GENERALIZED OPTIMAL MULTI-MICROPHONE SPEECH ENHANCEMENT USING SEQUENTIAL MINIMUM VARIANCE DISTORTIONLESS RESPONSE(MVDR) BEAMFORMING AND POSTFILTERING

Lae-Hoon Kim, Mark Hasegawa-Johnson

Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
Urbana, IL, USA

Koeng-Mo Sung

School of Electrical Engineering
Seoul National University
Seoul, Korea

ABSTRACT

A theoretical basis for optimal multichannel speech enhancementis presented, sufficient, flexible to be used with any assumed statistical model and optimality criterion. Any Bayesian optimal one-channel estimator for speech enhancement can be generalized to the multichannel case as a sequentially constructed minimum variance distortionless response (MVDR) beamformer followed by an optimal one-channel postfilter. We present experimental results using the minimum mean-square error log-spectral amplitude (MMSE-logSA) optimality criterion, applied to a statistical model with simplified channel but realistic inter-microphone noise coherence. Word error rate in the audio-visual speech in a car (AVICAR) corpus (moving car, windows open) is reduced from 18% to 9%.

1. INTRODUCTION

In recent years, many systems have used multi-microphone arrays for the task of speech enhancement and robust speech recognition. However few approaches have presented a theoretical basis for the optimal multichannel speech enhancement under assumed statistical models of source speech signal and noise. One of the few published systems that considers a theoretical basis is that of Simmer et al. [1], which extends the Wiener estimator for the case of multi-microphone estimation. They showed that multi-microphone minimum mean-square error (MMSE) spectral estimation can be factored into a minimum variance distortionless response (MVDR) beamformer followd by a single-microphone Wiener postfilter. Lotter et al. [2] tried to derive the spectral amplitude estimator using MMSE and maximum a posteriori (MAP) criteria, under the assumption that there is no correlation between noise spectral components of different microphones. In practice, at frequencies and microphone spacings of interest in practical systems, noises are correlated. Balan and Rosca [3] showed that multi-microphone MMSE spectral amplitude estimation can be factored into a beamforming-like sufficient statistic followed by a single-microphone postfilter. The sufficient statistic approach is based on the assumption that the whole impulse response from speaker to microphones can be estimated before applying the postfilter, but this blind channel identification problem is known to be very hard.

In this paper, we try to find a generalized optimal speech enhancement technique using the idea of a sufficient statistic, and as a result we introduce a sequentially cascaded MVDR beamforming technique followed by various well-known Bayesian optimal postfilters. From our approach, any optimal one-channel speech enhancement technique can be easily extended to the optimal multi-microphone speech enhancement technique. This one-channel optimal postfilter can be anything, including Wiener estimator, minimum mean-square error short-time spectral amplitude (MMSE-STSA) estimator [4], minimum mean-square error log spectral amplitude (MMSE-logSA) estimator [5], or maximum a posteriori spectral amplitude (MAP-SA) estimator [6] because we can assume that we know not only the *a posteriori* pdf of source signal and measured signals but also their *a priori* pdf. To give a realistic experimental result, we used data from the "audio-visual speech in car" (AVICAR) database [7] recorded in various noise conditions using 8 microphones in a real moving car. The result shows that our approach has value in all different noise conditions in the car.

2. PROBLEM FORMULATION

The observed signals of the N microphones $y_1, y_2, \cdots, y_N$ are given by

$$\begin{aligned} y_1(t) &= h_1(t) * s(t) + n_1(t) \\ y_2(t) &= h_2(t) * s(t) + n_2(t) \\ &\vdots \\ y_N(t) &= h_N(t) * s(t) + n_N(t) \end{aligned} \tag{1}$$

where $h_1, h_2, \cdots, h_N$ are the impulse responses from the source speaker to each microphone, s(t) is the source speech signal, * denotes convolution, and $n_1, n_2, \cdots, n_N$ are the noise signals measured by each microphone. The frequency domain representation of (1) is given by

$$\begin{aligned} Y_1(k) &= H_1(k)S(k) + N_1(k) \\ Y_2(k) &= H_2(k)S(k) + N_2(k) \\ &\vdots \\ Y_N(k) &= H_N(k)S(k) + N_N(k) \end{aligned} \tag{2}$$

where $Y_i(k)$, $H_i(k)$, $S(k)$, and $N_i(k)$ denote the k^{th} spectral component of $y_i(t)$, $h_i(t)$, $s(t)$, and $n_i(t)$ respectively, $i = 1, 2, \cdots, N$. (2) can be written conveniently as

$$Y = HS + N \tag{3}$$

where $Y = [Y_1 Y_2 \cdots Y_N]$, $H = [H_1 H_2 \cdots H_N]$, and $N = [N_1 N_2 \cdots N_N]$. Since we can assume that $S(k)$ and $N_i(k)$ are zero mean complex Gaussian random variables independent of $S(j)$ and $N_i(j)$ for $j \neq k$, "k" is omitted for readability. We can also assume that N

is a zero mean complex Gaussian random vector with the spectral covariance matrix R_n [3]. These assumptions about the source and noise signals give the posterior probability density function (PDF) $p(Y|S)$.

$$p(Y|S) = \frac{1}{\pi^N det(R_n)} exp\{-(Y-HS)^H R_n^{-1}(Y-HS)\} \quad (4)$$

where superscript H means conjugate transpose. The purpose of speech enhancement is to estimate the source speech signal in an optimal way based on the given information.
Given knowledge of H, the maximum likelihood (ML) estimate of S is

$$T(Y) = \frac{H^H R_n^{-1} Y}{H^H R_n^{-1} H} \quad (5)$$

(5) can be also derived by choosing the weights to minimize the expected value of output power as shown in (6), but only if we assume that we can estimate H before applying postfilter:

$$\underset{W}{min} \; E\{W^H X X^H W\} \; subject \quad to \; H^H W = 1 \quad (6)$$

In practice, we rarely know H. Without H, a reasonable estimate of S is the MVDR estimator,

$$T(Y) = \frac{d^H R_n^{-1} Y}{d^H R_n^{-1} d} \quad (7)$$

where $d = [d_1 d_2 \cdots d_N]$ and d_i is the delay response from speaker to microphones.

3. MULTI-CHANNEL MMSE USING FISHER-NEYMAN FACTORIZATION

It is possible to get better estimation results than the ML estimator by making use of an assumed prior pdf of speech spectral components, $p(A, \alpha)$, given by (8) based on the previously stated assumptions about source speech signals [4].

$$p(S) = p(A) \cdot p(\alpha) = \frac{A}{\frac{\lambda_s}{2}} exp\{-\frac{A^2}{\lambda_s}\} \cdot \frac{1}{2\pi} \quad (8)$$

where A is the spectral amplitude, α is the spectral phase, $\lambda_s = E\{|S|^2\}$. From (4) and (8), the *a posteriori* pdf $p(S|Y)$ can be obtained by

$$p(S|Y) = \frac{p(Y|S)p(S)}{\int p(Y|S)p(S)dS} \quad (9)$$

From (9), we can derive estimators that explicitly minimize the mean-square estimation error of the complex spectrum (Wiener filter), of the spectral amplitude (MMSE-STSA [5]), or of the log spectral amplitude (MMSE-logSA [4]). However the multi-channel $p(Y|S)$ is far more complex than the single-channel $p(Y_i|S)$ $i = 1, 2, \cdots, N$. Balan and Rosca [3] introduced a sufficient statistic to get a one-channel-summary of Y using the Fisher-Neyman Factorization Theorem [8]. Using their approach, (4) can be factored as

$$p(Y|S) = g(T(Y)|S) \cdot h(Y) \quad (10)$$

where g and h are some functions, and $T(Y)$ is given by

$$T(Y) = \frac{H^H R_n^{-1} Y}{H^H R_n^{-1} H} \quad (11)$$

$T(Y)$ contains all the information in Y that is useful for estimating S, therefore

$$p(S|Y) = p(S|T(y)) \quad (12)$$

Balan an Rosca assume that if they can replace the source S by the signal received by first microphone which is the contaminated source due to the channel response, one can find H in a recursive manner based on an estimate of the noise spectral power R_n and an estimate of the signal spectral power R_s which is obtained by spectral subtraction. They also assume that the noise signals are uncorrelated, but in practice, at frequencies and microphone spacings of interest in practical systems, noises are correlated. The goal of this paper is to correct these two assumptions in the method of Balan and Rosca, thus enabling us to use the time structure of channel responses, and to model correlated noises. From (4) and Fisher-Neyman factorization theorem the sufficient statistic has the form given by (15)

$$T(Y) = \rho \cdot \frac{H^H R_n^{-1} Y}{H^H R_n^{-1} H} \quad (13)$$

where ρ is some scalar.

4. FISHER-NEYMAN FACTORIZATION WITH THE TIME STRUCTURE OF CHANNEL RESPONSES

If we assume that the channel responses have finite impulse responses

$$H = d + r_1 + r_2 \cdots + r_M \quad (14)$$

where d is same as eq. (7), $r_i = [r_{i1} r_{i2} \cdots r_{iN}]$ for i=1,2,$\cdots$,M is the delay response from i^{th} reflection to microphones, and M is the number of echoes, then

$$\begin{aligned} T(Y) &= \rho \cdot \frac{H^H R_n^{-1} Y}{H^H R_n^{-1} H} \quad (15) \\ &= \rho \cdot \frac{d^H R_n^{-1} Y + r_1^H R_n^{-1} Y + \cdots + r_M^H R_n^{-1} Y}{H^H R_n^{-1} H} \end{aligned}$$

and if $\rho(\cdot) = \frac{H^H R_n^{-1} H}{d^H R_n^{-1} d}(\cdot)$ then

$$\begin{aligned} T(Y) &= \frac{H^H R_n^{-1} Y}{d^H R_n^{-1} d} \quad (16) \\ &= \frac{d^H R_n^{-1} Y}{d^H R_n^{-1} d} + \frac{r_1^H R_n^{-1} r_1}{d^H R_n^{-1} d} \cdot \frac{r_1^H R_n^{-1} Y}{r_1^H R_n^{-1} r_1} \\ &\cdots + \frac{r_M^H R_n^{-1} r_M}{d^H R_n^{-1} d} \cdot \frac{r_M^H R_n^{-1} Y}{r_M^H R_n^{-1} r_M} \end{aligned}$$

M is the number of dominant reflections, is dependent on how the room is reverberant, the relative response $\frac{r_i^H R_n^{-1} r_i}{d^H R_n^{-1} d}$ makes this number quite small. As can be seen in (16), this sufficient statistic $T(Y)$ can be interpreted as a sequentially constructed ML beamformer, where each sequential MVDR beamformer is normalized by the direct term. If we assume that the direct and a few dominant reflections can be found, $T(Y)$ may be approximated without knowledge of the whole H. Because we can get the one-channel-summary using the sequentially constructed MVDR beamformer without loss of the desired source information present in a multi-channel signal, we can simply apply any classical MMSE speech enhancement technique based on the speech *a priori* pdf. From the previously stated sufficient statics, we can get a simpler *a priori* pdf $p(S|Y)$ than (9) as shown in (17).

$$p(S|Y) = p(S|T(Y)) = \frac{p(T(Y)|S)p(S)}{\int p(T(Y)|S)p(S)dS} \quad (17)$$

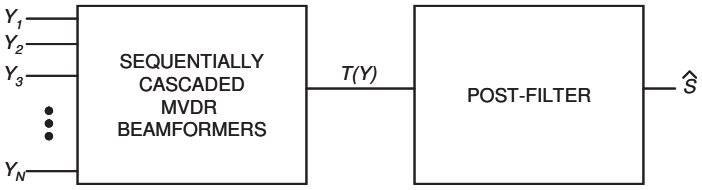

Fig. 1. Block diagram of optimal multi-microphone speech enhancement

where T(Y) can be anything with the form of (15) including the sequentially constructed MVDR beamformer as shown in (16). From this beamformer and postfiltering as shown in Fig. 1, we can derive any desired multi-channel speech estimator. The MMSE spectral component estimator (the Wiener filter) based on multi-channel input, is given by

$$\begin{aligned}\widehat{S}_{MMSE} &= E\{S|Y\} = E\{S|T(Y)\} \\ &= \rho^{-1}([\frac{R_s}{R_s + R_{n_{out}}}] \cdot T(Y)) \end{aligned} \quad (18)$$

where $R_{n_{out}} = \frac{\rho^2}{H^H R_n^{-1} H}$. Wiener postfiltering of an MVDR output has been previously derived using the matrix inversion lemma [1].

There has been no theoretical basis for generalized postfiltering after MVDR beamformer. MMSE-STSA postfiltering of the output signal of the sequentially cascaded MVDR beamformer provides an optimal MMSE solution for the spectral amplitude enhancement, which can be also formulated by (19).

$$\begin{aligned}\widehat{|S|}_{MMSE-STSA} &= E\{|S||Y\} = E\{|S||T(Y)\} \\ &= \rho^{-1}(\Gamma(1.5)\frac{\sqrt{\nu}}{\gamma} exp(-\frac{\nu}{2}) \\ &\quad \cdot[(1+\nu)I_0(\frac{\nu}{2}) + \nu I_1(\frac{\nu}{2})]) \\ &\quad \cdot|T(Y)| \end{aligned} \quad (19)$$

where $\Gamma\{\cdot\}$ denotes the gamma function, with $\Gamma(1.5) = \frac{\sqrt{\pi}}{2}$, I_0 and I_1 denote the modified Bessel functions of zero and first order respectively. ν is defined by

$$\nu = \frac{\xi}{1+\xi}\gamma \quad (20)$$

where ξ and γ are called as a priori and a posteriori signal-to-noise-ratio (SNR) of T(Y) respectively. Similarly multi-microphone MMSE-logSA estimator can be described as (21).

$$\begin{aligned}\widehat{|S|}_{MMSE-logSA} &= E\{log(|S|)|Y\} = E\{log(|S|)|T(Y)\} \\ &= \rho^{-1}(\frac{\xi}{1+\xi} exp\{\frac{1}{2}\int_v^\infty \frac{e^{-t}}{t} dt\}) \\ &\quad \cdot|T(Y)| \end{aligned} \quad (21)$$

As a postfilter for spectral amplitude enhancement, the MAP estimator (22), (23) of Wolfe and Godsill [6] can be used.

$$\begin{aligned}\widehat{|S|}_{MAP} &= \arg\max_{|S|}\ p(|S||Y) \\ &= \arg\max_{|S|}\ p(T(Y)||S|)p(|S|) \end{aligned} \quad (22)$$

$$\widehat{|S|}_{MAP} = \rho^{-1}(\frac{\xi + \sqrt{\xi^2 + (1+\xi)\frac{\xi}{\gamma}}}{2(1+\xi)} \cdot |T(Y)|) \quad (23)$$

5. EXPERIMENT

We applied MVDR beamforming ($M = 0$ in eq. (16)) followed by MMSE-logSA postfiltering and conducted a single digit recognition test using AVICAR corpus [7]. We chose $M = 0$ because the reflections are assumed to be weak relative to the direct sound in the noisy situation. If we consider more reflections, we may get slightly better performance. To do the digit recognition test, firstly, we trained 11 HMMs (from "oh" to "nine" and "silence (which is for the noise only period)") using data from 25 male talkers and tested using the other twenty five. We used mel-frequency cepstral coefficients (MFCC), energy, delta coefficients, and acceleration coefficients, total feature vector of size 39 and used the HTK toolkit [9] for building an isolated digit recognition system, and testing recognition accuracy. Number of states per word was 8, and the number of mixtures was 1 to 14. Secondly, we used 55 talkers from TIDIGITS' training corpus for training the 11 HMMs, and tested on the AVICAR corpus and enhanced AVICAR corpus.

The AVICAR corpus [7]is data recorded in a real car environment using a multi-sensory array consisting of eight microphones on the sun visor and four video cameras on the dashboard. The script for the corpus consists of four categories: isolated digits, isolated letters, phone numbers, and phonetically balanced sentences. Speakers from various language backgrounds are included, 50 male and 50 female. Each script has five different noise condition: idling (IDL), driving at 35mph with windows up (35U) and down (35D), and driving at 55mph with windows up (55U) and down (55D).

We chose the LMS-GSC as an MVDR beamformer with 7 filter taps per channel [10]. We chose this adaptive filtering algorithm because it can update the change of noise covariance in real time and this real time update is thought to be appropriate for in-car recordings like the AVICAR corpus. Before applying GSC, we obtained the delays between microphones to align direction of arrival to the source location using the SRP-PHAT algorithm [1]. SRP-PHAT is known to be a robust localizing algorithm even in reverberant rooms. As a post-filtering method, we chose the MMSE-logSA algorithm, because the cepstral features used in speech recognition are a linear transform of the log spectrum. Fig. 2 shows the time response of utterance "three"- original (upper), after MVDR beamforming (middle), and MMSE-logSA postfiltering after MVDR (lower). Proposed algorithm shows well enhanced results even though the noisy speech data were recorded in a real moving car with windows down at 55 miles-per-hour (mph). Informal listening found a remarkable enhancement in signal quality.

Table 1 is the result when HMMs are trained using half of the speakers in the AVICAR corpus and tested on the other half. Both traing set and test set include data from all five noise conditions in equal proportion. Table 2 is the result when TIDIGITS' training corpus is used as the train corpus, and models are tested on half of the AVICAR corpus, or of the enhanced AVICAR corpus respectively. In all of the experiments, we can confirm the success of our proposed approach. The proposed approach gives better recognition accuracy than when the HMMs were trained using speech from 5 different noise condition, and tested even on the same 5 different noise conditions. This result can be interpreted to mean that this enhancing algorithm makes the enhanced training and test data have more similar acoustic feature characteristics after the algorithm.

Training set	Test set	Number of Gaussian Mixtures							
		1	2	4	6	8	10	12	14
Original	Original	81.09	82.14	82.98	83.47	83.42	83.39	83.69	83.85
Enhanced	Enhanced	89.74	90.98	90.88	90.31	90.79	91.07	90.88	90.79

Table 1. Digit recognition accuracy(%) using half of the AVICAR corpus or enhanced AVICAR corpus for training

Training set	Test set	Number of Gaussian Mixtures							
		1	2	4	6	8	10	12	14
TIDIGIT	Original	30.49	37.57	33.84	30.16	30.13	29.67	30.84	32.35
TIDIGIT	Enhanced	62.3	58.88	55.94	53.85	54.61	52.14	52.99	53.09

Table 2. Digit recognition accuracy(%) using TIdigits training corpus for training

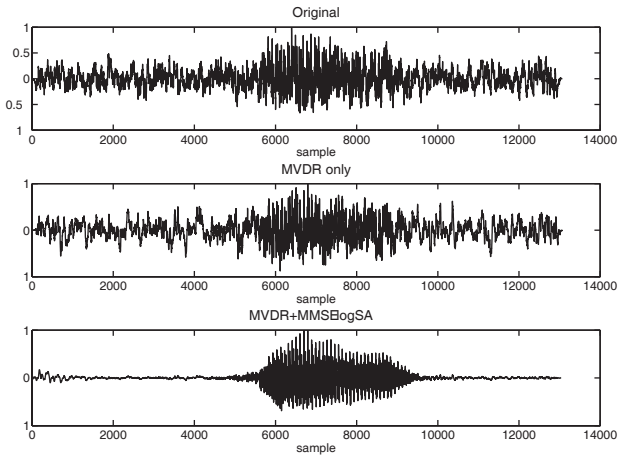

Fig. 2. Time response of utterance "three" in 55D condition: original (upper), after MVDR beamforming (middle) after MVDR beamforming and MMSE-logSA postfiltering (lower)

6. CONCLUSION

We demonstrated that any optimal multi-microphone speech enhancement algorithm can be expressed as the sum of several MVDR beamformers (one per dominant echo, plus one for the direct sound), followed by an one-channel postfilter. This theoretical background can guide the choice of a beamfomer and postfilter for optimal enhancement of speech in multi-microphone situations.

7. REFERENCES

[1] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techinques and Applications.* Berlin: Springer Verlag, 2001.

[2] T. Lotter, C. Benien, and P. Vary, "Mutichannel Speech Enhancement Using Bayesian Spectral Amplitude Estimation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Proces.* (ICASSP), 2003, pp. I880-I883.

[3] 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. Workshop* Aug. 2002, pp. 209-213.

[4] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Amplitude Estimator," *IEEE Trans. Acous., Speech, and Signal process.*, vol. ASSP-32, no. 6, pp. 1109-1121, Dec. 1984.

[5] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator," *IEEE Trans. Acoust., Speech, and Signal process.*, vol. ASSP-33, no. 2, pp. 443-445, Apr. 1985.

[6] P. J. Wolfe and S. J. Godsill, "Simple alternatives to the Ephraim and Malah suppression rule for speech enhancement," in *Proc. 11th IEEE Signal Process. Workshop* Aug. 6-8, 2001, pp. 496-499.

[7] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu and T. Huang, "AVICAR: An Audiovisual Speech Corpus in a Car Environment," in *Proc. Int. Conf. Spoken Language Processing* 2004.

[8] H. V. Poor, *An Introduction to Signal Detection and Estimation*, New York: Springer Verlag, 1994.

[9] `http://htk.eng.cam.ac.uk`

[10] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propagat.*, vol. AP-30, no. 1, pp. 27-34, Jan. 1982.