ELO-SPHERES consortium system description

Alastair H. Moore[1], *Sina Hafezi*[1], *Rebecca Vos*[1], *Mike Brookes*[1], *Patrick A. Naylor*[1], *Mark Huckvale*[2], *Stuart Rosen*[2], *Tim Green*[2], *Gaston Hilkhuysen*[2]

[1]Imperial College London, UK [2]University College London, UK

alastair.h.moore@imperial.ac.uk

Abstract

The Clarity Challenge provides an excellent opportunity to stimulate novel, performant machine learning approaches to hearing aid signal enhancement. It is important that these methods are compared with classical methods which are well understood. Here, an adaptive beamformer based on the minimum-variance distortionless response design approach is proposed as a superior baseline against which machine learning approaches can be benchmarked. The design exploits documented characteristics of the Challenge rules to identify noise-only segments and the direction-of-arrival of the target. Hearing-aid specific modifications include automatic gain control and listener-specific hearing loss compensation. On the *dev* dataset the proposed method obtains a mean MBSTOI metric of 0.61 compared to the baseline system which achieves 0.41.

Index Terms: MVDR beamformer, adaptive beamforming, direction-of-arrival estimation, hearing aids

1. Introduction

Binaural hearing aids (HAs) allow microphone signals to be passed between devices, enabling a pair of devices to be treated as a single array. Recent work in binaural beamforming for HAs has focused on binaural cue preservation for interfering sources and post-filtering approaches [1]. However, in this submission to the 2021 Clarity Enhancement Challenge [2], we employ classical minimum variance distortionless response (MVDR) beamforming [3] and prioritise linear signal processing approaches in the hope that, by minimising signal distortion, intelligibility can be improved.

During initial investigations it was noted that, since the spatial arrangement is static within a scene, a good estimate of the spatial properties of the noise could be obtained within about 0.5 s. It was also determined that steering the beam in the correct direction, or at least within ±7.5°, is critical. Using the correct HRIR was comparatively unimportant. Nevertheless, our system attempts to select the correct one.

2. Formulation

Working in the short time Fourier transform (STFT) domain, where ℓ and k are time and frequency indices, respectively, the target source signal, $S(k,\ell)$, is received at M microphones. Expressed in vector notation, the clean target microphone signals, $\mathbf{x}^{(d)}(k,\ell)$, are given by

$$\mathbf{x}^{(d)}(k,\ell) = \mathbf{H}(k)^T S(k,\ell) \tag{1}$$

where $\mathbf{H}(k)$ is the stacked Fourier transform of the direct path impulse responses between the target and the array and $(\cdot)^T$ denotes the transpose. In accordance with the Challenge definition, the first two channels of $\mathbf{x}^{(d)}(k,\ell)$ are the front microphone reference signals used by the MBSTOI metric.

The observed microphone signals, $\mathbf{y}(k,\ell)$, are expressed as

$$\mathbf{y}(k,\ell) = \mathbf{x}^{(d)}(k,\ell) + \mathbf{x}^{(r)}(k,\ell) + \mathbf{v}(k,\ell) \tag{2}$$

where $\mathbf{x}^{(r)}(k,\ell)$ is reflected sound due to the target and $\mathbf{v}(k,\ell)$ is the reverberant signal due to the masker. An estimate of the desired signal at the mth reference microphone, $Z_m(k,\ell)$, is obtained using

$$Z_m(k,\ell) = \mathbf{w}_m(k)^H \mathbf{y}(k,\ell) \tag{3}$$

where $(\cdot)^H$ is the conjugate transpose. The beamformer weights, $\mathbf{w}_m(k)$, are obtained at each k according to [3]

$$\mathbf{w}_m = \mathbf{R}_\varepsilon{}^{-1}\mathbf{d}_m \left[\mathbf{d}_m^H \mathbf{R}_\varepsilon{}^{-1}\mathbf{d}_m\right]^{-1} \tag{4}$$

where $\mathbf{R}_\varepsilon = \mathbf{R} + \varepsilon\mathbf{I}$, $\mathbf{I}$ is the identity matrix and $\varepsilon \geq 0$ is set to limit the condition number of $\mathbf{R}_\varepsilon$ to $\leq$1000. The frequency index is omitted from (4) for clarity.

The choice of steering vector, $\mathbf{d}_m(k)$, and covariance matrix, $\mathbf{R}(k)$, determine the extent of noise reduction. Ideally $\mathbf{d}_m(k)$ is the anechoic relative transfer function (RTF) of the array for a source in the target direction with respect to the mth microphone. In the context of the Challenge, measured array responses are available for 19 heads (or HRIRs) over a grid of directions, one of which corresponds to the true response. Our system estimates the correct HRIR from the received signals.

For maximum noise reduction one can choose $\mathbf{R}(k) = \mathbb{E}\{\mathbf{y}(\ell)\mathbf{y}^H(\ell)\}$, or equivalently, $\mathbf{R}(k) = \mathbb{E}\{(\mathbf{x}^{(r)}(k,\ell) + \mathbf{v}(k,\ell))(\mathbf{x}^{(r)}(k,\ell) + \mathbf{v}(k,\ell))^H\}$. However, the inclusion of coherent reflections in the covariance matrix can lead to signal cancellation. We therefore define our oracle noise covariance matrix (NCM) as

$$\mathbf{R}(k) = \mathbb{E}\{\mathbf{v}(\ell)\mathbf{v}^H(\ell)\}. \tag{5}$$

3. Implementation

Processing is implemented in MATLAB using a frame length of 200 samples (4.54 ms at 44.1 kHz) with 50 % overlap. The FFT size is also 200 so that algorithmic delay is <5 ms. Open source code is available[1].

On each file, the beamformer is designed assuming the HRIR is 'BuK', the direction of arrival (DOA) is 0° and $\mathbf{R}(k) = \mathbf{I} \forall k$, indicating spatially white noise. Adaptation is achieved by regularly updating parameter estimates. During the first 2 s, $\mathbf{R}(k)$ is estimated every 200 ms using the ensemble average of available frames. Between 2.1 s and 2.5 s, the DOA of the target and HRIR are estimated every 100 ms, as described below. After each update a new beamformer is designed and linear cross-fading used to switch in the new beamformer.

[1] https://github.com/alastairhmoore/clarity-challenge-2021-enhancer

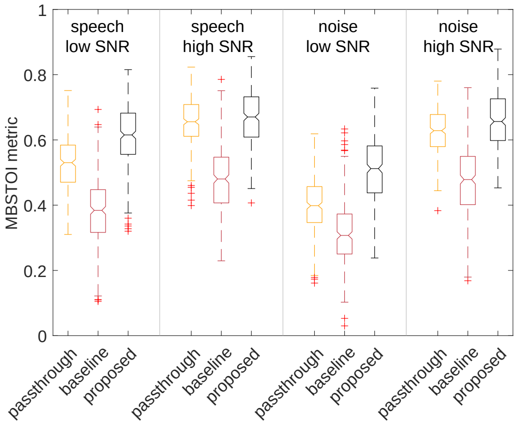

Figure 1: MBSTOI *distribution for noisy signals (passthrough), baseline approach and our proposed system. Results are shown for four subsets of the train dataset selected according to masker type (speech vs noise) and signal to noise ratio (SNR).*

3.1. Novel DOA and HRIR selection

At each update, the steering vector is selected according to the estimated DOA of the target and HRIR. Based on our recent work on model-based beamforming [4], we approximate the noisy signal covariance as the sum of the noise covariance and the covariance of the anechoic target, thus neglecting the contribution of the target's reverberation. The chosen DOA at each frequency is that which minimises the Frobenius norm of the difference between the sample covariance matrix and the modelled covariance. Steering vectors for all HRIRs are included in this optimisation. The final estimate of the DOA is obtained as the maximum of a histogram of DOAs selected at frequencies between 500 Hz and 16 kHz. Using only those frequencies at which the narrowband estimate of DOA equals the final estimated DOA, the most frequently selected HRIR is taken as the estimated HRIR. Estimates are updated at 0.1 s intervals between 2.1 s and 2.5 s using an STFT with 50 ms frames overlapping by 50 %, respecting the 5 ms look ahead constraint.

3.2. Levels and hearing loss correction

Conventionally, dynamic range compression is used to maximise the audible energy. However, introducing non-linearities may reduce intelligibility and does distort the envelope correlations used in MBSTOI. Accordingly, the proposed system employs automatic gain control (AGC) to limit the energy in the input signal to 65 dB SPL. The AGC has an effective release time of infinity, i.e. having been lowered to accommodate a peak the gain does not increase. The AGC gain is computed per frame in the STFT domain from the input signal but applied as a smoothed gain (time constant: 0.1 s) in the time domain as a post process, after beamforming and hearing loss (HL) compensation, taking care to respect causality constraints. Additionally, to avoid transients during initial adaptation, the output is muted for the first 200 ms and faded in over the following 500 ms.

Based on the assumption that the input signal is at 65 dB SPL, frequency-dependent hearing loss compensation is applied using the gain tables obtained using the Challenge baseline system. This is applied as scalar gains in the STFT domain.

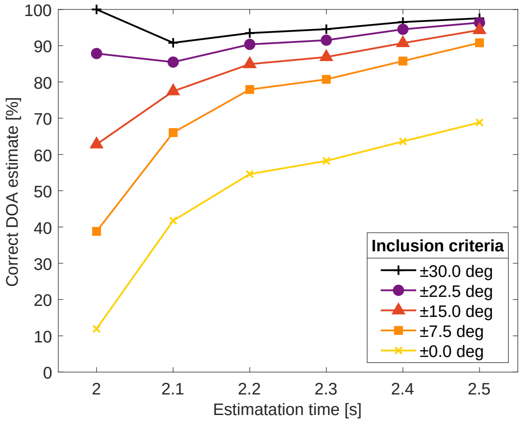

Figure 2: *DOA estimation accuracy as a function of update time for dev dataset.*

	Baseline	Proposed
Mean MBSTOI	0.41	0.61

Table 1: *Performance metric on dev dataset*

4. Results

Figure 1 shows the distribution of MBSTOI scores obtained for 4 subsets of the *train* dataset. Each subset contains the first 100 scenes in which masker is speech (or noise) and the SNR is within 1 dB of the lowest (or highest) SNR for that masker type. Curiously, the baseline system actually seems to degrade performance. The proposed method is substantially better than both the original signals and the baseline system. The benefit is greatest in the low SNR cases where there is more room for improvement.

Figure 2 shows the proportion of files in the *dev* dataset in which the DOA is correctly estimated as a function of the estimation time, where 'correct' is defined according to the absolute error in estimated angle. At time 2 s the estimated DOA is always 0° which limits the maximum error, but is rarely correct. By 0.5 s after the target starts, the estimated DOA is within ±7.5° of the true DOA in 90.8 % of scenes.

The final MBSTOI metric for the *dev* dataset is shown in Table 1. Running on a 2.4 GHz Quad-Core Intel Core i5 MacBook Pro with 16 GB of RAM, the averaged elapsed time for the proposed enhancement algorithm is approximately 14 s per file. Of this, over 9 s is taken by the DOA estimation and HRIR selection processing. With a little optimisation to avoid computing unused intermediate results this could be substantially reduced.

5. Conclusion

The proposed sytem follows a conventional MVDR beamforming paradigm and attempts to avoid excessive signal modulations. The final performance score on the *dev* dataset, according to the Challenge-provided MBSTOI is 0.61.

6. Acknowledgement

This work was supported by the Engineering and Physical Sciences Research Council [grant number EP/S035842/1].

7. References

[1] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic beamforming for hearing aid applications," S. Haykin and K. J. R. Liu, Eds. John Wiley & Sons, Inc., 2010, pp. 231–268.

[2] S. Graetzer, T. Barker, T. J. Cox, M. Akeroyd, J. F. Culling, G. Naylor, E. Porter, and R. Viveros Muñoz, "Clarity-2021 challenges: Machine learning challenges for advancing hearing aid processing," in *Proc. Conf. of Int. Speech Commun. Assoc. (INTERSPEECH)*, Brno, Czech Republic, 2021.

[3] J. Capon, "High resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.

[4] A. Moore, P. Naylor, and M. Brookes, "Improving robustness of adaptive beamforming for hearing devices," in *Proc. Int. Symp. on Auditory and Audiological Research. (ISAAR)*, vol. 7, Nyborg, Denmark, Jul. 2019, pp. 305–316.