A SUBBAND SPACE CONSTRAINED BEAMFORMER INCORPORATING VOICE ACTIVITY DETECTION

Alan Davis, Siow Yong Low, Sven Nordholm

WA Telecomms. Research Institute (WATRI)
35 Stirling Highway, Crawley
WA 6009, Australia
davisa,siowyong,sven@watri.org.au

Nedelko Grbić

Blekinge Institute of Technology
Dept. of Telecomms. and Signal Processing
SE-372 25 Ronneby, Sweden
nedelko.grbic@bth.se

ABSTRACT

This paper introduces a new subband adaptive space constrained beamforming structure for use in hands-free speech enhancement applications. The scheme incorporates a space constrained source model and voice activity information through the integration of a voice activity detector (VAD). The VAD information is used to estimate noise covariance information during non-speech periods and to optimally estimate the source power spectral density (PSD), which is used to provide a spectrally optimized constraint on the source. The proposed structure is evaluated in a real car environment, yielding results which compare well to the optimal Wiener solution where full knowledge of the source is known.

1. INTRODUCTION

Speech enhancement for use in hands-free scenarios has attracted much interest in recent times, and is primarily driven by recent explosive growth in communications. The main benefit of hands-free systems is that no close-range microphone is required to capture the desired speech signal. However, this freedom comes at a price of increased distortion from both room reverberation and additive noise. Microphone array techniques have shown promise for speech enhancement applications in hands-free situations [1, 2, 3, 4, 5]. The benefit of such systems is the ability to jointly spatially and temporally discriminate signals.

This paper presents a new adaptive subband beamforming structure that incorporates a voice activity detector (VAD) and a spatial model of the source of interest. The incorporation of a VAD not only allows estimation of noise statistics during non-speech periods, it also allows for estimation of the source power spectral density (PSD) to weigh the pre-calculated source spatial information.

The structure utilizes the model presented in [4], whereby the source spatial location is modeled as a number of clustered points in space located within a certain pre-defined constraining region. Given this region, the source auto-covariance and cross-covariance information is pre-calculated as per [4] and used to spatially discriminate the received signal from the interference noise sources.

Under the assumption that the noise and desired source are spatially independent, we then combine the pre-calculated source auto- and cross- covariance information, the estimated noise statistics and source PSD estimate to calculate the Wiener solution in each subband.

WATRI is a joint venture between Curtin University of Technology and the University of Western Australia. This research is supported by the Australian Research Council under grant number A00105530 and the Australian Telecommunications CRC.

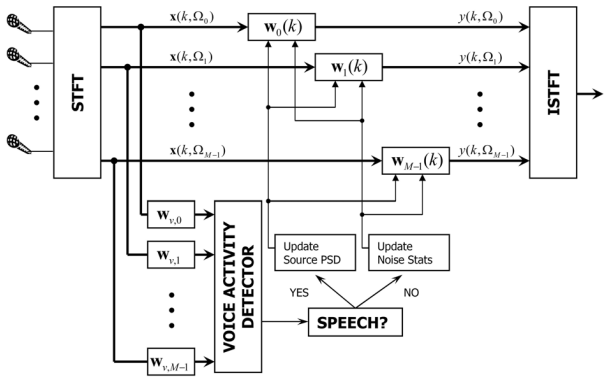

Fig. 1. *Proposed structure*

The proposed scheme is evaluated in a real car environment and compared to the optimal Wiener solution where full knowledge of the source is utilized. The evaluation shows promising results with the proposed scheme indicating a noise suppression level of more than 14dB over the whole test set.

2. PROPOSED STRUCTURE

The proposed structure utilizes a uniform over-sampled short-term discrete Fourier transform filter bank to decompose L microphone signals into M subbands with a decimation factor of $\frac{M}{4}$. Figure 1 shows the proposed structure. Initially, the L microphone signals are decomposed into subband signals. The subband signals are then used to compute the presence or absence of speech activity, and thus estimate the source PSD or update the estimated noise statistics respectively. Based on the estimated noise statistics, pre-calculated source auto- and cross- covariance matrices and noise covariance matrix, the Wiener solution is then calculated in each subband. Finally, the subband signals are recombined by transforming back to the time domain.

We consider a received noisy speech vector $\mathbf{x}(k, \Omega_m)$ at normalized frequency Ω_m and time instant k as,

$$\mathbf{x}(k, \Omega_m) = [x_0(k, \Omega_m), ..., x_{L-1}(k, \Omega_m)]^T , \qquad (1)$$

where $x_l(k,\Omega_m)$ is the received signal at microphone l and $(.)^T$ represents the vector transpose operation. We are interested in the case when a desired clean speech signal is corrupted by spatially independent interfering noise sources. We therefore model this received vector as a linear combination of the desired clean speech and the multiple interfering noise sources,

$$\mathbf{x}(k,\Omega_m) = \mathbf{s}(k,\Omega_m) + \mathbf{n}(k,\Omega_m), \tag{2}$$

where $\mathbf{s}(k,\Omega_m)$ is the desired received clean speech vector at frequency Ω_m (in a similar fashion to (1)) and $\mathbf{n}(k,\Omega_m)$ is the received noisy vector at frequency Ω_m, which includes all interfering noise sources.

2.1. Subband Wiener Solution

The problem at hand is that given this received noisy speech signal and an assumed speech source location, determine a set of optimal weights $\mathbf{w}_{m,opt}$ such that,

$$d(k,\Omega_m) = \mathbf{w}_{m,opt}^H \mathbf{x}(k,\Omega_m), \tag{3}$$

where

$$\mathbf{w}_{m,opt} = [w_{0,opt}, ..., w_{L-1,opt}]^T, \tag{4}$$

$d(k,\Omega_m)$ is the desired speech signal and $(.)^H$ represents Hermitian transpose. A solution to this optimization in the least mean square error sense is,

$$\mathbf{w}_{m,opt} = \arg\min_{\mathbf{w}_m}[\sigma_{d,m}^2 + \mathbf{w}_m^H \mathbf{R}_{\mathbf{xx,ms}} \mathbf{w}_m + \ldots \\ 2Re\left\{\mathbf{w}_m^H \mathbf{r}_{\mathbf{dx,m}}\right\}], \tag{5}$$

where $\sigma_{d,m}^2$ is the variance of the desired speech signal in the mth subband, $\mathbf{R}_{\mathbf{xx,m}}$ is the noisy signal covariance matrix in the mth subband and $\mathbf{r}_{\mathbf{dx,m}}$ is the cross-covariance of the received signal and the desired speech in the mth subband. The optimal weights may be found as,

$$\mathbf{w}_{m,opt} = [\mathbf{R}_{\mathbf{xx,m}}]^{-1} \mathbf{r}_{\mathbf{dx,m}}, \tag{6}$$

which is the well know optimal Wiener solution.

Under the earlier assumption that the interfering noise signals and the desired speech signal are spatially independent the subband spatial covariance matrix may be decomposed as,

$$\mathbf{R}_{\mathbf{xx,m}} = \mathbf{R}_{\mathbf{ss,m}} + \mathbf{R}_{\mathbf{nn,m}}, \tag{7}$$

where $\mathbf{R}_{\mathbf{nn,m}}$ is the subband covariance matrix of the multiple interfering noise sources and $\mathbf{R}_{\mathbf{ss,m}}$ is the subband source covariance matrix. During non-speech periods (7) reduces to,

$$\mathbf{R}_{\mathbf{xx,m}} = \mathbf{R}_{\mathbf{nn,m}}, \tag{8}$$

therefore it is possible to estimate the noise statistics during non-speech periods. In order to solve (6), the problem now becomes one of how to estimate the source covariance $\mathbf{R}_{\mathbf{ss,m}}$ and the cross-covariance between the source and received signal $\mathbf{r}_{\mathbf{dx,m}}$.

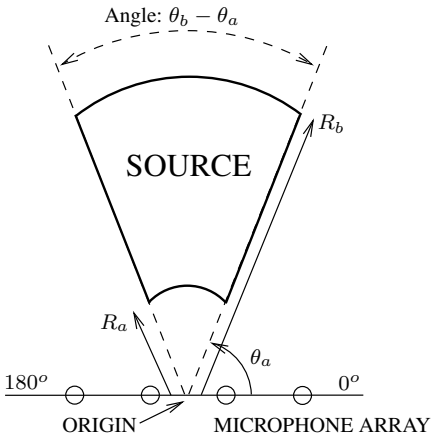

Fig. 2. *Source constraint region*

2.2. Space Constrained Model

In order to address the problem of source covariance estimation, we employ the model given in [4] whereby the desired source is modeled as a distributed source within a radius $[R_a, R_b]$ and angular region $[\theta_a, \theta_b]$ (see Figure 2) as,

$$\mathbf{R}_{\mathbf{ss,m}} = \int_{\theta_a}^{\theta_b} \int_{R_a}^{R_b} S(\Omega_m) \mathbf{d}(R,\theta,\Omega_m) \mathbf{d}^H(R,\theta,\Omega_m) dR d\theta, \tag{9}$$

where $S(\Omega_m)$ is the source PSD and $\mathbf{d}(R,\theta,\Omega_m)$ is the array response vector for a point source located at radius R and angle θ from the origin. The array response vector is represented as,

$$\mathbf{d}(R,\theta,\Omega_m) = \left[\frac{1}{R_1} e^{-j\Omega_m \tau_1(R,\theta)}, ..., \frac{1}{R_L} e^{-j\Omega_m \tau_L(R,\theta)}\right]^T, \tag{10}$$

where R_l represents the distance from sensor l to the point source and $\tau_l(R,\theta)$ represents the time delay from the point source to sensor l, for the given R and θ. The cross-covariance $\mathbf{r}_{\mathbf{dx,m}}$ is modeled in a similar manner,

$$\mathbf{r}_{\mathbf{dx,m}} = \int_{\theta_a}^{\theta_b} \int_{R_a}^{R_b} S(\Omega_m) \mathbf{d}(R,\theta,\Omega_m) dR d\theta. \tag{11}$$

Both (9) and (11) indicate that knowledge of the source PSD $S(\Omega_m)$ is required. This is commonly neglected and set to unity [5]. In the proposed scheme, we estimate the source PSD during speech active periods. We may represent (9) as,

$$\mathbf{R}_{\mathbf{ss,m}} = S(\Omega_m) \mathbf{R}_{\mathbf{dd,m}}, \tag{12}$$

where,

$$\mathbf{R}_{\mathbf{dd,m}} = \int_{\theta_a}^{\theta_b} \int_{R_a}^{R_b} \mathbf{d}(R,\theta,\Omega_m) \mathbf{d}^H(R,\theta,\Omega_m) dR d\theta. \tag{13}$$

Likewise, $\mathbf{r}_{\mathbf{dx,m}}$ may be rewritten as,

$$\mathbf{r}_{\mathbf{dx,m}} = S(\Omega_m) \mathbf{r}_{\mathbf{dd,m}}, \tag{14}$$

where

$$\mathbf{r}_{\mathbf{dd,m}} = \int_{\theta_a}^{\theta_b} \int_{R_a}^{R_b} \mathbf{d}(R,\theta,\Omega_m) dR d\theta. \tag{15}$$

In order to estimate the source PSD $S(\Omega_m)$, a least square error estimate was developed. We define our cost function J_m as,

$$J_m = \left|\left|\hat{\mathbf{R}}_{\mathbf{ss,m}}(k) - S(k,\Omega_m)\mathbf{R}_{\mathbf{dd,m}}\right|\right|_F^2, \quad (16)$$

where,

$$\hat{\mathbf{R}}_{\mathbf{ss,m}}(k) = \mathbf{R}_{\mathbf{xx,m}}(k) - \hat{\mathbf{R}}_{\mathbf{nn,m}}, \quad (17)$$

$\hat{\mathbf{R}}_{\mathbf{nn,m}}$ indicates the estimated noise covariance matrix found during non-speech periods, $\mathbf{R}_{\mathbf{xx,m}}(k)$ is the auto-covariance matrix of the received signal $\mathbf{x}(k,\Omega_m)$ at time instant k and $||.||_F$ indicates Frobenius norm. Hence, we find a least square estimate of the source PSD as,

$$\hat{S}(k,\Omega_m) = \frac{\sum_{i=0}^{L-1}\sum_{j=0}^{L-1} Re\left\{\hat{R}^*_{ss,m}(i,j) - R_{dd,m}(i,j)\right\}}{\sum_{i=0}^{L-1}\sum_{j=0}^{L-1} R_{dd,m}(i,j)R^*_{dd,m}(i,j)}, \quad (18)$$

where (i,j) indicates the element of the matrix located at the ith column and jth row and $Re\{.\}$ indicates the real part. In practise this is evaluated during speech active periods and recursively smoothed with a short forgetting factor,

$$\bar{S}(k,\Omega) = \alpha\bar{S}(k-1,\Omega_m) + (1-\alpha)\hat{S}(k,\Omega_m), \quad (19)$$

where α is the forgetting factor and $\bar{S}(k,\Omega)$ is the smoothed source PSD estimate. In this evaluation a value of $\alpha = 0.2$ was found to give good results.

Finally, we may find a set of weights based on the source model, estimated source PSD and estimated noise covariance matrix. Considering (6), we find the weights for the beamformer as,

$$\mathbf{w}_m(k) = \left[\bar{S}(k,\Omega_m)\mathbf{R}_{\mathbf{dd,m}} + \hat{\mathbf{R}}_{\mathbf{nn,m}}\right]^{-1}\bar{S}(k,\Omega_m)\mathbf{r}_{\mathbf{dd,m}}. \quad (20)$$

The final output subband signals are produced by spatially filtering with the previously developed weights,

$$y(k,\Omega_m) = \mathbf{w}_m^H(k)\mathbf{x}(k,\Omega_m). \quad (21)$$

2.3. Voice activity detector

A voice activity detector is employed to determine when to estimate the noise covariance information. We modify the approach given in [6] whereby the variance of the background noise is estimated and an optimal threshold determined, based on a signal-to-noise ratio (SNR). Rather than employing the Welch method of overlapping windows to generate a reduced variance spectrum estimate we average over adjacent subbands which has a similar effect. A summing beamformer with a look direction towards the source is utilized to increase the SNR before the VAD,

$$x_v(k,\Omega_m) = \mathbf{w}_{v,m}^H\mathbf{x}(k,\Omega_m), \quad (22)$$

where $\mathbf{w}_{v,m}$ are fixed beamformer weights with look direction towards the source and $x_v(k,\Omega_m)$ is the input to the VAD scheme.

A reduced resolution, reduced variance spectrum estimate is evaluated as,

$$P_{xx,v}(k,\Omega_u) = \frac{U}{M}\sum_{m=a_u}^{b_u}\left|x_v(k,\Omega_m)\right|^2, u = 0,1,...,U-1, \quad (23)$$

where U indicates the number of subbands in the reduced resolution estimate and $|.|$ indicates absolute value. The sets a_u and b_u are each a set of U linearly spaced coefficients over the whole frequency range indicating the start and stop band regions for the summation respectively,

$$a \in \left\{0, \frac{M}{U}, ..., \frac{(U-1)M}{U}\right\}, \quad (24)$$

and similarly,

$$b \in \left\{\frac{M}{U} - 1, \frac{2M}{U} - 1, ..., M-1\right\}, \quad (25)$$

where the ratio $\frac{M}{U}$ is constrained to be an integer value. Typically values of $M = 256$ and $U = 16$ are used. All other aspects of the algorithm presented in [6] remain the same, with the exception of the hang-over scheme, which is not utilized in this system. By incorporating the VAD scheme in this manner, all processing may be undertaken utilizing the same structure.

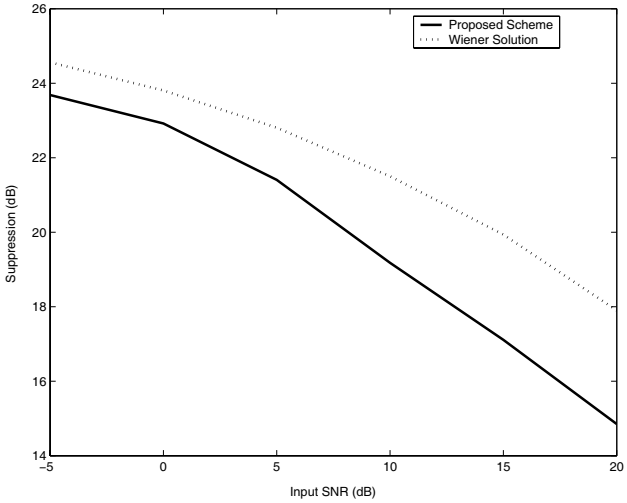

Fig. 3. *Noise suppression level for various SNRs*

3. EVALUATION

A performance evaluation of the proposed scheme was made in a real car hands-free situation. A four-sensor microphone array was mounted on the visor at the passenger side in a Volvo station wagon. Data was gathered on a multi-channel DAT-recorder with a sampling rate of 8kHz with the car moving at a constant speed of 110km/h. The desired target signal was recorded while the car was stationary, and the noise was recorded while the car was in motion. In order to evaluate the scheme, a set of noisy speech files was generated with differing SNRs from -5 to 20dB. To evaluate the noise suppression of the scheme, we define,

$$Supp = 10\log_{10}\left[\frac{\sum_{m=0}^{M-1} P_{xx,r}(\Omega_m)}{C\sum_{m=0}^{M-1} P_{yy}(\Omega_m)}\right], \quad (26)$$

where $P_{xx,r}(\Omega_m)$ is the PSD of the reference microphone signal, $P_{yy}(\Omega_m)$ is the PSD of the output signal, C is a scaling factor to account for the overall system gain and $Supp$ indicates the suppression level in dB.

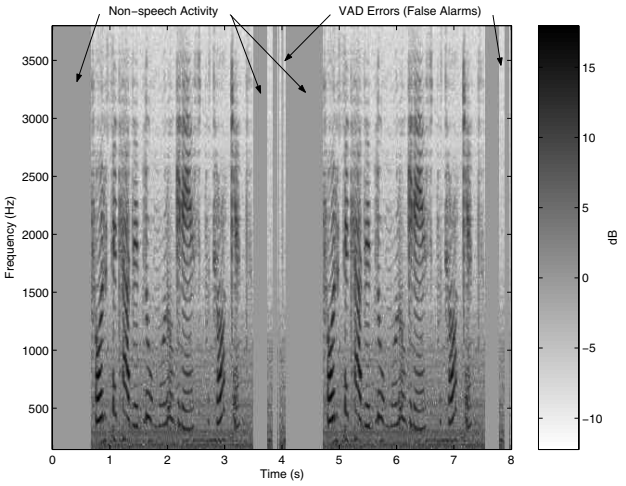

Fig. 4. *Estimated source PSD used to weight source autocovariance*

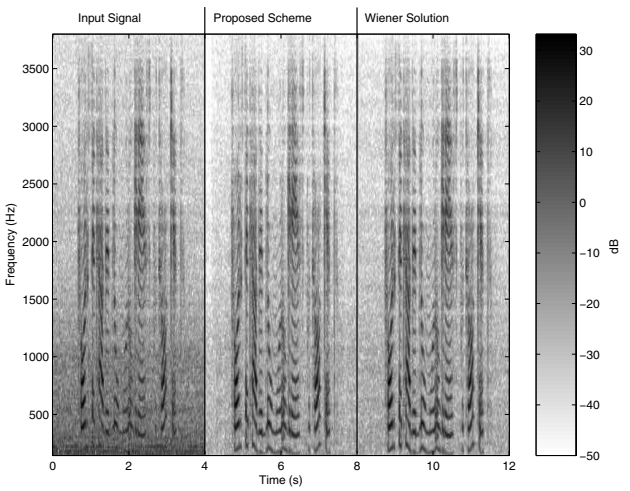

Fig. 5. *Spectogram showing original signal, proposed scheme output and optimal Wiener solution output*

The evaluation was conducted by first processing the noisy speech in the proposed manner and then recording the resulting weights for each subband and each time instance. The clean speech and noise were then individually processed using these recorded weights and the resulting outputs stored. The noise output and speech output were then compared to their original inputs in order to evaluate the effectiveness of the scheme. A subband optimal Wiener solution based scheme was also implemented using full knowledge of the source and was also tested in the same manner.

Figure 3 shows the suppression level as calculated by processing the noise only with the recorded weights, and comparing to the input noise. As can clearly be seen, the proposed scheme compares well to the optimal Wiener solution, with around 3dB less suppression at 20dB input SNR, and approaches the optimal Wiener solution as the SNR falls.

Figure 4 shows the source PSD $\bar{S}(\Omega_m)$ used to weight the pre-calculated source covariance matrix $\mathbf{R_{dd,m}}$. As shown in the figure, the source PSD is not estimated during non-speech activity and is forced to unity during these periods. The proposed scheme was found to be relatively insensitive to false-alarm VAD errors, however many consecutive miss-detection errors would result in source cancelation. It is also clear there is some residual noise in the source PSD estimate, which helps to account for the difference between the optimal Wiener solution and the proposed scheme.

Figure 5 shows a spectogram for an input signal, output from the proposed scheme and output from the subband optimal Wiener solution with 5dB average SNR at the reference sensor. It is clear there is very little speech distortion and high suppression of the background noise. Subjective listening tests confirm this.

4. CONCLUSION

We have presented a new subband adaptive beamforming structure. The structure incorporates a voice activity detector and predefined source constraining region. The scheme was shown to perform well over a range of signal to noise ratios in a real car environment, with little distortion of the desired speech signal. The main drawback of the proposed structure is its reliance on a VAD. Failure of the VAD was found to produce source cancelation in the event of multiple miss-detections, the scheme however was found to be relatively insensitive to false-alarm errors. Evaluations indicated the scheme performed well as compared to the theoretical Wiener solution, where full knowledge of the source was assumed.

5. REFERENCES

[1] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE Acoust., Speech and Signal Processing Magazine*, vol. 5, pp. 4–24, Apr. 1988.

[2] W. Kellermann, "A self-steering digital microphone array," *Int. Conf. on Acoust., Speech, and Signal Processing*, vol. 5, pp. 3581–3584, Apr. 1991.

[3] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Trans. on Signal Processing*, vol. 47, no. 10, pp. 2677–2684, Jun. 1999.

[4] N. Grbić and S. Nordholm, "Soft constrained subband beamforming for handsfree speech enhancement," *IEEE Int. Conf. on Acoust., Speech and Signal Process.*, vol. 1, pp. 885–888, May 2002.

[5] S. Y. Low, S. Nordholm, and N. Grbić, "Subband generalized sidelobe cancellor - a constrained region approach," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 41–44, Oct. 2003.

[6] A. Davis and S. Nordholm, "A low complexity statistical voice activity detector with performance comparisons to itu-t/etsi voice activity detectors," *Joint Int. Conf. on Information, Communications and Signal Processing and Pacific Rim Conf. on Multimedia.*, vol. 1, pp. 119–123, Dec. 2003.