

Blind Beamforming Using Parallel Single-channel Speech Enhancers

Benny Sällberg, Nedelko Grbić and Ingvar Claesson

Department of Signal Processing, Blekinge Institute of Technology, SE-372 25 Ronneby, Sweden

E-mail: bsa@bth.se

Abstract - This paper presents an idea to extend a certain class of single channel speech enhancement algorithms to include the spatial domain. The resulting blind beamformer does not rely on a-priori knowledge of source and sensor positions and it enhances one or several speech sources based only on received data. The underlying principle in this approach is the fact that speech signals are short time stationary. Provided that the single channel speech enhancers attenuates unwanted sources and at the same time preserve the short time stationarity of speech signals, a summation of a small array of such single channel processors constitutes a coherent spatial speech enhancement. As opposed to traditional beamforming where the phase alteration is pre-specified, the phase alteration of the proposed structure is controlled by the received data. The evaluation uses a two microphone array and indicates that the Signal to Interference Ratio is increased for a variety of source positions using the proposed method with only an insignificant decrease in speech quality.

Keywords - Blind Speech Enhancement, Adaptive Gain Equalizer, Beamforming

1 INTRODUCTION

Methods for blind speech enhancement can be separated into two dominant groups; Single channel methods and multiple channel methods. The underlying idea of the two concepts of single and multiple channel techniques are somewhat different where single channel methods operates in the time-frequency domain, e.g. [1, 2]. Multiple channel techniques also take the spatial domain into consideration, e.g. [3]. The lack of spatial diversity in the single channel techniques makes them in some cases unsuitable. This contribution investigates the beamforming capabilities when using a set of single channel blind speech enhancers in parallel, forming a multiple channel blind speech enhancer. The key feature of the proposed structure is that the phase alteration is controlled by the received data. This technique is different from traditional beamforming which uses pre-specified phase alterations. Two versions of the proposed structure are evaluated and compared to a single channel method. The evaluation assesses the Signal to Interference Ratio (SIR) enhancement and speech quality using the Perceptual Evaluation of Speech Quality (PESQ) measure from the International Telecommunication Union [4].

The outline of this paper is as follows; A single channel speech enhancer is formulated in Section 2. Blind beamforming using several single channel methods is presented in Section 3. The setup for evaluating the proposed structures is presented in Section 4 and performance measures are introduced in Section 5. Measured results are presented in Section 6. A short summary and conclusions are given in Section 7.

2 SINGLE CHANNEL SPEECH ENHANCEMENT

In this paper the Adaptive Gain Equalizer (AGE) [5] is used for blind speech enhancement. The AGE

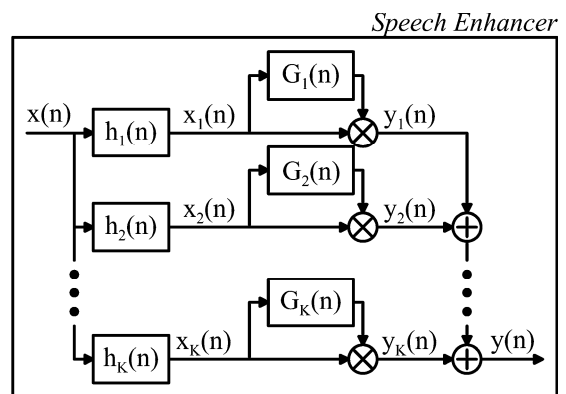


Figure 1: The Adaptive Gain Equalizer for blind speech enhancement.

is selected mainly due to its simplicity and good performance. Further, its robustness and scalability and the fact that it does not require supplementary structures like Voice Activity Detectors (VAD) makes the AGE a suitable method for speech enhancement. The AGE has also been shown to fit in a variety of implementations [6, 7]. The input-output signal assembly of the AGE is presented in Fig. 1.

2.1 AGE Input-Output Signal Assembly

The signal received at a single channel sensor is modelled as

$$x(n) = s(n) + v(n), \quad (1)$$

where $s(n)$ denotes the speech component and $v(n)$ denotes the noise component. A filter bank is used to transform the input signal into time-frequency domain, according to

$$x_k(n) = h_k(n) * x(n), \quad (2)$$

where $k \in [1, K]$ designates the subband index, $h_k(n)$ is the subband selective filter and $*$ denotes convolution. It should be noted that a more efficient fil-

ter bank would also employ decimation of the sub-band signals and incorporate a synthesis filter bank as well [8]. However, for the sake of readability a decimation scheme is not used here. A nonlinear gain function $G_k(n)$ is employed in the AGE. The gain function utilizes two exponential averages to track bursts of speech, $A_k(n)$, and the background noise level, $\underline{A}_k(n)$. The adaptive gain function is formulated as

$$A_k(n) = (1 - \alpha_k)A_k(n-1) + \alpha_k|x_k(n)|, \quad (3)$$

$$\underline{P}_k(n) = (1 - \beta_k)\underline{A}_k(n-1) + \beta_k|x_k(n)|, \quad (4)$$

$$\underline{A}_k(n) = \begin{cases} \underline{P}_k(n), & \text{if } \underline{P}_k(n) \leq A_k(n) \\ A_k(n), & \text{if } \underline{P}_k(n) > A_k(n) \end{cases}, \quad (5)$$

$$G_k(n) = \min\left(\frac{A_k(n)}{\underline{A}_k(n)}, L_k\right). \quad (6)$$

The variable $\underline{P}_k(n)$ is a prototype variable for temporary use, α_k and β_k are time constants controlling the integration time of the two averages. The function $\min(a, b)$ gives the minimum value of the two parameters a and b . The upper gain function limit L_k in combination with (5) makes the effective range of the gain function to become $1 \leq G_k(n) \leq L_k$ for $\forall n, k$. Thus, the AGE focuses on boosting of speech rather than suppression of noise. The output of the AGE method is the weighted combination of all sub-band signals, i.e.

$$y(n) = \sum_{k=1}^K G_k(n)x_k(n). \quad (7)$$

3 BLIND BEAMFORMING

The proposed approach is to use several single channel speech enhancers coupled in parallel to create a blind beamforming system. Three cases are presented:

- CASE I: A blind speech enhancer is attached to a single microphone. This corresponds to the single channel method described in Section 2 and is provided as a reference.
- CASE II: A fixed beamforming is carried out prior to the individual speech enhancers. The beamforming matrix is denoted \mathbf{U} and the input to the speech enhancers are $\mathbf{x}'(n) = \mathbf{U}\mathbf{x}(n)$ where $\mathbf{x}(n) = (x_1(n), \dots, x_M(n))^T$. The output of all speech enhancers are combined in the total output, see Fig. 2.
- CASE III: A speech enhancer is attached directly to each input signal. The outputs of all speech enhancers are combined in the total output, see Fig. 3.

4 EVALUATION SETUP

A two channel implementation of the Case II and Case III approaches described in Section 3 are compared to a single channel AGE, i.e. Case I. The test

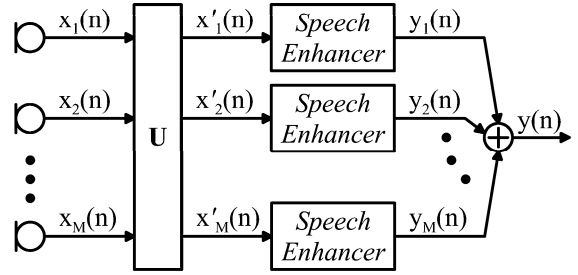


Figure 2: An M -channel speech enhancement system according to the Case II approach, here \mathbf{U} denotes a beamforming matrix.

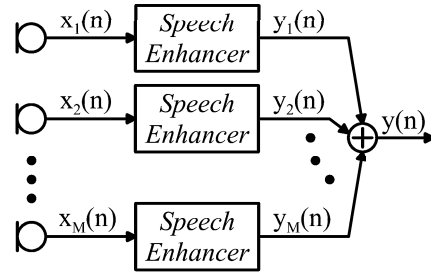


Figure 3: An M -channel speech enhancement system according to the Case III approach.

signals are recorded in a high-reverberant office room of size $2 \times 4 \times 2.5$ m where loudspeakers represent the speaker and the interference respectively. A sampling frequency of 8 kHz is used while it corresponds to standard telephone bandwidth.

4.1 Microphone and Source Positions

In the first part of the evaluation, only one speaker and one interference are present at each time. Two loudspeakers represent the speaker and interference respectively and five different loudspeaker positions are considered here. The setup is presented in Fig. 4 where \vec{p}_{S1} to \vec{p}_{S5} denotes possible loudspeaker positions and \vec{p}_{M1} , \vec{p}_{M2} designates the positions of the two microphones. The source and interference positions are $\vec{p}_{S_i} : (\cos(2\pi(i-1)/5), \sin(2\pi(i-1)/5))$ m, for $i = \{1, 2, 3, 4, 5\}$. The two microphones are situated 6 cm apart at positions $\vec{p}_{M1} : (-0.03, 0.00)$ m and $\vec{p}_{M2} : (+0.03, 0.00)$ m respectively.

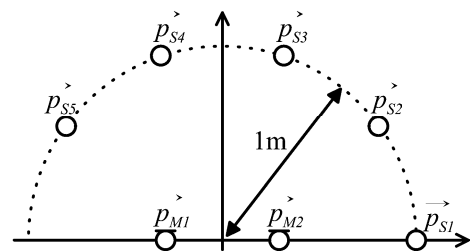


Figure 4: The points p_{S1} through p_{S5} denotes possible speaker and interference positions. The microphone positions are designated by p_{M1} and p_{M2} .

4.2 Test Signals

The speech signals used are two male and two female speakers from the TIMIT database concatenated into one sequence with half a second of separating silence. The noise is white Gaussian with zero mean. Three different Signal to Interference Ratios (SIR) are used in the evaluation corresponding to low, medium and high SIR according to $SIR = \{5dB, 15dB, 25dB\}$.

4.3 Speech Enhancer Settings

The time constant of the speech tracking average is set to 50 ms and the background noise tracking average time constant to 3 s. The maximal amount of speech enhancement was set to $L_k = 10^{15/20}$ to put a constraint on the maximum amount of artifacts introduced by the speech enhancer itself. Other recommended AGE settings can be found in [5]. This setting was used for all speech enhancers independently of configuration, i.e. CASE I, CASE II and Case III.

4.4 Case II Beamforming Matrix

The beamforming matrix \mathbf{U} in the Case II setup is set to

$$\mathbf{U} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \quad (8)$$

which forms two beams; A spatial low pass filter and an orthogonal spatial high pass filter. The sum of the two beamformers gives a spatial all pass characteristic.

5 PERFORMANCE MEASURES

Assessing the performance of speech enhancement algorithms is not a straightforward task. The measure should cover various aspects such as the amount of actual noise reduction as well as including the quality of speech. It is hard to find one measure that covers all desired performance aspects and two measures are used herein, motivated by [9, 10]. The first measure is the classical objective Signal to Interference Ratio (SIR) or rather a normalized version thereof where the SIR after enhancement is normalized with the SIR before enhancement. The second measure is the Perceptual Evaluation of Speech Quality from ITU-T (PESQ) [4]. The PESQ is also an objective measure but is based on cognitive models of the human hearing organ to form pseudo-subjective scores and it has high correlation with real subjective tests.

5.1 Differential SIR

The output of the speech enhancement structure is segregable into two components as

$$y(n) = y_s(n) + y_v(n), \quad (9)$$

where $y_s(n)$ includes enhanced components of speech and $y_v(n)$ includes components of noise and are both

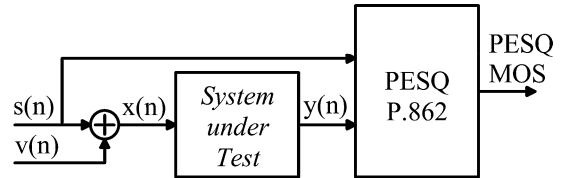


Figure 5: Setup for evaluating a system with environmental noise using the ITU-T PESQ P.862 measure.

of length N . The differential SIR, ΔSIR , is defined as the ratio of speech and noise in the signal before (the average SIR perceived by the set of microphones) and after the speech enhancement, according to

$$SIR_{before} = \frac{1}{M} \sum_{m=0}^{M-1} \frac{\widehat{Var}\{s_m\}}{\widehat{Var}\{v_m\}}, \quad (10)$$

$$SIR_{after} = \frac{\widehat{Var}\{y_s\}}{\widehat{Var}\{y_v\}}, \quad (11)$$

$$\Delta SIR = \frac{SIR_{after}}{SIR_{before}}. \quad (12)$$

Here, $\widehat{Var}\{\}$ is an estimator of variance, i.e.

$$\widehat{Var}\{x\} = \frac{1}{N} \sum_{n=0}^{N-1} |x(n) - m_x|^2, \quad (13)$$

$$m_x = \frac{1}{N} \sum_{n=0}^{N-1} x(n). \quad (14)$$

5.2 Differential PESQ

According to the ITU-T recommendation P.862 [4] it is possible to evaluate the speech performance using PESQ even for signals with noise. However, the reference signal must be clean. The setup for evaluating a system with environmental noise using the PESQ P.862 standard is presented in Fig. 5.

The PESQ of Case I is used as a reference to form a differential PESQ for Case II and Case III, i.e.

$$\Delta PESQ_X = PESQ_X - PESQ_I. \quad (15)$$

6 MEASUREMENT RESULTS

The following measured results are obtained by using the evaluation setup presented in Section 4 and the performance measures differential SIR and differential PESQ presented in Section 5.

6.1 Differential SIR

Differential SIR measures are plotted as a function of different speaker and interference positions for three input SIR levels in Fig. 6. This analysis indicates that we have approximately the same level of speech enhancement in the Case II as in the Case I approach. The Case III approach yields an overall gain of approximately 2 dB SIR enhancement compared to the single microphone approach in Case I.

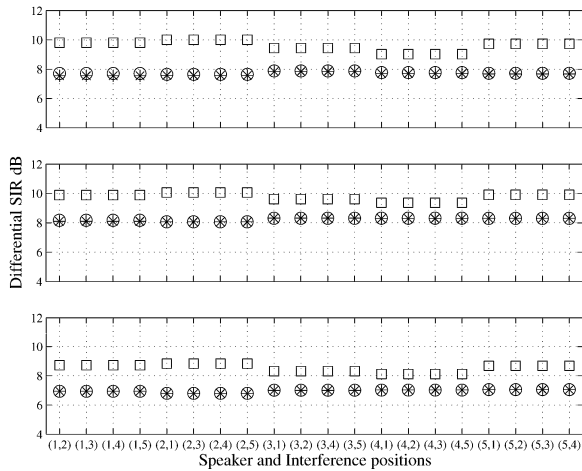


Figure 6: *Differential SIR enhancement for Case I (stars), Case II (circles) and Case III (boxes) as a function of different speaker and interference positions when the input SIR is 5 dB (upper plot), 15 dB (mid plot), and 25 dB (lower plot).*

6.2 Differential PESQ

Measures from the PESQ analysis are presented in Fig. 7. The results indicate that the maximal PESQ MOS deviation never exceeds ± 0.15 MOS when comparing the Case II to Case I and Case III to Case I. This in turn highlights that the speech distortion is minimal when introducing the proposed approaches (Case II and Case III) compared to the Case I approach.

6.3 Several Sources

An experiment with several sources (two speakers at positions \vec{p}_{S2} , \vec{p}_{S3} and one interference at position \vec{p}_{S5}) is evaluated using the differential SIR and differential PESQ. Measurements show that the differential SIR enhancement was approximately 1.5 dB per speaker with insignificant PESQ degradation for the Case III approach. The structure in Case II does not show any increase in SIR enhancement, consistent with previous results.

7 SUMMARY AND CONCLUSIONS

This paper presents an alternative approach to classical beamforming. Spatial diversity is achieved by using several single channel speech enhancers coupled in parallel. As opposed to classical beamforming where the phase alteration is pre-specified, the structure provided in this paper alters the phase based on the received data. Two different approaches are evaluated; one with a preprocessing fixed beamformer and one with a direct summation of the outputs of the single channel processors. While the first structure fails to provide a significant increase in performance, the other proposed structure provides approximately 2 dB of SIR improvement with only an insignificant decrease in speech quality, i.e. the PESQ measure, at all evaluated source positions.

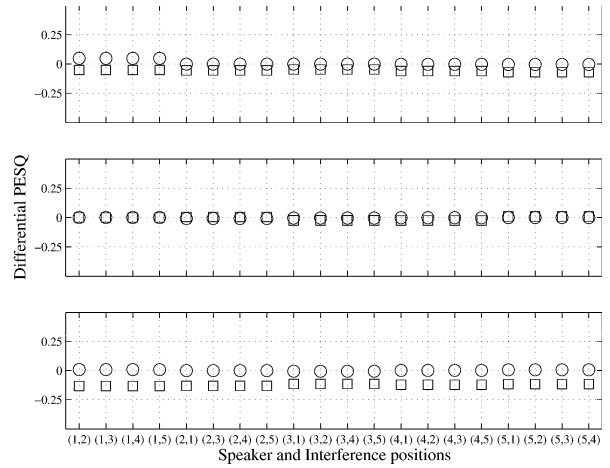


Figure 7: *Differential PESQ MOS results for Case II (circles) and Case III (boxes) related to the Case I PESQ MOS results as a function of different speaker and interference positions when the input SIR is 5 dB (upper plot), 15 dB (mid plot), and 25 dB (lower plot).*

The two sources evaluation in Section 6.3 indicates that the structure in Case III is not only capable of blindly enhancing a single source but also two sources simultaneously.

REFERENCES

- [1] S. F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust. Speech and Sig. Proc.*, 27:113–120, 1979.
- [2] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Trans. Acoust. Speech and Sig. Proc.*, 32:1109–1121, 1984.
- [3] W. Kellerman. A self-steering digital microphone array. *IEEE ICASSP*, 5:3581–3584, 1991.
- [4] ITU-T p.862. *Perceptual evaluation of speech quality (PESQ)*.
- [5] N. Westerlund, M. Dahl, and I. Claesson. Speech enhancement for personal communication using an adaptive gain equalizer. *Elsevier Signal Processing*, 85(6):1089–1101, 2005.
- [6] B. Sällberg, H. Åkesson, N. Westerlund, M. Dahl, and I. Claesson. Analog circuit implementation for speech enhancement purposes. *IEEE 38th Asilomar Conference*, 2004.
- [7] B. Sällberg, H. Åkesson, M. Dahl, and I. Claesson. A mixed analog - digital hybrid for speech enhancement purposes. *IEEE ISCAS*, 2005.
- [8] R. Crochiere and L. Rabiner. *Multirate Digital Signal Processing*. Prentice-Hall, 1983.
- [9] T. Rohdenburg, V. Hohmann, and B. Kollmeier. Objective perceptual quality measures for the evaluation of noise reduction schemes. *IWAENC*, pages 169–172, 2005.
- [10] A. Rix. Perceptual speech quality assessment - a review. *IEEE ICASSP*, (3):1056–1059, 2004.