase Estimation over Noisy Reference Channel Noisy

The results using the newly derived phase estimator exceed the results using noisy reference channel phase by a substantial margin. In the noisiest case (-20 dB SNR) the benefit is less pronounced, gaining less than 1 dB in the 32 microphone case, whereas in the least noisy case (+10 dB SNR) the gain is quite pronounced, reaching about 5.8 dB at 32 microphones. As with the overall enhancement results, the benefit due to using the multichannel phase estimator does not asymptote but continues to increase with additional microphones.

VII. CONCLUSION

In this work, a minimum mean-square error phase estimator of the source signal has been derived for speech enhancement in the distributed microphone scenario. Results show significant performance gains compared to baseline approach using noisy phase from a reference channel. Based on the results for unity attenuation factors, the STSA and phase estimators improve speech quality over the STSA and standard single channel phase estimators with SSNR improvements ranging from 0.8 dB (-20 dB) to 5.8 dB (10 dB SNR) for 32 microphones.

VIII. ACKNOWLEDGEMENTS

The authors would like to thank the National Science Foundation (Grant No. IIS-0326395) and U.S. Department of Education (GAANN Grant P200A010104) for supporting this work and Thomas Lotter and Christian Benien for providing invaluable insights into their multichannel speech enhancement research.

IX. REFERENCES

[1] B. D. V. Veen and K. M. Buckley, "Beamforming: A Versatile Approach to Spatial Filtering," in *IEEE ASSAP Magazine*, 1988.

[2] S. Doclo and M. Moonen, "GSVD-Based Optimal Filtering for Single and Multimicrophone Speech Enhancement," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 50, pp. 2230-2244, 2002.

[3] P. W. Shields and D. R. Campbell, "Improvements in Intelligibility of Noisy Reverberant Speech using a Binaural Sub-Band Adaptive Noise-Cancellation Processing Scheme," *Journal of American Acoustical Society*, vol. 110, pp. 3232-3242, 2001.

[4] T. H. Dat, K. Takeda, and F. Itakura, "Multichannel Speech Enhancement based on Speech Spectral Magnitude Estimation using Generalized Gamma Priori Distribution," presented at International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2006.

[5] R. Balan and J. Rosca, "Microphone Array Speech Enhancement by Bayesian Estimation of Spectral Amplitude and Phase," presented at International Conference on Acoustics, Speech, and Signal Processing, 2002.

[6] L.-H. Kim, M. Hasegawa-Johnson, and K.-M. Sung, "Generalized Optimal Multi-Microphone Speech Enhancement Using Sequential Minimum Variance Distortionless Response (MVDR) Beamforming and Postfiltering," presented at International Conference on Acoustics, Speech, and Signal Processing, 2006.

[7] M. Brandstein and D. Ward, *Microphone Arrays*. New York, NY: Springer-Verlag, 2001.

[8] I. A. McCowan, "Robust Speech Recognition using Microphone Arrays," Queensland University of Technology, 2001.

[9] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-32, pp. 1109-1121, 1984.

[10] Y. Ephraim and D. Malah, "Speech Enhancement using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, pp. 443-445, 1985.

[11] T. Lotter, C. Benien, and P. Vary, "Multichannel Direction-Independent Speech Enhancement Using Spectral Amplitude Estimation," *EURASIP Journal on Applied Signal Processing*, pp. 1147-1156, 2003.

[12] J. Garofolo, L. Lamel, and W. Fisher, "TIMIT Acoustic-Phonetic Continuous Speech Corpus." Linguistic Data Consortium, 1993.