EMPLOYING STOCHASTIC CONSTRAINED LMS ALGORITHM FOR ASR FRONTEND PROCESSING

Michael Stadtschnitzer, Daniel Stein, Rolf Bardeli

Fraunhofer IAIS, Sankt Augustin, Germany
name.surname@iais.fraunhofer.de

ABSTRACT

In scenarios with multiple input single output systems, the stochastic constrained least mean-squares (LMS) algorithm has been proven to be an effective approach. However, when only two input channels are available, it is unclear whether this approach still yields improvements. In this paper, we investigate the stableness and the robustness of the constrained LMS algorithm on "Track 1" of "2^{nd} CHiME Challenge" [1] and show that it leads to small yet consistent improvements on all signal-to-noise settings.

Index Terms— Stochastic constrained LMS algorithm, automatic speech recognition, two input channels

1. INTRODUCTION

The stochastic constrained least mean-squares (LMS) algorithm [2], also known as Frost's Beamformer, has been proven to be useful in multiple sensor scenarios (e.g. [3], [4]).

While designed for large quantities of sensors, the question remains whether this approach is useful for a two channel setting as well. In this paper, we employ the constrained LMS algorithm as an automatic speech recognition (ASR) frontend for the speech recognizer and test its performance on "Track 1" of the "2^{nd} CHiME Speech Separation and Recognition Challenge" [1]. The data is provided for two channels with realistic living-room noise and artificially mixed in speech. We show that constrained LMS leads to small yet consistent improvements on all signal-to-noise (SNR) settings.

2. ALGORITHM

The Frost's Beamformer [2] is a constrained LMS algorithm that is able to adapt an array of sensor weights to respond to a certain direction while attenuating noise from other directions. The algorithm requires only that the direction of arrival and a frequency band of interest is specified a priori. The algorithm is summarized as:

$$\min W^T R_{xx} W \quad \text{subject to } C^T W = \mathcal{F}\,,$$

where W is the adaptive filter weight vector, R_{xx} is the autocorrelation matrix of the input vector, C is the constraint matrix, and $\mathcal{F}$ is the J-dimensional vector of weights in the look-direction equivalent tapped delay line.

The optimum filter weights W_{opt} are then obtained by:

$$W_{\text{opt}} = R_{xx}^{-1} C \left[C^T R_{xx}^{-1} C\right]^{-1} \mathcal{F}\,.$$

The adaptive stochastic constrained LMS algorithm [2] is given by:

$$W(0) = F$$
$$W(k+1) = P\left[W(k) - \mu y(k) X(k)\right] + F\,,$$

where the $KJ \times KJ$-dimensional matrix P is defined by:

$$P := I - C\left(C^T C\right)^{-1} C^T\,,$$

and the KJ-dimensional vector F is defined by:

$$F := C\left(C^T C\right)^{-1} \mathcal{F}\,.$$

The positive scalar μ is the step-size parameter and thus a trade-of between convergence time and misadjustment from the optimum solution. The choice of μ and the convergence behavior is discussed in [2]. A computable upper bound for μ is given by:

$$\mu < \frac{2}{3E\left[X^T(k)\, X(k)\right]}\,.$$

3. EXPERIMENTAL SETUP

The number of sensors K is set to $K = 2$ and is determined by the number of audio channels of the provided signals. We did not account for the target speaker movements and hence, the look-direction was set perpendicular to the line of sensors towards the target speaker. To let the desired signals pass from look-direction without distortion, the weights of the look-direction-equivalent tapped delay line $\mathcal{F}$ are set for odd numbers of J to

$$\mathcal{F}^T = \left[\delta(-(J-1)/2), \ldots, \delta(0), \ldots, \delta((J-1)/2)\right]\,,$$

where $\delta(n)$ is the discrete-time unit impulse function. For a given β, the step-size parameter μ is set to

$$\mu = \frac{2\beta}{3E\left[X^T(k) X(k)\right]}.$$

Table 1. Performance in terms of keyword accuracies [%] on development set (left) and test set (right). The results for $\beta = 4.9 \cdot 10^{-3}$ are listed for comparison reasons only, since they do not reflect the optimal development setting.

method	β	-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB	-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB
left	—	50.17	56.08	64.67	73.75	77.33	80.92	48.17	57.93	67.17	73.33	78.50	82.58
right	—	42.58	47.67	58.00	68.08	74.00	77.92	42.25	49.58	59.92	68.25	73.33	78.42
DS/BL	—	49.67	57.92	67.83	73.67	80.75	82.67	49.33	58.67	67.50	75.08	78.83	82.92
s.c.LMS	$5.6 \cdot 10^{-3}$	50.67	59.58	67.42	75.00	82.00	82.83	50.00	60.25	68.67	75.67	80.17	82.92
s.c.LMS	$4.9 \cdot 10^{-3}$	50.08	59.08	68.08	74.17	80.92	83.50	50.75	60.17	68.75	76.83	80.00	83.00

4. EVALUATION

The provided waveforms were processed by the algorithm and new models were trained using the processed waveforms. The parameter J was empirically set to the value of 7 (out of $3, 5, 7, 9, 11$, and 13 tested), since it lead to consistent improvement for all SNR settings on the development corpus.

Figure 1 shows the mean improvement (calculated over all SNRs) in the keyword accuracy compared to the baseline (BL) configuration, based on the step size β. All presented values of β improve the mean keyword accuracy of both development and test set. For $\beta = 5.6 \cdot 10^{-3}$, there is a local optimum for the development set which translates reasonably on the test set (here, the optimum would be at $\beta = 4.9 \cdot 10^{-3}$). Table 1 shows the results for the individual SNR settings. We also provide the results for the single channel configurations (left channel, right channel). The BL configuration, which takes the sum (or more correctly the average) of the two input channels, correponds to the delay-and-sum (DS) beamformer in that special case. The target speaker is situated in the line perpendicular to the microphone array axis, since we do not account for target speaker movements, and the target speaker signals are assumed to imping coherently (i.e. without any delay, which otherwise has to be compensated by the DS beamformer) on the microphones. The results of the DS configuration are, as expected, an improvement compared to the single channel configurations. The performance is furthermore improved for every SNR condition of the test set by Frost's Beamformer, when compared to the DS/BL system.

5. CONCLUSION

In this paper, we applied the constrained LMS algorithm for the extreme case of two input channels. On "Track 1" of "2^{nd} CHiME challenge", we showed that the algorithm leads to small yet consistent improvements compared to the BL configuration. We therefore conclude that this algorithm serves as a meaningful frontend addition when robustifying the ASR. As future work, we plan to employ an additional source localization algorithm so that target speaker movements can be incorporated into this workflow.

Acknowledgements: This work has been partly funded by the European Community's Seventh Framework Programme (FP7-ICT) under grand agreement n° 269980 AXES.

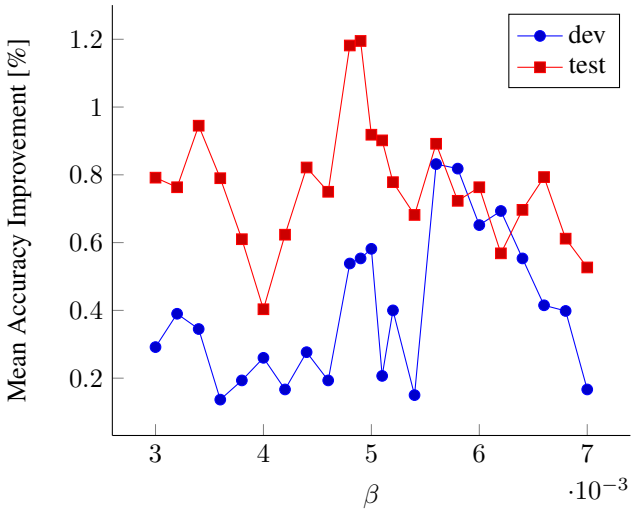

Fig. 1. Mean keyword accuracy improvement, for $J = 7$ and various β compared to the baseline configuration

6. REFERENCES

[1] Emmanuel Vincent, Jon Barker, Shinji Watanabe, Jonathan Le Roux, Francesco Nesta, and Marco Matassoni, "The Second 'CHiME' Speech Separation and Recognition Challenge: Datasets, Tasks and Baselines," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, May 2013.

[2] Otis Lamond Frost, "An Algorithm for Linearly Constrained Adaptive Array Processing," *Proceedings of the IEEE*, vol. 60, no. 8, pp. 926–935, 1972.

[3] Yong Zhao, Wei Liu, and Richard J. Langley, "Adaptive Wideband Beamforming with Response Variation Constraints," *Proc. of EUSIPCO 2010*, pp. 2077–2081, 2010.

[4] Jacob Benesty, Jingdong Chen, Yiteng (Arden) Huang, and Jacek Dmochowski, "On Microphone-Array Beamforming From a MIMO Acoustic Signal Processing Perspective," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 1053–1065, Mar. 2007.