JOINT ESTIMATION OF DOA AND SPEECH BASED ON EM BEAMFORMING

Lae-Hoon Kim[1,†], *Mark Hasegawa-Johnson,*[1] *Gerasimos Potamianos,*[2] *Vit Libal*[*]

[1] Department of Electrical and Computer Engineering, University of Illinois, Urbana, IL 61801, USA
[2] Institute of Informatics and Telecommunications, NCSR "Demokritos", Athens, 15310 Greece

ABSTRACT

In this paper, we propose a multi-microphone joint optimal estimation of the direction of arrival (DOA) and the source speech signal through newly introduced EM beamforming. This produces a posterior PDF for the DOA, based only on the reliable speech spectrum. By maximizing over the posterior PDF of the DOA, we achieve maximum *a posteriori* DOA estimation. After convergence, the estimated source spectrum through weighted sum in the Bayesian sense is a maximum likelihood estimate (MLE). This is a sufficient statistic for minimum mean square error (MMSE) optimal estimation using a subsequent single channel MMSE filter.

***Index Terms*—** optimal speech enhancement, direction of arrival, multi-microphone, EM beamforming, Bayesian beamforming.

1. INTRODUCTION

Beamforming has been a versatile technology in various fields of applications including radar, sonar, and speech enhancement. However, due to the fact that everyday speech is a non-stationary broad-band signal contaminated with reverberation and noise, the application of beamforming for speech enhancement has been a unique issue. Recently, minimum mean square error (MMSE) multi-microphone optimal speech enhancement has been theoretically shown to be achieved using minimum variance distortionless response (MVDR) beamforming, which is a maximum likelihood estimation (MLE) of the source speech [1], followed by an MMSE post-filter, under the assumption of perfect knowledge of channel responses from source position to the positions of multiple microphones [1, 2]. However, the underlying assumption is not realistic, because perfect blind channel response identification is infeasible.

Standard beamforming follows a two-step procedure: (i) direction of arrival (DOA) estimation; and (ii) beamforming based on the best estimation of DOA [3]. However, mismatch between the "real DOA" and "estimated DOA" causes challenging robustness problems. In this paper, we derive an expectation-maximization (EM) algorithm, where we consider the DOA as an invisible observation with a posterior PDF given the array measurements and source estimate in the previous EM step. Due to the fact that the speech signal is sparse in the short-time spectral domain, a large percentage of frequency bins are corrupted by the coexisting noise and reverberation. We only use the speech-dominant reliable frequency bins to update the DOA posterior. Because we avoid using the contaminated bins to estimate the DOA posterior PDF, the proposed approach provides more robust DOA estimation than the conventional methods based on the whole band. In fact, after convergence, the estimated source spectrum is the maximum likelihood estimate (MLE). This is a sufficient statistic for an MMSE estimation using the subsequent single channel MMSE filter such as the Wiener filter. The proposed EM beamforming turns out to have a similar form with the Bayesian beamformer [4], but takes only the reliably estimated frequency bins into account to update the posterior PDF. Note that through iterative updates of both estimates of source DOA and signal, we introduce an explicit link between the source signal estimate and the DOA estimate.

To validate the proposed method, we performed a simple experiment using real recordings inside the IBM smart room from four microphones arranged into a T-shape array configuration. In this setting, the distance between the microphones is 0.26 m, which means that spatial aliasing is unavoidable above 654 Hz. However, unlike the radar application, for speech inside smart rooms we often have video as a supplemental modality that can be used for DOA estimation with limited time and spatial resolution. For example, [5] reports 3D tracking of the speaker's head position every second. Therefore, we may have an acoustic viewfinder with narrow beam angles as candidate DOAs, and need only refine them to accommodate higher temporal resolution in the audio side (e.g. 16 ms frame rate). We obtained maximum *a posteriori* estimation of DOA and compared it with two baselines (GCC-PHAT and MUSIC based) by plotting DOA estimation per frame with the ground truth at every second. The proposed method produces smoothly changing maximum *a posteriori* estimation of the DOA around the ground truth, which yields better spectral enhancement at the high frequency range.

[†] Part of this work was conducted over a summer internship of Lae-Hoon Kim with the IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10593, USA, under the guidance of Gerasimos Potamianos and Vit Libal.

[*] Vit Libal is currently with the Data-Centric Technology Group, Honeywell Prague Laboratory, 14800 Prague 4, Czech Republic.

2. MULTI-MICROPHONE SPEECH ENHANCEMENT

2.1. Problem formulation

Multi-microphone speech measurements in a reverberant acoustic space with background noise can be modeled in the time domain by

$$x^i[n] = h^i_{d,\theta}[n] * s[n] + h^i_r[n] * s[n] + d^i[n] \ , \qquad (1)$$

where $x^i[n]$ is the measured signal, $h^i_{d,\theta}[n]$ and $h^i_r[n]$ denote the early room impulse response (RIR) and late RIR from the speech source to the i^{th} microphone, $s[n]$ is the source signal, $d^i[n]$ is the noise, and i varies from 1 to N, where N represents the number of microphones. Note that only $h^i_{d,\theta}[n]$ is assumed to be dependent on the DOA parameter θ. In this paper, we use the direct path response as the early RIR, as in previous research where the statistical reverberation model has been successfully employed [6]. The direct path can be modeled as

$$h^i_{d,\theta}(t) = \frac{1}{d^i_\theta}\delta(t - t^i_\theta) \ , \qquad (2)$$

where d^i_θ and t^i_θ denote the distance and time delay between the source and the i^{th} microphone, respectively. The time domain representation of (1) yields a frequency domain representation as

$$\begin{aligned} X^i[\omega] &= H^i_{d,\theta}[\omega]S[\omega] + H^i_r[\omega]S[\omega] + D^i[\omega] \\ &= H^i_{d,\theta}[\omega]S[\omega] + N^i[\omega] \ , \end{aligned} \qquad (3)$$

where $N^i[\omega]$ is the noise. Hereafter, we omit ω for convenience. Our objective is to jointly estimate the source speech and the DOA given the multi-microphone measurements together with the statistical assumptions about the source speech, late room impulse response, and the additive noise. We make the following statistical assumptions:

- The spectrum of the source speech follows a normal distribution with zero mean, i.e., $S(\omega) \sim N(0, \sigma^2_\omega)$ [7].
- The spectrum of the late room impulse response follows a normal distribution with zero mean, i.e., $\underline{H}(\omega) \sim N(\underline{0}, \Sigma_{\underline{H}(\omega)})$ [6].
- The spectrum of the additive noise follows a normal distribution with zero mean, i.e., $\underline{D}(\omega) \sim N(\underline{0}, \Sigma_{\underline{D}(\omega)})$ [7].

The underlined variables denote vectors of N variables, and Σ is an N by N covariance matrix.

2.2. Background: Optimal multi-microphone speech enhancement with known RIR

Balan and Rosca showed that MVDR beamforming is a sufficient statistic for obtaining an optimal MMSE estimate [2]:

$$E\left[f(S)\,|\underline{X}, \underline{H}\right] = E\left[f(S)\,\middle|S_{ML|\underline{X},\underline{H}}\right] \ , \qquad (4)$$

where

$$S_{ML|\underline{X},\underline{H}} = \frac{\underline{H}^*\Sigma^{-1}_{\underline{D}}\underline{X}}{\underline{H}^*\Sigma^{-1}_{\underline{D}}\underline{H}} \ , \qquad (5)$$

and $f(\cdot)$ is an arbitrary function. However, it is not a practical assumption to assume *a priori* RIR knowledge, because it is difficult to blindly estimate such RIRs, and naturally RIRs are time-varying. Nevertheless, note that MLE followed by MMSE is the way of achieving MMSE optimality in multi-microphone scenarios.

In this paper, we decompose RIRs into two parts: a deterministically treatable direct path and a statistically treatable late response. Based on this model, we develop EM beamforming as an alternative sufficient statistic to (5), with $\Sigma_{\underline{X}}$ instead of $\Sigma_{\underline{D}}$. This is a minimum power distortionless response (MPDR) beamformer, which can be shown to be the same with MVDR under the assumption of known DOA [8].

3. SUFFICIENT STATISTIC: EM BEAMFORMING

Our goal is to iteratively estimate S given the complete observation of $\underline{X}_k$ and incomplete observation θ with the updated PDF $p(\theta|\underline{X}_1, \cdots, \underline{X}_K, S)$. This situation fits well into the EM setup by introducing the functional

$$Q(S^i_k, S^{i-1}_k) = E_{p(\theta|\underline{X}_1,\cdots,\underline{X}_K,S^{i-1}_k)}[\log p(\underline{X}_k, \theta|S^i_k)], \quad (6)$$

where S^{i-1}_k and S^i_k denote the k^{th} frequency bin estimate at the previous and current step, respectively. Hereafter without the subscript k, all the terms are assumed to be in the same frequency bin. $p(\underline{X}_k, \theta|S^i_k)$ is decomposed as product of a likelihood PDF given θ and *a priori* PDF of θ.

$$p(\underline{X}, \theta|S^i) = p(\underline{X}|\theta, S^i)p(\theta|S^i), \qquad (7)$$

where the likelihood PDF is a Gaussian PDF under the assumption of (3),

$$\begin{aligned} p(\underline{X}|\theta, S^i) \quad &\sim \quad \frac{1}{\pi^N|\Sigma_{\underline{H}} + \Sigma_{\underline{D}}|S^i|^2|} \\ &\cdot \quad \text{exp}^{-[\underline{X}-\underline{H}_{d,\theta}S^i]^*[\Sigma_{\underline{H}}+\Sigma_{\underline{D}}|S^i|^2]^{-1}[\underline{X}-\underline{H}_{d,\theta}S^i]}. \end{aligned} \qquad (8)$$

In the maximization step, we estimate S^i_k of the current step:

$$\begin{aligned} \hat{S_k}^i \quad &= \quad \text{argmax}_{S^i_k} Q(S^i_k, S^{i-1}_k) \\ &\simeq \quad \int_\Theta p(\theta|\underline{X}_1, \cdots, \underline{X}_K, S^{i-1}_k) \\ &\cdot \quad \frac{\underline{H}^*_{d,k,\theta}(\Sigma_{\underline{D}_k} + \Sigma_{\underline{H}_k}|S^{i-1}_k|^2)^{-1}\underline{X}_k}{\underline{H}^*_{d,k,\theta}(\Sigma_{\underline{D}_k} + \Sigma_{\underline{H}_k}|S^{i-1}_k|^2)^{-1}\underline{H}_{d,k,\theta}} d\theta \ . \end{aligned} \qquad (9)$$

Note that we have used $|S^{i-1}_k|^2$ instead of $|S^i_k|^2$ in the covariance term in (9), which avoids solving the original intractable maximum likelihood estimation. Eq. (9) is a weighted average of MPDR beamformers, weighted by the posterior PDF of

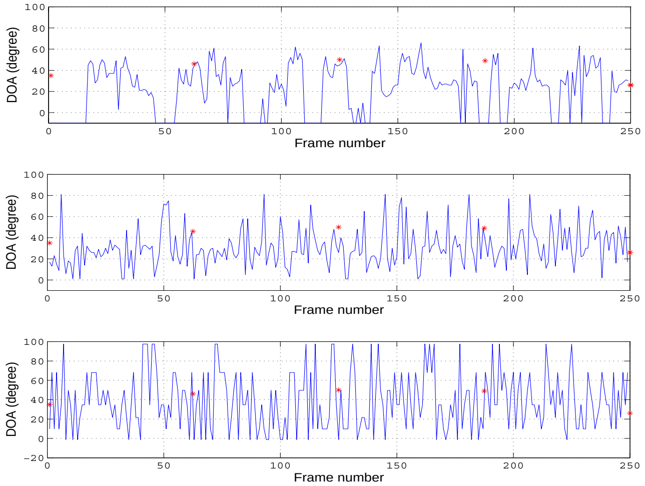

Fig. 1. DOA estimation with ground truth (every 1 sec marked as red star): Proposed method (top), MUSIC (middle), GCC-PHAT (bottom)

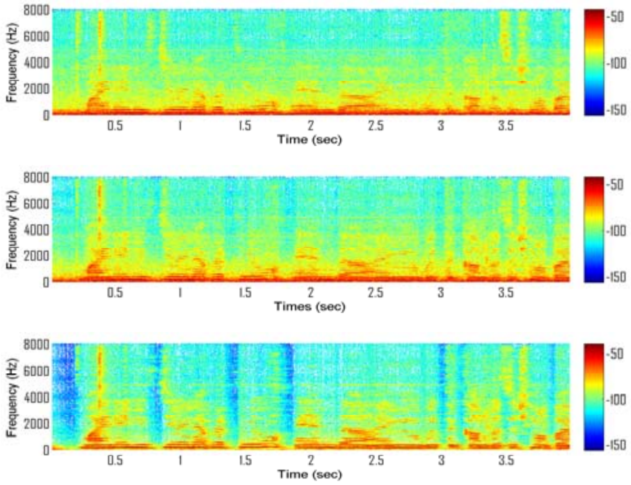

Fig. 2. Spectrogram: unprocessed channels (top), EM beamforming with one iteration (middle), EM beamforming + MMSE-logSA (bottom)

θ rather than a single specific MPDR beamformer with single DOA estimation, where the weights are given by the *a posteriori* PDF of θ. Therefore, (9) constitutes a more general formulation of the conventional MPDR beamformer. Note that if we can assume no inter-channel correlation for the noise covariance matrix, which is not achievable especially in the low frequency region, with a single DOA estimate it becomes the well-known delay-and-sum beamformer. The proposed approach is similar to the case of the previous Bayesian beamformer [4], except that we are currently taking the current estimate of the source spectrum into consideration, and we are dealing with a broad-band sparse signal. In fact, the covariance term in (8), (9) can be estimated using the sample covariance matrix $\Sigma_{\underline{X}}$. Statistical reverberation H_r plays the role of increasing the covariance from noise only to noise plus reverberation, and it also affects the *a posteriori* PDF of θ.

The *a posteriori* PDF follows Bayes' rule:

$$p(\theta|\underline{X}_1,...,\underline{X}_K,S_k^{i-1}) \sim p(\underline{X}_1,...,\underline{X}_K|\theta,S_k^{i-1})p(\theta|S_k^{i-1}), \tag{10}$$

where $p(\theta|S_k^{i-1})$ is the *a priori* PDF of θ, reflecting our prior knowledge of the possible source DOAs. In the expectation step, we calculate the likelihood

$$p(\underline{X}_1,...,\underline{X}_K|\theta,S_k^{i-1}) \sim \prod_{k=1}^{K} N(H_{d,k,\theta}S_k^{i-1},\Sigma_{\underline{D}_k}+\Sigma_{\underline{H}_k}|S_k^{i-1}|^2), \tag{11}$$

because regardless of frequency, the source should have the same DOA. However, note that because of speech sparsity, we only use the reliable bins to obtain (11) based on the flatness of posterior PDF of each frequency bin and by excluding bins having the maximum at the boundary. For probability multiplications, and to prevent underflow, we apply a heuristic rescaling per each bin. For prior $p(\theta|S^{i-1})$ we might be able to simply use real prior information about θ based on the application scenario. Note that because we are implementing the proposed algorithm in the short-time Fourier transform domain, we use the overlap-addition method to synthesis back to a time domain signal with a hanning window and 50% overlap.

4. EXPERIMENTS

In our experiment, we used the real 3 horizontal channels of the 4-microphone measurements, which come from one of the four T-shape microphone arrays of the IBM smart room [5]. 5-second long speech is arbitrarily selected with the 1 second rate of head tracking results from video data taken as the ground truth. We set the search DOA angle from 31 to 71 degrees with 0.5 degree increase based on the head tracking result. Initial estimation on speech spectrum has been performed by average MPDR based on uniform distribution among the candidate DOAs. The sampling rate was 16 kHz and the room geometry was roughly 7 m × 6 m × 3 m. The detailed configuration can be found in [5].

Fig. 1 shows the DOA tracking results every 16 ms with the ground truth marked as a red star every 1 second. Figure 2 shows the spectrograms of one of the 3 channel measurements, the output of the EM beamforming based on the result of Fig. 1, and the MMSE-logSA [7] estimate on the EM beamforming output. Finally, Fig. 3 depicts the corresponding time-domain signals.

We first observe that the DOA tracking result of the proposed method produces a reasonably smooth trajectory connecting the ground truth angles, compared to conventional methods. Instantaneous update of DOA based on MUSIC produces severe perturbations due to the unreliable covariance

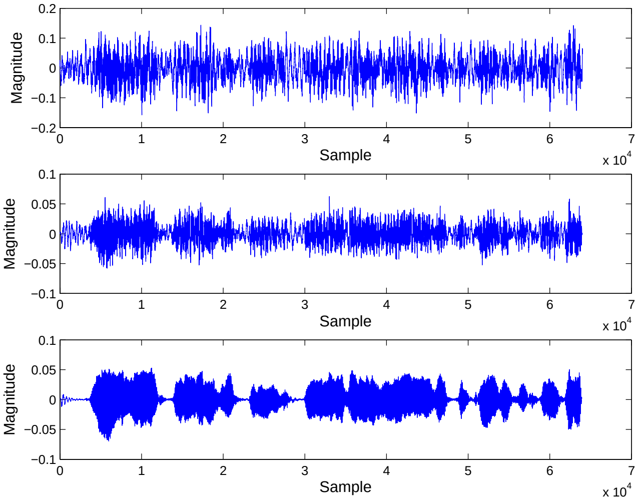

Fig. 3. Time-domain results: unprocessed channels (top), EM beamforming with 1 iteration (middle), EM beamforming + MMSE-logSA (bottom)

matrix estimation at noise-dominant frequency bins. To perform GCC-PHAT, the nearer two channels have been used. Again we observe the more severe jittering due to the same reason, namely that we do not have the ability to exclude the unreliable frequency portion in the short-time spectral representation. We do not introduce a conventional trajectory smoothing method such as Kalman filtering, but the proposed method can be combined with it easily because we can obtain the DOA observation likelihood by default.

The DOA estimate performance can be evaluated not only by smoothness of the trajectory but also by the enhanced spectrogram in the high frequency range. As we see in the spectrogram, the estimated DOA actually boosts the speech spectrum (darker yellow) up to the relatively high frequency range around 4 kHz and attenuates the region of silence (darker blue) effectively, which is ultimately further suppressed by the MMSE post-filtering. Note that the wavelength of 4 kHz is around 8.5 cm, therefore enhancement in the high frequency can only be achieved with more accurate DOA estimates. The speech spectrum has been boosted and this boosted spectra remain after the MMSE process. Note also that during non-speech regions, the DOAs have been reported as -10, which is an arbitrarily assigned number when there exist no reliable frequency bins at a specific frame, without relying on conventional voice activity detection (VAD) algorithms.

5. CONCLUSIONS AND FUTURE WORK

In this paper, joint optimal estimation of the DOA and the source spectrum is proposed through a newly derived EM beamforming algorithm. The EM beamformed spectrum constitutes a realistic sufficient statistic for MMSE post-filter. The maximum *a posteriori* PDF of DOA is obtained after convergence by taking the maximum of the updated PDF. An experiment on real data shows that the proposed algorithm represents a realistic way to achieve good speech signal estimation with reasonable instantaneous DOA estimation even at the high frequency range. The proposed method is a general approach which can be accommodated with any microphone array configuration. Combination with trajectory smoothing based on temporal dynamics will be studied in the near future.

6. REFERENCES

[1] L. Kim, M. Hasegawa-Johnson, and K. Sung, “Generalized optimal multi-microphone speech enhancement using sequential minimum variance distortionless response (MVDR) beamforming and postfiltering,” in *Proc. Int. Conf. Acoust., Speech, and Signal Process.*, 2006.

[2] R. Balan and J. Rosca, “Microphone array speech enhancement by Bayesian estimation of spectral amplitude and phase,” in *Proc. Sensor Array and Multichannel Signal Process. Works.*, 2002.

[3] B.D. Van Veen and K.M. Buckley, “Beamforming: A versatile approach to spatial filtering,” *IEEE Acoust. Speech, Signal Process. Mag.*, vol. 5, pp. 4–24, 1988.

[4] K.L. Bell, Y. Ephraim, and H.L. Van Trees, “A Bayesian approach to robust adaptive beamforming,” *IEEE Trans. Signal Process.*, vol. 48, pp. 386–398, 2000.

[5] CHIL, “Computers in the Human Interaction Loop,” [Online] Available at: `http://chil.server.de`

[6] T. Gustafsson, B.D. Rao, and M. Trivedi, “Source localization in reverberant environments: Modeling and statistical analysis,” *IEEE Trans. on Speech and Audio Process.*, vol. 11, pp. 791–803, 2003.

[7] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,” *IEEE Trans. on Acous., Speech, and Signal Process.*, vol. 33, pp. 443–445, 1985.

[8] H.L. Van Trees, *Optimum Array Processing Part IV of Detection, Estimation, and Modulation Theory,* Wiley, New York, 2002.