A SPEECH PRESENCE MICROPHONE ARRAY BEAMFORMER USING MODEL BASED SPEECH PRESENCE PROBABILITY ESTIMATION

Tao Yu, John H.L. Hansen

CRSS: Center of Robust Speech Systems, University of Texas at Dallas
800 West Campbell Road, Richardson, TX 75080
E-mails: {txy073000, john.hansen} @ utdallas.edu

ABSTRACT

The purpose of this study is to investigate the performance of speech presence (SP) microphone array beamforming. When the presence uncertainty of the desired speech is considered, noise reduction is greatly achieved while preserving low speech distortion level. Furthermore, we propose a novel model based speech presence probability (SPP) estimator, exploring both the sinusoid structure of speech and signal-to-noise ratio (SNR). Finally, experiments verify the effectiveness of the proposed SP-beamformer, resulting in a better trade-off between speech distortion and noise leakage, and a corresponding higher output segmental SNR, when compared with the classical beamformers.

Index Terms—Microphone array, speech presence probability, speech enhancement

1. INTRODUCTION

Distance based speech acquisition via microphone array is a viable approach for speech recognition. In most applications of microphone array beamforming systems, speech detection and estimation are treated distinctly and separately. Generally, time domain voice-activity-detection (VAD) is performed whenever the system is engaged. If speech is detected, the array input[1] is decomposed into the short-time Fourier transform (STFT) domain and every frequency bin coefficient is further processed by a subsequent narrowband beamformer, such as minimum mean square error (MMSE) or minimum variance distortion-less response (MVDR), in order to decrease the ambient noise and enhance the desired speech; otherwise, beamforming will not be activated and the entire system will have a null output.

One main problem with this overall approach may be stated as: the speech signal is generally sparse in the frequency domain, which means that the speech signal is significantly condensed within a limited range of frequency components. Performing beamforming in a speech absent frequency bin is not wise because the ambient noise of that very frequency will be incorporated by the beamformer.

Moverover, the decision on activity for the beamformer is based on the binary decision of the VAD. A miss detection, which happens in real noise environments, leads to unrecoverable/incorrect results from the beamformer output. Hence, a 'soft' decision is needed from the VAD and the corresponding soft-decision oriented beamformer should be designed to taken into account both the presence and absence of the speech signal.

In this study, we present a novel speech presence beamformer (SP-Beamformer) which incorporates a speech sinusoid model based speech presence probability estimation and soft-decision orientated beamforming.

2. SPPECH PRESENCE BEAMFORMER

2.1. Classical Beamformer

To illustrate the proposed idea, we consider an array of M microphones located in the far-field. We assume that there is only one desired signal and treat the remaining signals as noise or interference. Taking the STFT of the array received signals; the following data model is obtained:

$$Y(l,k) = A(\theta,k)s(l,k) + N(l,k) \quad (1)$$

where (l,k) denotes the time-frequency bin, l=0,1,$\cdots$ is the time frame index and $k = 0,1,\cdots K-1$ is the frequency bin index. Hereby, $Y(l,k) \in C^{M\times 1}$ is the array observed data. $A(\theta,k) \in C^{M\times 1}$ is the array steering vector for the desired speech $s(l,k) \in C$ with a direction-of-arrival(DOA) θ, and $N(l,k) \in C^{M\times 1}$ is a noise-plus-interference vector. In this section, we assume that the desired signal, noise and interference are all Gaussian i.i.d; hence we have the simplify notification:

$$Y = A(\theta)s + N. \quad (2)$$

The classical optimal narrowband beamformer is a linear processor for the array observations [1]. The output of the beamformer is given by:

$$\hat{s} = W^H Y, \quad (3)$$

This project was funded by AFRL under a subcontract to RADC Inc. under FA8750-05-C-0029, and the University of Texas at Dallas under Project EMMITT.

where $(\cdot)^H$ stands for the complex conjugate transpose. The weights W are chosen according to some optimization criteria, such as MMSE or MVDR [1].

2.2. Optimum Signal Presence (SP) Beamformer

In general, classical beamforming is developed with the assumption that the desired signal is always present; however, this is not practical in real world applications. Here, an optimal beamformer in the MMSE sense is developed that takes the presence and absence of the desired signal into consideration.

Let $H=\{H^0, H^1\}$ denote the hypothecs space, with H^0 and H^1 respectively indicting the presence and absence of the desired signal in the time-frequency bin,

$$\begin{cases} H^1 : Y = A(\theta)s + N \\ H^0 : Y = N \end{cases} . \tag{4}$$

The MMSE estimator of the desired signal is the conditional mean of $s(t,k)$, given Y and can be written as:

$$\begin{aligned} \hat{s}_{SP-mmse} &= E\{s \mid Y\} = E\{E\{s \mid Y, H\}\} \\ &= p(H^0 \mid Y)E\{s \mid Y, H^0\} + p(H^1 \mid Y)E\{s \mid Y, H^1\} \end{aligned} . \tag{5}$$

If we further assume the signal's DOA θ is known a priori, the expectation $E\{s \mid Y, H^1\}$ is the MMSE estimator when the desired signal is present at DOA θ, leading to the classical spatial Wiener filter point to θ [1], defined as:

$$E\{s \mid Y, H^1\} = W_{WF}^H Y . \tag{6}$$

Moreover, if we define $E\{s \mid Y, H^0\} = 0$, which means zero output from the beamformer when the desired single is absent, the entire MMSE estimator is simply the scaled version of the classical spatial Wiener filter, given as:

$$\hat{s}_{SP-WF} = p(H^1 \mid Y) W_{WF}^H Y , \tag{7}$$

which is scaled by the presence probability of desired signal.

2.3. Relation to Distortion-weighted Wiener Filter

Next, we explore the relationship between the speech presence beamformer and Speech-Distortion-Weighted Wiener Filter [2]. If we require the same estimator for both the case of speech presence and absence, that is, in the speech present case, we require that the estimator provide a minimal distortion of the desired speech while in the speech absent case, we want the same estimator to provide a minimal noise output power. With these goals, the filter is given by solving the following relation:

$$\begin{aligned} W_{PA} &= \arg\min_W \{p(H^0 \mid Y)E\{|W^H N|^2\} \\ &\quad + p(H^1 \mid Y)E\{|W^H A(\theta)s - s|^2\}\} \\ &= \arg\min_W \{\frac{p(H^0 \mid Y)}{p(H^1 \mid Y)} \underbrace{W^H R_N W}_{\varepsilon_n^2} + \underbrace{\delta_s^2 |1 - W^H A(\theta)|^2}_{\varepsilon_s^2}\} \end{aligned} \tag{8}$$

where ε_n^2 is related to the noise reduction, and ε_s^2 is related to speech distortion, which is the same as that in the standard spatial Wiener Filter.

However, the probability ratio $\mu = p(H^0 \mid Y)/p(H^1 \mid Y)$ which serves as a weighting term here, controls the tradeoff between the noise reduction and speech distortion terms, whereas the standard Wiener filter assigns equal importance to both terms, as $\mu = 1$. If the ratio $\mu > 1$, the residual noise level is reduced at the expense of increased signal distortion. On the contrary, if ratio $\mu < 1$, signal distortion is decreased while the remaining residual noise level is increased. This is actually the same as the so-called speech-distortion-weighted Wiener filter [2], which can be incorporated into the speech presence beamforming framework.

3. MICROPHONE ARRAY SPP ESTIMATION

3.1. Sufficient Statistics Space

The statistical hypotheses under signal presence uncertainty employed here can be formulated as:

$$\begin{cases} p(Y \mid H^1) = \dfrac{1}{\pi^M \det(R_N)} \exp\{-(Y - As)^H R_N^{-1}(Y - As)\} \\ p(Y \mid H^0) = \dfrac{1}{\pi^M \det(R_N)} \exp\{-Y^H R_N^{-1} Y\} \end{cases} \tag{9}$$

Instead of directly working on the array observations, we employ the sufficient statistics $z \in C^1$ of s instead of using Y for SPP estimator design, which can be written as [3]:

$$z = \frac{A^H R_N^{-1} Y}{A^H R_N^{-1} A} = s + \frac{A^H R_N^{-1} N}{A^H R_N^{-1} A} = s + n \tag{10}$$

where $n = A^H R_N^{-1} N / (A^H R_N^{-1} A)$ is also a random Gaussian variable, with power $\sigma_n^2 = 1 / A^H R_N^{-1} A$. As shown in [3], z in is the sufficient statistics in the Bayes sense for any function of s. Noting that z is a single channel signal containing both the source speech term s and effective noise term n, sufficient statistics allow us to perform single channel SPP estimation over direct multichannel observations, as:

$$p(H^1 \mid Y) = p(H^1 \mid z) \tag{11}$$

3.2. Classical Speech Presence Probability Estimation

Under the stochastic speech model assumption, the speech STFT coefficients are complex zero-mean Gaussian random variables. The noisy speech distribution conditioned on the speech presence/absence hypothesis is given as:

$$\begin{cases} p(z \mid H^1) = \dfrac{1}{\pi(\sigma_s^2 + \sigma_n^2)} \exp\{-|z|^2 / (\sigma_s^2 + \sigma_n^2)\} \\ p(z \mid H^0) = \dfrac{1}{\pi\sigma_n^2} \exp\{-|z|^2 / \sigma_n^2\} \end{cases} \tag{12}$$

The probability of speech presence given the observation $z(l,k)$ can be written as:

$$p(H^1 \mid z) = \Lambda / (1 + \Lambda) . \tag{13}$$

The general likelihood ratio (GLR) Λ at time-frequency bin

$z(l,k)$ is defined as the weighted ratio of the likelihood of speech presence and the likelihood of speech absence [4]:

$$\Lambda = \frac{p(H^1 \mid z)}{p(H^0 \mid z)} = \frac{p(H^1)}{p(H^0)} \frac{p(z \mid H^1)}{p(z \mid H^0)} = \frac{q}{1-q} \frac{1}{1+\xi} \exp\{\frac{\xi \cdot \gamma}{1+\xi}\} \qquad (14)$$

where $q = p(H^1)$ denotes the *a priori* probability of speech presence. $\xi = \sigma_s^2 / \sigma_n^2$ and $\gamma = (|z|^2 / \sigma_n^2)$ are defined as the *a priori* and *a posteriori* SNRs. In general, ξ is unknown and can be estimated by using the *decision-directed* approach [4].

There exists an intrinsic drawback when the decision directed approach is applied to SPP estimation [6]. Here, we want the SPP transition curve to be steep enough to give a clear decision, yielding a larger value for speech presence and a smaller value for speech absence. Thereby, noise is effectively reduced and/or speech is preserved. However, since speech is highly non-stationary and has a wide dynamic range of amplitude even within a short time, a steeper transition curve may result in both higher false-alarm rate and miss-hit rate. For example, Fig.1.(a) with $q = 0.3$ gives the steepest transition curve among the three, resulting in a cleaning out of the weak speech part. However, a flatter transition as in (c) with $q = 0.7$ preserves a highly noise level even with an a priori SNR= –40 dB.

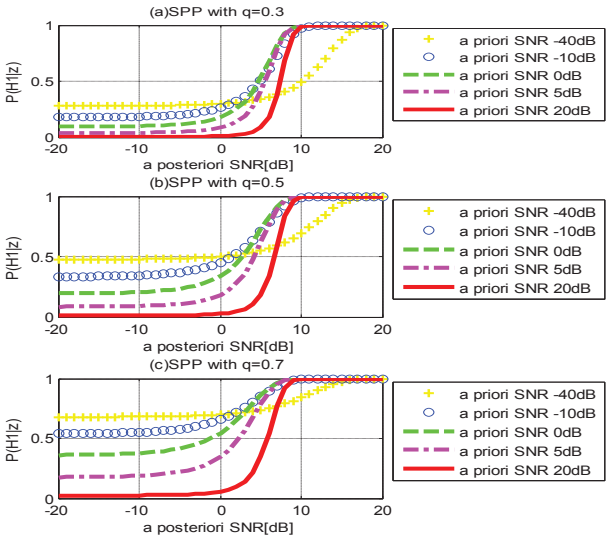

Fig. 1. *SPP Curves for different q, where q is the a priori SPP.*

To address this problem, Cohen and Berdugo[5] suggest to update q according to both local and global averages of ξ. Next, averaged values are non-linearly mapped between 0 and 1, to represent the *a priori* SPP. Although this SNR based *a priori* SPP estimation approach may be efficient for stationary or slowly time-varying background noise situations, the false detection rate is extremely high when non-stationary or intransient noise is present. The next section proposes an alternative method.

3.3. Proposed SPP Estimation

In this section, we propose to estimate the *a priori* SPP by exploring the structure of speech itself rather than SNR, resulting in a speech structure model and SNR combined GLR. As suggested in [7], speech can be approximately represented by a sum of sinusoids on a frame-by-frame basis, in the time domain as:

$$\hat{z}_t(l) = \sum_{i=1}^{I_l} a_{i,l} \cos(\omega_{i,l} t + \phi_{i,l}) \qquad (15)$$

where $\hat{z}_t(l)$ is the approximated speech for frame l in the time domain, I_l is the number of sinusoids used; and $a_{i,l}$, $\omega_{i,l}$ and $\phi_{i,l}$ are the i^{th} sinusoidal amplitude, angular frequency and phase. For voiced speech frames, this model is effective because of the harmonic structure of the voice speech; and for the unvoiced frames, relatively more sinusoidal components are needed.

Let Ω_l represent the discrete frequency-bin sets that correspond to all sinusoid frequencies $\{\omega_{1,l}, \omega_{2,l}, \cdots \omega_{I_l,l}\}$ at frame l. Hence in the STFT domain, the time- frequency bin $z(l,k)$ with k belongs to set Ω_l will have a higher *a priori* speech presence probability than others,

$$q_M(t,k) = \begin{cases} \beta & \text{if } k \in \Omega_l \\ 1-\beta & \text{otherwise} \end{cases} \qquad (16)$$

where q_M denotes the model based *a priori* SPP and $\beta \in [0,1]$ is typically set to 0.8 in this study. Therefore, we can modify GLR as,

$$\Lambda_{MS} = \left(\frac{q_M}{1-q_M} \right)^{\lambda} \frac{p(z \mid H^1)}{p(z \mid H^0)} \quad , \qquad (17)$$

where Λ_{MS} denotes the modified GLR, and λ is a weighting factor. This modified GLR integrates the information from both speech structure and SNR, and should be more reliable than SPP estimated only from SNR.

The notation used here is that, *a priori* SPP q_M and q is conceptually the same (both are $p(H^1)$), but measured from different perspectives. In [5], the authors believe that higher SNR corresponds to higher probability that hypothesis H^1 is true. In our model, a higher probability is assigned to the one that 'looks' more like speech. Hence, with $\lambda = 1$, Λ_{MS} still has the same meaning as a likelihood ratio. However, if $\lambda \neq 1$, Λ_{MS} cannot be interpreted as likelihood ratio: for $\lambda > 1$ we accentuate the *a priori* information and for $\lambda < 1$, we give more credit to the observation.

To obtain reliable estimation of the sinusoid frequencies, we also employ the same direction-directed idea: perform sinusoid decomposition at the output of the estimator, through the method proposed by Jensen and Hansen [7].

4. EXPERIMENTS AND DISCUSSION

The performance evaluation consists of two parts: both single channel and microphone array are considered. We

employ the same measurement approach described in [6] to evaluate the SPP estimators, in terms of speech distortion (SD), and noise leakage (NL). The SD measure indicates the ratio of speech energy that the SPP estimator neglects to the entire speech energy; NL measures the ratio of the noise energy from the noise-only bins that are not decreased. Finally, we compute the segmental SNR improvement when SPP estimator is applied to MMSE filter[4] in the signal channel case, and MMSE/MVDR beamformer in the array case, respectively. The improvement of the segmental SNR [6] can be written as: $\Delta SNR = SNR_{seg}\{\hat{s}\} - SNR_{seg}\{y\}$ with $\hat{s}$ as the estimated output and y the observation signal. Noise power is estimated and updated by Martin's method [8].

4.1. Single Channel SPP Performance

Table 1 compares performance with three different SPP estimators. The results given are averaged over ten phonetically balanced sentences from the TIMIT database, and five noise sources taken from NoiseX92 database. All acoustic data are down-sampled to 8 kHz.

Table 1: results of single channel SPP performance

	-5dB			0dB			5dB		
	SD %	NL %	ΔSNR [dB]	SD %	NL %	ΔSNR [dB]	SD %	NL %	ΔSNR [dB]
	SPP with fixed *a priori* [4] , $q = 0.5$								
white	3.6	41.2	7.8	1.2	41.6	5.5	0.4	42.6	4.2
pink	1.3	45.3	7.1	0.7	45.0	5.2	0.4	44.5	4.0.
car	0.3	72.0	8.7	0.3	73.0	7.9	0.2	71.0	6.2
factory	1.4	53.3	5.1	0.6	51.3	3.6	0.4	49.1	2.8
babble	1.1	61.2	4.7	0.5	56.7	3.2	0.3	50.5	2.4
	SPP with *a priori* SPP updated according to [5]								
white	6.9	6.2	8.5	2.7	8.3	6.3	1.0	10.6	5.4
pink	7.3	45.3	8.2	2.1	16.0	6.0	1.2	15.3	5.1
car	1.2	24.1	11.3	0.9	30.0	9.2	0.6	36.0	7.4
factory	5.4	38.8	6.1	1.7	40.9	5.0	0.7	42.6	3.7
babble	5.7	40.2	5.0	1.6	41.5	4.1	0.7	42.8	3.2
	Proposed SPP with $\beta = 0.8$, $\lambda = 1.5$ for frequency >1.4k, $\lambda = 1$ for frequency ≤ 1.4k								
white	7.2	13.1	8.7	3.3	14.0	6.0	1.1	16.1	5.2
pink	4.0	16.2	8.2	2.1	15.1	5.7	1.0	16.4	4.8
car	1.0	18.1	10.0	0.9	18.5	8.2	0.7	17.3	6.8
factory	3.1	23.8	7.7	2.0	21.6	5.3	1.3	20.7	4.2
babble	2.0	25.9	6.3	1.6	24.5	4.6	1.2	21.3	3.9

From Table 1, the proposed SPP has a relatively smaller noise leakage and larger speech distortion; but in many cases, especially in non-stationary noises (babble, factory), proposed SPP has the highest segmental SNR improvement. SD may be introduced by the speech modeling error, but from informal listening evaluations we believe that speech quality is not tremendously affected. NL is significantly reduced over the previous SPP method for all the tested non-white background noises.

4.2 SP-Beamformer in Real Car Noise Environment

Table 2 compares the segmental SNR improvement of our SP-beamformer versus classical beamformers under in-car noises taken from noise portions of CU-Move database [9]. The microphone array used for CU-Move is a linear five-channel array, with 4.25cm distance between consecutive microphones, to avoid spatial aliasing at sampling frequency of 8 kHz. The experiment is conducted in a semi-real version: 10 clean sentences (with DOA $\theta = 70^{\circ}$) from the TIMIT database degraded by 5 channel CU-Move noise recordings. Because in-car noise is diffused, diagonal loading (DL) is used. ΔSNR is computed over signal channel response.

Table 2: ΔSNR comparison between classical and SP beamformers

	-5dB ΔSNR	**0dB** ΔSNR	**5dB** ΔSNR
MVDR-DL	2.2dB	1.7dB	1.1dB
SP-MVDR-DL	**3.4dB**	**2.8dB**	**2.2dB**
MMSE-DL	5.7dB	4.4dB	4.0dB
SP-MMSE-DL	**7.6dB**	**6.3dB**	**5.3dB**

From Table 2, the proposed SP-beamformer has the highest segmental SNR improvement over classical beamformers. In practice, speech is sparse in the STFT domain, leaving much room for noise reduction and SNR improvement.

5. SUMMARY

In this study, a speech presence (SP) microphone array beamforming algorithm was proposed. The model based speech presence estimator integrated with beamforming is shown to measurably improve performance over traditional MMSE and MVDR based systems. An absolute 1.1 to 1.9dB improvement in segmental SNR is obtained for real in-car speech beamforming application.

6. REFERENCES

[1] H.Van Trees, *Optimum Array Processing*. New York, NY: Wiley,2002.

[2] A. Spriet, M. Moonen, and J. Wouters, "Spatially pre-processed speech distortion weighted multi-channel Wiener filtering for noise reduction," *Signal Proc.*, vol. 84, no. 12, pp.2367-2387, Dec. 2004.

[3] R. Balan and J. Rosca, "Microphone array speech enhancement by Bayesian estimation of spectral amplitude and phase," *IEEE Sensor Array and Multichannel Signal Proc. Workshop*, Rosslyn VA, Aug. 2002

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

[5] I.Cohen and B.Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Proc.*, vol.81, no.11, pp.2403-2418,Nov. 2001.

[6] T. Gerkmann, C. Breithaupt and R. Martin, "Improved *a posteriori* speech presence probability estimation based on a likelihood ratio with fixed priors," *IEEE TASLP*, vol. 16, no. 5, pp. 910-919, Jul. 2008.

[7] J. Jensen and J. Hansen, "Speech enhancement using a constrained iterative sinusoidal model ," *IEEE Trans. Acoust., Speech, Signal Proc.*, vol. 9, no. 7, pp. 731-740, Oct. 2001.

[8] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Acoust., Speech, Signal Proc.*, vol. 9, no. 5, pp. 504-512, Jul. 2001.

[9] J. Hansen, etc., "CU-Move: analysis & corpus development for interactive in-vehicle Speech Systems", *Eurospeech'01*, Sept. 2001