Interspeech 2022
18-22 September 2022, Incheon, Korea

Listening with Googlears: Low-Latency Neural Multiframe Beamforming and Equalization for Hearing Aids

Samuel J. Yang[1], *Scott Wisdom*[1], *Chet Gnegy*[1], *Richard F. Lyon*[1], *Sagar Savla*[1]

[1]Google Research

dicklyon@google.com

Abstract

Understanding speech in the presence of noise with hearing aids can be challenging. Here we describe our entry, submission E003, to the 2021 Clarity Enhancement Challenge Round 1 (CEC1), a machine learning challenge for improving hearing aid processing. We apply and evaluate a deep neural network speech enhancement model with a low-latency recursive least squares (RLS) adaptive beamformer, and a linear equalizer, to improve speech intelligibility in the presence of speech or noise interferers. The enhancement network is trained only on the CEC1 data, and all processing obeys the 5 ms latency requirement. We quantify the improvement using the CEC1 provided hearing loss model and Modified Binaural Short-Time Objective Intelligibility (MBSTOI) score (ranging from 0 to 1, higher being better). On the CEC1 test set, we achieve a mean of 0.644 and median of 0.652 compared to the 0.310 mean and 0.314 median for the baseline. In the CEC1 subjective listener intelligibility assessment, for scenes with noise interferers, we achieve the second highest improvement in intelligibility from 33.2% to 85.5%, but for speech interferers, we see more mixed results, potentially from listener confusion.

Index Terms: speech enhancement, beamforming, hearing aids, deep learning

1. Introduction

The intelligibility of speech degrades significantly in the presence of noise or other interfering speakers, especially for hearing-impaired users. One way of overcoming this challenge is to integrate processing into hearing aids that removes interfering sounds before applying hearing loss compensation. Neural networks have been shown to be quite adept at both removing noise from speech [1, 2, 3] and separating multiple speakers [4, 5]. If multiple microphones are available, beamforming is also an effective approach for enhancing a desired speech signal, and its linear processing generally introduces less distortion than a nonlinear neural network [6].

The Clarity Enhancement Challenge Round 1 (CEC1) [7] was a machine learning challenge for applying machine learning to improve hearing aid processing, and in particular, for the scenario of a listener attending to a target speaker in the presence of either a noise or speech interferer.

Motivated by the benefit of mask-based separation for hearing-impaired users [1] and the effectiveness of neural beamforming [6], we propose a complete low-latency and low-complexity hearing aid processing chain that uses a neural multiframe beamformer to improve intelligibility. Our approach is trained and evaluated on CEC1 data, and evaluated with both objective and subjective measures of intelligibility. We achieve competitive performance for both speech and noise interferers in terms of objective intelligibility metrics, and ranked at second place in the CEC1 challenge in terms of subjective listening tests for noise interferers. Our results for speech interferers were not as competitive, but they reveal an interesting property of the challenge. In particular, our mixed results seem due to confusion of listeners who were instructed to transcribe the *second* speaker they heard, when our system completely zeroed out the interfering *first* speaker at the beginning of test clips.

1.1. Related Work

Compared to most other entries to the CEC1 challenge, our approach is unique. Two submissions [8, 9] used a sequence of beamforming across all 6 microphones using an estimated direction of arrival (DOA) of the target speaker followed by post-filtering (either with an automatic gain control module or a neural network), before performing listener compensation. In contrast, as shown in Figure 1, we *first* apply a neural enhancement network individually to each of two binaural microphones. For each of the two binaural microphones, the resulting target speech estimate is used to derive a linear beamformer across all microphones. Another submission [10] is somewhat similar this setup, in that it uses single-microphone neural network enhancement to produce a target speech estimate to derive a multi-microphone weighted prediction error (WPE) module to remove reverberation. However, this WPE module is only intended to remove reverberation from a single reverberant speech signal, and cannot remove interference.

Three other CEC1 submissions [11, 12, 13] followed a somewhat similar approach to ours in that an initial neural denoising network is used with multi-microphone [11, 12] or single-microphone [13] inputs. However, none of these approaches use linear beamforming across microphones, which we think introduces less distortion than the output of a neural network (especially for small neural networks). Instead of relying on a specified DOA, this class of approach relies on cues about the target and interference signals themselves. In certain scenarios, this can provide an advantage over DOA-based beamforming, such as when target and infererer signals originate from nearly the same direction. Our method has this advantage, while also introducing less distortion than a mask-based suppression neural network does.

2. Hearing aid model

Our hearing aid model contains three components, as illustrated in Figure 1: a parallel bank of two single-channel target speech enhancement models, a recursive least-squares (RLS) beamformer, and a linear equalizer. The speech enhancement model is used to predict left and right channels of a stereo target signal for the RLS beamformer, and was trained on the provided CEC1 dataset [7] only. No other existing data or trained models were used. The enhancement, beamforming, and linear equalizer all operate on 16 kHz audio, which is then upsampled to the CEC1-required 44.1 kHz. The enhancement model utilizes

10.21437/Interspeech.2022-10783

samples no more than 5 ms into the future, and the beamformer and linear equalizer add no additional latency, so the entire solution strictly obeys the 5 ms causal requirement.

2.1. Enhancement

We assume the following signal model for a single microphone:

$$y_n = s_n + v_n, \tag{1}$$

where y_n is an input mixture waveform, s_n is a target reverberant speech waveform, and v_n is a reverberant interferer waveform. For single-channel enhancement, we use a causal Conv-TasNet masking network [5]. Rather than a learnable basis, we use a STFT with 5 ms (80 samples at 16 kHz) square-root Hann analysis window, 2.5 ms (40 samples at 16 kHz) hop, and FFT size 256, where the analysis frame is zero-padded on the right from 80 to 256 samples before computing the FFT. This ensures that we satisfy the 5 ms latency requirement, and allows enhanced STFT frames to be passed directly to the RLS beamformer. The convolutional masking network takes 0.3-power-compressed magnitude STFT as input, and predicts a single real-valued mask $\hat{\mathbf{M}}$ through a sigmoid activation. This mask is multiplied with the complex input STFT $\mathbf{Y}$ to yield a complex estimated target STFT: $\hat{\mathbf{S}} = \hat{\mathbf{M}} \odot \mathbf{Y}$. Power-law compression with power 0.3 approximates a log function while avoiding $-\infty$ at 0, partially equalizing the importance of quieter sounds relative to loud ones [14, chapter 3], [2].

The enhancement model was trained with a multi-resolution spectral loss, which is mean-squared error between compressed magnitude and compressed complex consistent STFTs [3] at several different window sizes. At training time, to get consistent STFTs to pass to the loss function, an estimate of the time-domain target $\hat{s}_n$ is computed by applying an inverse STFT to the predicted target speech STFT $\hat{\mathbf{S}}$. An estimate of the interferer waveform $\hat{v}_n$ is also created by subtracting the time-domain target speech estimate from the input mixture waveform: $\hat{v}_n = y_n - \hat{s}_n$.

The loss for a given window size between a reference STFT $\mathbf{X}$ and an estimated STFT $\hat{\mathbf{X}}$ is as follows:

$$L(\mathbf{X}, \hat{\mathbf{X}}) = \left\| |\mathbf{X}|^{0.3} - |\hat{\mathbf{X}}|^{0.3} \right\|_F^2 + 0.2 \cdot \left\| \tilde{\mathbf{X}}^{0.3} - \hat{\tilde{\mathbf{X}}}^{0.3} \right\|_F^2, \tag{2}$$

where $\|\mathbf{Z}\|_F^2 := \sum_{t,f} |Z_{t,f}|^2$ is the squared Frobenius norm, $[|\mathbf{X}|^{0.3}]_{t,f} := |X_{t,f}|^{0.3}$ is the compressed magnitude STFT, and $[\tilde{\mathbf{X}}^{0.3}]_{t,f} := |X_{t,f}|^{0.3} e^{j \angle X_{t,f}}$ is the compressed complex STFT. For the STFTs used in the loss function, we use square-root Hann windows of 64 ms, 32 ms, 16 ms, 8 ms, and 5 ms, with 75% overlap. Since past work has found that applying the loss to both reference signals is beneficial [3], the loss is applied equally to target and interferer signals. Thus, the total loss $L_{\text{tot}}(s_n, v_n, \hat{s}_n)$ is

$$\sum_{r \in \mathcal{R}} L\Big(\mathcal{S}_r\{s_n\}, \mathcal{S}_r\{\hat{s}_n\}\Big) + L\Big(\mathcal{S}_r\{v_n\}, \mathcal{S}_r\{y_n - \hat{s}_n\}\Big), \tag{3}$$

where $\mathcal{R}$ is the set of STFT window sizes and $\mathcal{S}_r$ is the forward STFT operator with window size r. Note that the inverse and forward STFTs that preserve consistency do not violate the strict latency requirement of the model itself, since these operations are only performed at training time as part of the loss function. Also, only the initial estimated target STFT $\hat{\mathbf{S}} = \hat{\mathbf{M}} \odot \mathbf{Y}$, before computing $\hat{s}_n$, is passed to the downstream beamformer.

For training data, we use the CEC1 training set of 6000 scenes. We found on-the-fly augmentation to be advantageous, leading to better validation metrics on the CEC1 development set. This augmentation was done by remixing targets with interferers from other examples in the batch. Note that this avoids needing to generate additional training scenes. Though this remixing does not respect the acoustic consistency between sources, this inconsistency does not seem to prevent learning. This may be because the enhancement model is single-channel, and thus not as sensitive to acoustic spatial inconsistencies.

Since the target always begins two seconds after the interferer in the training data examples, the enhancement model implicitly learns this cue and can apply this at test time. Note that the model likely utilizes timing cues from zero-padding of the ground-truth target reference, and that no additional modifications to the architecture or training were required to achieve the exploitation of this timing cue.

The model is implemented in TensorFlow, and is trained on 32 Google Cloud TPU v3 cores with Adam [15], batch size 256, and learning rate 0.001.

2.2. Causal RLS beamformer

The single-channel enhancement model separately processes the front-left and front-right microphones, producing a stereo complex estimated target speech STFT. This STFT is used to derive a causal RLS adaptive filter [16, 17, 18] that performs beamforming, which introduces no additional latency.

Mathematically, at step t, given a new target vector $\mathbf{x}_t \in \mathbb{R}^N$ and corresponding input vector $\mathbf{y}_t \in \mathbb{R}^M$, the RLS filter computes a linear filter $\mathbf{W}_t \in \mathbb{R}^{M \times N}$ given the previous filter $\mathbf{W}_{t-1}$, previous estimated inverse input covariance matrix $\mathbf{P}_{t-1}$, and averaging weight λ:

$$\mathbf{g}_t = \mathbf{P}_{t-1}\mathbf{y}_t \,/\, (\lambda + \mathbf{y}_t^T \mathbf{P}_{t-1} \mathbf{y}_t), \tag{4}$$

$$\mathbf{P}_t = (\mathbf{P}_{t-1} - \mathbf{g}_t \mathbf{y}_t^T \mathbf{P}_{t-1})/\lambda, \tag{5}$$

$$\mathbf{W}_t = \mathbf{W}_{t-1} + \mathbf{g}_t (\mathbf{x}_t^T - \mathbf{y}_t^T \mathbf{W}_{t-1}). \tag{6}$$

$\mathbf{P}_0$ is initialized to $\mathbf{I}/\delta$, where δ is a diagonal loading factor and $\mathbf{I}$ is the identity matrix. We use averaging weight of $\lambda = 1.0$ (which assumes the listener is stationary, and thus all observations up to the present time are used) and diagonal loading factor of $\delta = 0.001$. For nonstationary scenarios, a smaller averaging weight $\lambda < 1.0$ could be used to adapt to time-varying conditions.

In addition to the 6 microphones on the hearing aids, we also use the past 4 frames of context [6] as additional virtual microphones. Furthermore, we use the real and imaginary parts of the stereo target STFT and multichannel input STFT as additional dimensions, which corresponds to a widely-linear RLS filter [19]. Thus, the size of target and input vectors for each frequency are $N = 2 \cdot 2 = 4$ and $M = 2 \cdot 4 \cdot 6 = 48$, respectively. A separate RLS filter is used for each frequency. For each time frame t, the predicted RLS filter $\mathbf{W}_t$ is applied to the input $\mathbf{y}_t$ as $\mathbf{W}_t^T \mathbf{y}_t$ to yield the real and imaginary values of a stereo beamformed target STFT. This beamformed STFT is then passed to the linear equalizer.

2.3. Linear equalizer

For each provided listener binaural audiogram, we utilize a linear equalizer to adjust the final audio. In the beamformer's STFT space, each coefficient is multiplied by a gain corresponding to 0.65 times the hearing level (HL) in dB specified in the audiogram, interpolated between the audiogram frequencies. Results are near optimal for a range of factors from about 0.4

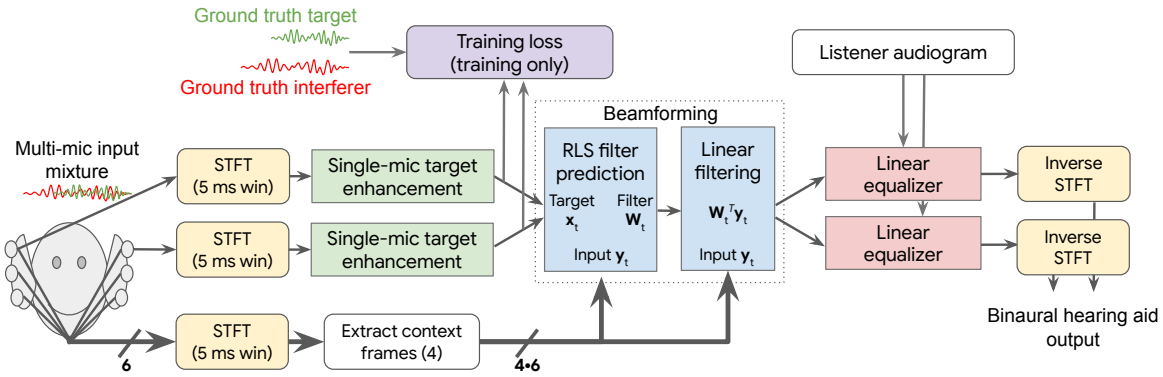

Figure 1: *Block diagram of our proposed system.*

to 0.75, consistent with the audiologist's "one-half gain rule" [20] to provide gain to compensate about half the loss for a linear aid. Ideally, as in a multiband compressor, a higher fraction of the HL would be compensated at lower speech levels, and a lower fraction at higher levels, but there is not so much dynamic range in the challenge material that that seemed necessary.

Finally the gain is reduced by -30 dB to be at the level that works best through the hearing loss model followed by MBSTOI evaluation. The -30 dB and 0.65 fraction were jointly optimized. For output for listening, we use -20 dB instead which appeared best from brief qualitative testing.

To reconstruct the time-domain estimate of enhanced speech, the inverse STFT is applied to the output of the linear equalizer. The inverse STFT uses a 5 ms square-root Hann synthesis window with 2.5 ms hop, which obeys the strict 5 ms latency requirement of the challenge.

3. Results

3.1. Model-based intelligibility evaluation

The CEC1 provided a baseline hearing aid solution which achieved on the development dataset a mean and median MBSTOI score of 0.41 [7]. During development of our system, we also used an additional baseline: a simple solution where the front two microphones were directly used as the output of our baseline hearing aid (so the other 4 microphone inputs are ignored). On the CEC1 development dataset, and using the CEC1 hearing loss model, this baseline yielded MBSTOI mean of 0.559 and median of 0.569.

In comparison, our proposed solution with our beamformer but no equalizer achieved on the development set an MBSTOI mean of 0.596 and median of 0.605. By adding the linear equalizer, we achieved MBSTOI mean of 0.632 and median of 0.642. Lastly, we note on the test set, we achieve a mean of 0.644 and median of 0.652 compare to 0.310 mean and 0.314 median CEC1 baseline.

Table 1 shows several ablations for our approach compared to our proposed system, and the two baselines (front two microphones and CEC1 baseline). Removing both the beamformer and the equalizer results in the worst degradation (reduction of 0.073 mean MBSTOI), indicating the relative importance of these components. Adding either of these components back in boosts MBSTOI by nearly 0.03. Training the enhancement model without augmentation led to overfitting quicker (in only about 20000 training steps), and degrades MBSTOI by 0.024. Finally, using only 1 context frame instead of 4 for the beamformer produces a drop of 0.019. Note that it is not possible to ablate the enhancement model, because the beamformer depends on it for a target signal.

Table 1: *Ablations for MBSTOI on development set.*

Ablation/Model	MBSTOI mean	MBSTOI median
Proposed	0.632	0.642
1 context frame	0.613	0.624
No augmentation	0.608	0.618
No beamformer	0.596	0.607
No equalizer	0.596	0.605
No b.f., no eq.	0.567	0.577
Front 2 mics	0.559	0.569
CEC1 baseline	0.41	0.41

3.2. Example submission processed audio

We plot sample audio waveforms for a speech interferer example in Figure 2, and include the ground truth target waveform along with the baseline of using just the front two microphones. For both types of interferers, our submission demonstrates an emergent phenomenon, whereby the initial waveform is complete silence right up until the detected target speech onset, which varies from $2.0s$ to $3.0s$. As we note in Section 2.1, the enhancement model likely learns this pattern from the training data. A review of additional examples suggests the enhancement model is quite accurate at identifying speech onset with sub-second precision. Lastly, we note that beyond the initial period before target speech onset, the interferer is only attenuated, and not entirely eliminated, as is evident in the last $1s$ of each example.

3.3. Listener intelligibility evaluation

As a part of the CEC1, real listeners listened to submitted audio samples and the intelligibility of those samples were evaluated quantitatively. For each utterance, the correctness, the number of words identified correctly as a percentage of the total number of words, was assessed. The results (shown in Figure 3), reveal our submission performed among the best for scenes with noise interferers, with an average increase in correctness over the baseline from 33.2% to 85.5% (second only to an entry achieving 86.7%), but among the worst for scenes

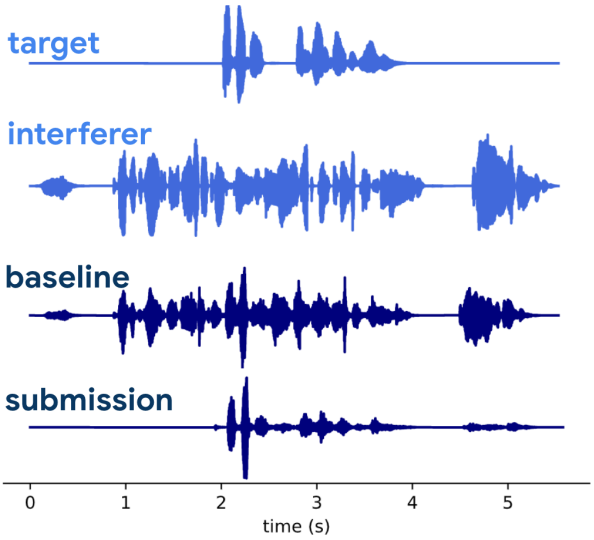

Figure 2: *Sample audio waveforms (average of left and right channels) for scene S07458, a target speaker with an interfering speaker, from the development dataset (withheld from model training). Our submission consists of silence up until about* 2.0*s, approximately the onset of the target speaker. The baseline here is just the front two microphones.*

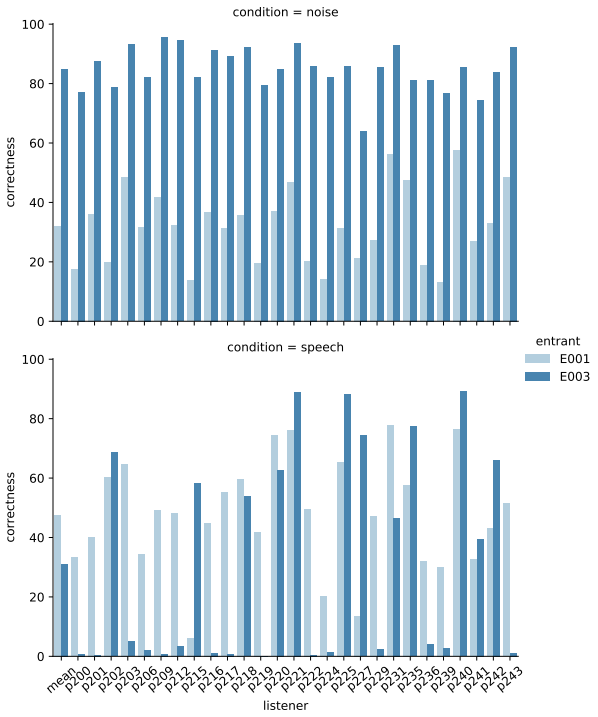

Figure 3: *Listener intelligibility evaluation results for baseline (E001) and our submission (E003) for noise and speech interferers from a subset of listeners. Correctness is number of words identified correctly as a percentage of the total number of words.*

with speech interferers, where correctness decreasing from the baseline's 51.2% to 4.4%.

After investigating these lower scores with the speech interferers, it appears many listeners may have experienced confusion with identifying the incorrect speaker as a target in many or all utterances. Some listeners (e.g. p219) had multi-word transcripts for each utterance that were a 0% match with the ground truth transcript – with the exception of the highest signal-to-noise ratio utterances where they had blank transcripts, suggesting they were transcribing the wrong (i.e. interfering) speaker. Listeners were instructed "You will hear two talkers speaking at the same time. One talker will start later than the other. You must repeat what this 2nd talker is saying." However, as illustrated in Figure 2, in our submission, both speaker's voices appear simultaneously after an initial period of silence, likely causing confusion.

3.4. Computational resources

The enhancement model was trained for 76360 steps on 32 Google Cloud TPU v3 cores, which took about 10 hours wall-clock time. This model has 2.9M trainable parameters. Since it operates on a STFT with hop 2.5 ms, the enhancement model requires approximately 1.16B multiply-and-accumulate (MAC) operations per second for each stereo channel. The RLS beamformer computes a new beamforming filter every 2.5 ms for 129 frequencies, each of which requires only a few matrix multiplies. The linear equalizer only needs to compute a 129-dimensional vector of gains from the audiogram once for a given listener, and this filter is applied every 2.5 ms.

4. Discussion

We optimized our approach using MBSTOI scores only, without conducting quantitative listening tests, out of the convenience of the CEC1-provided evaluation framework. We used linear equalization because it seemed good enough for this challenge with relatively limited dynamic range of speech loudness. We note that multiband compression could be applied using the same STFT frames without additional algorithmic latency.

Our speech enhancement model implicitly learned to identify which speaker was the target from the fact that in all training data, the target begins speaking only after 2–3 seconds. Realistically when one cannot rely on this cue, some conditioning information (e.g speaker identity, distance, or azimuth) indicating the target speaker could be used. Or, a model could separate all the sources (individual speech sources and noise), and a user could select one via some user interface.

Lastly, our submission completely silenced the interferer during the initial period of each scene before target speech onset, a phenomenon that may have confused listeners. This was not an intentional decision, but an artifact of the way the enhancement model was trained on the training data. Realistically, allowing an attenuated version of the interferer to be audible might allow listeners to adapt to the noise source, and thereby achieve better intelligibility; or listeners could be given control over the intereferer suppresion level.

5. Acknowledgements

We thank the Clarity Challenge organizers for organizing the challenge and responding quickly to problems along the way.

6. References

[1] E. W. Healy, S. E. Yoho, J. Chen, Y. Wang, and D. Wang, "An algorithm to increase speech intelligibility for hearing-impaired listeners in novel segments of the same noise type," *The Journal of the Acoustical Society of America*, vol. 138, no. 3, pp. 1660–1669, 2015.

[2] K. Wilson, M. Chinen, J. Thorpe, B. Patton, J. Hershey, R. A. Saurous, J. Skoglund, and R. F. Lyon, "Exploring tradeoffs in models for low-latency speech enhancement," in *Proc. IWAENC*, 2018, pp. 366–370.

[3] S. Wisdom, J. R. Hershey, K. Wilson, J. Thorpe, M. Chinen, B. Patton, and R. A. Saurous, "Differentiable consistency constraints for improved deep speech enhancement," in *Proc. ICASSP*, 2019, pp. 900–904.

[4] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. ICASSP*, 2016, pp. 31–35.

[5] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[6] Z.-Q. Wang, H. Erdogan, S. Wisdom, K. Wilson, and J. R. Hershey, "Sequential multi-frame neural beamforming for speech separation and enhancement," in *Proc. SLT*, 2021.

[7] S. Graetzer, J. Barker, T. J. Cox, M. Akeroyd, J. F. Culling, G. Naylor, E. Porter, and R. V. Muñoz, "Clarity-2021 challenges: Machine learning challenges for advancing hearing aid processing," in *Proc. Interspeech*, 2021.

[8] K. Zmolikova and J. H. Cernocky, "BUT system for the first clarity enhancement challenge," in *Proc. Clarity Workshop on Machine Learning Challenges for Hearing Aids*, 2021.

[9] A. H. Moore, S. Hafezi, R. Vos, M. Brookes, P. A. Naylor, M. Huckvale, S. Rosen, T. Green, and G. Hilkhuysen, "A binaural mvdr beamformer for the 2021 clarity enhancement challenge: ELO-SPHERES consortium system description," in *Proc. Clarity Workshop on Machine Learning Challenges for Hearing Aids*, 2021.

[10] X. Chen, Y. Shi, W. Xiao, M. Wang, T. Wu, S. Shang, Q. Meng, and N. Zheng, "A cascaded speech enhancement for hearing aids in noisy-reverberant conditions," in *Proc. Clarity Workshop on Machine Learning Challenges for Hearing Aids*, 2021.

[11] Z. Tu, J. Zhang, N. Ma, and J. Barker, "A two-stage end-to-end system for speech-in-noise hearing aid processing," in *Proc. Clarity Workshop on Machine Learning Challenges for Hearing Aids*, 2021.

[12] P. Kendrick, "Hearing aid speech enhancement using u-net convolutional neural networks," in *Proc. Clarity Workshop on Machine Learning Challenges for Hearing Aids*, 2021.

[13] T. Gajecki and W. Nogueira, "Binaural speech enhancement based on deep attention layers," in *Proc. Clarity Workshop on Machine Learning Challenges for Hearing Aids*, 2021.

[14] R. F. Lyon, *Human and machine hearing*. Cambridge University Press, 2017.

[15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

[16] R. L. Plackett, "Some theorems in least squares," *Biometrika*, vol. 37, no. 1-2, pp. 149–157, 1950.

[17] A. H. Sayed, *Fundamentals of adaptive filtering*. John Wiley & Sons, 2003.

[18] S. S. Haykin, *Adaptive filter theory*. Pearson Education India, 2008.

[19] S. C. Douglas, "Widely-linear recursive least-squares algorithm for adaptive beamforming," in *Proc. ICASSP*, 2009, pp. 2041–2044.

[20] K. W. Berger, E. N. Hagberg, and R. L. Rane, "A reexamination of the one-half gain rule," *Ear and Hearing*, vol. 1, no. 4, pp. 223–225, 1980.