MICROPHONE ARRAY POST-FILTER USING INCREMENTAL BAYES LEARNING TO TRACK THE SPATIAL DISTRIBUTIONS OF SPEECH AND NOISE

Michael L. Seltzer, Ivan Tashev, and Alex Acero

Microsoft Research
Redmond, WA 98052 USA

ABSTRACT

While current post-filtering algorithms for microphone array applications can enhance beamformer output signals, they assume that the noise is either incoherent or diffuse, and make no allowances for point noise sources which may be strongly correlated across the microphones. In this paper, we present a novel post-filtering algorithm that alleviates this assumption by tracking the spatial as well as spectral distribution of the speech and noise sources present. A generative statistical model is employed to model the speech and noise sources at distinct regions in the soundfield, and incremental Bayesian learning is used to track the model parameters over time. This approach allows a post-filter derived from these parameters to effectively suppress both diffuse ambient noise and interfering point sources. The performance of the proposed approach is evaluated on multiple recordings made in a realistic office environment.

***Index Terms*—** microphone arrays, beamforming, speech enhancement, post-filtering

1. INTRODUCTION

The use of microphone arrays has been extensively studied in the literature as a means of improving the quality of sound capture in scenarios where the use of a close-talking microphone is undesirable [1]. Microphone array algorithms jointly process the signals from all microphones to create a single-channel output signal with increased directivity and thus higher SNR compared to a single microphone. The output signal can be further enhanced by the use of a single-channel *post-filter*. The post-filtering algorithms in [2–4] demonstrate that applying a post-filter to the beamformer output can result in significantly higher SNR compared to the beamformer alone.

While these post-filtering algorithms have been shown to be effective in several environments, they assume that the noise is either incoherent or diffuse, and make no allowance for point noise sources which may be strongly correlated across the microphones. In this paper, we present a novel post-filtering algorithm that removes this assumption by tracking the *spatial* as well as spectral distribution of the speech and noise sources present.

In the proposed method, the soundfield spanned by the microphone array is divided into sectors and the speech and noise in each sector are modeled by individual probability distributions. The model parameters in each sector are tracked and updated online using incremental Bayesian learning. This gives two significant benefits. First, noise sources that are spatially distinct are modeled separately and thus, more accurately. In addition, by incorporating spatial information into our model, we can distinguish between speech that comes from a desired target direction, and speech that comes from other directions, which should be treated as interference. The performance of the proposed approach is evaluated on multiple real recordings made in a realistic office environment.

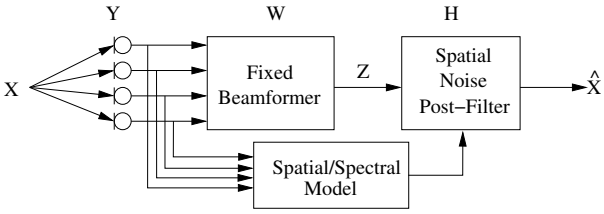

Fig. 1. Block diagram of the proposed post-filtering algorithm.

2. BEAMFORMING AND POSTFILTERING

2.1. System architecture

In this work, we perform all processing in the short-time spectral domain. We assume a speech signal $X(\omega, t)$ is captured by an array of M microphones. The M signals captured by the microphones $\mathbf{Y}(\omega, t) = \{Y_1(\omega, t), \ldots, Y_M(\omega, t)\}$ are then processed by an fixed beamformer, e.g. delay-and-sum or MVDR, to produce a single-channel output signal $Z(\omega, t)$. This output signal is then processed by an adaptive post-filter $H(\omega, t)$ to generate an enhanced output signal $\hat{X}(\omega, t)$. This paper focuses on the design of an adaptive post-filter $H(\omega, t)$ that exploits the spectral information in the beamformer output $Z(\omega, t)$ and the spatial information contained in the array signals $\mathbf{Y}(\omega, t)$. The overall architecture of the proposed approach in shown in Figure 1.

2.2. Instantaneous direction of arrival as a feature vector

For a microphone array, the phase differences at a particular frequency bin between the signals received at a pair of microphones give an indication of the *instantaneous* direction of arrival (IDOA) of a given sound source. Thus, the IDOA given by microphones i and j is computed as

$$r_{ij}(\omega, t) = \angle Y_i(\omega, t) - \angle Y_j(\omega, t) \tag{1}$$

For an array of M microphones, we can construct a vector of IDOA estimates using the phase differences of $M-1$ pairs of microphones

$$\mathbf{r}_{\omega t} = [r_{12}(\omega, t), r_{13}(\omega, t), \ldots, r_{1M}(\omega, t)]^T \tag{2}$$

We note that the presence of both ambient noise and sensor noise makes it impossible to ascertain the direction corresponding to a particular IDOA vector with absolute certainty.

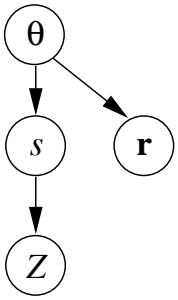

Fig. 2. The graphical model that captures the spatial distributions of speech and noise. The arrows indicate the conditional dependencies between variables.

Note that because we assume that all frequency subbands can be processed independently, the frequency variable ω is removed from all subsequent derivations for simplicity.

3. A MODEL FOR THE SPATIAL DISTRIBUTIONS OF SPEECH AND NOISE

In order to perform effective speech enhancement, we would like to model the speech and noise at all points in the working space of the array. Because this corresponds to an infinite number of locations, we quantize the sound field into a number of non-overlapping regions or *sectors*. The goal then becomes to accurately model the speech and noise in each sector.

To do so, we use the generative graphical model shown in Figure 2. In this model, a discrete random variable θ is used to represent the sector of soundfield. A distribution $p(\theta)$ indicates the prior probability of each sector in the soundfield. Associated with each sector is 1) a probability distribution $p(\mathbf{r}|\theta)$ that models the IDOA vectors in that sector, and 2) a binary random variable s that takes on one of two values associated with the speech state, i.e. s = {speech, non-speech}, also governed by an associated distribution $p(s|\theta)$. In turn, the speech states (speech and non-speech) in a given sector have associated pdfs that model the spectral observations, $p(Z|s,\theta)$.

The total likelihood of this model is given by the joint probability

$$p(Z,\mathbf{r},s,\theta) = p(Z,\mathbf{r}|s,\theta)p(s,\theta) \quad (3)$$
$$= p(Z|s,\theta)p(\mathbf{r}|\theta)p(s|\theta)p(\theta) \quad (4)$$

where we note the conditional independence of Z and $\mathbf{r}$ given θ. The distributions of the spectral observations Z and the IDOA vectors $\mathbf{r}$ are assumed to be Gaussian.

$$p(Z|s,\theta) = \mathcal{N}(Z;0,\sigma^2_{s,\theta}) \quad (5)$$
$$p(\mathbf{r}|\theta) = \mathcal{N}(\mathbf{r};\mu_\theta,\Phi_\theta) \quad (6)$$

where we have assumed the spectral observations Z are zero mean.

If we define λ to be the set of all parameters of this model, i.e. $\lambda = \{\sigma^2_{s\theta},\mu_\theta,\Phi_\theta,\ \forall\ s,\theta\}$, our goal is to estimate λ based on the spectral values observed at the beamformer output $\mathcal{Z} = \{Z_1,\ldots,Z_T\}$ and IDOA values $\mathcal{R} = \{\mathbf{r}_1,\ldots,\mathbf{r}_T\}$ derived from the array signals themselves. In the absence of any prior information about the model parameters, the optimal estimate for their values can be obtained through Maximum Likelihood (ML) parameter estimation. If some knowledge about the model parameters is available in the form of a prior distribution $p(\lambda)$, then Maximum A Posteriori (MAP) estimation can be performed. In either case, however, learning the model parameters is not straightforward because for each observation pair $\{Z_t,\mathbf{r}_t\}$, the sector θ and speech state s that generated these observations are unknown and must be inferred. Inference in hidden variable problems such as this one is typically performed using the Expectation-Maximization (EM) algorithm [5]. EM operates by iteratively maximizing the conditional log likelihood of the *complete data* (observations plus hidden variables), given the observed data.

While EM has been successfully applied in many hidden variable problems, it has the significant drawback that it is a batch-mode algorithm. This causes two related problems for adaptive speech processing algorithms. First, because it requires a sequence of frames to be accumulated before parameter estimation can be performed, it is inherently unsuitable for online applications. In addition, because the prior distribution is assumed fixed over that time period, it cannot accurately model time-varying parameters.

4. INCREMENTAL BAYES LEARNING OF THE MODEL PARAMETERS

To remedy the batch-mode processing requirement of conventional EM solutions, we employ *incremental Bayes learning* [6]. This method allows MAP estimation using EM to be performed in an online manner using a time-varying prior distribution over the model parameters. At each time step, this adaptive prior distribution is updated recursively using the posterior distribution over the hidden variables computed at the previous time step.

In order to use this approach, we first need to define a prior distribution over model parameters $p(\lambda)$. In this work, we will restrict ourselves to the online updating of the speech and noise variances, and thus need to define priors for these parameters only. It will prove mathematically convenient to model precisions (inverse variances), rather than the variances directly. Following [6], we model the precisions using gamma distributions. Thus, we define the prior distribution of $\nu_{s\theta} = 1/\sigma^2_{s\theta}$ as

$$p(\nu_{s\theta}|\phi_{s\theta}) = \nu_{s\theta}^{(\alpha_{s\theta}-\frac{1}{2})}\exp(-\beta_{s\theta}\nu_{s\theta}) \quad (7)$$

where $\phi_{s\theta} = \{\alpha_{s\theta},\beta_{s\theta}\}$ are the hyperparameters that characterize the gamma distribution for sector θ and speech state s. The prior over all model parameters λ can then be defined as

$$p(\lambda|\phi) = \prod_s\prod_\theta p(\nu_{s\theta}|\phi_{s\theta}) \quad (8)$$

We can now define the MAP EM algorithm with incremental Bayes learning. If we define $V_t = \{Z_t,\mathbf{r}_t\}$ to be the observed data at frame t we can express the time-varying EM likelihood function as

$$Q(\lambda,\lambda^{(t-1)}) = E[\log\{p(V_t,s,\theta|\lambda)\}|V_t,\lambda^{(t-1)}] \quad (9)$$
$$= \sum_s\sum_\theta \log(p(V_t,s,\theta|\lambda))p(s,\theta|V_t,\lambda^{(t-1)}) \quad (10)$$

The MAP estimate of λ at time t can be computed by combining $Q(\lambda,\lambda^{(t-1)})$ with the prior distribution $p(\lambda|\phi)$ to obtain the following EM algorithm

E-step:

$$R(\lambda,\lambda^{(t-1)}) = Q(\lambda,\lambda^{(t-1)}) + \rho\log p(\lambda|\phi^{(t-1)}) \quad (11)$$

M-step:

$$\lambda^{(t)} = \underset{\lambda}{\text{argmax}}\ R(\lambda,\lambda^{(t-1)}) \quad (12)$$

where ρ is a forgetting factor with $0 < \rho \leq 1$. The forgetting factor controls the influence of new input data relative to the past observations.

To perform the E-step in (11), we need to compute the posterior probability $\gamma_{s\theta}^{(t)} = p(s, \theta | V_t) = p(s, \theta | Z_t, \mathbf{r}_t)$. This expression can be computed exactly from the distributions in the model using Bayes rule. However, in this work, we assume for simplicity that the speech state and the sector are independent. While not strictly true, we have found this to be a reasonable assumption in practice. Thus, we can approximate the posterior as

$$\gamma_{s\theta}^{(t)} \approx p(\theta | \mathbf{r}_t) p(s | Z_t) \tag{13}$$

where $p(\theta|\mathbf{r}_t)$ can be computed from (6) using Bayes rule as

$$p(\theta | \mathbf{r}_t) = p(\mathbf{r}_t | \theta) p(\theta) / \sum_{\theta'} p(\mathbf{r}_t | \theta') p(\theta') \tag{14}$$

and the speech state posterior $p(s|Z_t)$ can be computed from a voice activity detector (VAD) that outputs probability of speech activity in each frequency bin, such as [7].

After computing $\gamma_{s\theta}^{(t)}$, the model parameters are updated by performing the M-step in (12). This is done by taking the derivative of $R(\lambda, \lambda^{(t-1)})$ with respect to λ, setting the result equal to zero,and solving for λ in the usual manner. This leads to the following update expression for the speech and noise variances in each sector.

$$\sigma_{s\theta}^{2(t)} = \frac{1}{\nu_{s\theta}^{(t)}} = \frac{2\rho\beta_{s\theta}^{(t-1)} + \gamma_{s\theta}^{(t)} |Z_t|^2}{\rho(2\alpha_{s\theta}^{(t-1)} - 1) + \gamma_{s\theta}^{(t)}} \tag{15}$$

As expected, the variances for speech and noise in all sectors are updated as a linear combination of the previously seen data (represented by the hyperparameters $\alpha_{s\theta}$ and $\beta_{s\theta}$) and the current spectral observation Z_t. However, as shown in (15), not all model parameters are updated uniformly. The observed data in the current frame will influence the model parameters in a particular sector and speech state in proportion to its posterior probability.

At each time step, the hyperparameters ϕ of the prior distribution $p(\lambda|\phi)$ are also updated using the same maximization procedure. This generates the following updated hyperparameters

$$\alpha_{s\theta}^{(t)} = \rho \left(\alpha_{s\theta}^{(t-1)} - 0.5 \right) + 0.5 + 0.5\gamma_{s\theta}^{(t)} \tag{16}$$

$$\beta_{s\theta}^{(t)} = \rho\beta_{s\theta}^{(t-1)} + 0.5\gamma_{s\theta}^{(t)} |Z_t|^2 \tag{17}$$

which define a new prior distribution $p(\lambda|\phi^{(t)})$ for the next time step.

5. CONSTRUCTING A POST-FILTER FROM THE SPATIAL DISTRIBUTIONS OF SPEECH AND NOISE

The learning algorithm described in the previous section generates online MAP estimates of the variances of the speech and noise in every frequency bin and every sector. To create a post-filter, the spatially-distinct parameter estimates are first merged into a single speech variance and a single noise variance. This is done by marginalizing the speech and noise distributions over all sectors. Thus, for frame t, the global speech variance is computed as

$$\sigma_{s=1}^{2(t)} = \sum_{\theta} \gamma_{s=1,\theta}^{(t)} \sigma_{s=1,\theta}^{2(t)} / \sum_{\theta} \gamma_{s=1,\theta}^{(t)} \tag{18}$$

Similarly, the global noise variance is computed as

$$\sigma_{s=0}^{2(t)} = \sum_{\theta} \gamma_{s=0,\theta}^{(t)} \sigma_{s=0,\theta}^{2(t)} / \sum_{\theta} \gamma_{s=0,\theta}^{(t)} \tag{19}$$

In our post-filter, we assume that the sector that contains the desired target signal is known *a priori*. We want the post-filter to suppress both noise that comes from any direction as well as speech that originates from a direction (sector) other than our target sector θ_T. In order to do so, we define $\eta^{(t)}$ as the total posterior probability that the observed signal was speech *and* came from a direction other than the desired target direction. This term can be computed as

$$\eta^{(t)} = \sum_{\theta \neq \theta_T} \gamma_{s=1,\theta}^{(t)} \tag{20}$$

Using $\eta^{(t)}$, we can compute the final noise estimate for the post-filter as

$$\sigma_N^{2(t)} = \eta^{(t)} \sigma_{s=1}^{2(t)} + (1 - \eta^{(t)}) \sigma_{s=0}^{2(t)} \tag{21}$$

Thus, if the current frame has a high probability of being either speech that originated from the target sector or noise from any sector, $\eta^{(t)}$ will be close to 0, and the noise estimate will be dominated by the noise variance $\sigma_{s=0}^{2(t)}$. On the other hand, if the posteriors indicate that the current frame is speech that originates from an interfering sector, $\eta^{(t)}$ will approach 1, and the noise estimate will be dominated by that sector's speech model.

The final noise estimate in (21) can then be used to create a post-filter using any of the gain-based suppression rules in the literature, e.g. [8]. In this work, we employ the Wiener noise suppression rule based on *a priori* SNR estimation. The *a priori* SNR is estimated as

$$\xi^{(t)} = |Z_t|^2 / \sigma_N^{2(t)} - 1 \tag{22}$$

and is used to generate the final post-filter as

$$H_t = \xi^{(t)} / (1 + \xi^{(t)}) \tag{23}$$

Of course, the decision-directed approach proposed in [8] can be used to smooth the estimates of $\xi^{(t)}$ if desired. Finally, the filter H_t is applied to the array output to generate the final output signal as

$$\hat{X}_t = H_t Z_t \tag{24}$$

6. EXPERIMENTAL EVALUATION

To evaluate the performance of the proposed post-filtering algorithm, we performed a series of experiments on microphone array data recorded in an office environment. We used a 4-element linear microphone array with a length of 190 mm. The microphones in the array are electret directional elements with a cardioid directivity pattern. Incoming audio was sampled at 16 kHz and segmented into 20 ms frames with a 10 ms overlap. The frames were then converted to the frequency domain using an MCLT [9]. The arrays signals were processed by a delay-and-sum beamformer. The output of the beamformer was then processed by the proposed post-filter.

6.1. Training the IDOA distributions

In order to compute the sector posteriors using (14), the Gaussian parameters in (6) must be estimated. To train these parameters, synthetic training data was generated using acoustic propagation principles and common noise models. In these experiments, the working area of the array spanned from -90^o to 90^o with 0^o defined to be broadside, directly in front of the array. This spatial region was divided to 18 sectors with a 10^o sector width. In each sector, 100 locations were randomly generated from a uniform distribution of positions within that sector. Each sample location is defined by its

position using a radial coordinate system, i.e. $c_l = \{\phi_l, \theta_l, \rho_l\}$. For given frequency ω, the signal gain (and delay) to each microphone m is:

$$G_m(l,\omega) = U_m(\omega, c_l)\frac{\exp(-j2\pi\omega v \parallel c_l - p_m \parallel)}{\parallel c_l - p_m \parallel} \quad (25)$$

where p_m are the microphone coordinates, $\parallel c_l - p_m \parallel$ is the Euclidean distance between the microphone and the sound source, and v is the speed of sound. $U_m(\omega, c_l)$ is a model of the microphones response obtained from acoustical measurements. To model correlated noise gain, the response of a signal with random amplitude and position in the space is modeled as :

$$NC_m(l,\omega) = \mathcal{N}(0,\psi^{-1})U_m(\omega, c_l)\frac{\exp(-j2\pi\omega v \parallel c_r - p_m \parallel)}{\parallel c_r - p_m \parallel} \quad (26)$$

where c_r is a random position in the space and ψ is the signal to correlated noise ratio. Finally, uncorrelated noise gain is modeled for each microphone as $NN_m(l,\omega) = \mathcal{N}(0,\zeta^{-1})$, where ζ is the signal to non-correlated noise ratio. The signal model for each microphone is the superposition of all three components:

$$Y_m(l,\omega) = G_m(l,\omega) + NC_m(l,\omega) + NN_m(l,\omega) \quad (27)$$

For a given value of ψ and ζ, a set of samples were generated for the microphone array and converted to IDOA vectors using (1) and (2). Multiple sets of training data were created using combinations of ψ = 20, 10 and 5 dB and ζ = 30, 25 and 20 dB. All sets were then merged together to train the means and covariances of the IDOA pdfs in (6).

6.2. Experiments

We evaluated the proposed post-filter using recordings made in three office scenarios. All recordings contained a high degree of ambient noise due to the presence of several computers and an air conditioner. In the first recording, the user is located directly in front of the array at a distance of 1 m. In the second recording, the user remained in the same position and a radio playing music was placed at -55^o. A third recording was made with two male talkers both approximately 1 m from the array, one at -20^o and the other at 40^o. The talkers alternated reading short passages and their speech did not overlap. In this evaluation, the talker at -20^o was considered the target speaker.

For each of these recordings, we compared the performance of the proposed post-filter to a conventional single channel noise suppressor, obtained by running the proposed algorithm under the assumption that the working area of the area is considered a single sector. We maintain a single model for speech and a single model for noise and the models are updated strictly on the basis of the speech state posterior generated by the VAD. Employing this noise estimation process results in a single-channel noise-suppression algorithm comparable to algorithms such as [10]. By using this algorithm, it enables us to directly evaluate the benefit of maintaining spatially distinct speech and noise models, while keeping all other aspects of the algorithms consistent. Table 1 compares the SNRs obtained at the output of the beamformer, the output of the single channel post-filter, and the output of the proposed post-filter. As the table shows, incorporating the spatial distributions of the speech and noise results in improved performance over a conventional single-channel approach.

Recording condition	DS only	DS+PF	DS+SPF
Single speaker in an office	16.2	25.8	27.5
Single speaker + off-axis radio	16.6	22.9	24.7
Single talker + off-axis talker	10.9	11.4	15.2

Table 1. SNR (dB) obtained using a delay-and-sum beamformer alone (DS), with a traditional single-channel post-filter (DS+PF) and the proposed spatial post-filter (DS+SPF).

7. CONCLUSIONS

In this paper, we have presented a novel probabilistic model that can be used to track both the spectral and spatial distributions of speech and noise using a microphone array. The parameters of the model are learned and adapted using an online implementation of the EM algorithm called incremental Bayes learning. We have shown how this model can be used to derive an adaptive post-filter that can be applied to a beamformer output. Unlike previous post-filter algorithms, the proposed model makes no assumptions about the nature of the noise, and as a result, it can accurately model and then suppress both diffuse and directional noise sources as well as interfering speech sources. The benefit of the proposed approach over a conventional single channel noise suppressor was demonstrated using real recordings made in multiple office scenarios.

8. REFERENCES

[1] M. Brandstein and D. Ward, Eds., *Microphone Arrays*, Springer, New York, 2001.

[2] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *Proc. ICASSP-88*, 1988, vol. 5, pp. 2578–2581.

[3] J. Bitzer K. U. Simmer and C. Marro, *Microphone Arrays*, chapter Post-filtering techniques, pp. 36–60, Springer, New York, 2001.

[4] I. A. McCowan and H. Boulard, "Microphone array post-filter based on noise field coherence," *IEEE Trans. Speech Audio Proc.*, vol. 11, no. 6, pp. 709–716, Nov. 2003.

[5] A. P. Dempster, N. M. Laird, D. B., and Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[6] Qiang Huo and C.-H. Lee, "On-line adaptive learning of the continuous density hidden Markov model based on approximate recursive Bayes estimate," *IEEE Trans. Speech Audio Proc.*, vol. 5, no. 2, pp. 161–172, Mar. 1997.

[7] N. S. Kim J. Sohn and W. Sung, "A statistical model-based model-based voice activity detection," *IEEE Sig. Proc. Lett.*, vol. 6, no. 1, Jan. 2000.

[8] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Sig. Proc.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[9] H. S. Malvar, "A modulated complex lapped transform and its applications to audio processing," in *Proc. ICASSP-99*, Phoeniz, AZ, Mar. 1999, pp. 1421–1424.

[10] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Sig. Proc.*, vol. 28, no. 2, pp. 137–145, Apr. 1980.