AN INTRODUCTION TO THE SPEECH ENHANCEMENT FOR AUGMENTED REALITY (SPEAR) CHALLENGE

Pierre Guiraud *[†], *Sina Hafezi* *[†], *Patrick A. Naylor*[†], *Alastair H. Moore*[†],
Jacob Donley[‡], *Vladimir Tourbabin*[‡], *Thomas Lunner*[‡]

[†] Electrical and Electronic Engineering, Imperial College London, London, UK
[‡] Meta Reality Labs Research, Redmond, Washington, USA

ABSTRACT

It is well known that microphone arrays can be used to enhance a target speaker in a noisy, reverberant environment, with both spatial (e.g. beamforming) and statistical (e.g. source separation) methods proving effective. Head-worn microphone arrays inherently sample a sound field from an egocentric perspective — when the head moves the apparent direction of even static sound sources change with respect to the array. Traditionally, enhancement algorithms have aimed at being robust to head motion but hearable devices and augmented reality (AR) headsets/glasses contain additional sensors which offer the potential to adapt to, or even exploit, head motion. The recently released EasyCom database contains microphone array recordings of group conversations made in a realistic restaurant-like acoustic scene. In addition to egocentric recordings made with AR glasses, extensive metadata, including the position and orientation of speakers, is provided. This paper describes the use and adaptation of EasyCom for a new IEEE SPS Data Challenge.

***Index Terms*—** augmented reality, data challenge, microphone array, head-worn array, head movement

1. INTRODUCTION

Augmented Reality (AR) systems render virtual objects which appear to be present in the user's local environment. This requires a certain amount of knowledge about the local scene which, in practice, requires appropriate sensors, such as head tracking, cameras, etc. It is also necessary for the user to retain their sense of presence in the local scene. In [1], in-ear microphones were equalised to compensate for the effect of blocking the ear canals but, by being in the ears, the spatial cues of the environment were inherently maintained. The effect of microphone placement and compensating for it was studied in [2], where microphones were placed at various position in or behind the ear. Obtaining correctly spatialised binaural signals from a circular array on the head has also been studied [3–5].

In many situations, it is desirable to enhance speech from one (or more) target talkers which exist in the local environment. For example, to help a normal-hearing or hard-of-hearing listener to understand more words in a conversation [6]. With head-tracking capabilities, AR devices could adapt to head motion [7, 8] or exploit it [9, 10]. Moreover, non-acoustic sensor data may provide more reliable direction of arrival information, e.g. during periods of source inactivity.

Speech enhancement for augmented reality can be seen as a logical evolution of existing microphone array signal processing. Along with the rest of the field, there is potential for machine learning approaches to find innovative solutions to fusing multi-modal sensor data. To stimulate research in this emerging field, we are running an IEEE SPS data challenge. The goal of the SPeech Enhancement for Augmented Reality (SPEAR) challenge is to obtain the "best possible" binaural signals given noisy array signals, head orientation and direction of arrival of a single target. In contrast to existing microphone array systems, it is assumed that these data can all be obtained from a single AR device using the available sensor modalities. Since speech enhancement for AR is a new application area, it is not yet clear what "best possible" means. Accordingly, challenge entrants are free to optimise the enhancement in whatever way they think listeners will appreciate most. Entries will be scored in terms of a wide variety of intrusive metrics and in listening tests. All enhanced audio submitted by entrants, along with the evaluation results, will be shared with the research community after the challenge so that the data can be fully and openly interpreted. Full and up-to-date details of the evaluation can be found on the challenge website [11].

In Sec.2 the datasets used in the challenge are presented while Sec. 3 develops our approach to improve the reliability of intrusive metrics computed on the first of these.

2. SCENES AND DATASETS

The EasyCom database (hereafter EasyCom) contains several realisations of a scenario (see Fig. 1) in which a small group of people have a natural conversation in a noisy envi-

[*] Both authors contributed equally to this work.

978-1-6654-6867-1/22/$31.00 ©2022 IEEE

ronment whilst sitting around a table. The strength of EasyCom comes from its realism and rich metadata. However, since the recordings are made “live”, it is impossible to define the ground truth signals at the array that an ideal enhancement would produce. Moreover, the time required to make such a set of recordings means that the possible diversity is unavoidably limited. To be able to evaluate intrusive metrics over a variety of acoustic conditions, SPEAR will use three simulated datasets, D2–D4, in addition to a dataset, D1, of real recordings taken from EasyCom. All four datasets are detailed below and summarised in Table. 1.

2.1. Dataset 1: EasyCom Original

In EasyCom, participant ID 2 is wearing a pair of spectacles fitted with a 6-microphone array, representing an AR device, where 4 microphones are integrated into the frames and 2 microphones are placed in the ear canals, as shown in Fig. 2. It is the enhancement of this array audio which is the focus of SPEAR. Participant ID 1 adopts the role of “waiter” and is exceptional in not being tracked or individually recorded. The remaining participants each wear a close-talking headset microphone. For SPEAR, this is approximately time- and level-aligned to the array signals to provide a reference for computing intrusive metrics (see Sec. 3).

An optical tracking system provides position and orientation data for all but participant ID 1. For SPEAR, enhancement algorithms will have access to the orientation of the AR device and the direction with respect to the AR device of a single target talker. Acoustic transfer functions for the array measured on a mannequin under anechoic conditions may also be used. Metadata, such as human-labelled voice activity for each participant, is provided to assist with training but will not be provided in the evaluation.

The purpose of D1 is to train/evaluate algorithms with signals that are as representative as possible of a real-world scenario. Of course, this implies some limitations. For example, the close-talking microphones are noisy references and might lead to imperfect calculation of intrusive metrics, especially in the case of binaural metrics. Also, it has been observed that the group conversations in EasyCom contain very little overlapping speech (less than 7 % overall). This means it is not well suited to evaluate the effectiveness of algorithms to suppress undesired speech, as encountered in multi-dialogue situations.

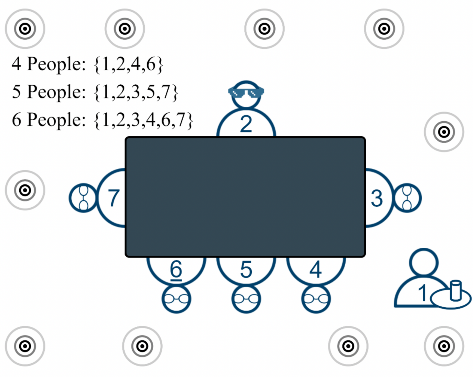

Fig. 1. Schematic of the scene used in EasyCom [12].

Table 1. Summary of dataset properties.

	Scene Audio	Acoustic Scene	Head Movements
D1	Original	Original	Original
D2	Denoised	Original	Original
D3	Denoised	Modified	Original
D4	New	Modified	Synthesised

2.2. Dataset 2: EasyCom Reproduced

The purpose of D2 is to reproduce using simulations the same scenes as encountered in D1. This is useful for several reasons, namely (i) as a proof of concept of the simulation procedure, (ii) to compute intrusive metrics on EasyCom where the reference signal(s) are known precisely, and (iii) to allow a comparison of algorithm performance on real and simulated audio for both metrics and listening tests.

Simulated scenes are rendered using the TASCAR software platform [13]. Each scene is defined in terms of the “acoustic scene”, meaning the geometry and properties of the environment, the “scene audio”, meaning the signals emitted by each source, and the “head movements”, meaning the trajectories of each participant relative to their nominal position and orientation. The acoustic scene is created using a shoebox room with reflective walls and a horizontal surface representing a table. Since the aim of D2 is to reproduce D1, the geometry of the scene is matched as closely as possible. Distributed across the room are 10 fixed sources, representing loudspeakers, each emitting restaurant-like, babble noise to mimic the creation of diffuse noise as employed in EasyCom. The required number of sources, representing participants, are arranged around the table according to the positions and head movement trajectories in EasyCom. Each participant source emits a denoised version of the associated close-talking head-

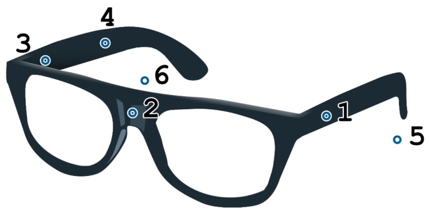

Fig. 2. Glasses microphone array used in EasyCom [12]. Mic 5 and 6 are in ears.

set signal from EasyCom. This denoising, performed using the CEDAR DNS Two plugin[1], was found in informal listening to greatly improve the naturalness of the simulations whilst causing little, if any, distortion to the target speech.

In addition to the complete simulation, by simulating the scene separately for each source without any reflections or reverberation, the reference audio at the array, required for intrusive metrics, is obtained. It should be noted that, should any distortion be introduced by denoising the headset signals, this becomes part of the ground truth input signal and so will not affect the calculation of intrusive metrics.

2.3. Dataset 3: EasyCom Augmented

Using the same simulation approach as in D2, D3 introduces variations in the acoustic scene in order to increase the diversity. It is expected that this will help to avoid machine learning-based approaches from over-fitting to the specific acoustic properties of D1/D2.

The particular parameters that are varied are (i) the acoustic scattering coefficient of the table, (ii) the reverberation time of the room, (iii) the dimensions of the room, (iv) the position of the table and, consequently, of the participants within that room (v) the sound level of the 10 loudspeakers used for diffuse noise, (vi) the type of diffuse noise used, (vii) the positions of the 10 loudspeakers used for diffuse noise, and (viii) the nominal head positions of each participant (slightly moved). These modifications are randomly perturbed for each minute-long segment of each session.

2.4. Dataset 4: Synthetic Dialogues

The purpose of D4 is to increase further the diversity of the training materials by removing all dependence on the specifics of EasyCom. In particular it uses (i) clean speech from an independent corpus [14], (ii) synthesised head movements, and (iii) competing dialogues which significantly overlap with the target speech.

Competing dialogues are synthesised by concatenating utterances for each talker with random intervals of silence between each utterance. To synthesise head movement and head rotation, the look direction of each participant is oriented approximately towards each of the other participants in a randomly generated sequence. Jitter is introduced to both positions and orientations to simulate real life non-stationarity.

3. INTRUSIVE METRICS USING CLOSE-TALKING MICROPHONES AS REFERENCE

Calculation of intrusive metrics require that the clean target signal at the array's reference microphone(s) be available. This is ensured, by design, for the simulated datasets (D2-4). However, for D1 only the close-talking headset microphone signals are available. In this section we briefly present a simulation-based study in which the effect of background noise and interferer leakage on the calculation of intrusive metrics is investigated and describe the approach used in SPEAR to limit the error.

3.1. Method

Room Impulse Responses (RIRs) for a shoebox room with T_{60}=645 ms and similar dimensions to the room used in EasyCom are obtained using the image-source method [20]. As shown in Fig. 4, the array is placed at $(4.0, 2.5, 1.0)$ m. A target and an interferer are located $1.5\,\mathrm{m}$ from the array, in the same horizontal plane, at azimuth angles $0°$ and $40°$, respectively. Ten source positions representing loudspeaker locations for generating ambient noise are arranged with almost uniform spacing around a circle of radius $3\,\mathrm{m}$, again in the same horizontal plane as the array. RIRs from each source to the array were obtained by convolving each reflection with the nearest measured array impulse response from [12]. Additionally the RIRs from each source to an omnidirectional microphone co-incident with the target were used to simulate the response of a close-talking headset microphone.

The target and interferer consist of $8\,\mathrm{s}$ anechoic speech utterances taken from [21] where each source is selectively muted such that they are simultaneously active for approximately $0.6\,\mathrm{s}$. This matches the overlapping duration ratio of approximately $7.5\,\%$ observed in EasyCom. The loudspeaker sources have the same restaurant-like signals used for EasyCom. Signal levels are adjusted in order to achieve signal-to-interferer ratio of $0\,\mathrm{dB}$ and signal-to-babble ratio of $5\,\mathrm{dB}$ at the array. Spatially white Gaussian noise with signal-to-noise ratio of $40\,\mathrm{dB}$ is also added to the sensors. Ten trials are performed using independent realisations of speech activity and different speech utterances. The data is processed at sample rate of $8\,\mathrm{kHz}$.

To assess the effect of noise on a selection of intrusive metrics, three possible references are considered; "Headset": the raw noisy headset signal; "Headset-VAD": Headset signal that is muted during target inactivity, as indicated using Voice Activity Detector (VAD) labels obtained according to [22] using clean target at the headset; "Oracle-Array": the ground truth anechoic target signal at the array. To ensure that only the effect of noise is evaluated, the Headset and Headset-VAD signals are time- and level-aligned to the Oracle-Array signal using `sigalign` from [23].

Metrics are computed for both the noisy array signal, "Passthrough", and the "Perfect Enhancement" case, where the enhanced signal is identical to "Oracle-Array".

[1] `www.cedar-audio.com/products/cedarstudio/cedarstudiodns.shtml`. Thanks to CEDAR Audio for their generous assistance.

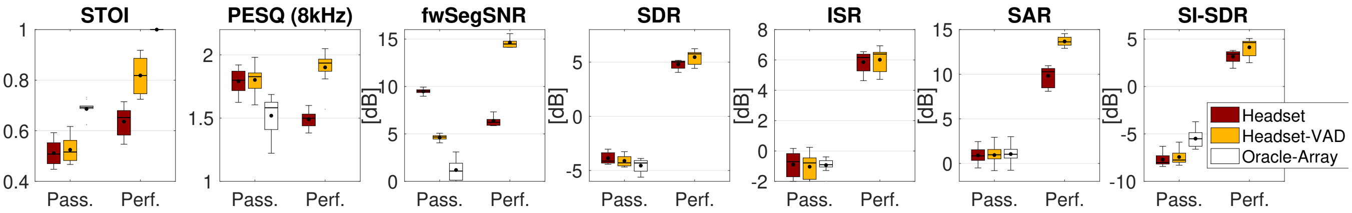

Fig. 3. Absolute metric scores for Passthrough (Pass.) and Perfect-Enhancement (Perf.) using Headset, Headset-VAD and Oracle-Array each as the reference signal based on simulated data. Metrics are STOI [15], PESQ [16], fwSegSNR [17], SDR [18], ISR [18], SAR [18] and SI-SDR [19]. Note that the infinite values in case of Perfect-Enhancement (method) and Oracle-Array (reference) where both signals are identical are excluded in the visualisation. Boxes, horizontal black lines and black dots show the inter-quartile range, median and mean, respectively.

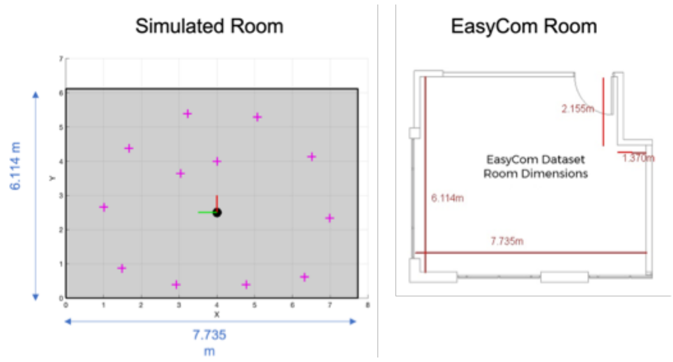

Fig. 4. Simulated setup for analysis of Reference signal pre-processing.

3.2. Results and Analysis

Figure 3 shows the distribution of the scores for the metrics for different combination of processed signal (Passthrough or Perfect-Enhancement) and Reference signal (Headset, Headset-VAD, Oracle-Array). For Passthrough, there are several metrics (PESQ, fwSegSNR and SDR) where Oracle-Array (white) reference leads to a worse metric value than Headset (red), suggesting that noise in the reference makes it more similar to the noisy array signal. For both PESQ and fwSegSNR using Headset reference, Perfect-Enhancement scores lower than Passthrough, indicating a complete failure of the metric. Comparing Headset (red) and Headset-VAD (yellow) it can be seen that muting the reference signal during signal inactive periods improves the reference and so makes the metric value closer to the true value obtained with Oracle-Array (white). For STOI, PESQ, SAR and SI-SDR the effect is most pronounced for the Perfect-Enhancement signals whereas for fwSegSNR both Passthrough and Perfect-Enhancement are substantially impacted.

These results suggest that, whilst metrics computed using close-talking headset microphones as reference should be treated with some caution, they can be made more reliable by muting the reference during periods of speech inactivity.

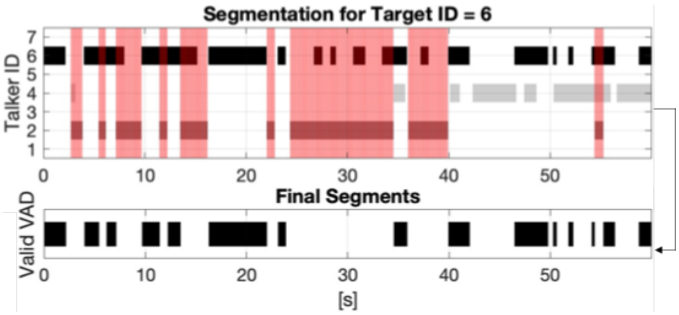

Fig. 5. For target ID = 6, excluding active own-voice moments resulting in final segments used to calculate the metrics.

3.3. Implications for SPEAR

In light of the results presented in Sec. 3.2, metrics for SPEAR will be computed individually for each period of target activity, using ground truth VAD labels provided in EasyCom (D1-3) or from the anechoic source signals (D4). Additionally, it is observed that during periods in which the wearer of the array (participant ID 2) is talking (own-voice condition) the received signal level at the array is substantially louder than during listening-only periods. This makes A-B comparisons of passthrough and processed signals challenging and so these own-voice periods are also removed from the metrics calculations. For a given target talker, segments in which metrics are computed is visualised in Fig. 5.

4. CONCLUSIONS

Speech enhancement for augmented reality is a new application of microphone array processing. The SPEAR challenge seeks to benchmark existing algorithms and encourage more researchers to get involved. More information can be found on the challenge website [11].

5. REFERENCES

[1] A. Härmä, J. Jakka, et al., "Augmented Reality Audio for Mobile and Wearable Appliances," *J. Audio Eng. Soc. (AES)*, vol. 52, no. 6, pp. 618–639, June 2004.

[2] F. Denk, S. M. A. Ernst, et al., "Adapting hearing devices to the individual ear acoustics: Database and target response correction functions for various device styles," *Trends Hear.*, vol. 22, Jan. 2018.

[3] P. Calamia, S. Davis, et al., "A conformal, helmet-mounted microphone array for auditory situational awareness and hearing protection," in *Proc. IEEE Workshop on Appl. of Signal Process. to Audio and Acoust. (WASPAA)*, Oct. 2017, pp. 96–100.

[4] J. Ahrens, H. Helmholz, et al., "A head-mounted microphone array for binaural rendering," in *2021 Immersive and 3D Audio: From Architecture to Automotive (I3DA)*, Sept. 2021, pp. 1–7.

[5] J. Fernandez, L. McCormack, et al., "Enhancing binaural rendering of head-worn microphone arrays through the use of adaptive spatial covariance matching," *J. Acoust. Soc. Am.*, vol. 151, no. 4, pp. 2624–2635, Apr. 2022.

[6] S. Doclo, S. Gannot, et al., "Acoustic beamforming for hearing aid applications," in *Handbook on Array Processing and Sensor Networks*, S. Haykin and K. Ray Liu, Eds. John Wiley & Sons, Inc., 2008.

[7] M. Zohourian and R. Martin, "Binaural speaker localization and separation based on a joint ITD/ILD model and head movement tracking," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Mar. 2016, pp. 430–434.

[8] A. H. Moore, L. Lightburn, et al., "Binaural mask-informed speech enhancement for hearing aids with head tracking," in *Proc. Int. Workshop on Acoust. Signal Enhancement (IWAENC)*, Tokyo, Japan, Sept. 2018, pp. 461–465.

[9] V. Tourbabin and B. Rafaely, "Direction of arrival estimation using microphone array processing for moving humanoid robots," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 11, pp. 2046–2058, Nov. 2015.

[10] A. H. Moore, W. Xue, et al., "Noise covariance matrix estimation for rotating microphone arrays," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 3, pp. 519–530, Mar. 2019.

[11] "SPEAR challenge website," `https://imperialcollegelondon.github.io/spear-challenge`.

[12] J. Donley, V. Tourbabin, et al., "EasyCom: An Augmented Reality Dataset to Support Algorithms for Easy Communication in Noisy Environments," *arXiv:2107.04174 [cs, eess]*, Oct. 2021.

[13] G. Grimm, J. Luberadzka, and V. Hohmann, "A toolbox for rendering virtual acoustic environments in the context of audiology," *Acta Acustica united with Acustica*, vol. 105, no. 3, pp. 566–578, 2019.

[14] I. Demirsahin, O. Kjartansson, et al., "Open-source Multi-speaker Corpora of the English Accents in the British Isles," in *Proceedings of The 12th Language Resources and Evaluation Conference (LREC)*, Marseille, France, May 2020, pp. 6532–6541, European Language Resources Association (ELRA).

[15] C. H. Taal, R. C. Hendriks, et al., "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2125–2136, Sept. 2011.

[16] A. Rix, J. Beerends, et al., "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Salt Lake City, UT, USA, May 2001, vol. 2, pp. 749–752.

[17] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[18] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 4, pp. 1462–1469, July 2006.

[19] J. L. Roux, S. Wisdom, et al., "SDR – Half-baked or well done?," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, May 2019.

[20] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[21] "Open speech repository," `https://www.voiptroubleshooter.com/open_speech`.

[22] "Objective measurement of active speech level," Recommendation, Int. Telecommun. Union (ITU-T), Mar. 1993.

[23] D. M. Brookes, "VOICEBOX: A speech processing toolbox for MATLAB," `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`, 1997.