DOI: 10.61782/fa.2023.0502

forum acusticum 2023

THE SPEAR CHALLENGE - REVIEW OF RESULTS

Vladimir Tourbabin[*1] **Pierre Guiraud**[2] **Sina Hafezi**[2] **Patrick A. Naylor**[2]
Alastair H. Moore[2] **Jacob Donley**[1] **Thomas Lunner**[1]
[1] Meta Reality Labs Research, Redmond, Washington, USA
[2] Electrical and Electronic Engineering, Imperial College London, London, UK

ABSTRACT

Verbal communication can be challenging in the presence of acoustic noise. To tackle this problem, microphone arrays coupled with numerous processing methods have been studied in the past few decades. Recent interest in Augmented Reality (AR) applications gives rise to head-worn microphone arrays. This highlights additional important aspects such as motion of the capture device with respect to the scene, the need to preserve spatial characteristics in the processed sound, and the tightened constraints on latency and computational budgets. The SPeech Enhancement for Augmented Reality (SPEAR) Challenge, endorsed by the IEEE Challenges and Data Collection initiative, was organized in order to further ignite interest in this important problem and to obtain a better sense of the remaining technological gaps. The challenge is based on an adaptation of the recently published EasyCom dataset that contains noisy conversation recordings from a glasses form-factor AR device with 6 microphones along with the positional information and additional labeled modalities. A competitive evaluation of the challenge entrant algorithms was carried out by using both objective metrics and subjective listening tests. The current contribution is focused on providing an overview of the SPEAR challenge and highlighting some of the most important findings and outcomes.

Keywords: *augmented reality, speech enhancement, microphone array processing, cocktail party problem, multimodal data.*

*__Corresponding author__: *vtourbabin@meta.com.*

1. INTRODUCTION

Acoustic noise accompanies people in many everyday situations, be it a busy street, a subway station, or a social gathering [1]. The acoustic noise hinders verbal communication between people thereby hurting social connection. Significant efforts have been put into understanding the acoustics of the problem [2], harvesting the power of microphone arrays and development of computational methods to enhance the desired voice while controlling the noise [3].

The rising new augmented reality (AR) technology is promised to bring strong computational and sensing capabilities closer to the user's sensory system, for example, in the form of AR glasses. This has the potential to facilitate new exciting applications assisting people to communicate effortlessly in even the most acoustically challenging situations by capturing the sound, processing and manipulating it in a desired way, and then delivering the enhanced version to the listener seamlessly and in real-time.

This application also brings up new technological dilemmas related to how dynamic the potential scenarios can get, how tight the latency constraints are, and what the fidelity of spatialization requirement would be to deliver seamless and natural interactions. Tackling these challenges today is even more exciting because of the rising power of artificial intelligence (AI) systems due to increased computational capacities and data diversity.

In order to gain a better insight into this problem, the Speech Enhancement for Augmented Reality (SPEAR) Challenge [4] was proposed, and was endorsed by the IEEE Challenges and Data Collection initiative. The main aim of SPEAR is to foster excitement about the potential of AR technology in helping people to communicate and stay connected. We hope to bring the relevant scientific communities, such as acoustics, spatial hearing, signal processing, and AI, closer together and facilitate a tighter

collaboration.

This paper summarizes some of the main aspects of the challenge by explaining the task, the data, and the evaluation procedures in Section 2, giving an overview of the participating algorithms in Section 3, presenting and discussing the results in Section 4, and a brief conclusion and proposed future directions in Section 5.

2. SETUP

This section describes the challenge including the task, the various datasets employed, and the evaluation approach.

2.1 Task

The setting of the challenge revolves around an indoor restaurant-like environment. Several participants are seated around a table, and a glasses-form factor microphone array is worn by one of the participants. The microphones capture partially overlapping voices of the participants as well as the restaurant noise. The task is to produce the *best possible* estimate of the voice of the desired participant in the ears of the device wearer using the microphone array signals as the input. The term *best possible* is not yet clearly defined in the context of augmented reality applications; it may involve aspects like improved intelligibility, being free from processing artifacts, reduced reverberation and ambient noise, and preservation of spatial cues. In order to compete in this challenge the algorithm latency of a proposed solution must not exceed $50\ ms$ and no external data can be used. Some pre-trained models were allowed provided they do not exceed the latency limitation.

2.2 Data

To maximize diversity of the data, four distinct datasets labeled D1, D2, D3, and D4 were provided, as described in more detail below. All four datasets are based to a various extent on the EasyCom [5] dataset that contains recordings of 3-5 people seated around a table and having natural conversations with a restaurant-like ambient noise at $\approx 70\ dB(A)$ produced by a number of uncorrelated loudspeakers distributed throughout the room. All four sets contained 6-channel audio from the glasses' form-factor microphone array depicted in Fig. 1. Channels 5 and 6 are binaural microphones placed at the entrance to the ear canal. The audio sample rate was $48\ kHz$. In addition to the microphone array recording, the participants were provided with azimuth and elevation of the desired source direction with respect to the array center. Head orientation has also been provided as quaternions. Both, the direction of the source and orientation of the array were sampled at a rate of $20\ Hz$. This data was included to remove the desired source localization problem and allow the participants to focus on the speech enhancement task.

Figure 1. Glasses form-factor microphone array that was used to record and/or simulate the four datasets provided with the challenge.

Table 1 summarizes some of the most important aspects of the four datasets, D1-4. Each dataset is roughly the same duration of $5\ h$ and is split into 15 sessions (9 train | 3 dev | 3 eval) with roughly equal duration of $\approx 20\ min$ each. The first dataset, D1, contains the actual audio and orientation taken directly from EasyCom. The target signals provided with D1 were the close-talk microphone recordings attached to each participant delayed to roughly match the microphone array recordings. The second dataset, D2, is a reproduction of D1 using the actual room's geometry with the table added as a reflector. Source signals for the simulation were obtained by further denoising the close-talk microphones of each participant using the CEDAR DNS 2 plugin [6]. The target signals for this dataset are the signals in the binaural microphones (channels 5 and 6) simulated with only the desired source present and using only the direct path propagation. The third dataset, D3, was introduced to improve diversity by varying the simulated room dimensions, reverberation level, and position of the table in the room. Everything else was kept the same as in D2. Finally, D4 is a simulated dataset with artificially created dialogue. Here, the individual participant voices were created by concatenating clean speech utterances with randomly introduced pauses. The main distinction of D4 from D1-3 is the much more significant overlap between the voices of the participants as shown in Table 2.

Table 1. Summary of some of the most important aspects of the four provided datasets, D1-4.

	D1	D2	D3	D4
microphone audio	real	simulated	simulated	simulated
target audio	close-talk mic	binaural mics	binaural mics	binaural mics
head movements	real	real	real	synthetic
acoustic diversity	fixed conditions	fixed conditions	multiple rooms/locations	multiple rooms/locations
voices overlap	little	little	little	substantial

Table 2. Percentage of audio content as a function of the number of concurrent voices.

# voices	1	2	3	4
D1-3	62%	32%	5%	1%
D4	2%	26%	43%	29%

2.3 Evaluation

The performance of the competing algorithms was evaluated using both objective metrics and subjective listening tests.

The objective metrics included energetic metrics like SNR, fwSegSNR [7], and SI-SDR [8], speech quality such as PESQ [9], speech intelligibility such as STOI [10], and the binaural intelligibility metric MBSTOI [11]. In total, 12 objective metrics were used. Additional details can be found on the challenge webpage [4].

The subjective listening tests were administered using the GoListen platform [12]. The participants were recruited and paid using Prolific [13], with the corresponding consents obtained through Qualtrics [14]. The process included an equipment check based on audibility of tones at different frequencies and a headphone check by comparing levels of stereo tones with and without phase inversion. Passing the two checks was strictly required to proceed to the actual listening tests. The test included pairwise comparisons between audio examples produced by the various algorithms. In addition, free-form feedback was also collected. Each session included 20 pairwise comparisons with a total of over 400 recorded sessions. In this test only up to two algorithms from each participating research team were selected for full round robin comparison. Participants who submitted more than 2 entries were asked to nominate their preferred algorithms. The organisers used this preference together with results of pilot tests and the system descriptions to select the best performing and most diverse systems for evaluation in the listening test.

3. PARTICIPATING ALGORITHMS

Overall, 5 research teams from around the world have participated in the challenge. Most teams have submitted multiple entries which, together with the passthrough and the baseline, resulted in 16 distinct algorithms. For convenience, the various algorithms (abbreviated as alg. below) were labeled using capital letters as detailed below.

Alg. A - passthrough, i.e. the binaural microphone signals as captured by the array.

Alg. B - the baseline. This is the widely used maximum directivity beamformer (a.k.a isotropic Minimum Variance Distortionless Response (MVDR)) obtained by assuming stationary diffuse noise as described in more detail, for example, in Sec 2.1 of [15].

Alg. C, D - these entries were made by Audifon GmBH, Kölleda, Germany. Alg. C consists of a 3-mic beamformer followed by a single channel subtraction. The beamformer points a fixed beam towards the front by summing up the left and right binaural microphone signals with a 2-sample delayed version of the frontal microphone. Alg. D is obtained by adding a compressor and a limiter to Alg. C. See [16] for additional details. An important advantage of these two entries over all other participants in the challenge is its very low computational and power consumption footprint - a version of this algorithm was shown to run successfully on ARM Cortex M4 microcontroller.

Alg. E - this contribution was made by a team from the Institute of Electronic Music and Acoustics University of Music and Performing Arts, Graz, Austria. The proposed algorithm uses the maximum directivity beamformer and matched filters to form the input features and encode the direction of the desired source. Two consequtive U-net structures, sub-band and full-band, are then

applied with gated RNN layers in the bottleneck. The output is respatialized by convolving with the binaural microphone transfer functions in the direction of the desired source. Additional details can be found in [17].

Alg. F, G, H, I - submitted by a research team from Sogang University, Seoul, South Korea. The first stage in all systems is independent monaural enhancement of the array signals using a DNN based on LSTM-ResUNet. Algs. F and G then use the baseline beamformer to obtain binaural outputs whereas H and I use the authors' own beamformer. Alg. I additionally uses a TRUNet post filter on each of the binaural outputs. Alg. G differs from the others in the training parameters of the DNN. Additional details can be found in [18].

Alg. J, K, L, M, N - submitted by a research team from the University of Illinois Urbana-Champaign, Champaign, IL, USA. The overall method proposed here consists of three stages. First is the challenge baseline beamformer followed by a monaural speech enhancement based on DeepFilterNet 2 similar to Alg. P. The monaural filter was further fine-tuned using the baseline beamformer output. In the final stage, a speaker separation network with temporal convolutions and a causal transformer was trained to estimate a mask to be applied to the monaural postfilter output. The 5 submissions differ by the parameter of the mask thresholds and the frequency range of the speaker separation module. Additional details can be found in [19].

Alg. O, P - submitted by the research team from Electrical and Electronic Engineering Department, Imperial College London, London, UK (challenge organizers). The general approach proposed is called the subspace hybrid MVDR beamformer; it is aiming to benefit from an adaptive beamformer while minimizing the robustness issues involved with voice activity detection and adaptation. The algorithm consists of two stages. The first stage produces two outputs; one using the Iso MVDR beamformer exactly as in baseline Alg. B, the other using a hybrid MVDR beamformer that utilizes a dictionary of precomputed noise covariance matrices (NCM) and selects the one that minimizes output power at any given time-frequency bin. In the second stage, both beamformer's outputs are projected into the subspace spanned by the first eigenvector of the inter-beamformer correlation matrix to reduce musical noise generated by frequent switching of the NCM in the hybrid MVDR beamformer. The reader is referred to [15] for further details. The challenge entry Alg. O is an extension of the subspace hybrid MVDR beamformer by duplicating the process with the left binaural mic and the right binaural mic as a reference. Alg. P is a extension of Alg. O by adding a post filtering step on each channel separately using the single-channel denoising approach called DeepFilterNet2 [20].

Table 3 compares some of the important computational aspects of the various algorithms.

4. RESULTS AND DISCUSSION

Results of the objective evaluation for three selected metrics are plotted in Figs. 2, 3, and 4. Although the results vary significantly between metric, a few common observations can be made. First, although the baseline appears to be quite effective according to the objective metrics, most competing algorithms were able to result in more significant benefits. Second, submissions by the different research teams seem to be consistently grouped and perform similarly with respect to the other teams. Loosely, the performance according to all three metrics can be ordered from high to low as follows: Alg. E, Algs. J-N, Algs F-I, Algs O and P, and Algs C and D.

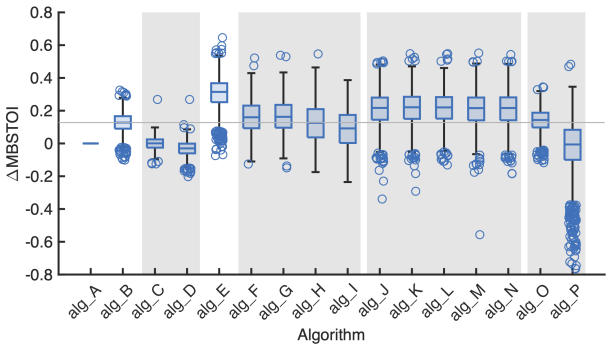

Figure 2. Improvement in binaural intelligibility as measured by MBSTOI relative to the passthrough condition, Alg. A. The blue horizontal line indicates the baseline performance for convenience.

Results of the subjective pairwise comparison in the listening tests are shown in Fig. 5. It can be seen that overall alg. E is strongly preferred over all other entries. Surprisingly, alg E is also the only one that outperforms the baseline in the listening test.

An aggregate percentage of wins for each algorithm was computed and is presented in a ranked order in Fig. 6. This picture also shows that alg. E is expected to be preferred over another randomly selected algorithm in most

Table 3. Summary of important computational characteristics of the competing algorithms.

Alg.	spatial filter	single-channel/postfilter	computational complexity
B	model(DSP)-based	none	moderate/low
C, D	model(DSP)-based	model(DSP)-based	low
E	learning(NN)-based	learning(NN)-based	high
F,G,H,I	model(DSP)-based	learning(NN)-based	high
J,K,L,M,N	model(DSP)-based	learning(NN)-based	high
O,P	model(DSP)-based	learning(NN)-based	moderate

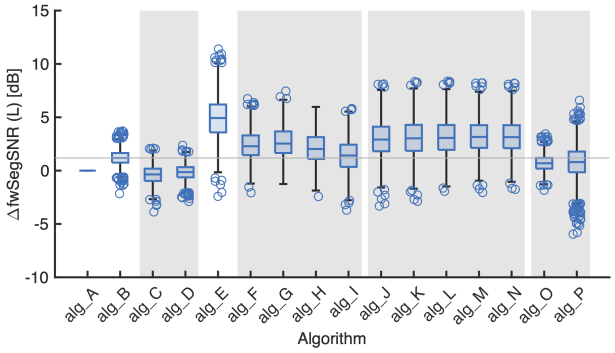

Figure 3. Improvement in signal to noise ratio as measured by fwSegSNR relative to the passthrough condition, Alg. A. The blue horizontal line indicates the baseline performance for convenience.

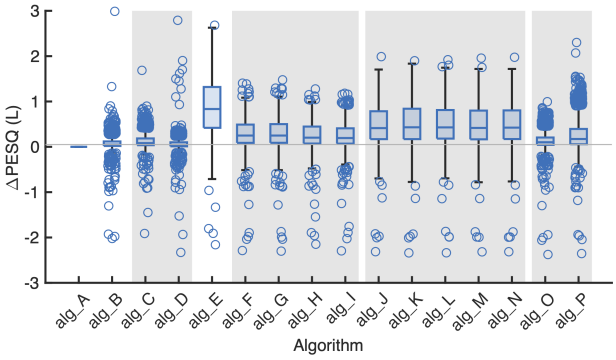

Figure 4. Improvement in speech quality as measured by PESQ relative to the passthrough condition, Alg. A. The blue horizontal line indicates the baseline performance for convenience.

cases. Algorithm E is followed with a significant gap by the baseline alg. B, which is closely followed by alg. M.

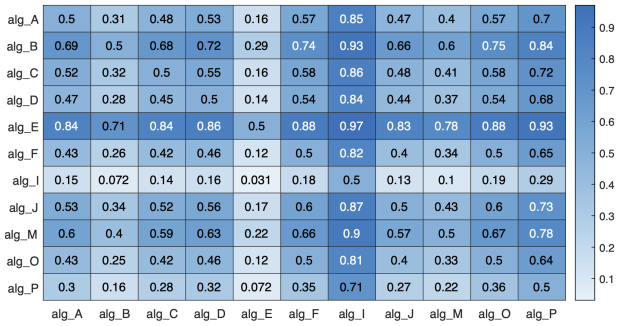

Figure 5. Proportion of the times in which the row index algorithm was preferred over the column index algorithm in perceptual listening tests.

A couple important observations can be made by analyzing the results of evaluations with the algorithm characteristics summarized in Table 3. First, it is interesting to observe that, according to the perceptual evaluation, alg. C is ranked in the middle and is expected to be preferred in roughly half of the cases. This is an important result both because alg. C came after most of the other algorithms according to the objective metrics and because the algorithm is by far more efficient computationally than the rest of the entries. Secondly, it is interesting to note that the best performing alg. E is the only approach which is end-to-end learning-based including the spatial filter. This might suggest that learning-based approaches have the potential to utilize spatial information more effectively than even the most sophisticated adaptive model-based beamformers.

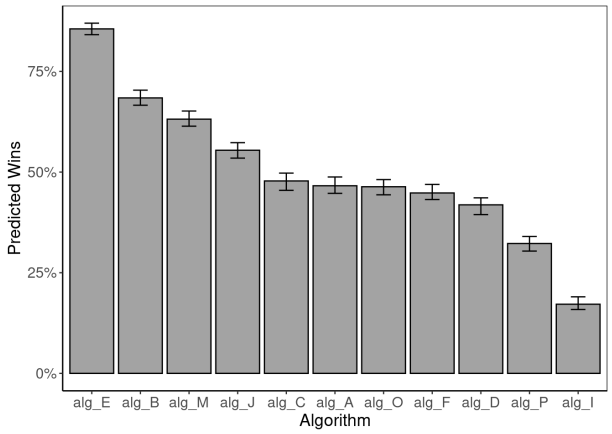

Figure 6. Overall predicted percentage of times in which each of the algorithms is expected to win versus another randomly selected one.

5. CONCLUSION

An international challenge on SPeech Enhancement for Augmented Reality (SPEAR) was organized and successfully run with a total of 14 submissions from research teams around the world including Germany, Austria, UK, USA, and South Korea. The submissions included extremely low power algorithms ready to run on miniaturized hearing aid devices, hybrid solutions combining advanced adaptive beamforming with DNN-based pre/post-processing, and end-to-end learning-based systems. All the submissions were evaluated using both objective metrics and perceptual listening tests. The results show that, while most algorithms can outperform the baseline in objective evaluation, beating the baseline in a perceptual test is rather hard. This can be related to the trade-off between interference suppression and the desired speech distortion that is inevitably introduced by most enhancement algorithms. Notably, the best performing algorithm consistently outperformed the rest of the competitors and the baseline in both objective and subjective evaluation. Interestingly, this solution from the Institute of Electronic Music and Acoustics, Austria, is the only submission that was learning-based end-to-end. Looking forward, several important directions can be explored. First, the algorithms explored in this challenge haven't utilized the head rotation information, which can be beneficial for better spatial tracking of interference. Second, diversity of the data can be improved to better generalize to a wider variety of scenarios and applications. Third, multi-modal approaches can be explored that benefit from visual and other domains much like humans are able to do in acoustically challenging conditions.

6. ACKNOWLEDGEMENTS

The authors would like to thanks CEDAR for providing access to their denoising software.

7. REFERENCES

[1] G. Farber and L. Wang, "Analyses of crowd-sourced sound levels of restaurants and bars in New York City," *Proc. Mtgs. Acoust 31*, 02 2018.

[2] J. Rindel, "Verbal communication and noise in eating establishments," *Applied Acoustics - APPL ACOUST*, vol. 71, pp. 1156–1161, 12 2010.

[3] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017.

[4] "SPEAR webpage." `https://imperialcollegelondon.github.io/spear-challenge/`. Accessed: Jan 24, 2023.

[5] J. Donley, V. Tourbabin, J.-S. Lee, M. Broyles, H. Jiang, J. Shen, M. Pantic, V. K. Ithapu, and R. Mehra, "EasyCom: An Augmented Reality Dataset to Support Algorithms for Easy Communication in Noisy Environments," *arXiv*, 2021.

[6] "CEDAR webpage." `https://www.cedar-audio.com/products/dns2/dns2.shtml`. Accessed: Apr 26, 2023.

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

[8] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – half-baked or well done?," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 626–630, 2019.

[9] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new

method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 2, pp. 749–752 vol.2, 2001.

[10] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[11] A. Heidemann Andersen, J. de Haan, Z.-H. Tan, and J. Jensen, "Refinement and validation of the binaural short time objective intelligibility measure for spatially diverse conditions," *Speech Communication*, vol. 102, pp. 1–13, Sept. 2018.

[12] "GoListen platform webpage." https://golisten.ucd.ie/. Accessed: Apr 26, 2023.

[13] "Prolific website." https://www.prolific.co/. Accessed: Apr 26, 2023.

[14] "Qualtrics website." https://www.qualtrics.com/. Accessed: Apr 26, 2023.

[15] S. Hafezi, A. H. Moore, P. Guiraud, P. A. Naylor, J. Donley, V. Tourbabin, and T. Lunner, "Subspace hybrid beamforming for head-worn microphone arrays," *SPEAR challenge submission*, 2023.

[16] I. Pieper, A. J. Hintermaier, and T. Harczos, "Noise reduction for audio in real time and with low power consumption," *SPEAR challenge submission*, 2023.

[17] B. Stahl and A. Sontacchi, "Hybrid subband-fullband gated convolutional recurrent neural network for multichannel speech enhancement," *SPEAR challenge submission*, 2023.

[18] J. H. Kim, B. H. Ku, S. H. Kim, J. H. Ko, and U. H. Shin, "Cascade of LSTM-RES-UNET-based enhancement and beamformer for target speech extraction on wearable glasses," *SPEAR challenge submission*, 2023.

[19] Z. Xu, D. Dutta, X. Fan, M. Hasegawa-Johnson, and R. R. Choudhury, "Multi-channel speech enhancement for SPEAR challenge: A three stage approach," *SPEAR challenge submission*, 2023.

[20] H. Schröter, A. Maier, A. Escalante-B, and T. Rosenkranz, "Deepfilternet2: Towards real-time speech enhancement on embedded devices for full-band audio," in *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 1–5, 2022.