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

Samuel J. Yang, Scott Wisdom, Chet Gnegy, Richard F. Lyon, Sagar Savla

Google Research

dicklyon@google.com

Abstract

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, as submission E003 to the 2021 Clarity Enhancement Challenge Round 1 (CEC1). 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. On the development set we achieve a mean of 0.632 and median of 0.642, compared to the mean and median of 0.41 for the baseline baseline. On the 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 real listener intelligibility assessment, for scenes with noise interferers, we see an average improvement in intelligibility from 32% to 85%, but for speech interferers, we see more mixed results, potentially from listener confusion.

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

1. Introduction

This is a technical report for our submission to the Clarity Enhancement Challenge Round 1 (CEC1) [1].

2. Hearing aid model

Motivated by the benefit of mask-based separation for hearing-impaired users [2] and the effectiveness of neural beamforming [3], our hearing aid model contains three components, as illustrated in Figure 1: a parallel bank of 2 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 [1] 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 44.1 kHz. The enhancement model utilizes 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 [4]. 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 [5, chapter 3], [6].

The enhancement model was trained with a multi-resolution spectral loss, which is mean-squared error between compressed magnitude and compressed complex consistent STFTs [7] 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:

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

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 [7], 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.

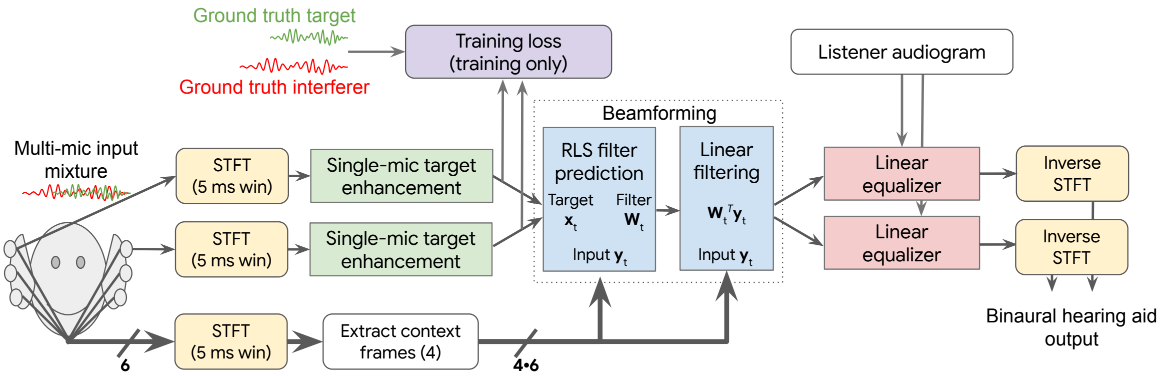

Figure 1: *Block diagram of our proposed system.*

For training data, we use the Clarity Challenge training set of 6000 scenes. We found on-the-fly augmentation to be advantageous, leading to better validation metrics on the Clarity Challenge 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 [8], 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 [9, 10, 11] 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 [3] 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 [12]. 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 to 0.75, consistent with the audiologist's "one-half gain rule" [13] 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 [1]. 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.

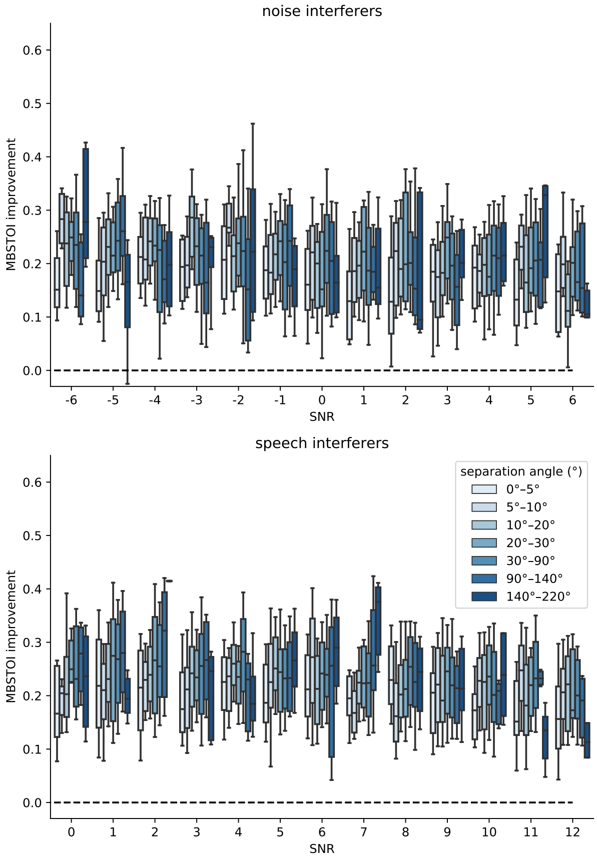

Figure 2: *Improvement in MBSTOI score of proposed hearing aid model, compared with the Clarity Challenge baseline system [1], based on interferer type, SNR and angle between target and interferer. Dashed line denotes no improvement. Boxes represent quartiles and median; whiskers denote 10%–90%.*

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.

To understand in more detail the particular scenarios where the proposed solution performed significantly better than the CEC1 baseline, we computed, for each scene and listener, the difference in MBSTOI score, with a positive values reflecting an improvement of the proposed solution over the CEC1 baseline. Figure 2 shows this difference broken down by interferer type, the absolute azimuth angle difference between the target and interferer (i.e. 0 degrees reflects the target and interferer sounds coming from the same direction, while 180 degrees reflects the sounds coming from entirely opposite directions) and input signal-to-noise ratio (SNR). Our proposed solution generally improves in all scenarios, with the improvements tending to be higher for larger separation angles, and uniform improvement over input SNR.

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

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 baseline.

3.2. Example submission processed audio

We plot sample audio waveforms or a noise and speech interferer examples in Figure 3 and Figure 4, respectively, 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 in these examples varies from $2.0s$ to $2.7s$. 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.

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 in a listening panel 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. Based on preliminary results from a subset of the listeners presented, shown in Figures 5 and 6, our submission performed among the best for scenes with noise interferers, with an average increase in correctness over the baseline from 32% to 85%, but among the worst for scenes with speech interferers, where correctness decreasing from 48% to 31%. In fact, for speech interferers, many listeners had near-zero scores; although some listeners did achieve scores much higher than baseline, an improvement over baseline more consistent with that achieved with the noise interferers.

After investigating these lower scores with the speech interferers, it appears many listeners may have experienced con-

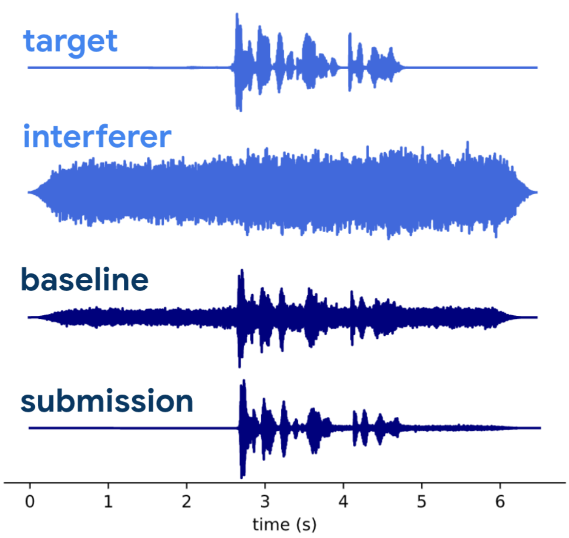

Figure 3: *Sample audio waveforms (average of left and right channels) for scene S08143, a target speaker with a noise interferer, from the development dataset (withheld from model training). Our submission consists of silence up until about* 2.7*s, approximately the onset of the target speaker. The baseline is here is just the front two microphones.*

fusion with identifying the incorrect speaker as a target, as evidenced by the following. 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.” and during the study, were reminded ”You will hear two talkers. Repeat the 2nd talker.” However, as illustrated in Figure 4, in our submission, both speaker's voices appear simultaneously after an initial period of silence. The target speaker finishes speaking first, whereas the interfering speaker finishes speaking last, and we believe this may have led some listeners to mistake the interferer as the target. Furthermore, a review of the preliminary listener transcripts shows some listeners (e.g., p219) with a near-zero overall score gave no response at all for only the highest SNR examples (examples where the interferer may not be audible at all in our submissions), suggesting they were (incorrectly) listening for the interferer. In addition, there were some listeners (e.g., p218) who have nearly perfect utterance-level scores on some of the utterances, but completely incorrect transcripts for other utterances, suggesting that they correctly identified the target speaker in some utterances but not others.

3.4. Computational resources

The enhancement model was trained for 76360 steps on 32 Google Cloud TPU v3 cores, which took about 10 hours wallclock 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.

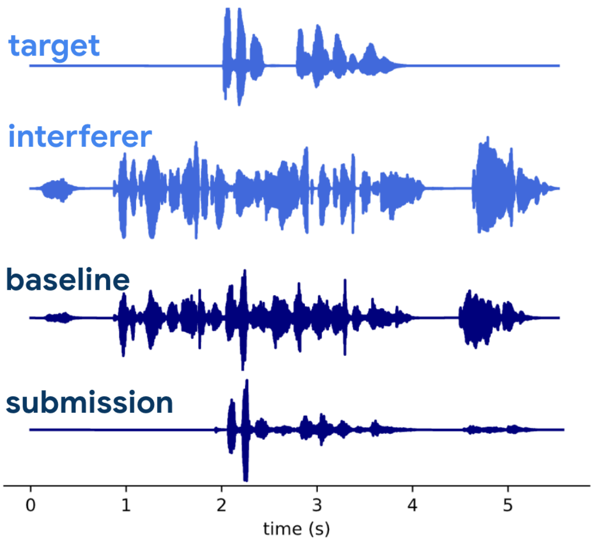

Figure 4: *Same as Figure 3 but for scene S07458, a target speaker with an interfering speaker. Our submission consists of silence up until about* 2.0*s, slightly before the onset of the target speaker.*

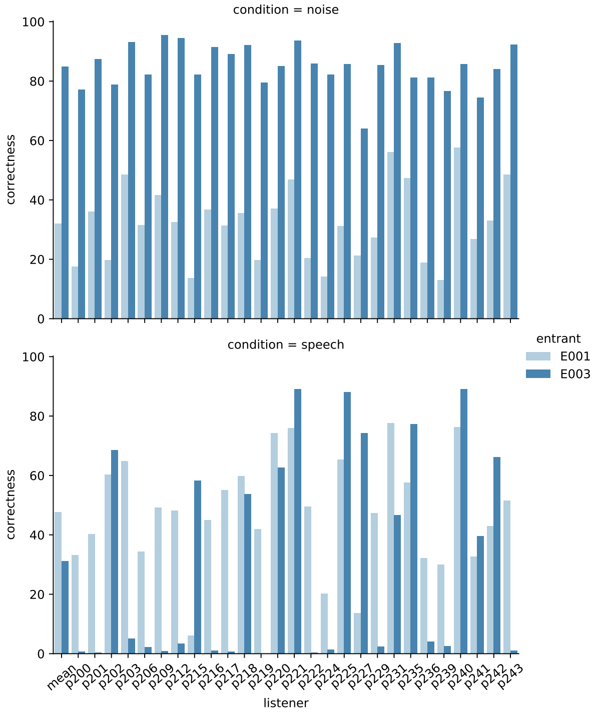

Figure 5: *Preliminary listener intelligibility evaluation results for baseline (E001) and our submission (E003) for noise and speech interferers. Correctness is number of words identified correctly as a percentage of the total number of words.*

4. Discussion

Though our proposed solution yields a higher MBSTOI score, and qualitatively sounds less noisy than our baseline of just using the front two microphones, we did not conduct quantitative

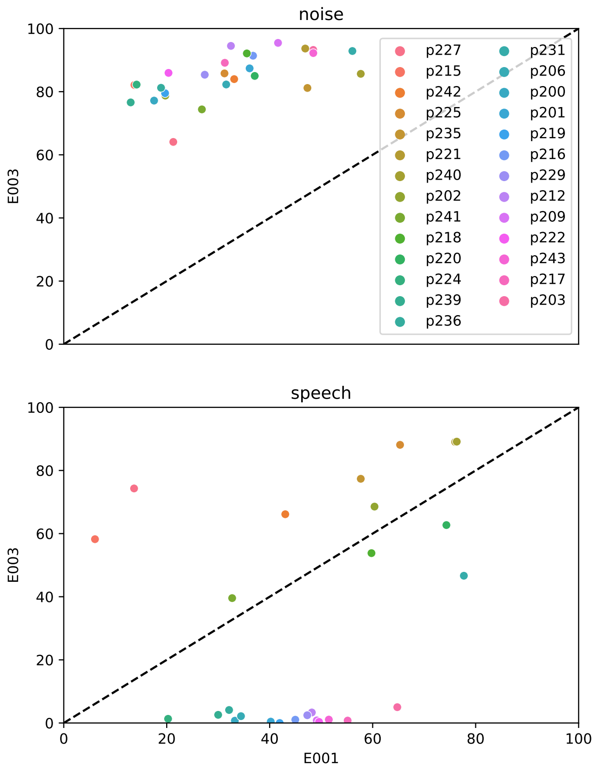

Figure 6: *Same as Figure 5 but plotted as a scatter plot, with listeners (colors) sorted by improvement in speech interferers over baseline.*

listening tests.

The fact that we used linear equalization instead of the usual multiband compression (MBC) does not mean we believe that is a better solution for a hearing aid, but it seemed good enough for this challenge that has a relatively limited dynamic range of speech loudness. We note that MBC could be applied using the same STFT frames such that no additional algorithmic latency would be introduced.

Our speech enhancement model implicitly learned to identify the target speaker from the speech interferer from the fact that in all training (and test) data, the target begins speaking after 2 seconds. To get this model working in a real scenario where it cannot rely on this cue, some conditioning information indicating the target speaker could be used, e.g speaker identity, distance, or azimuth. Or, a model could separate all the sources (individual speech sources and noise), and a user could select via some user interface which one to focus on.

Lastly, our submission completely silenced the interferer during the initial period of each scene before target speech onset. This may have created confusion for the listener panel evaluation since the listeners were instructed to ignore the first voice and transcribe the second. This was not an intentional decision on our part, but an artifact of the way the enhancement model was trained and the characteristics of the training data. It would be interesting to explore if allowing an attenuated version of the interferer to be audible would allow listeners to adapt to the noise source, and thereby achieve better intelligibility. In addition, in real scenarios, it may be helpful to explore ways to give listeners control over the interferer suppression level.

5. Acknowledgements

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

6. References

[1] 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.

[2] 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. [Online]. Available: https://doi.org/10.1121/1.4929493

[3] 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.

[4] 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.

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

[6] 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.

[7] 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.

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

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

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

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

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

[13] 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.