Microphone Array Post-filter based on Spatially-Correlated Noise Measurements for Distant Speech Recognition

Kenichi Kumatani[1,2]*, Bhiksha Raj*[1]*, Rita Singh*[1]*, John McDonough*[1]

[1] School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA.
[2] Disney Research, Pittsburgh (DRP), PA, USA.

Abstract

This paper presents a new microphone-array post-filtering algorithm for distant speech recognition (DSR). Conventionally, post-filtering methods assume static noise field models, and using this assumption, employ a Wiener filter mechanism for estimating the noise parameters. In contrast to this, we show how we can build the Wiener post-filter based on actual noise observations without any noise-field assumption. The algorithm is framed within a state-of-the-art beamforming technique, namely maximum negentropy (MN) beamforming with super directivity. We investigate the effectiveness of the proposed post-filter on DSR through experiments on noisy data collected in a car under different acoustic conditions. Experiments show that the new post-filtering mechanism is able to achieve up to 20% relative reduction of word error rates (WER) under the represented noise conditions, as compared to a single distant microphone. In contrast, super-directive (SD) beamforming followed by Zelinski post-filtering achieves a relative WER reduction of only up to 11%. Other post-filters evaluated perform similarly in comparison to the proposed post-filter.

Index Terms: Microphone array, Post-filter, Distant speech recognition, Automotive speech application

1. Introduction

Microphone array processing has received much attention for distant speech recognition (DSR) [1] due to the potential to relieve users from the necessity of wearing intrusive devices such as a head-set microphone. A main advantage of the microphone array against single channel techniques is that array processing can use spatial information about sound sources. The spatial directivity for a sound wave can be realized by beamforming, which is further improved by post-filtering [2, 3, 4, 5, 6]. Simmer et al. showed in [7, §3] that the optimal multi-channel filter in the sense of the minimum mean-square error (MMSE) can be decomposed into the minimum variance distortionless response (MVDR) beamformer followed by a single channel Wiener post-filter.

However, since the second-order statistics of the target and noise signals are unknown in many applications, it is difficult to realize the Wiener filter in practice. Accordingly, various methods have been developed for estimating the post-filter [2, 3, 4, 6] [7, §3]. Among those techniques, Zelinski's algorithm [2] is one of the most popular methods. It assumes that noise signals among sensors are spatially uncorrelated. However, such an assumption does not hold in some situations. Accordingly, McCowan and Bourlard used a more accurate noise field model for post-filter design. They showed in [3] that speech recognition performance can be improved by applying the diffuse noise field model to the post-filter. A generalized approach of those model-based post-filters was investigated by Lefkimmiatis and Maragos in terms of speech enhancement in [4].

In contrast to prior work, we use actual noise observations for estimating the post-filter without any assumption of the static noise field model. In order to separate the noise signal, we first find the direction of arrival (DOA) of the noise signal with the source localization method based on the maximum steering response power (SRP) [7, §8.2.1]. Then, we construct a null-steering beamformer which places a null point on the direction for the target signal and maintain the *distortionless constraint* for the noise direction. By doing so, we can extract the dominant spatially-correlated noise signal. Moreover, for clean signal estimates, we use outputs of the *super-directive maximum negentropy* (SD-MN) beamformer [8]. The SD-MN beamformer is configured in the *generalized sidelobe canceller* (GSC) structure [1, §13.3.7]. The *quiescent vector* of the SD-MN beamformer consists of the weight of the super-directive beamformer [7, §2]. The *active weight vector* is adjusted so as to achieve the maximum negentropy of the beamformer's output subject to the distortionless constraint for the *look direction*. In [8, 9], it was shown that the SD-MN beamformer is able to suppress noise sources and reverberation effects without the signal cancellation problem encountered in conventional adaptive beamforming.

We perform speech recognition experiments on real data captured with a microphone array with two sensors in a car. The effects of the post-filtering methods are investigated through a set of the DSR experiments.

The rest of this paper is organized as follows. Section 2 briefly reviews basic formulae of beamforming and post-filtering. In Section 2, we also describe representative post-filtering algorithms, Zelinslki and McCowan post-filter. Section 3 describes our post-filtering method. In Section 4, the performance of the beamforming and post-filtering methods is evaluated in terms of automatic speech recognition. In Section 5, we conclude our work and describe our future plan.

2. Beamforming with post-filtering

Let us consider a situation where a desired sound wave is propagating from a point to S microphones of an array in a noisy environment. We can denote a vector consisting of the signals observed at each sensor in the frequency domain as

$$\mathbf{X}(\omega, t) = [X_0(\omega, t), \cdots, X_s(\omega, t), \cdots, X_{S-1}(\omega, t)]^T,$$

where t indicates the frame index and ω represents the angular frequency index. We also define the *array manifold vector* as

$$\mathbf{v}(\omega, t) = \left[e^{-i\omega\tau_0}, \cdots, e^{-i\omega\tau_s}, \cdots, e^{-i\omega\tau_{S-1}}\right]^T,$$

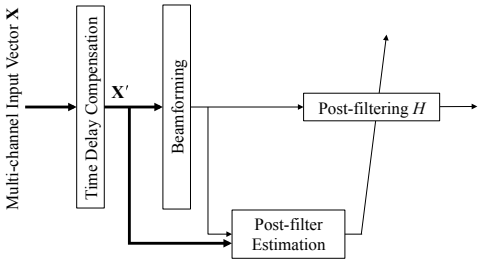

Figure 1: Block chart of a beamformer with a post-filter.

where i indicates the imaginary unit and τ_s is the time delay of arrival for each sensor s. With the desired signal $F(\omega, t)$ and additive noise vector $\mathbf{N}(\omega, t)$, the most popular model of the observed signal can be formulated as

$$\mathbf{X}(\omega, t) = \mathbf{v}(\omega, t)F(\omega, t) + \mathbf{N}(\omega, t). \quad (1)$$

Notice that the observation vector contains the delayed and attenuated replicas of the desired signal in a reverberant environment.

Under the assumption that the target and noise signals are uncorrelated, Simmer et al. in [7, §3] showed that the optimal minimum mean square error (MMSE) filter solution can be factorized into the single-channel Wiener filter and the classical minimum variance distortionless response (MVDR) beamformer. Such a solution can be expressed as

$$\mathbf{w}_{\text{MMSE}} = \left[\frac{\phi_{FF}}{\phi_{FF} + \phi_{NN}}\right] \frac{\Sigma_{\mathbf{N}}^{-1}\mathbf{v}}{\mathbf{v}^H \Sigma_{\mathbf{N}}^{-1}\mathbf{v}}, \quad (2)$$

where ϕ_{FF} is the power spectral density (PSD) of the desired signal, ϕ_{NN} is the noise PSD and $\Sigma_{\mathbf{N}}$ is the noise coherence matrix. We suppress the frequency and time indices for the sake of simplicity. The problem of the microphone post-filter is to estimate the single-channel Wiener filter in (2), that is,

$$H_{\text{opt}} = \frac{\phi_{FF}}{\phi_{FF} + \phi_{NN}}. \quad (3)$$

2.1. Zelinski post-filter

The most popular post-filter is perhaps the Zelinski method; see [2] for details of the performance analysis. Here, we briefly review the Zelinski post-filter. Figure 1 shows a block chart of the beamforming with the post-filter. As shown in Figure 1, the time delay of arrival (TDOA) for the target signal is first compensated. Those time delays can be, for example, estimated through the phase transform (PHAT) [1, §10.1][7, §8.3.2]. Once the TDOA estimates are obtained, we multiply the conjugate of each component of the array manifold vector with the input signal on the corresponding channel to obtain the time-aligned signal.

In order to derive the Zelinski post-filter, let us now describe the auto- and cross-spectral densities of the time-aligned signals at sensor m and n. Denoting the time-aligned versions of the input and noise signals as X' and N', we can compute their auto- and cross-spectral densities as

$$\begin{aligned} \phi_{X'_m X'_m} &= \phi_{FF} + \phi_{N'_m N'_m} + 2\Re\left\{\phi_{FN'_m}\right\} & (4)\\ \phi_{X'_m X'_n} &= \phi_{FF} + \phi_{N'_m N'_n} + \phi_{FN'_n} + \phi_{N'_m F}. & (5) \end{aligned}$$

Under the assumptions that:

1. the target and noise signals are uncorrelated, $\phi_{FN'_m}\ \forall m$,
2. the noise PSD is the same among all the channels, $\phi_{N'_m N'_m} = \phi_{NN}\ \forall m$, and
3. the noise signals are uncorrelated between different channels, $\phi_{N'_m N'_n} = 0\ \forall m \neq n$,

equations (4) and (5) are simplified to

$$\begin{aligned} \phi_{X'_m X'_m} &= \phi_{FF} + \phi_{NN} & (6)\\ \phi_{X'_m X'_n} &= \phi_{FF}. & (7) \end{aligned}$$

Normally, the auto- and cross-spectral densities are recursively updated at each frame [3] as

$$\begin{aligned} \hat{\phi}_{X'_m X'_n}(t) &= \alpha\hat{\phi}_{X'_m X'_n}(t-1) + (1-\alpha)\phi_{X'_m X'_n} & (8)\\ \hat{\phi}_{X'_n X'_n}(t) &= \alpha\hat{\phi}_{X'_n X'_n}(t-1) + (1-\alpha)\phi_{X'_n X'_n}, & (9) \end{aligned}$$

where α is the forgetting factor [3]. Based on substituting (8) and (9) into (3) and averaging the spectral densities over all the possible channel combinations, we obtain the Zelinski post-filter:

$$H_{\text{Z}} = \frac{\frac{2}{S(S-1)}\text{R}\left\{\sum_{m=0}^{S-2}\sum_{n=n+1}^{S-1}\hat{\phi}_{X'_m X'_n}\right\}}{\frac{1}{S}\sum_{n=0}^{S-1}\hat{\phi}_{X'_n X'_n}}, \quad (10)$$

where the use of the real operator $\text{R}\{.\}$ is justified by the fact that the PSD of the desired signal is real and positive. For the real operator, we take the absolute value since it leads to the most robust result in our preliminary experiments. It should be noted that the denominator provides an overestimate of the noise PSD at the beamformer since it is calculated with the input signals.

2.2. McCowan post-filter

Although Zelinki post-filtering has been shown to provide reasonable recognition performance in various conditions, the performance can be improved if the noise field is accurately modeled. In fact, the noise signals between different sensors are correlated in many cases. McCowan and Bourlard considered the coherence matrix of the diffuse noise field; each component of that coherence matrix can be expressed as

$$\Gamma_{mn} = \text{sinc}\left(\omega d_{mn}/c\right), \quad (11)$$

where d_{mn} is the distance between sensors m and n and c is the sound of speed. Under the assumption of the diffuse noise field, the auto- and cross-spectral densities of the time-aligned signals can be written as

$$\begin{aligned} \phi_{X'_m X'_m} &= \phi_{FF} + \phi_{NN} & (12)\\ \phi_{X'_m X'_n} &= \phi_{FF} + \Gamma_{mn}\phi_{NN}. & (13) \end{aligned}$$

The PSD of the desired signal can be then estimated as

$$\hat{\phi}_{FF}^{(mn)} = \frac{\Re\left\{\phi_{X'_m X'_n}\right\} - \frac{1}{2}\Re\left\{\Gamma_{mn}\right\}\left(\phi_{X'_m X'_m} + \phi_{X'_n X'_n}\right)}{1 - \Re\left\{\Gamma_{mn}\right\}} \quad (14)$$

in the same manner as the Zelinski's method, the denominator of the post-filter can still be estimated and auto- and cross spectral densities are recursively updated at each frame. The robustness of estimate can be improved by averaging the results over all unique sensor combinations. The resultant post-filter can be expressed as

$$H_{\text{M}} = \frac{\frac{2}{S(S-1)}\sum_{m=0}^{S-2}\sum_{n=m+1}^{S-1}\hat{\phi}_{FF}^{(mn)}}{\frac{1}{S}\sum_{n=0}^{S-1}\phi_{X'_n X'_n}}. \quad (15)$$

We refer to the post-filter designed with (15) as the McCowan post-filter. If $\Gamma_{mn} = 1$, $m \neq n$, the McCowan post-filter leads to an indeterminate solution. It is normally avoided by applying a maximum threshold on the coherence model.

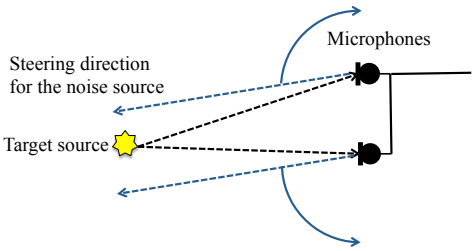

Figure 2: Schematic view of noise source localization.

3. New post-filter design algorithm based on noise observations

In this work, we use the actual noise observations separated with the null-steering beamformer for the Wiener post-filter. In order to do it, we have to localize the dominant active sources including the target and noise signals. We first find the peaks of cross-correlation values between reference and target channels through the PHAT [1, §10.1][7, §8.3.2]. In our application, speech recognition in a car, we can roughly know where the speaker is. Thus, we select the peak associated with the direction closest to a passenger's position and use the corresponding time delay of arrival (TDOA) for beamforming. Then, we steer the delay-and-sum beamformer over the areas except for the region of the target signal and seek the direction which provides the maximum response power [7, §8.2.1]. In order to prevent the desired signal into the noise estimate, we set the margin of $\pm 30^\circ$ from the direction of interest for the search space of the noise. Figure 2 shows the scheme of localizing the dominant noise source.

3.1. Separation of target signal

Given the position estimate of the target speaker, we compute the weights of the super-directive (SD) beamformer with diagonal loading as

$$\mathbf{w}_{\text{SD}} = \frac{(\mathbf{\Gamma} + \sigma \mathbf{I})^{-1}\mathbf{v}}{\mathbf{v}^H (\mathbf{\Gamma} + \sigma \mathbf{I})^{-1}\mathbf{v}}, \tag{16}$$

where each component of $\mathbf{\Gamma}$ is (11). The SD beamformer can provide the better directivity at the low frequencies than delay-and-sum beamforming and the sensitivity against mismatches between actual and theoretical conditions can be controlled by adjusting an amount of diagonal loading σ. For experiments described in Section 4, we set $\sigma = 0.01$. With the SD beamformer's weight, we build a beamformer in generalized sidelobe canceler (GSC) configuration. The output of such a GSC beamformer can be expressed as

$$Y_{\text{SDMN}} = [\mathbf{w}_{\text{SD}} - \mathbf{B}\mathbf{w}_{\text{a}}]^H \mathbf{X}. \tag{17}$$

The *blocking matrix* $\mathbf{B}$ is computed so as to satisfy the orthogonal condition $\mathbf{w}_{\text{SD}}^H \mathbf{B} = \mathbf{0}$, which implies that the target signal arriving from the look direction will not be distorted. In contrast to normal practice, we adjust the active weight vector $\mathbf{w}_{\text{a}}$ to achieve the maximum negentropy of the beamformer's output [8]. As demonstrated in [8, 9], such a beamformer can suppress interference signals as well as reverberation effects without signal cancellation encountered in traditional MVDR beamforming. The maximum negentropy beamformer is illustrated in a box with a broken line of Figure 3.

3.2. Separation of noise signal

For our microphone array post-filter, the noise signal has to be separated. It can be effectively accomplished by the null-steering beamforming technique [10]. The null-steering beamformer's weight $\mathbf{w}_{\text{null}}$ can be computed by solving the following linear equation:

$$[\mathbf{v}_N \ \mathbf{v}]^H \mathbf{w}_{\text{null}} = [1 \ 0]^T, \tag{18}$$

where $\mathbf{v}_N$ is the array manifold vector for the noise source. Equation (18) indicates that $\mathbf{w}_{\text{null}}$ nulls the target source corresponding to the array manifold vector $\mathbf{v}$ and preserves the noise signal associated with $\mathbf{v}_N$. Now, we can express the null-steering beamformer's output

$$Y_{\text{null}} = \mathbf{w}_{\text{null}}^H \mathbf{X}. \tag{19}$$

For the post-filter, we use the beamformers' outputs (17) and (19) as the target and noise signal estimates, respectively. Then, the PSDs of the target and noise signals are recursively updated at each frame t as

$$\hat{\phi}_{FF}(t) = \alpha \hat{\phi}_{FF}(t-1) + (1-\alpha)\,|Y_{\text{SDMN}}(t)|^2 \tag{20}$$

$$\hat{\phi}_{NN}(t) = \alpha \hat{\phi}_{NN}(t-1) + (1-\alpha)\,|Y_{\text{null}}(t)|^2, \tag{21}$$

where α is set to 0.6 for the experiments described later.

Upon substituting (20) and (21) into (3), we can compute the post-filter as

$$H_{\text{p}} = \frac{\hat{\phi}_{FF}(t)}{\hat{\phi}_{FF}(t) + \hat{\phi}_{NN}(t)}. \tag{22}$$

Now, with (17) and (22), we can write the final output of our beamformer and post-filter as $\hat{F} = H_{\text{p}}\, Y_{\text{SDMN}}$. Figure 3 shows our entire beamforming system. For the experiments described in Section 4, subband analysis and synthesis were performed with a uniform DFT filter bank based on the modulation of a single prototype impulse response [11], which was designed to minimize each aliasing term individually. Adaptive processing in the subband domain has the considerable advantage that the filter coefficients can be optimized for each subband independently, which provides a tremendous computational saving with respect to time-domain processing with the filters of the same length.

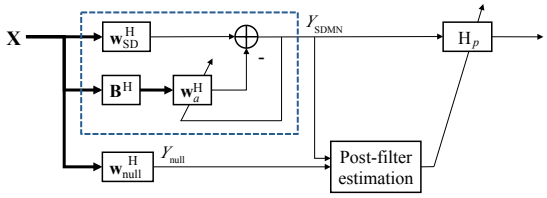

Figure 3: Block chart of our entire system.

4. Speech Recognition Experiment

Test data for speech recognition experiments were recorded with a microphone array with two sensors in a car under eight different operating conditions that were some combination of the following states: engine running in a stationary state (Idle), moving on a highway at speeds of 35 mph and 65 mph, with a fan on (Fan), turning signal on (Turn) and keeping passenger-side window open (Wind). The recording setup consisted of two microphones placed 3.8 cm apart, mounted on the passenger-side sun shield. Speakers were seated in the passenger seat beside a driver, broadside to the microphone array. The passenger seat was adjusted so that the distance between the speakers and the microphone array was approximately 25 cm. The

Post-filtering	WER (%WER)
Close-talking microphone (CTM)	14.5 %
Single distant microphone (SDM)	33.7 %
D&S BF with Zelinski PF	30.1 %
Super-directive (SD) BF + Zelinski PF	30.0 %
SD-MN BF + Zelinski PF	28.9 %
SD-MN BF + McCowan PF	29.9 %
SD-MN BF + the new PF	26.7 %

Table 1: Averages of word error rates (WERs) for each post-filtering algorithm with a matched model.

speaker additionally wore a headset mounted (close-talking) microphone. All three channels were digitized at a sampling rate of 48 kHz. The same setup was used for recording speech uttered by the multiple speakers. Each speaker read out sentences from the Wall-Street Journal-0 (WSJ0) corpus. The test data consisted of 1000 utterances from the recorded data. Distant speech recognition was performed on the data processed through the proposed algorithm (and it's comparators). For this, the CMU Sphinx-3 ASR system was used. Acoustic models were trained using the WSJ1 corpus, and the language model was trained using the WSJ1 transcriptions, with an extended 27,000 word vocabulary. The baseline acoustic models consisted of 8 Gaussian/state, the left-to-right HMMs with 6000 tied states. In order to improve robustness, acoustic models were trained on data containing digitally added noise of various types (recorded from different car states in different car types) at various SNRs ranging from -20 to 20 dB. It was found that acoustic model trained with these corrupted data provided the best performance in general for real data recorded in car-noise environments for this experiment.

Table 1 shows the averages of word error rates (WERs) over different operating conditions for each beamformer and post-filtering algorithm. As a reference, the WER obtained with a close-talking microphone (CTM) is also described. It is clear from Table 1 that the WER of 33.7 % obtained with the single distant microphone (SDM) can be reduced by any beamforming algorithm with post-filtering. It is also clear that super-directive maximum negentropy beamforming (SD-MN BF) can achieve the better recognition performance than the other traditional beamformers, delay-and-sum beamforming (D&S BF) and super-directive beamforming (SD BF). We can also see from Table 1 that our post-filter achieves the best recognition performance, the WER of 26.7 %.

We further investigate the effects of the post-filters in different acoustic conditions. Figure 4 shows the WERs obtained under the different operating conditions. As references, the WER obtained with the CTM and SDM are also depicted in Figure 4. In order to plot results of the post-filters for Figure 4, SD-MN beamforming was performed before post-filtering. It is clear from Figure 4 that our post-filter can further improve the recognition performance of the SD-MN beamformer. We consider that this is because our post-filter design method does not make any static noise field assumption which can cause a mismatch between the theoretical and actual noise fields. Notice that our post-filter can also obtain spatially-uncorrelated noise estimates and filter them out even if the assumption of the Zelinski post-fitler holds. Thus, our post-filter method is more robust than the other post-filters.

5. Conclusion

In this work, we proposed the new post-filter method based on noise measurements separated with the null-steering beamformer. We also investigated the effects of the microphone array post-filters on speech recognition. It was demonstrated through the speech recognition experiments in the car that our post-filter with maximum negentropy beamforming with super directivity could achieve the best recognition performance.

We plan to apply our post-filtering method to the larger size of the microphone array with more than two sensors.

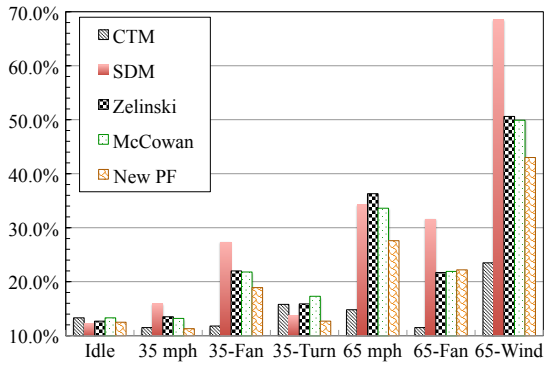

Figure 4: Word error rates for each post-filtering method in different acoustic conditions.

6. References

[1] M. Wölfel and J. McDonough, *Distant Speech Recognition*. New York: Wiley, 2009.

[2] C. Marro, Y. Mahieux, and K. U. Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering," *IEEE Trans. Speech Audio Process.*, vol. 6, pp. 240–259, 1998.

[3] I. A. McCowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Trans. Speech Audio Process.*, vol. 11, pp. 709–716, 2003.

[4] S. Lefkimmiatis and P. Maragos, "A generalized estimation approach for linear and nonlinear microphone array post-filters," *Speech Communication*, vol. 49, no. 7-8, pp. 657–666, 2007.

[5] J. McDonough, K. Kumatani, T. Gehrig, E. Stoimenov, U. Mayer, S. Schacht, M. Wölfel, and D. Klakow, "To separate speech!: A system for recognizing simultaneous speech," in *Proc. MLMI*, 2007.

[6] T. Wolff and M. Buck, "A generalized view on microphone array postfilters," in *Proc. International Workshop on Acoustic Signal Enhancement*, Tel Aviv, Israel, 2010.

[7] M. Brandstein and D. Ward, Eds., *Microphone Arrays*. Heidelberg, Germany: Springer Verlag, 2001.

[8] K. Kumatani, L. Lu, J. McDonough, A. Ghoshal, and D. Klakow, "Maximum negentropy beamforming with superdirectivity," in *European Signal Processing Conference (EUSIPCO)*, Aalborg, Denmark, 2010.

[9] K. Kumatani, J. McDonough, D. Klakow, P. N. Garner, and W. Li, "Adaptive beamforming with a maximum negentropy criterion," *IEEE Trans. Audio, Speech, and Language Processing*, August 2008.

[10] H. L. Van Trees, *Optimum Array Processing*. New York: Wiley-Interscience, 2002.

[11] K. Kumatani, J. McDonough, S. Schacht, D. Klakow, P. N. Garner, and W. Li, "Filter bank design based on minimization of individual aliasing terms for minimum mutual information subband adaptive beamforming," in *Proc. ICASSP*, Las Vegas, Nevada, U.S.A, 2008.