A SIGNAL-SEPARATION-BASED ARRAY POSTFILTER FOR DISTANT SPEECH RECOGNITION

Rita Singh[1]*, Kenichi Kumatani*[2]*, John McDonough*[1]*, Liu Chen*[3]

1. Carnegie Mellon University, Pittsburgh, PA, USA.
2. Disney Research, Pittsburgh, PA, USA.
3. Spansion Inc., Sunnyvale, CA, USA.

ABSTRACT

In standard microphone array processing for distant speech recognition, the beamformed output is *postfiltered* to reduce residual noise. Postfiltering is usually performed through a Weiner filter whose parameters are estimated from both the beamformer output and the signals captured at the microphones themselves. Conventional postfiltering methods assume diffuse or incoherent noise at the various microphones in order to estimate these parameters. When the noise does not conform to this assumption they perform poorly. We propose an alternate postfiltering mechanism that attenuates noise by estimating and separating out the contributions of speech and noise explicitly. Experiments on a corpus of in-car two-channel recordings show that the proposed postfiltering algorithm outperforms conventional postfilters significantly under many noise conditions.

Index Terms: Microphone arrays, postfiltering, beamforming, compositional models, signal separation.

1. INTRODUCTION

Microphone arrays are often used to capture speech signals in distant-speech recognition scenarios, where speech and interfering noise sources are spatially separated. Array processing aims to spatially filter incoming signals to selectively enhance signals from the target (speaker's) location through beamforming. The beamformer output, however, continues to include noise, albeit at an attenuated level. The output must be *postfiltered* to further reduce this residual noise. Conventional postfiltering approaches are, in essence, Wiener filters that attempt to estimate the clean speech signal from the potentially noisy one that emerges from the beamformer. The estimate of the power spectral density (PSD), or alternately, the autocorrelation of clean speech, that is required by the Wiener filter is derived from assumptions about the noise, *e.g.* that the noise is diffuse [1] or uncorrelated at the individual microphones [2]. The effectiveness of Wiener filters generally depends on the accuracy of the estimate of the clean speech PSD – imprecise estimates result in the attenuation of speech components along with noise [3]. This holds true for microphone array postfilters as well – if the assumptions about the noise are not correct, the estimated clean speech PSD is incorrect and filter performance degrades. Even if the assumptions are reasonable, robust estimation of filter parameters typically requires averaging over multiple microphones, and filter performance can degrade for small arrays with few microphones, for instance if there are only two microphones.

In this paper we propose an alternate mechanism that treats the problem of postfiltering as one of semi-supervised signal separation, rather than one of noise filtering. For the separation, we employ a compositional model traditionally associated with monaural source separation [4]. The technique assumes that the beamformer output is a *composition* of unit elements of speech and noise. The compositional units of the noise are estimated through an alternate channel in the beamformer, while the units of speech are learned from the beamformed signal itself. The filter uses these to suppress the contributions of noise to achieve a cleaner signal.

The proposed method has multiple advantanges over conventional postfiltering methods. Firstly, it requires no assumptions about the noise to be suppressed, since it explicitly characterizes the dominant noise. Secondly, it also requires no simplifying assumptions about noise in order to estimate clean speech PSDs. Finally, since it only employs beamformer outputs, we do not depend on averaging across microphone pairs to estimate filter parameters. Consequently, there is no direct dependence of filter performance on array size.

Speech recognition experiments on in-car stereo recordings show that the proposed method consistently outperforms other forms of postfiltering by a significant margin.

The rest of this paper is arranged as follows: In Section 2 we briefly review conventional beamforming and postfiltering methods. In Section 3 we explain our proposed compositional postfiltering method in detail. In Section 4 we present our experimental results and in Section 5 we present our conclusions.

2. BEAMFORMING AND POSTFILTERING

In distant-speech recognition scenarios, the effect of interfering noise can be particularly deleterious. Microphone array techniques are predicated on the idea that in these situations, the audio signals from the desired target speaker emanate from a specific location in space. Interfering noises are typically unlikely to arrive from exactly the same location. If the audio signals are captured using an *array* of microphones and appropriately combined, the captured audio can be *spatially filtered* to selectively enhance signals from the specific direction of the target source. The processing typically comprises two components: *beamforming*, to perform the spatial filtering, and *postfiltering*, to further suppress residual noise after beamforming.

2.1. Beamforming

Beamforming is the process of combining the signals captured by multiple microphones such that signals arriving from desired target locations in space are enhanced, while interferences from other locations are suppressed. This is usually implemented as a multi-channel filter. Let $x_i[t]$ represent the signals captured by i-th microphone, and $X_i[t, f]$ its f-th frequency component at time t. Let $\mathbf{X}[t, f]$ be a vector composed by stacking $X_1[t, f], X_2[t, f], \cdots$. The f-th frequency component $Y[t, f]$ of the array output $y[t]$ is obtained as

$$Y[t, f] = \mathbf{w}(f)^{\mathrm{H}}\mathbf{X}[t, f] \tag{1}$$

where $\mathbf{w}(f)$ is a vector of frequency-dependent array weights. The superscript H represents the Hermitian operator. The spatial response of the microphone array can be controlled by manipulating $\mathbf{w}(f)$. We will henceforth drop the explicit notation of frequency f for brevity; it is nonetheless implicit everywhere.

Most commonly, $\mathbf{w}$ is computed to achieve *minimum variance distortionless response* (MVDR), which explicitly requires that signals from the target source are not distorted, while the variance of interfering signals is minimized. If we represent the total interference arriving at the i-th microphone as $n_i[t]$, the cross-spectral density between $n_i[t]$ and $n_j[t]$ as $\Phi_{n_i n_j}$, and the coherence between $n_i[t]$ and $n_j[t]$ as Γ_{ij},

$$\Gamma_{ij} = \frac{\Phi_{n_i n_j}}{\sqrt{\Phi_{n_i n_i}\Phi_{n_j n_j}}}$$

and define $\mathbf{\Gamma}$ as the noise cross-coherence matrix with elements Γ_{ij}, MVDR beamformers are generically given by

$$\mathbf{w} = \frac{(\mathbf{\Gamma} + \sigma\mathbf{I})^{-1}\mathbf{v}}{\mathbf{v}^H(\mathbf{\Gamma} + \sigma\mathbf{I})^{-1}\mathbf{v}} \tag{2}$$

where $\mathbf{v}$ is a *array manifold vector* that accounts for phase differences in the signals from the target source that are captured by the individual microphones, due to differences in the length of the direct acoustic path from the target source. $\sigma\mathbf{I}$ is a diagonal-loading factor often introduced for stability.

The matrix $\mathbf{\Gamma}$ is unknown and usually difficult to estimate, so simplifying assumptions are often made. If we assume that the interferences at the various microphones are zero-mean and uncorrelated with one another, then $\mathbf{\Gamma} = \mathbf{I}$, the identity matrix, and we obtain the well-known "delay-and-sum" beamformer. If the interference is assumed to be from diffuse, isotropic noise, then Γ_{ij} is a *sinc* function of the distance between the i-th and j-th microphones, and the resulting beamformer is a *superdirective* beamformer.

When the noise cannot be assumed to be merely uncorrelated or isotropic, then additional processing is required. This is usually implemented as a "Generalized Sidelobe Canceller" (GSC) [5, §13], such as the one shown within the rectangular box in Figure 1. Here, in the upper branch the array signals are processed using a "fixed" distortionless-response beamformer such as the delay-and-sum or superdirective beamformers, which do not require an explicit noise estimate. In the lower channel, signals from the target source are *blocked* by a blocking matrix $\mathbf{B}$ such that only the interfering signals are allowed through. For M microphones, the blocking matrix can generate up to $M-1$ such interference-only signals. These are adaptively filtered, combined, and subtracted out from the signal in the upper branch. The adaptive array filter $\mathbf{w}_a$ is optimized for desired characteristics of the overall output $Y[t,f]$ of the beamformer. Optimization criteria include minimum variance of beamformer's outputs [5, §13.3.1], the kurtosis of the output signal [6] and its negentropy [7].

In this paper, we specifically employ a super-directive maximum-negentropy (MNES) GSC [7] shown in Figure 1, as this has been previously demonstrated to outperform other forms of beamforming for speech signals, particularly when used to enhance speech for recognition. In our experiment, we set $\sigma = 0.01$ in Equation 2 for the superdirective beamformer in the upper channel. The adaptive filter parameters in the lower channel of the GSC are optimized to maximize the negentropy of the output of the beamformer. As demonstrated in [7], such a beamformer can suppress interference signals as well as reverberation effects without the signal cancellation encountered in traditional MVDR beamforming.

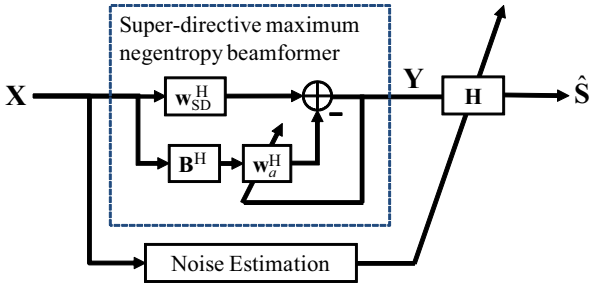

Fig. 1. Block chart of super-directive maximum negentropy beamformer with postfiltering.

2.2. Postfiltering

Even after the spatial filtering performed by the beamformer, the output of the array typically contains significant residual noise. A *postfilter* must be applied to the beamformer output to reduce the noise. The postfilter usually takes the form of a Wiener filter:

$$H(f) = \frac{\Phi_{ss}}{\Phi_{ss} + \Phi_{nn}} \tag{3}$$

The denominator of Equation 3 is the PSD of the signal to be filtered and is easily obtained. The more difficult term is the numerator, Φ_{ss}, which represents the PSD of the desired target signal $s[t]$.

It can be shown [1] that if the target signal $s[t]$ is captured identicaly at all microphones, *i.e.* that $x_j[t] = s[t] + n_j[t]$ (assuming all $x_i[t]$ are aligned appropriately), then the power spectral density of the clean signal can be estimated from the signals captured by the i-th and j-th microphones as:

$$\Phi_{ss}^{ij} = \frac{\mathcal{R}\{\Phi_{x_i x_j}\} - \frac{1}{2}\left(\Phi_{x_i x_i} + \Phi_{x_j x_j}\right)\mathcal{R}\{\Gamma_{ij}\}}{1 - \mathcal{R}\{\Gamma_{ij}\}} \tag{4}$$

where $\mathcal{R}$ is the *real* operator that extracts the real component of a complex argument. However, once again, in order to estimate this term Γ_{ij} must be known, a difficult requirement. Once again, therefore, it is obtained based on assumptions about the noises at the individual microphones. The *Zelinski* postfilter [2] assumes that the noises at the different microphones have the same power, but are incoherent, *i.e.* that $\mathbf{\Gamma} = \mathbf{I}$. The postfilter proposed by McCowan [1] assumes that the noise is diffuse and isotropic, which, as we saw earlier, implies that Γ_{ij} is a sinc function of the distance between the i-th and j-th microphones. Other such assumptions result in other postfilter formulations [8].

We note that these assumptions about the noise are identical to those made by various beamforming algorithms. The Zelinksi postfilter makes the same assumption as the delay-and-sum beamformer. The assumption of the isotropic noise field in [1] is also made by the superdirective beamformer. In effect, these postfilters can only deal with the kinds of noises that fixed beamformers (that do not require explicit knowledge of the noise) can already deal with. Moreover, unlike the beamformers which have a distortionless response constraint, the postfilter, having no such constraint can in fact *degrade* the output of the beamformer when the assumptions are inaccurate.

A second issue that arises is that of the *robustness* of the estimates of the numerator and denominator terms in Eq. 3. In order to derive robust estimates it is customary to average the estimates across all microphones and microphone pairs. For the estimates to be sufficiently accurate, a sufficient number of microphones are required. For small arrays, such as those that employ only two microphones, the estimates are often noisy and result in poor postfiltering.

3. A COMPOSITIONAL POSTFILTER

The compositional postfilter proposed in this paper takes a different approach. Instead of attempting to derive a postfilter from knowledge (or estimates) of *clean-speech* and noisy-speech parameters as in Equation 3, it treats postfiltering as a semi-supervised signal separation problem to be performed from estimated characterizations of *noise* and the noisy speech.

Before we explain exactly where the initial noise and noisy-speech estimates come from, we first briefly describe the technique employed for separation. We follow the mechanism based on probabilistic latent component analysis (PLCA) described in [4]; however, we must modify it in order to account for the fact that clean speech parameters are unknown.

3.1. Semi-supervised source separation

Consider a problem where we are given examples of a mixed signal $y[t] = s[t] + n[t]$ along with examples of one of the two components of the mixture, $n[t]$. The goal is to estimate $s[t]$ from $y[t]$. Representing the short-time Fourier transforms of $y[t]$, $s[t]$ and $n[t]$ as $Y[t,f]$, $S[t,f]$ and $N[t,f]$ respectively, we have the approximate relation $|Y[t,f]| = |S[t,f]| + |N[t,f]|$. The problem can then be restated as determining $S[t,f]$ from $Y[t,f]$.

The PLCA model treats each magnitude spectral vector $S[t] = [\ |S[t,f]|\ \forall f\ ]$ as a histogram of draws from a mixture-multinomial distribution:

$$|S[t,f]| \sim P_S(t,f) = \sum_z P_t^S(z) P_S(f|z)$$

The component multinomials $P_S(f|z)$ are assumed to be the *compositional units* that are common to all spectral vectors from $s[t]$, while the mixture weights $P_t^S(z)$ represent the manner in which the units must be combined to compose the t-th spectral vector $S[t]$. Similarly, $N[t] = [\ |N[t,f]|\ \forall f\ ]$ is assumed to be a histogram of draws from a mixture-multinomial distribution: $N[t,f] \sim P_N(t,f) = \sum_z P_t^N(z) P_N(f|z)$, where $P_N(f|z)$ are the compositional units that compose all spectral vectors of $n[t]$ and mixture weights $P_t^N(z)$ represent the manner in which these units must be combined to compose the t-th spectral vector $N[t]$.

Each spectral vector $Y[t] = [\ |Y[t,f]|\ \forall f\ ]$ of the mixed signal $y[t]$ can now directly be characterized as a histogram drawn from a distribution $|Y[t,f]| \sim P_Y(t,f)$, where

$$P_Y(t,f) = P_t(S) \sum_z P_t^S(z) P_S(f|z) + P_t(N) \sum_z P_t^N(z) P_N(f|z)$$

and $P_t(S)$ and $P_t(N)$ represent the relative proportions of $S[t]$ and $N[t]$ in $Y[t]$.

Given the component multinomials $P_S(f|z)$ and $P_N(f|z)$ and the magnitude spectrogram $|Y[t,f]|$ of the mixed signal, $P_t(S)$,$P_t(N)$,$\{P_t^S(z)\ \forall z\}$ and $\{P_t^N(z)\ \forall z\}$ can all be estimated using the EM algorithm [9]. Thereafter, $S(t,f)$ is estimated using the Wiener filter formulation:

$$\hat{S}[t,f] = Y[t,f] \otimes \frac{P_t(S) P_S(t,f)}{P_Y(t,f)}$$

The estimate for the separated signal $s[t]$ can then be obtained by computing the inverse short-time Fourier transform of $\hat{S}[t,f]$. We note that this formulation is very successful at separating mixed signals in a variety of scenarios [4, 9, 10].

The problem, of course, is that of accurately learning the component multinomials $P_S(f|z)\ \forall z$ and $P_N(f|z)\ \forall z$ for $s[t]$ and $n[t]$. We assume that we have "training" examples of $n[t]$ from which instances of $N[t,f]$ can be computed. $P_N(f|z)\ \forall z$ can be estimated from training instances of $|N[t,f]|$ using the EM algorithm [9]. In practice, it is found to be effective to just derive them by normalizing randomly drawn magnitude spectral vectors from the signal [10].

In our setting we do not have exemplars of $s[t]$ to learn $P_S(f|z)$, however. These must be learned from the magnitude spectrogram of the mixed signal, $|Y[t,f]|\ \forall t,f$ itself. This can be performed through the following iterative update rule. The superscript k represents the iteration. Note that all terms, including $P_S(f|z)$, $P_t(S)$,$P_t(N)$,$P_t^S(z)$, $P_t^N(z)$ and therefore $P_Y(t,f)$ are estimated iteratively:

$$\begin{aligned} \hat{P}_S^{k+1}(f|z) &= C^{k+1} P_S^k(f|z) \left(\sum_t \frac{|Y[t,f]| P_t^k(S) P_t^{S,k}(z)}{P_Y^k(t,f)} \right) \\ P_S^{k+1}(f|z) &= \alpha \hat{P}_S^{k+1}(f|Z) + (1-\alpha) P_S^k(f|Z) \end{aligned} \quad (5)$$

where C^{k+1} is a normalization constant. α is an update factor that controls the rate of divergence of the learned $P_S(f|z)$ terms from their initial values. For our setup, low α values are preferred.

3.2. The compositional postfilter

We now specify the design of the actual postfilter using the separation approach described previously. We assume that in the above discussion, $y[t]$ is the output of the GSC beamformer, which carries residual noise. $n[t]$ is the noise in $y[t]$ and $s[t]$ is the desired signal.

The mixed signal $y[t]$ is available from the beamformer output. However, $n[t]$ must be separately obtained. We employ a variant of the null-steering beamformer [11, §3] to estimate noise. Our implementation of the null-steering beamformer "scans" the recording environment to identify the region other than the target location from which the maximum signal energy is captured, while simultaneously also ensuring that signals from the target location are completely nulled out. This 'null-beamformer" is shown by the lower channel outside the rectangle in Figure 1. In reality, although we attempt to find the most energetic interference, the output of this channel combines all interfering noises, with the only guarantee that it does not contain the target signal.

Given $y[t]$ and the signal $n[t]$ from the null-steered beamformer, we can now apply the semi-supervised separation procedure of Section 3.1. The current implementation of our post-filter is as a batch-processing module. The entire recording $n[t]$ corresponding to any $y[t]$ is used to learn the noise multinomials, $P_N(f|z)$. The speech multinomials $P_S(f|z)$ and the time-dependent mixture weights $P_t(S)$, $P_t^S(z)$, $P_t(N)$ and $P_t^N(z)$ are all learned from $y[t]$.

Finally, a *time-varying* postfilter $H_t(f)$ is computed as:

$$H_t(f) = \frac{P_t(S) P_S(t,f)}{P_Y(t,f)} \quad (6)$$

4. EXPERIMENTAL RESULTS

The proposed method was tested on data recorded via a two-microphone array in a car under eight different operating conditions that were some combination of the following states: engine running in stationary state (Idle), moving on the highway at speeds of 35 mph and 65 mph, with the fan on (Fan), turning signal on (Turn) and passenger-side window open (Wind). The recording setup consisted

	Baselines			Post-Filtering methods				
State	Close-talking	Farfield Channel 2	MNES woPF	Leukim	McCowan	Zelinski	Comp. PF	Comp. PF Iterated
Idle	13.3	12.3	11.4	16.2	13.3	12.7	11.6	12.1
35 mph	11.5	15.9	12.1	22.0	13.2	13.5	10.8	12.2
35 mph, Fan	11.8	27.3	20.2	32.7	21.8	22.0	20.1	18.8
35 mph, Turn	15.8	13.7	12.7	22.0	17.3	15.9	12.2	13.7
35 mph, Wind	13.8	65.9	65.9	77.6	67.9	58.4	64.8	69.8
65 mph	14.8	34.3	28.4	49.7	33.6	36.3	28.0	27.6
65 mph, Fan	11.5	31.6	21.9	35.0	21.9	21.7	21.7	21.5
65 mph, Wind	23.5	68.6	46.7	63.4	49.9	50.6	45.6	43.9

Table 1. A comparison of different postfiltering methods. The table shows word error rates (in %) for recognition performed with the same set of acoustic and language models across all methods tested.

of two microphones placed 3.8 cm apart, mounted on the passenger-side sun shield. The speaker was seated in the front passenger seat, broadside to the microphone array. The distance of the speaker to the microphone array was approximately 25 cm. The speaker additionally wore a headset mounted (close-talking) microphone. All three channels were digitized at a sampling rate of 48 kHz. Speech was recorded by multiple speakers on the same setup. Each speaker read out sentences from the Wall-Street Journal-0 (WSJ0) corpus. The test data consisted of 1000 utterances from the data thus recorded.

The CMU Sphinx-3 ASR system was used for speech recognition. Acoustic models were trained on 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, 3-state HMMs with 6000 tied states. The 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 these corrupted data resulted in the best performance in general for the in-car recordings used for this experiment.

The beamforming method used was the super-directive maximum negentropy (MNES) beamformer [7] in all cases. Table 1 shows the WER % obtained under different conditions: close-talking recordings, recordings from a single distant microphone, beamforming without postfiltering, three well-known postfiltering algorithms (Zelinski [2], McCowan [1] and Leukim [8]) and with the proposed compositional postfilter. For the compositional postfiltering we used $max(3000, no\ spectral\ vectors\ in\ utt.)$ multinomial bases for the clean speech, initialized on an utterance-by-utterance basis by random selection from the magnitude spectral vectors obtained from the baseline MNES beamformed output. Correspondingly, the same number of multinomial noise bases were obtained on an utterance-by-utterance basis from the output of the null-steering beamformer, using the exemplar-based mechanism of [10]. A natural question is whether an iterative application of the compositional postfilter would result in further improvements, since the postfiltered signal still contains residual noise. This is not the case, as we see in Table 1.

5. CONCLUSIONS

The proposed postfiltering clearly outperforms other postfilters. The proposed compositional postfilter has two very desirable properties: it is adaptive (bases are different by design on a utterance-by-utterance basis), and is completely unsupervised. Unlike other postfiltering mechanisms that attempt to eliminate noise, thereby also eliminating some useful components of speech, the proposed method tends to be conservative, actively preserving speech while suppressing the noise. This tends to preserve more information in speech at the cost of keeping a little more noise than conventional methods. This reversal of balance pays off in lower speech recognition error rates. It must be noted that the compositional models used in this work are quite basic, and can be developed further in many ways. Future work will involve developing more intelligent sparsity constraints in the use of overcomplete bases representations, and better update rules for estimating the weights.

6. REFERENCES

[1] 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.

[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] R. Singh, "Compensating for denoising artifacts," in *ICASSP*, Kyoto, Japan, 2012.

[4] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semisupervised separation of sounds from single-channel mixtures," in *Intl. Conf. on ICA ans Signal Separation*, 2007.

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

[6] K. Kumatani, J. McDonough, and B. Raj, "Maximum kurtosis beamforming with a subspace filter for distant speech recognition," in *Proc. ASRU*, 2011.

[7] K. Kumatani, L. Lu, J. McDonough, A. Ghoshal, and D. Klakow, "Maximum negentropy beamforming with superdirectivity," in *EUSIPCO*, Aalborg, Denmark, 2010.

[8] S. Leukimmiatis, D. Dimitriadis, and P. Maragos, "An optimum microphone array post-filter for speech applications," in *Intl. Conf. on spoken language processing*, 2006.

[9] M. Shashanka, B. Raj, and P. Smaragdis, "Probabilistic latent variable models as non-negative factorizations," *Computational intelligence and Neuroscience*, 2008.

[10] P. Smaragdis, M. Shashanka, and B. Raj, "A sparse non-parameteric models for single-channel separation of known sounds," in *Neural Info. Processing Systems (NIPS)*, 2009.

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