LIMITS IN FIR SUBBAND BEAMFORMING FOR SPATIALLY SPREAD NEAR-FIELD SPEECH SOURCES

N. Grbić, S. Nordholm, A. Cantoni

Western Australian Telecommunications Research Institute
University of Western Australia
35 Stirling Highway, Crawley
Western Australia

ABSTRACT

This paper analyses optimal subband beamforming performance mainly aimed at speech enhancement and acoustic echo suppression for personal communication devices, personal computers and wireless cellular telephones. The focus is on theoretical limits of finite impulse response (FIR) beamformers for spatially spread sources in the array near-field. Performance of the Wiener solution is compared to the direct maximization of the array gain for different lengths of the FIR filters and different source interference spreads. The evaluation is performed individually in subbands with constant increasing logarithmic bandwidth. Results show that the difference between the Wiener solution and the direct array gain maximization is less than 2 dB in the measure of Signal-to-Noise plus Interference Ratio (SNIR), for small interference spread. With increasing interference spread the difference in SNIR performance increases, in favor of the array gain maximization.

1. INTRODUCTION

The increased use of multimedia applications in personal communication devices, personal computers and wireless cellular telephones enables the development of new inter-personal communication systems. The convergence between computers and telephony technologies brings up the demand for convenient hands-free communications. In such systems the user wishes to lead a conversation in much the same way as in a normal person-to-person conversation. The advantages of hands-free communication are safety, convenience and greater flexibility.

Unfortunately, by installing the microphone far away from the user a number of disadvantages are introduced. These problems are mainly caused by room reverberation, noise and acoustic feedback. Several FIR beamforming techniques have been proposed to reduce the negative effects of hands-free communication, [1], [2], [3]. Most of these techniques estimate the multichannel Wiener filter or minimize a least square (LS) error criterion. In [4] a combination of Wiener filtering and a subspace tracking is proposed in an adaptive structure, to allow for source movement. A subspace algorithm is used for elimination of the

non-directional noise components in [5], followed by a minimum variance (MV) beamformer used in a second stage, for the reduction of directional components. Furthermore, a single channel Wiener filter was used for the low frequency noise reduction.

Independent on the algorithm used, performance is restricted by physical properties of the acoustical environment.

This paper evaluates theoretical limits of FIR filter beamforming in a scenario where the desired source and one interfering source are present in the array near-field. The sources are modeled as spatially spread sources, uniformly contained within an area segment. The surrounding noise is modeled as a spherically isotropic noise field. The FIR beamforming solutions are analyzed by subdividing the traditional telephone bandwidth into six regions with linearly increasing logarithmic bandwidth. Since the spatial resolution increases with increasing frequency, the number of FIR filter taps needed becomes approximately constant across the subbands, when implemented in this type of subbands. The Wiener solution and the direct maximization of array gain are evaluated for different FIR filter lengths. Analysis with varying interference spread and angles are also included.

2. PROBLEM FORMULATION

We consider a wide band source located in the near-field of a uniform linear array with N microphones. Since the source is assumed to be a person speaking, it is modeled as an infinite number of point sources clustered closely within a region in space $\mathbf{A}_1$. Let $\mathbf{x}$ represent a received stacked array data vector, situated in an isotropic noise field, receiving a number M of stationary, independent and spatially spread sources within spatial regions $\mathbf{A}_m$, having power spectral densities, PSD, $S_m(\Omega)$ with energy contained in the spectral band $[\Omega_a, \Omega_b]$. The spatio-temporal covariance matrix for the received signal is then given by

$$\mathbf{R}_x = \int_{\Omega_a}^{\Omega_b} \left(\sum_{m=1}^M \frac{S_m(\Omega)}{C_m} \right) \left(\iint_{\mathbf{A}_m} \mathbf{E}(\Omega) \mathbf{E}^H(\Omega) \otimes \mathbf{G}(r, \theta, \Omega) \mathbf{G}^H(r, \theta, \Omega) d\mathbf{A} \right) d\Omega + \mathbf{R}_n \quad (1)$$

The authors are with Western Australian Telecommunication Research Institute, which is a joint venture of The University of Western Australia and Curtin University. The work has also been sponsored by ARC under grant no. A00105530. (e-mail: ngr@bth.se, sven@watri.uwa.edu.au, cantoni@watri.uwa.edu.au)

where $\otimes$ denotes the Kronecker product and

$$C_m = \iint_{\mathbf{A}_m} d\mathbf{A} \quad (2)$$

is a spatial spreading normalization constant. Here, we have assumed that the spectral densities, $S_m(\Omega)$, are uniformly distributed inside the spatial regions, $\mathbf{A}_m$. The extension to a nonuniform distribution is straight forward by simply including a weighting of the power spectral density of the source within each spatial region. The temporal response vector is given by

$$\mathbf{E}(\Omega) = e^{j\frac{L}{2}\Omega T} \left[1 e^{-j\Omega T} \dots e^{-j(L-1)\Omega T} \right]^T \quad (3)$$

which is normalized to the center lag of the FIR filters by the constant $e^{j\frac{L}{2}\Omega T}$, where T is the sampling period and where L denotes the FIR filter lengths. The spatial response vector is given by

$$\mathbf{G}(r, \theta, \Omega) = r e^{j\Omega r/c} \left[\frac{1}{r_1} e^{-j\Omega \tau_1(r, \theta)} \frac{1}{r_2} e^{-j\Omega \tau_2(r, \theta)} \dots \frac{1}{r_N} e^{-j\Omega \tau_N(r, \theta)} \right]^T \quad (4)$$

where $\tau_n(r, \theta)$ denotes the time delay from a point source at radius r and angle θ to sensor n , and r_n is the distance between the source and sensor n . Parameter c is the speed of wave propagation. The response vector includes a constant, $r e^{j\Omega r/c}$, which normalizes the response to unity at origin of coordinates. In the calculation model, a spherical propagation in a free-field and homogeneous medium has been assumed. Figure 1 illustrates the array geometry and model setup, with one desired spread source and one interfering spread source.

Fig. 1. Array geometry and model setup.

The spatio-temporal spherically isotropic noise covariance matrix is given by, [6],

$$\mathbf{R}_n(i \cdot k, j \cdot l) = \frac{1}{4\pi^3} \int_{\Omega_a}^{\Omega_b} S_n(\Omega) \frac{\sin\left(\frac{\Omega d_{ij}}{c}\right)}{\frac{\Omega d_{ij}}{c}} e^{-j\Omega(k-l)T} d\Omega \quad (5)$$

at position $(i \cdot k), (j \cdot l)$ where i and j denotes sensor index positions, and k and l denotes time lag positions, for row and column, respectively. Parameter c is the speed of wave propagation, d_{ij} is the distance between sensors i and j and $S_n(\Omega)$ is the noise power spectral density.

2.1. Finite Length Signal-to-Noise plus Interference Beamformer

The output signal-to-noise plus interference power ratio (SNIR) is defined as

$$\text{SNIR} = \frac{\text{average signal output power}}{\text{average noise-plus-interference output power}} \quad (6)$$

and the maximum FIR Signal-to-Noise plus Interference Beamformer (SNIB) is defined as the set of FIR filters which maximizes the power ratio SNIR. We define the filter vector $\mathbf{w}$ as a vector with stacked FIR weights from each microphone input as

$$\mathbf{w}^T = [\mathbf{w}_1^T \ \mathbf{w}_2^T \ \dots \ \mathbf{w}_N^T] \quad (7)$$

where

$$\mathbf{w}_n^T = [w_n(0) \ w_n(1) \ \dots \ w_n(L-1)] \quad (8)$$

is the FIR weight vector for microphone channel n . We may express the SNIR as, [7]

$$\text{SNIR} = \frac{\mathbf{w}^H \mathbf{R}_s \mathbf{w}}{\mathbf{w}^H (\mathbf{R}_n + \mathbf{R}_i) \mathbf{w}} \quad (9)$$

where the spatio-temporal covariance matrix in (1) is split up as

$$\mathbf{R}_x = \mathbf{R}_s + \mathbf{R}_i + \mathbf{R}_n \quad (10)$$

where $\mathbf{R}_s$, $\mathbf{R}_i$ and $\mathbf{R}_n$ denotes covariance matrices received by contributions of the target source, undesired (jammer) sources and noise, respectively. Without loss of generality we assume that source number one is the target source.

A weight vector that maximizes the SNIR is given by, [6], [7]

$$\mathbf{w}_{opt} = (\mathbf{R}_n + \mathbf{R}_i)^{-1/2} \tilde{\mathbf{v}}_{max} \quad (11)$$

where $\tilde{\mathbf{v}}_{max}$ is an eigenvector, corresponding to the largest eigenvalue, complying with

$$\left((\mathbf{R}_n + \mathbf{R}_i)^{-1/2} \right)^H \mathbf{R}_s (\mathbf{R}_n + \mathbf{R}_i)^{-1/2} \tilde{\mathbf{v}}_{max} = \lambda \tilde{\mathbf{v}}_{max}. \quad (12)$$

The measure of SNIR is scale invariant and any constant scaling of the weight vector given in (11) also maximizes the SNIR.

2.2. Finite Length Wiener filter Beamformer

The Wiener filter is the solution of the linear mean-square waveform estimation problem, provided the noise and the signals are stationary random variables. The finite length Wiener filter is the best (in mean square sense) approximation of the infinite length Wiener filter, and the weight vector may be found by expressing the orthogonality between the output error and the received array vector [8]. The solution is given by

$$\mathbf{w}_{opt} = \mathbf{R}_x^{-1} \mathbf{r}_s = (\mathbf{R}_s + \mathbf{R}_n + \mathbf{R}_i)^{-1} \mathbf{r}_s \quad (13)$$

where the covariance matrix $\mathbf{R}_x$, is given in (1) and where the covariance vector is given by

$$\mathbf{r}_s = \int_{\Omega_a}^{\Omega_b} \frac{S_1(\Omega)}{C_1} \iint_{\mathbf{A}_1} \mathbf{E}(\Omega) \otimes \mathbf{G}(r, \theta, \Omega) d\mathbf{A} d\Omega \quad (14)$$

since the covariance between all sources and noise are assumed to be zero. It follows from the definition of the response vectors in (3) and (4) that the resulting beamformer is temporally delay-normalized to the center lag of the FIR filters, at the spatial origin of coordinates. Other choices of normalization may be used as it will affect the final solution. By normalizing at the center lag of the FIR filters, a fair compromise is achieved since this fact allows for equal length approximations of both the causal and noncausal parts of the infinite length Wiener solution. While performance generally increases by this normalization, it results in a constant delay of $L/2$ samples at the output.

Fig. 2. Output SNIR vs. FIR length for different subbands showing the SNIB (solid line) and the Wiener solution (dashed line). Linear array with $N = 6$, sensor distance is 0.05 m, source SIR=0 dB, source SNR=30 dB, Angle of interference $\theta = 30$ degrees, angle of interference spread, $\beta = 5$ degrees, Angle of source $\phi = 0$ degrees.

3. EVALUATION

Studies have been carried out with one desired source and one interfering source in the near-field of a six sensor linear array with 0.05 m sensor spacing, as illustrated in Fig. 1. The interference source may be a person speaking, a hands-free loudspeaker or a spatially large disturbing source, e.g. a fan or an air conditioner. The evaluation is performed individually in six frequency bands with constant increasing logarithmic bandwidth across the traditional telephone bandwidth. The source is located at broad side of the array throughout the evaluation, i.e. $\phi = 0$. The source signal-to-interference ratio SIR is defined and set as

$$SIR = 10 \log \left(\frac{\int_{\Omega_a}^{\Omega_b} S_1(\Omega) d\Omega}{\int_{\Omega_a}^{\Omega_b} S_2(\Omega) d\Omega} \right) = 0 \text{ dB}$$

Fig. 3. Output SNIR vs. Angle of interference for different subbands showing the SNIB (solid line) and the Wiener solution (dashed line). Linear array with $N = 6$, FIR filter length is 15 taps, sensor distance is 0.05 m, source SIR=0 dB, source SNR=30 dB, Angle of source $\phi = 0$ degrees, angle of interference spread $\beta = 5$ degrees.

where $S_1(\Omega)$ and $S_2(\Omega)$ are source and interference power spectral densities, respectively. The source signal-to-noise ratio SNR is defined as

$$SNR = 10 \log \left(\frac{\int_{\Omega_a}^{\Omega_b} S_1(\Omega) d\Omega}{\int_{\Omega_a}^{\Omega_b} S_n(\Omega) d\Omega} \right) = 30 \text{ dB}$$

where $S_n(\Omega)$ is the noise power spectral density as given in (5). All power spectral densities are temporally constant in this evaluation.

The covariance matrices, used in the solutions given by (11) and (13), are calculated by numerical integration where the sampling period is chosen such that critical sampling is used in all subbands, i.e. the sampling rate is twice the highest frequency in each subband. It should be noted that in a practical implementation using a filterbank realisation, the spatio-temporal properties of the filterbank should be included in Eq. (1).

3.1. SNIR vs. FIR filter length

The performance in terms of SNIR of the SNIB (solid line) and the Wiener solution (dashed line) is given in Fig. 2, as a function of FIR length for the individual frequency bands. The angle of the interfering source θ , is 30 degrees and the angle of interference spread β , is 5 degrees. It can be seen that the difference between the SNIB and the Wiener solution is smaller in the low- and the high-frequency band (~ 1 dB), while it is slightly larger in the middle bands (~ 2 dB). The number of needed FIR filter taps in order to reach optimum is between 10-20 and it is approximately constant for all subbands.

Fig. 4. Output SNIR vs. Angle of interference spread, β , for different subbands showing the SNIB (solid line) and the Wiener solution (dashed line). Linear array with $N = 6$, FIR filter length is 15 taps, sensor distance is 0.05 m, source SIR=0 dB, source SNR=30 dB, Angle of source $\phi = 0$ degrees, angle of interference $\theta = 30$ degrees.

3.2. SNIR vs. Angle of interference

The output SNIR is given in Fig. 3 as a function of angle of interference, with 15 parameters in the FIR filters. The angle of spread β , is 5 degrees, representing a human speaker or a hands-free loudspeaker as the interference source. The difference in performance between the two solutions is small ($\sim 1 - 3$ dB) for angles of interference below 70 degrees. As the angle increases above 70 degrees, the gain with the SNIB beamformer becomes large ($\sim 5 - 10$ dB). Low frequency bands, where the covariance matrices generally have reduced rank, exhibit smaller differences between the solutions.

3.3. SNIR vs. Angle of interference spread

Large interfering objects such as computer fans and air conditioners may be present in the array near-field and thus impact on the hands-free operation. The dependency of interference spread on beamformer performance is given in Fig. 4, where the angle of interference center is separated by 30 degrees from the desired source center. The length of the FIR filters is 15 taps. The results show convex curves with a peak at approximately 20 degrees. The somewhat surprising increase follows from the finite precision nulling in the spatial domain. As the source spread increases, the power per area unit decreases (this follows from the spatial normalization in (1)), and the finite precision spatial nulling is able to suppress a larger portion of the total interference power. The performance drops as the spread increases from 20 degrees to 60 degrees, where an angular overlap occurs. The difference between the solutions is small for small angular spread and for low frequency bands. The gain by using the SNIB is as much as 10 dB in comparison with the

Wiener beamformer, for large interference angular spread.

4. CONCLUSIONS

Performance of the Wiener solution is compared to the optimum signal-to-noise plus interference beamformer (SNIB) for different lengths of the FIR filters. The comparison includes different spatial spreading of the interference source. Results show that the difference in the measure of SNIR is small between the solutions in low frequency bands. It is also shown that the performance is close between the solutions when the spatial spread of the interference is small, i.e. the same size as the source. However, when the interference spread increases, the performance gain with the SNIB is significant, as much as 10 dB.

By subdividing the fullband signals into constant increasing logarithmic bandwidth subbands, the number of FIR filter parameters needed is approximately 10-20 taps and it is nearly the same across the subbands.

5. REFERENCES

- [1] D. A. Florêncio and H. S. Malvar, "Multichannel filtering for optimum noise reduction in microphone arrays," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2001, vol. 1, pp. 197–200.
- [2] S. Nordholm and H. Leung, "Performance limits of the generalized sidelobe cancelling structure in an isotropic noise field," *Journal of Acoustical Society of America*, vol. 107, no. 2, pp. 1057–1060, Feb. 2000.
- [3] N. Grbić and S. Nordholm, "Soft constrained sub-band beamforming for hands-free speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2002, vol. 1, pp. 885–888.
- [4] S. Affes and Y. Grenier, "A signal subspace tracking algorithm for microphone array processing of speech," *IEEE Trans. Acoust. Speech Signal Processing*, vol. 5, no. 5, pp. 425–437, Sep. 1997.
- [5] F. Asano, S. Hayamizu, T. Yamada, and S. Nakamura, "Speech enhancement based on the subspace method," *IEEE Trans. Acoust. Speech Signal Processing*, vol. 8, no. 5, pp. 497 – 507, Sep. 2000.
- [6] D. Johnson and D. Dudgeon, *Array Signal Processing - Concepts and Techniques*, Prentice Hall, 1993.
- [7] J. E. Hudson, *Adaptive Array Principles*, Peter Peregrinus, 1991.
- [8] B. Widrow and S. D. Stearns, *Adaptive Signal Processing*, Prentice Hall, 1985.