Comparison between the equalization and cancellation model and state of the art beamforming techniques

FREDRIK GRAN[1,*], JESPER UDESEN[1,*], TOBIAS PIECHOWIAK[1],
AND ANDREW B. DITTBERNER[2]

[1] *GN ReSound A/S, Lautrupbjerg 7, DK-2750 Ballerup, Denmark*

[2] *GN ReSound North America, 8001 Bloomington Freeway, Bloomington, MN 55420-1036, USA*

This paper investigates the performance of a selection of state-of-the-art array signal-processing techniques for the purpose of predicting the binaural listening experiments from the equalization and cancellation (EC) paper by Durlach written in 1963. Two different array signal-processing techniques are analyzed, 1) filter and sum beamforming (FS), and 2) minimum variance distortionless response (MVDR) beamforming. The theoretical properties of these beamformers for the specific situation of prediction of binaural masking level differences are analyzed in conjunction with the EC model. Also, the performance of the different beamformers on the data sets in the Durlach paper from 1963 is compared to the EC model.

INTRODUCTION

Some of the earliest work on binaural listening effects date back to the duplex theory presented by Lord Rayleigh (1876, 1907), where interaural time and level differences (ITDs and ILDs) characterized the localization of sound sources. Over four decades later, it was shown (Cherry, 1953) that the benefit of listening with two ears compared to monaural listening is especially pronounced in complex listening scenarios with several competing talkers. The binaural listening advantage in these adverse circumstances, also referred to as the ‘cocktail party problem’, was extensively studied in the fifties and sixties (Cherry and Taylor, 1954; Cherry and Sayers, 1956; Leaky and Cherry, 1957; Sayers and Cherry, 1957; Cherry and Bowles, 1960) where a cross correlation model was used to explain the binaural listening effect.

The Equalization and Cancellation (EC) model was proposed to model the binaural masking level differences (BMLD) of detecting tones in noise for dichotic vs diotic (Kock, 1950; Durlach, 1960, 1963) signal presentation.

This model was later modified and used to explain several data sets for more complicated listening experiments, such as modeling speech-intelligibility improvement for speech masked by a single noise source in an anechoic space (Zurek, 1992), speech-intelligibility improvements in multi-talker speech-shaped interference in an anechoic space (Culling *et al.*, 2004), speech-intelligibility tasks in anechoic and diffuse

[*] Corresponding author: fgran@gnresound.com

Proceedings of ISAAR 2013: Auditory Plasticity – Listening with the Brain. 4th symposium on Auditory and Audiological Research. August 2013, Nyborg, Denmark. Edited by T. Dau, S. Santurette, J. C. Dalsgaard, L. Tranebjærg, T. Andersen, and T. Poulsen. ISBN: 978-87-990013-4-7. The Danavox Jubilee Foundation, 2014.

conditions, both for hearing-impaired and normal-hearing listeners (Beutelmann and Brand, 2006). In Wan *et al.* (2010), an extended version of the EC model was used to explain the data sets acquired in Hawley *et al.* (2004).

In Durlach (1963), the EC model was compared to array processing, where the model tries to put a null at the location of the masker to suppress this component as much as possible. The purpose of this paper is to investigate how more generic beamforming techniques compare to the EC model, both from a theoretical stand point, but also in terms of predictive performance on the original data sets. In particular, fixed filter and sum beamforming is investigated (Johnson and Dudgeon, 1993), as well as the minimum variance distortionless response (MVDR) beamforming technique (Capon *et al.*, 1967).

GENERAL MODEL

The general data model assumes a binaural signal set consisting of a mixture of two signals, one representing the target and one the masker. The short-term spectrum of the target is denoted $X(f,t)$ and the corresponding spectrum of the masker is denoted $Y(f,t)$. Then the binaural signal set can be written as:

$$\underbrace{\begin{pmatrix} S_{\mathrm{l}}(f,t) \\ S_{\mathrm{r}}(f,t) \end{pmatrix}}_{\mathbf{s}(f,t)} = \underbrace{\begin{pmatrix} A_{\mathrm{l}}(f) \\ A_{\mathrm{r}}(f) \end{pmatrix}}_{\mathbf{a}(f)} X(f,t) + \underbrace{\begin{pmatrix} B_{\mathrm{l}}(f) \\ B_{\mathrm{r}}(f) \end{pmatrix}}_{\mathbf{b}(f)} Y(f,t), \qquad \text{(Eq. 1)}$$

where $S_{\mathrm{l}}(f)$ and $S_{\mathrm{r}}(f)$ are the signal mixtures, $A_{\mathrm{l}}(f)$ and $A_{\mathrm{r}}(f)$ are the left and right acoustical transfer functions for the target, respectively, and $B_{\mathrm{l}}(f)$ and $B_{\mathrm{r}}(f)$ are the left and right acoustical transfer functions for the masker, respectively. Note that the assumption here is that these transfer functions do not change over time. Furthermore, it is assumed that $X(f,t)$ and $Y(f,t)$ are independent stochastic variables and spectrally white. In this paper, the binaural signal is estimated via a beamforming approach where

$$\mathbf{b}(f,t) = \mathbf{w}^H(f)\mathbf{s}(f,t), \qquad \text{(Eq. 2)}$$

where $\mathbf{b}$ is the binaural spectrum and $\mathbf{w}^H$ is the complex conjugate transpose of the coefficients used to combine the right- and left-ear signals. The coefficients are defined as:

$$\mathbf{w}(f) = \mathbf{M}(f)\mathbf{h}(f), \qquad \text{(Eq. 3)}$$

where $\mathbf{M}$ can be interpreted as a process that models amplitude and timing jitters and is defined by:

$$\mathbf{M}(f) = \begin{pmatrix} (1-\varepsilon_1)e^{-j2\pi f\delta_1} & 0 \\ 0 & (1-\varepsilon_2)e^{-j2\pi f\delta_2} \end{pmatrix}, \qquad \text{(Eq. 4)}$$

where ε_1 and ε_2 are independent Gaussian-distributed variables with zero mean and a variance of 0.25, δ_1 and δ_2 are independent Gaussian-distributed variables with zero mean and a variance of 105 μs and $\mathbf{h}$ are the beamforming coefficients applied to minimize the masker and enhance the target. The experienced reader immediately realizes that if one chooses the beamforming coefficients to be:

$$\mathbf{h}_{\mathrm{EC}}(f) = \begin{pmatrix} B_r(f) & -B_l(f) \end{pmatrix}^T \tag{Eq. 5}$$

the equalization and cancellation model follows from Eq. 2 and $(\cdot)^T$ is the transpose of $(\cdot)$.

ARRAY SIGNAL PROCESSING TECHNIQUES

In this section the two beamformers are derived for the condition described in Eq. 1.

Filter and Sum beamformer

The Filter and Sum (FS) beamformer is an array signal-detection technique developed for optimal signal detection in white Gaussian-distributed noise in the maximum likelihood sense (Johnson and Dudgeon, 1993). The beamforming coefficients would in this case be:

$$\mathbf{h}_{\mathrm{FS}}(f) = \begin{pmatrix} A_l(f) & A_r(f) \end{pmatrix}^T = \mathbf{a}(f), \tag{Eq. 6}$$

Minimum Variance Distortionless Response beamformer

In the Minimum Variance Distortionless Response (MVDR) beamformer, the strategy is to suppress all noise sources as much as possible while maintaining the signal of interest. If the model in Eq. 1 is used this can be expressed mathematically as:

$$\begin{gathered} \mathbf{h}_{\mathrm{MVDR}}(f) = \arg\min_{\mathbf{h}} \mathbf{h}^H \mathbf{R}_{\mathbf{ss}}(f)\mathbf{h} \\ \text{subject to } \mathbf{h}^H \mathbf{a}(f) = 1 \end{gathered} \tag{Eq. 7}$$

where

$$\mathbf{R}_{\mathbf{ss}}(f) = E\left[\mathbf{s}(f,t)\mathbf{s}^H(f,t)\right] \tag{Eq. 8}$$

is the spatial auto correlation matrix of $\mathbf{s}(f,t)$ and E is the expectancy operator.

SIMULATION SETUP

The binaural signal-to-noise ratio (SNR) was evaluated using stochastic simulations. Once a given experimental setup E had been determined (i.e., determining $\mathbf{a}$ and $\mathbf{b}$) and a given set of beamforming coefficients $\mathbf{h}$ had been chosen, the binaural SNR was estimated as:

$$\mathrm{SNR}_E(\mathbf{h}, f) = \frac{\sum_{q=0}^{Q-1} \left|\mathbf{w}_q^H(\mathbf{h}, f)\mathbf{a}(f)\right|^2}{\sum_{q=0}^{Q-1} \left|\mathbf{w}_q^H(\mathbf{h}, f)\mathbf{b}(f)\right|^2}, \tag{Eq. 9}$$

where $\mathbf{w}_q$ is the qth realization of the stochastic process defined by Eq. 3 and Q is the total number of realizations used in the simulation. If the beamforming coefficients are chosen so that $\mathbf{h} = \mathbf{h}_{\text{EC}}$, this SNR estimate is actually equivalent to the variable denoted the EC factor in Durlach's paper from 1963, because the spectral amplitudes of the target and masker are the same and uniform over frequency. In this paper $Q =$ 10000 realizations were used, as this was found to be sufficient to generate a good approximation of the results presented in Durlach (1963). The ratio of binaural SNR between two different experimental conditions E and E' is then given by

$$R_{E/E'}(\mathbf{h}, f) = \frac{\text{SNR}_E(\mathbf{h}, f)}{\text{SNR}_{E'}(\mathbf{h}, f)}. \qquad \text{(Eq. 10)}$$

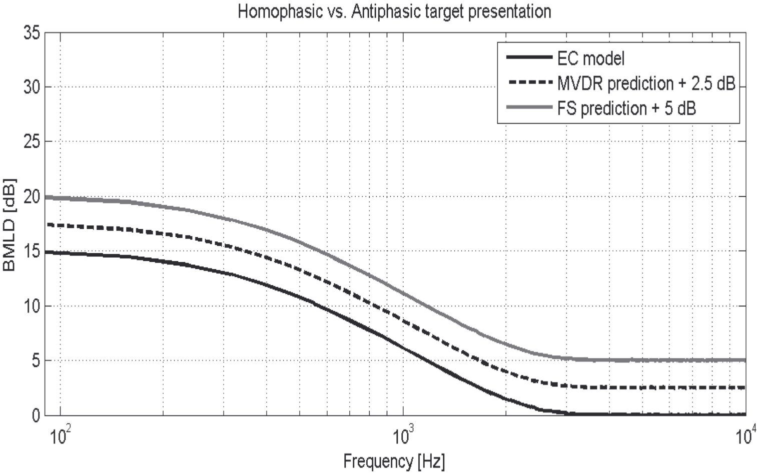

Fig. 1: BMLD for the antiphasic signal presentation compared to the homophasic signal presentation, where the masker is homophasic in both cases. The traditional EC model is given by the solid black curve, the MVDR prediction offset by 2.5 dB is given by the dashed black curve, and the FS beamformer prediction offset by 5 dB is given by the gray curve. Both beamformers are capable of accurately predicting the BMLD according to the original EC model.

SIMULATION RESULTS

In this section a selection of the experimental setups from Durlach (1963) will be reproduced and the BMLD predictions of the different beamformers are compared to the corresponding BMLD prediction of the EC model.

Antiphasic vs homophasic as a function of frequency

In the first simulation, the SNR for antiphasic target-signal presentation was compared to the homophasic target-signal condition. The masker was in both cases homophasic:

$$\text{Condition A} \;:\; \mathbf{a}(f) = \begin{pmatrix} 1 & -1 \end{pmatrix}^T, \mathbf{b}(f) = \begin{pmatrix} 1 & 1 \end{pmatrix}^T \qquad \text{(Eq. 11)}$$

$$\text{Condition H} \;:\; \mathbf{a}(f) = \begin{pmatrix} 1 & 1 \end{pmatrix}^T, \mathbf{b}(f) = \begin{pmatrix} 1 & 1 \end{pmatrix}^T \qquad \text{(Eq. 12)}$$

In Fig. 1, the quantity $R_{\mathrm{A/H}}(f)$ is plotted as a function of frequency. The traditional EC model is given by the solid black curve, the MVDR prediction offset by 2.5 dB is given by the dashed black curve and the FS beamformer prediction offset by 5 dB is given by the gray curve. Both beamformers are capable of accurately predicting the BMLD according to the original EC model.

Variations in the interaural time delays of the signal and noise

The following section describes various conditions where the interaural delay is varied for the masker or the target. The first condition describes a situation where the delay of the target is varied:

$$\begin{aligned} \text{Condition DT} : \mathbf{a}(f) &= \begin{pmatrix} 1 & e^{-j2\pi f\tau} \end{pmatrix}^T, \\ \mathbf{b}(f) &= \begin{pmatrix} 1 & 1 \end{pmatrix}^T. \end{aligned} \qquad \text{(Eq. 13)}$$

In Fig. 2, $R_{\mathrm{DT/H}}(\tau)$ is shown where the frequency is f = 167 Hz and τ is varied between -3 and 3 ms. The traditional EC model is given by the solid black curve, the MVDR prediction offset by 2.5 dB is given by the dashed black curve, and the FS beamformer is given by the gray curve. All beamformers have the correct predictions for ±3 ms and 0 ms. The filter and sum beamformer has the wrong shape in between these points and seems to be a shifted and inverted version of the EC model. The MVDR, however, seems capable of accurately predicting the BMLD.

In Fig. 3, the situation is reversed and the delay of the masker is varied:

$$\begin{aligned} \text{Condition DM} : \mathbf{a}(f) &= \begin{pmatrix} 1 & 1 \end{pmatrix}^T, \\ \mathbf{b}(f) &= \begin{pmatrix} 1 & e^{-j2\pi f\tau} \end{pmatrix}^T. \end{aligned} \qquad \text{(Eq. 14)}$$

$R_{\mathrm{DM/H}}(\tau)$ is shown where the frequency is f = 500 Hz and τ is varied between 0 and 4 ms. The traditional EC model is given by the solid black curve, the MVDR prediction offset by 2.5 dB is given by the dashed black curve, and the FS beamformer prediction is given by the gray curve. All beamformers have the correct predictions for 0 ms, 2 ms, and 4 ms. All predictions seem periodic, however, the filter and sum beamformer again has the wrong shape in between 0, 2, 4 ms compared to the EC model, whereas the MVDR accurately predicts the BMLD.

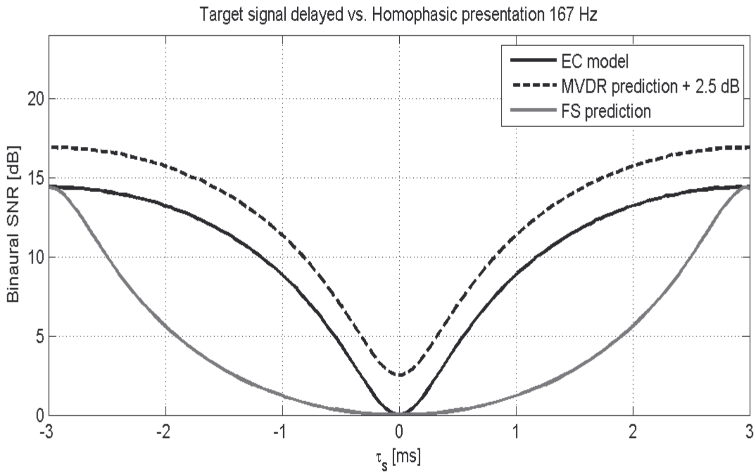

Fig. 2: BMLD for interaurally time-delayed target condition compared to homophasic target presentation. The center frequency was 167 Hz. The traditional EC model is given by the solid black curve, the MVDR prediction offset by 2.5 dB is given by the dashed black curve, and the FS beamformer prediction is given by the gray curve.

DISCUSSION

In this paper two different beamformers were analyzed for the purpose of predicting BMLDs: the filter and sum beamformer (FS) and the minimum variance distortionless response (MVDR). The work spawned from a statement in Durlach (1963) where the EC model was compared to a null-pointing array. Analogous to this, adaptive beamforming techniques automatically adjust the nulls of the array to correspond to the directions of the interferers. The mathematical details of the processing both for the static FS beamformer and for the adaptive MVDR showed large discrepancies in the beamforming coefficients compared to the EC model. However, when applying the beamformers to the examples in the original paper, it was shown that the MVDR was able to accurately predict the BMLD given by the EC model, whereas the static FS beamformer only accounted for the correct BMLD in the condition with the target signal in anti-phase and the masker signal in phase in the two ears. The MVDR has the advantage over the EC model that it does not need any a priori knowledge of the acoustic transfer function between the masker and the listener; instead, it only requires information about the target. This can simplify the use of the model when investigating complex listening environments with multiple interferers from different directions and/or diffuse-noise listening conditions.

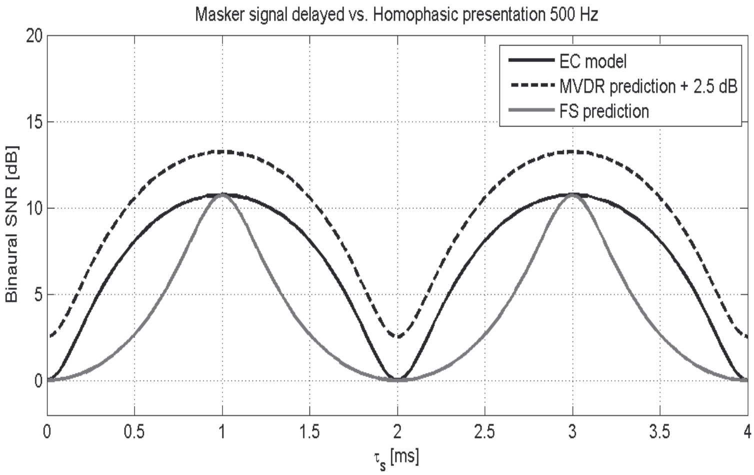

Fig. 3: BMLD for interaurally time-delayed masker condition compared to homophasic masker presentation, where the target is homophasic in both cases. The center frequency was 500 Hz. The traditional EC model is given by the solid black curve, the MVDR prediction offset by 2.5 dB is given by the dashed black curve, and the FS beamformer prediction is given by the gray curve.

REFERENCES

Beutelmann, R., and Brand, T. (**2006**). "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing impaired listeners," J. Acoust. Soc. Am., **120**, 331-342.

Capon, J., Greenfield, R., and Kolker, R. (**1967**). "Multidimensional maximum-likelihood processing of a large aperture seismic array," P. IEEE, **55**, 192-211.

Cherry, E.C. (**1953**). "Some experiments on the recognition of speech with one and two ears," J. Acoust. Soc. Am., **25**, 975-979.

Cherry, E.C., and Bowles, J.A. (**1960**). "Contribution to a study of the cocktail party problem," J. Acoust. Soc. Am., **32**, 884.

Cherry, E.C., and Sayers, B.M.A. (**1956**). "Human cross-correlator - a technique for measuring certain parameters of speech perception," J. Acoust. Soc. Am., **28**, 889-895.

Cherry, E.C., and Taylor, W.K. (**1954**). "Some further experiments upon the recognition of speech, with one and with two ears," J. Acoust. Soc. Am., **26**, 554-559.

Culling, J.F., Hawley, M.L., and Litovsky, R.Y. (**2004**). "The role of head induced interaural time and level differences in the speech reception threshold for multiple interferring sound sources," J. Acoust. Soc. Am., **116**, 1057-1065.

Durlach, N.I. (**1960**). “Note on the equalization and cancellation theory of binaural masking level differences,” J. Acoust. Soc. Am., **32**, 1075-1076.

Durlach, N.I. (**1963**). “Equalization and cancellation theory of binaural masking level differences,” J. Acoust. Soc. Am., **35**, 1206-1218.

Hawley, M.L., Litovsky, R.Y., and Culling, J.F. (**2004**). “The benfit of binaural hearing in a cocktail party: Effect of location and type of interferer,” J. Acoust. Soc. Am., **115**, 833-843.

Johnson, D.H., and Dudgeon, D.E. (**1993**). *Array Signal Processing* (Prentice Hall, Englewood Cliffs, NJ).

Kock, W.E. (**1950**). “Binaural localization and masking,” J. Acoust. Soc. Am., **22**, 801-804.

Leaky, D.M., and Cherry, E.C. (**1957**). “Influence of noise upon the equivalence of intensity differences and small time delays in two-loudspeaker systems,” J. Acoust. Soc. Am., **29**, 284-286.

Rayleigh, L. (**1876**). “Our perception of the direction of sound,” Nature, **14**, 32-33.

Rayleigh, L. (**1907**). “On our perception of sound direction,” Phil. Mag., **6**, 213-242.

Sayers, B.M., and Cherry, E.C. (**1957**). “Mechanism of binaural fusion in the hearing of speech,” J. Acoust. Soc. Am., **28**, 973-987.

Wan, R.W., Durlach, N.I., and Colburn, H.S. (**2010**). “Application of an extended equalization-cancellation model to speech intelligibility with spatially distributed maskers,” J. Acoust. Soc. Am., **128**, 3678-3690.

Zurek, P.M. (**1992**). “Binaural advantages and directional effects in speech intelligibility,” in *Acoustical Factors affecting Hearing Aid Performance*, 2nd ed. Edited by G.A. Studebaker and I. Hochberg (Allyn and Bacon, Boston), pp. 255-276.