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 beamforming filter, and a linear equalizer, to improve speech intelligibility in the presence of speech or noise interferers, as a submission 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.
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 ConvTasNet 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.

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.

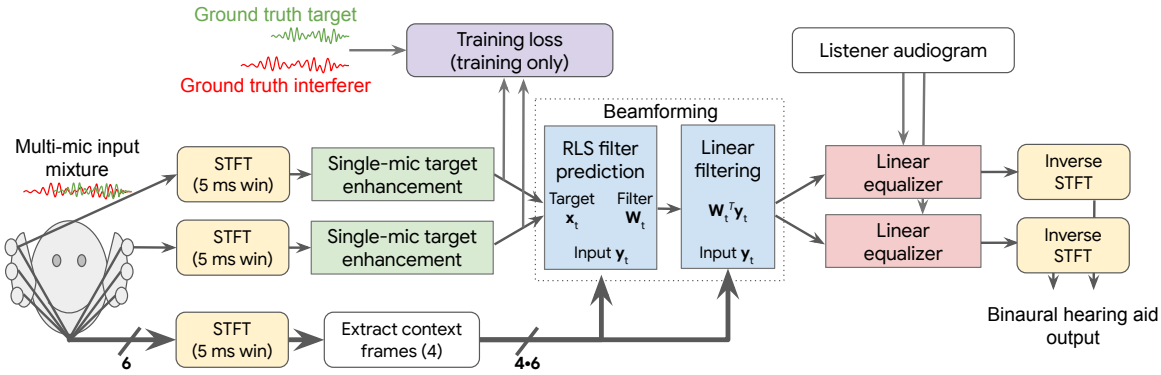

Figure 1: *Block diagram of our proposed system.*

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 4 Google Cloud TPUs (16 chips) 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.

In comparison, our proposed solution with our beamformer but no equalizer achieved on the development set an MBSTOI

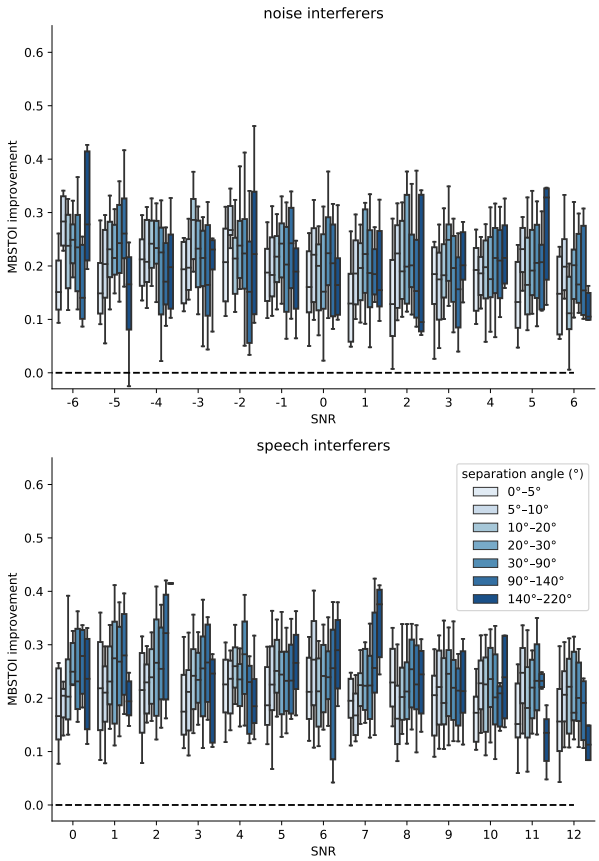

Figure 2: *Improvement in MBSTOI score of proposed hearing aid model, compared with the Clarity Challenge baseline system [1], based on interferer type, SNR and angle between target and interferer. Dashed line denotes no improvement. Boxes represent quartiles and median; whiskers denote 10%–90%.*

mean of 0.602 and median of 0.611. 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.

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. Computational resources

The enhancement model was trained for 76360 steps on 4 Google Cloud TPUs (16 chips), 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

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 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.

While we demonstrated a complete solution using an enhancement model trained with data augmentation, an RLS beamformer, and a linear equalizer, we note that several ablation studies could be conducted to confirm our solution is optimal, or to identify components of our solution that could be further improved. First, the enhancement model could be trained without data augmentation. Second, the output of the enhancement model could be directly used as input to the linear equalizer, instead of as an input to the RLS beamformer, eliminating the need for the RLS beamformer. Third, the RLS beamformer could be used with some sort of direction of arrival estimate, eliminating the need for the speech enhancement model altogether.

Lastly, we note that 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.

5. Conclusions

TBD after our submission is evaluated.

6. Acknowledgements

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

7. 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.