A binaural MVDR beamformer for the 2021 Clarity Enhancement Challenge: 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

An adaptive beamformer based on the minimum-variance distortionless response design approach is proposed in the context of the 2021 Clarity Enhancement Challenge. The beamformer aims to improve the signal-to-noise ratio of the target signal while broadband automatic gain control and linear filtering compensate for listener-specific hearing loss. The proposed system exploits a priori knowledge of the target onset time to estimate essential beamforming parameters with more certainty than might be expected in unconstrained listening conditions.

On the *eval* dataset the proposed method obtains a mean MBSTOI metric of 0.66 and, for the 21 hearing impaired listeners for whom data was available, a mean "correctness" of 81.1 %. This a substantial improvement over the baseline which achieves 0.31 and 39.8 %, respectively.

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

1. Introduction

The 2021 Clarity Enhancement Challenge [1], hereafter the Challenge, requires that entrants optimally process simulated stimuli containing a mixture of a single target and single noise or speech interferer, under mildly reverberant conditions. Performance is measured using both the modified binaural short-time objective intelligibility measure (MBSTOI) metric [2], modified to simulate hearing loss, and using speech intelligibility experiments with hearing impaired (HI) listeners. In this paper, we present the system contributed by the Environment and Listener-Optimised Speech Processing for Hearing Enhancement in Real Situations (ELO-SPHERES) project[1] consortium.

To be effective, hearing aids (HAs) must make a desired sound source both audible and intelligible. The former requires that the HAs provide sufficient, frequency-dependent gain to overcome a listener's raised hearing threshold, whilst avoiding discomfort-inducing over amplification. The latter requires that undesired signals which mask the desired target be selectively attenuated. The system presented here performs these two functions independently, as shown in Fig. 1. Improved audibility is achieved using broadband automatic gain control and listener-specific linear filtering, as described in Section 3. Improved intelligibility is achieved using binaural beamforming, as described in Section 2.

Many speech enhancement systems use non-linear processing to selectively attenuate time-frequency regions dominated by noise. This can introduce disturbing artefacts. Typically, HAs include non-linear processing in the form of frequency-dependent wide dynamic range compression (WDRC). This ensures audibility across a range of scenarios, where the signal level can vary drastically. However, in the Challenge, where the scenarios are relatively constrained, fast-acting WDRC is unlikely to be necessary and may actually reduce the signal to noise ratio (SNR) [3]. In light of the potential negative effects of non-linear processing, our system uses linear processing throughout.

[1]https://www.imperial.ac.uk/speech-audio-processing/projects/elo-spheres/

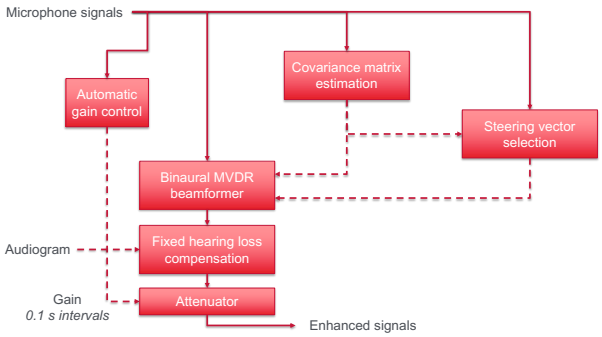

Figure 1: *System diagram showing main processing blocks.*

2. Beamformer

Beamforming, or spatial filtering, combines the signals from multiple microphones to enhance incident sound from a desired direction while attenuating sound from other directions. In general, using more microphones and spreading them over a larger volume improves the performance of beamforming. Binaural HAs allow microphone signals to be passed between devices, enabling a pair of devices to be treated as a single array. In binaural beamforming, two beamformers are implemented, leading to an output for the left and right ears. Depending on how the beamformers are designed, the binaural outputs will contain different spatial cues. Recent work in binaural beamforming for HAs has focused on binaural cue preservation for interfering sources and post-filtering approaches [4, 5, 6]. However, in this submission to the Challenge [1], we employ classical minimum variance distortionless response (MVDR) beamforming [7] and only attempt to maintain the spatial attributes associated with the direct path component of the target source. As a result, the residual noise is perceived to originate from the same position as the target.

2.1. Signal model

Working in the short time Fourier transform (STFT) domain, where k and ℓ are frequency and time indices, respectively, the monaural target source signal, $S(k, \ell)$, is received at M micro-

phones. The direct-path signal, $X_m^{(d)}(k,\ell)$, received at the mth microphone is

$$X_m^{(d)}(k,\ell) = H_m(k)S(k,\ell) \tag{1}$$

where $H_m(k)$ is the Fourier transform of the direct-path impulse response between the target and the mth microphone. We refer to $X_m^{(d)}(k,\ell)$ as the *desired* signals. Without loss of generality, $m = 1$ and $m = 2$ are used to denote the reference (front) microphones at the left and right ear, respectively.

The total received signal, $Y_m(k,\ell)$, at the mth microphone is

$$\begin{aligned} Y_m(k,\ell) =& X_m^{(d)}(k,\ell) + X_m^{(r)}(k,\ell) + V_m(k,\ell) && (2)\\ =& X_m(k,\ell) + V_m(k,\ell) && (3)\end{aligned}$$

where $X_m^{(r)}(k,\ell)$ is the reflected signal due to the target source, $X_m(k,\ell)$ is the total signal due to the target source and $V_m(k,\ell)$ is the sound due to the interferer. We refer to $Y_m(k,\ell)$ as the *received* signals.

Using vector notation and substituting (1), (2) can be written for all $m = 1 \dots M$ as

$$\mathbf{y}(k,\ell) = \mathbf{h}(k)S(k,\ell) + \mathbf{x}^{(r)}(k,\ell) + \mathbf{v}(k,\ell) \tag{4}$$

where $\mathbf{y}(k,\ell) = \left[Y_1(k,\ell) \cdots Y_M(k,\ell)\right]^T$, $\mathbf{h}(k)$, $\mathbf{x}(k,\ell)$ and $\mathbf{v}(k,\ell)$ are similarly defined and $(\cdot)^T$ denotes the transpose.

2.2. MVDR formulation

The beamformer output, $Z_m(k,\ell)$, which is an estimate of the desired signal at the mth reference microphone, $X_m^{(d)}(k,\ell)$, is obtained using

$$Z_m(k,\ell) = \mathbf{w}_m^H(k)\mathbf{y}(k,\ell) \tag{5}$$

where $(\cdot)^H$ is the conjugate transpose.

By design, the MVDR beamformer minimises the power in the output signal subject to a constraint that the desired signal should be passed undistorted. In this case, the beamformer weights, $\mathbf{w}_m(k)$, are obtained at each k according to [7]

$$\mathbf{w}_m(k) = \frac{{\mathbf{R}_\varepsilon}^{-1}(k)\mathbf{d}_m(k)}{\mathbf{d}_m^H(k){\mathbf{R}_\varepsilon}^{-1}(k)\mathbf{d}_m(k)} \tag{6}$$

where $\mathbf{R}_\varepsilon(k) = \mathbf{R}(k) + \varepsilon(k)\mathbf{I}$, $\mathbf{I}$ is the identity matrix and $\varepsilon(k) \geq 0$ is set to limit the condition number of $\mathbf{R}_\varepsilon(k)$ to $\leq$1000. Crucially, (6) defines the weights, and therefore the beamformer behaviour, in terms of a covariance matrix, $\mathbf{R}(k)$, and a steering vector, $\mathbf{d}_m(k)$. The choice of these quantities is discussed next.

2.3. Covariance matrix estimation

A covariance matrix quantifies the interchannel coherence between each pair of microphone signals.

The desired, interferer and received signal covariance matrices are defined respectively as

$$\begin{aligned} \mathbf{R}_{\mathbf{x}^{(d)}}(k) =& \mathbb{E}\{\mathbf{x}^{(d)}(k,\ell)(\mathbf{x}^{(d)})^H(k,\ell)\} && (7)\\ \mathbf{R}_{\mathbf{v}}(k) =& \mathbb{E}\{\mathbf{v}(k,\ell)\mathbf{v}^H(k,\ell)\} && (8)\\ \mathbf{R}_{\mathbf{y}}(k) =& \mathbb{E}\{\mathbf{y}(k,\ell)\mathbf{y}^H(k,\ell)\} && (9)\end{aligned}$$

where $\mathbb{E}\{\cdot\}$ denotes expectation over time.

In MVDR beamforming, $\mathbf{R}(k)$ defines the covariance matrix of the noise which should be attenuated. In real-world usage, it is common to assume the noise field has certain, fixed, spatial characteristics, such as isotropic noise [8]. In this case, $\mathbf{R}(k)$, is signal independent, which means it can be calculated a priori, but is suboptimal. Alternatively, one can obtain an adaptive estimate of the true, time-varying covariance matrix. However, if coherent reflections from the target are erroneously included in the estimate, this can lead to attenuation of the desired signal.

The Challenge scenario is peculiar in that the positions of all sources and microphones are static throughout a trial and there is always a 2 s interferer-only period before the target onset. During initial investigations it was noted that a good estimate of $\mathbf{R}_{\mathbf{v}}(k)$ could be obtained by approximating the expectation in (9) using the ensemble average over the first 0.5 s of frames and that extending the averaging period improves the estimate, up to 2.0 s seconds, when the target starts.

2.4. Steering vector selection

To obtain the best MBSTOI score, $\mathbf{d}_m(k)$ should be the relative transfer function (RTF) derived from the direct-path impulse response from the target source to the array, normalised with respect to the mth microphone

$$\mathbf{d}_m = \left[\frac{H_1}{H_m} \cdots \frac{H_M}{H_m}\right]^T \tag{10}$$

where the dependence on k is omitted for clarity.

Estimating the RTF from the received signal [9, 10] will generally yield a response which includes the early reflections and becomes unreliable at low SNRs. Therefore, it is common to derive $\mathbf{d}_m(k)$ from a database. Typically, impulse responses are measured from a spherical grid of source directions, indexed as $q \in 1 \dots Q$, to the ears of several real people and/or mannequins where the 'head' is indexed as $p \in 1 \dots P$. Using an individually measured impulse response which matches the true direct-path impulse response for the target direction leads to improved speech intelligibility [8].

In the context of the Challenge, the database used to simulate the microphone signals is available. It contains 19 'heads' and the target direction of arrival (DOA) is constrained to be one of 9 possible target directions, spaced 7.5° apart. Our system attempts to select the correct impulse response by estimating p and q.

In the frequency domain, the available direct-path impulse responses are denoted $\mathbf{h}_{p,q}(k)$.

Based on our recent work on model-based beamforming [11], we assume that the received signal covariance, $\mathbf{R}_{\mathbf{y}}(k)$ can be approximated as

$$\mathbf{R}_{\mathbf{y}}(k) \approx \mathbf{R}_{\mathbf{x}^{(d)}}(k) + \mathbf{R}_{\mathbf{v}}(k) \tag{11}$$

thus neglecting the contribution of the target's reverberation. Noting that $\mathbf{R}_{\mathbf{x}^{(d)}}(k)$ is a rank-1 matrix and is proportional to $\mathbf{h}(k)$, we model $\mathbf{R}_{\mathbf{y}}(k)$ as

$$\tilde{\mathbf{R}}_{\mathbf{y}}(k) = \alpha\mathbf{h}_{p,q}(k)\mathbf{h}_{p,q}^H(k) + \beta\hat{\mathbf{R}}_{\mathbf{v}}(k) \tag{12}$$

where α and β are scalar parameters and $\hat{\mathbf{R}}_{\mathbf{v}}(k)$ is an estimate of the interferer covariance matrix, obtained during the interferer-only interval.

Estimates of the DOA and 'head' are obtained independently at each k corresponding to frequencies between 500 Hz

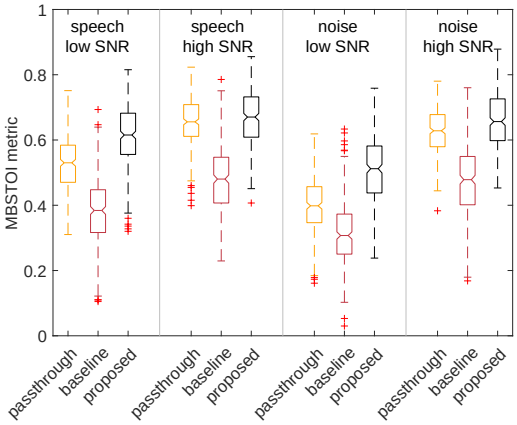

Figure 2: MBSTOI *distribution for noisy signals (passthrough), baseline approach and our proposed system. Results are shown for four subsets of the train dataset selected according to masker type (speech vs noise) and SNR.*

and 16 kHz by solving

$$\arg\min_{p,q,\alpha,\beta} \left\{ \left\| \mathbf{R}_{\mathbf{y}}(k) - \tilde{\mathbf{R}}_{\mathbf{y}}(k) \right\|_F^2 \right\} \tag{13}$$

where $\|\cdot\|_F^2$ denotes the Frobenius norm and α and β are scalar parameters which account for the relative powers in the target and interferer signals, respectively.

Since the simulated scenes use the same p and q at all k, an heuristic method is used to select a single impulse response; first the DOA is chosen as the mode of the per-frequency DOA estimates; then the 'head' is chosen as the mode of the per-frequency head estimates, considering only the subset of frequencies at which the estimated DOA matches the final estimate.

During initial investigations it was observed that selecting the correct q, or at least the neighbouring grid index, was crucial for good beamforming performance. In contrast, correctly selecting p was less important.

3. Hearing loss compensation

In the baseline system, and in many commercial hearing aids, multiband dynamic range compression is used to maximise the audible energy in different frequency regions. However, introducing non-linearities may reduce intelligibility and does distort the envelope correlations used in MBSTOI. Furthermore, we suspect that time-varying manipulations of the scene may be distracting to listeners. For the Challenge, we propose broadband level control combined with fixed linear filtering to partially compensate for the listener's audiogram.

3.1. Audiogram compensation

The baseline system [1] adopts the openMHA [12] implementation of the Camfit compressive fitting rule [13]. As a preprocessing stage, this generates a gain table in which the gain in each frequency band is stored for each possible input signal level. The proposed system uses the entries from this table corresponding to an input signal level of 65 dB. Frequency dependent gain is applied in the STFT domain by interpolating the gain table values to the frequency resolution of the Fourier transform. Since the signal level is assumed to be fixed, this filtering process is fixed over time.

Dataset	Baseline	Proposed
dev	0.41	0.61
eval	0.31	0.66

Table 1: *Mean* MBSTOI *performance metric*

3.2. Automatic gain control

The process described above assumes that there is sufficient headroom to apply the specified frequency-dependent gain. To ensure this assumption is valid, the level of the input signal is reduced, if required.

The root mean square energy is computed in each frame and recursively smoothed with a time constant of 200 ms. If the smoothed energy exceeds 65 dB SPL, the automatic gain control (AGC) gain is reduced, allowing 6 dB of headroom. The AGC has an effective release time of infinity, i.e. having been lowered to accommodate a peak the gain does not increase.

4. Implementation

Processing is implemented in MATLAB using a frame length of 220 samples (4.99 ms at 44.1 kHz) with 50 % overlap. The FFT size is also 220 so that algorithmic delay is <5 ms. Open source code is available[2].

On each file, the beamformer is initialised assuming the 'head' is the Brüel & Kjær mannequin labelled 'BuK', the 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. It remains fixed thereafter.
- Between 2.1 s and 2.5 s, the DOA of the target and 'head' are estimated every 100 ms as described in Sec 2.4. For this, an STFT with 50 ms frames overlapping by 50 % is used, taking care to respect the 5 ms look ahead constraint.

After each update a new beamformer is designed and linear crossfading used to switch in the new beamformer.

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.

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 'head' selection processing. With a little optimisation to avoid computing unused intermediate results this could be substantially reduced.

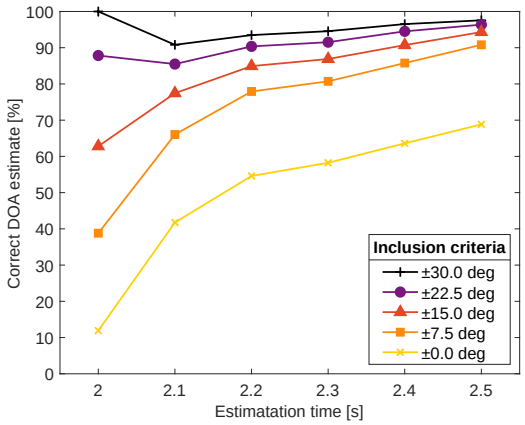

Figure 3: *DOA estimation accuracy as a function of update time for dev dataset.*

5. Results

Figure 2 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 3 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* and *eval* datasets are shown in Table 1.

At the time of writing, speech intelligibility evaluations are ongoing. Based on the results of 23 HI listeners, the baseline methods achieves 'correctness' score of 39.8 % whereas the proposed achieves 81.1 %.

6. Conclusion

The proposed sytem follows a conventional MVDR beamforming paradigm and attempts to avoid excessive signal modulations. It provides a substantial improvement over the baseline in terms of both the Challenge-provided MBSTOI metric and in speech intelligibility experiments using HI listeners.

7. Acknowledgement

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

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

8. References

[1] 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.

[2] A. H. Andersen, J. M. d. Haan, Z. H. Tan, and J. Jensen, "Refinement and validation of the binaural short time objective intelligibility measure for spatially diverse conditions," *Speech Commun.*, vol. 102, pp. 1–13, Sep. 2018.

[3] P. E. Souza, L. M. Jenstad, and K. T. Boike, "Measuring the acoustic effects of compression amplification on speech in noise," *J. Acoust. Soc. Am.*, vol. 119, no. 1, pp. 41–44, Jan. 2006.

[4] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic beamforming for hearing aid applications," in *Handbook on Array Processing and Sensor Networks*, S. Haykin and K. J. R. Liu, Eds. John Wiley & Sons, Inc., 2010, pp. 269–302.

[5] S. Doclo, W. Kellermann, S. Makino, and S. Nordholm, "Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones," *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 18–30, 2015.

[6] V. Best, E. Roverud, C. R. Mason, and G. Kidd, "Examination of a hybrid beamformer that preserves auditory spatial cues," *J. Acoust. Soc. Am.*, vol. 142, no. 4, pp. EL369–EL374, Oct. 2017.

[7] M. S. Brandstein and D. B. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*. Berlin, Germany: Springer-Verlag, 2001.

[8] A. H. Moore, J. M. de Haan, M. S. Pedersen, P. A. Naylor, M. Brookes, and J. Jensen, "Personalized signal-independent beamforming for binaural hearing aids," *J. Acoust. Soc. Am.*, vol. 145, no. 5, pp. 2971–2981, 2019.

[9] M. Taseska and E. A. P. Habets, "Relative transfer function estimation exploiting instantaneous signals and the signal subspace," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2015.

[10] S. Markovich-Golan and S. Gannot, "Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015, pp. 544–548.

[11] 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.

[12] H. Kayser, T. Herzke, P. Maanen, C. Pavlovic, and V. Hohmann, "Open Master Hearing Aid (openMHA)—An integrated platform for hearing aid research," *J. Acoust. Soc. Am.*, vol. 146, no. 4, pp. 2879–2879, Oct. 2019.

[13] B. C. Moore, J. I. Alcántara, M. A. Stone, and B. R. Glasberg, "Use of a loudness model for hearing aid fitting: II. Hearing aids with multi-channel compression," *British J. of Audiology*, vol. 33, no. 3, pp. 157–170, Jun. 1999.