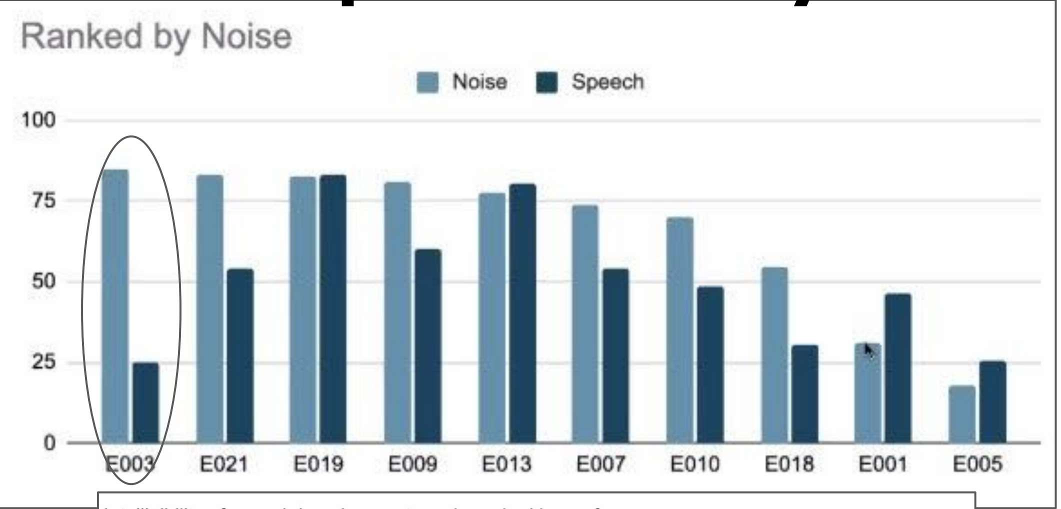
Listening with Googlears: Low-latency neural multiframe beamforming and equalization for hearing aids

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

System E003

Google Research

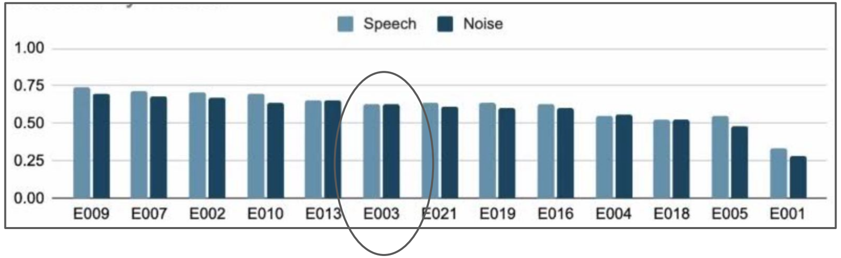
E003 (preliminary listening data)



Intelligibility of speech in noise, systems in ranked by performance

Entrant	Beamforming	DNN Noise Removal	Hearing Loss Compensation
E003	RLS	Conv-TasNet	Linear, fitting formula
E021	Weighted LCMP	DNN (Deep MFMBVDR)	MBDRC
E019	Weighted LCMP		MBDRC
E009		MC Conv-TasNet	Linear, NN optimised
E013	MVDR		Linear, fitting formula but AGC
E007	MVDR	Conv-TasNet	Linear, NN optimised
E010		U-Net CNN	Linear, fitting formula
E018		2D CNN + LSTM, WPE	Dynamic EQ
E001			Baseline
E005		Binaural Conv-Tasnet	

MBSTOI



Google Research

Agenda

01 Motivation

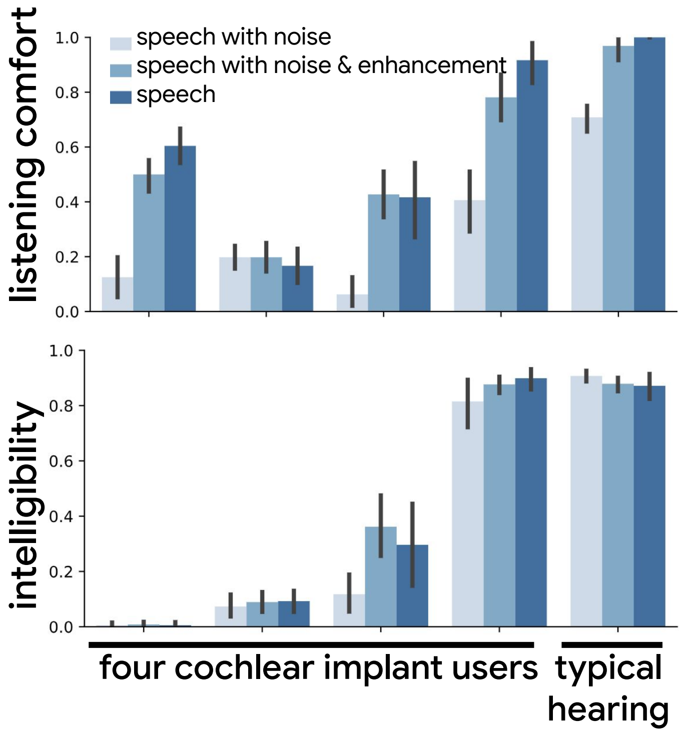
02 System description

03 Audio demos

04 MBSTOI results

05 Listening test results

Understanding speech in noise is hard (previous study with cochlear implants)



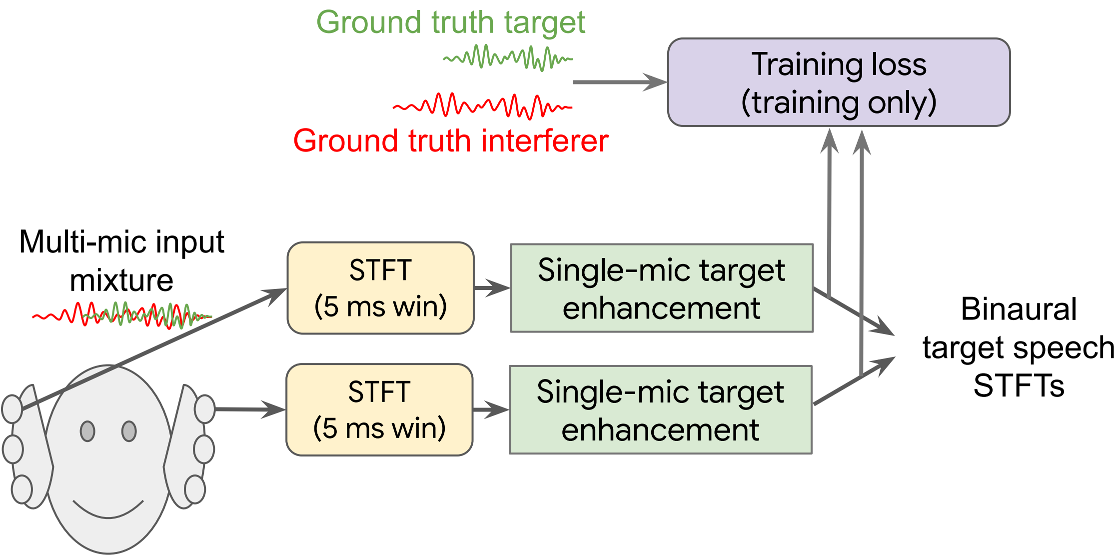
- In a small study, our application of speech enhancement helped cochlear implant (CI, a close relative of hearing aids) users' speech understanding
- See our Google AI blog post (https://ai.googleblog.com/2021/07/applying-advanced-speech-enhancement-in.html)

CI hackathon	Clarity Challenge
2-mic input	6-mic input
speech babble	single speech or noise interferer
simulated CI audio	audiogram-adjusted audio
speech enhancement only	enhance + beamform

Google Research

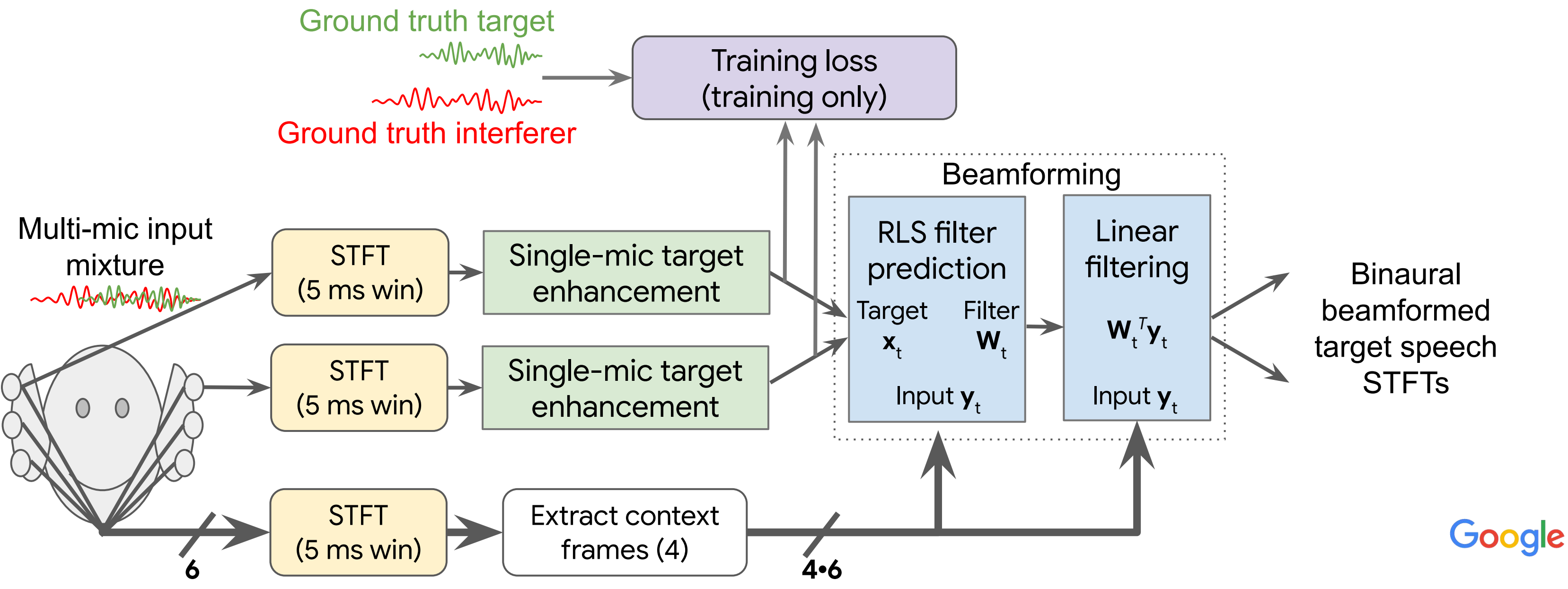
Our solution: overview

- 1) Separate single-microphone audio from left and right into target and interference signals.



Our solution: overview

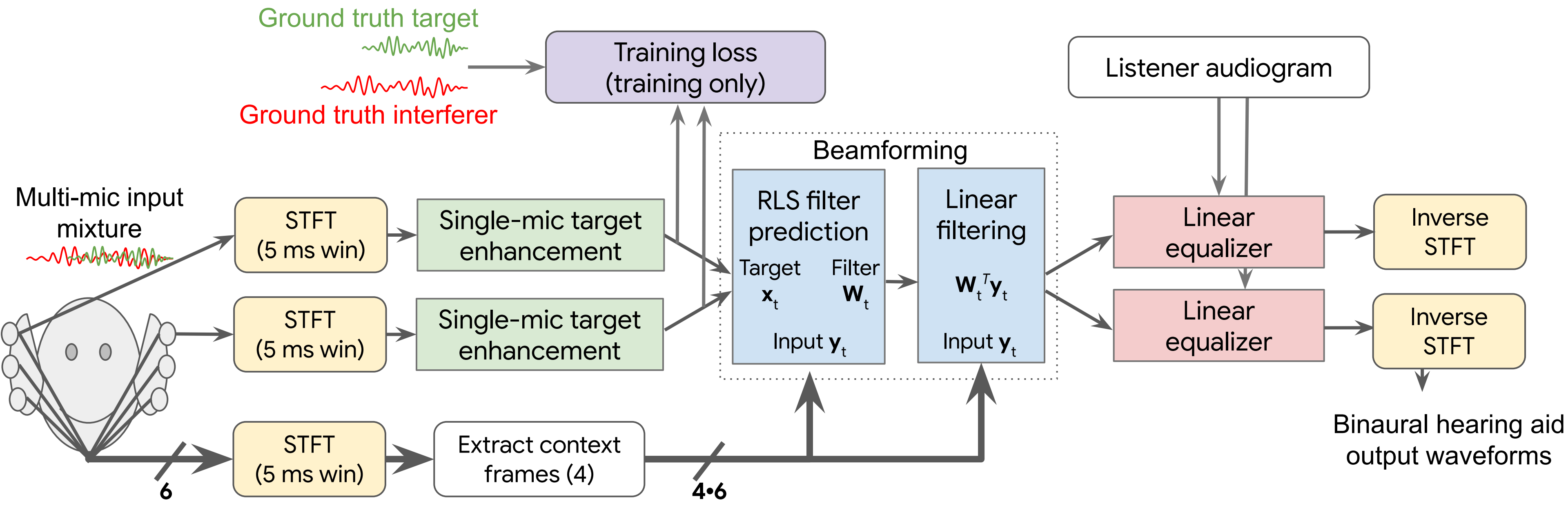
- 1) Separate single-microphone audio from left and right into target and interference signals.
- 2) Use estimate of target signal to beamform across all 6 mics with 4 context frames.



Google Research

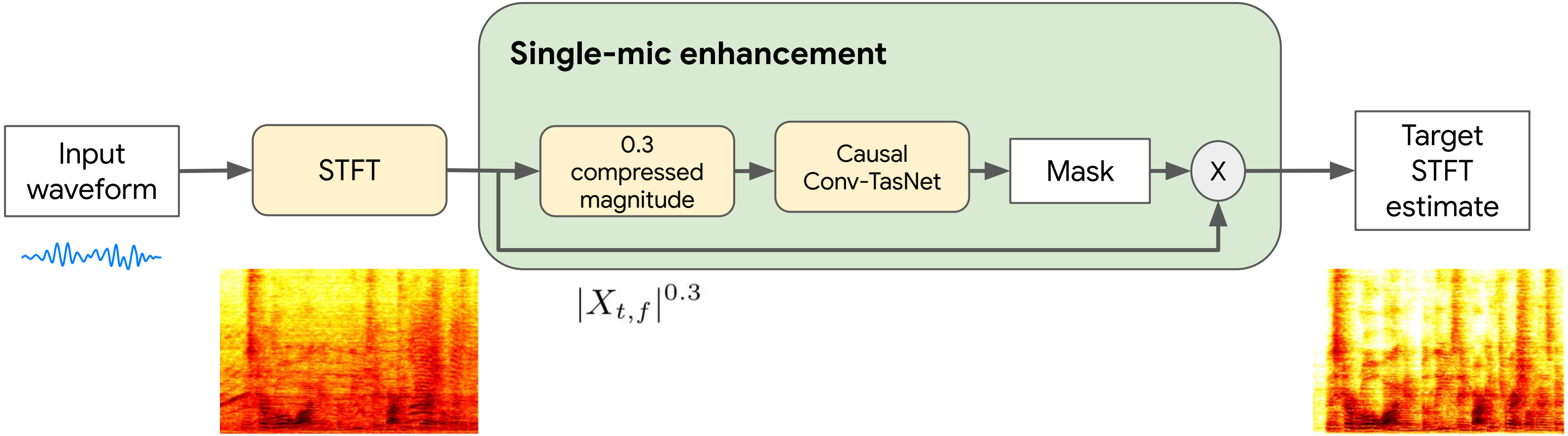
Our solution: overview

- 1) Separate single-microphone audio from left and right into target and interference signals.
- 2) Use estimate of target signal to beamform across all 6 mics with 4 context frames.
- 3) Apply linear equalizer using listener audiogram to compensate for hearing loss.



Single-mic enhancement

- Causal Conv-TasNet masking network [1] predicts a mask for input STFT.
- Trained on synthetic mixtures of target speech and interferer using TPU (next slide).

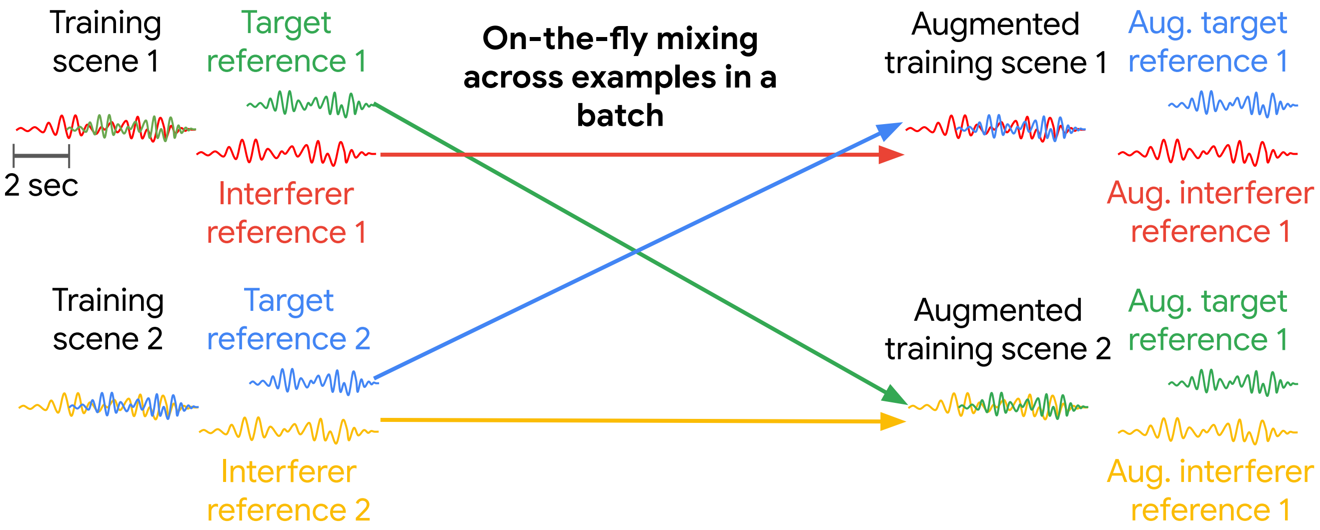


Google Research

[1] *Luo, Yi, and Nima Mesgarani. "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation." IEEE/ACM TASLP 2019*

Training for enhancement

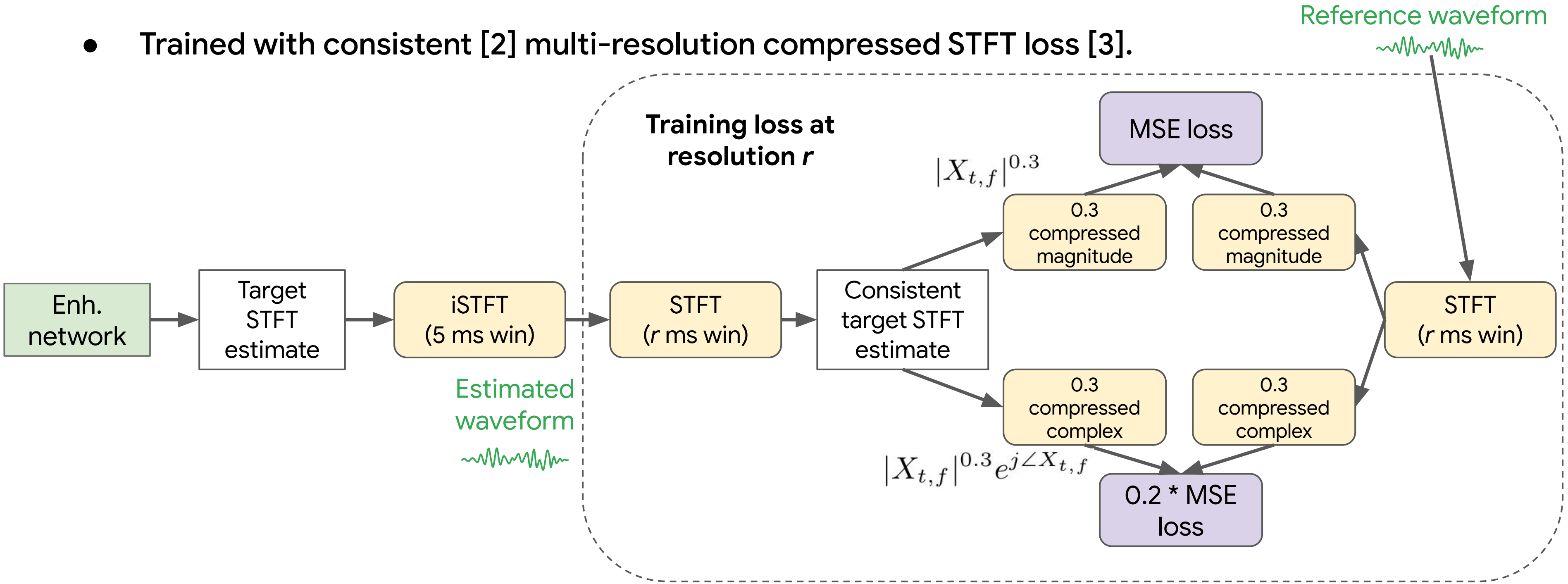
- Augmentation on single-microphone audio from Clarity Challenge scenes.
- Leverages cue that target starts after two seconds.



Google Research

Training for enhancement

- Trained with consistent [2] multi-resolution compressed STFT loss [3].

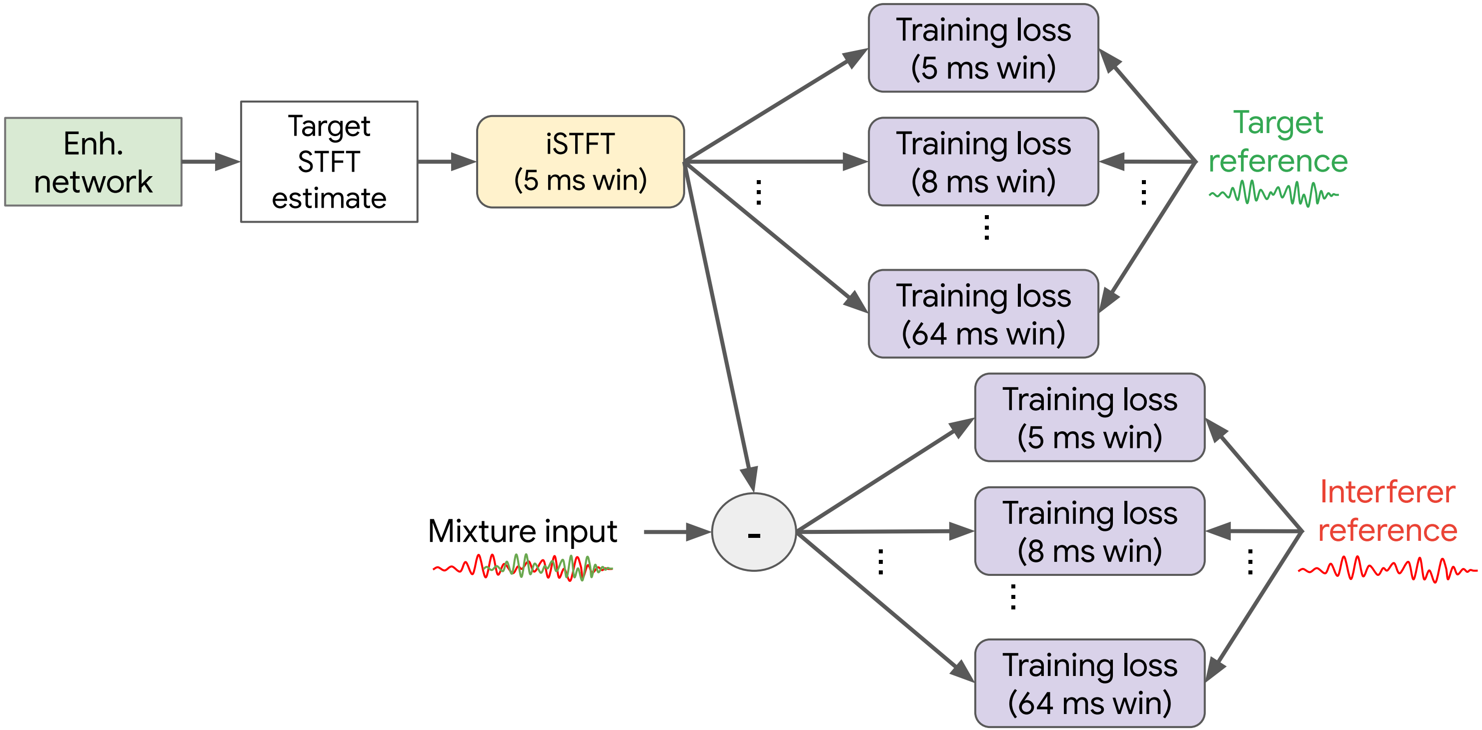


[2] Wisdom, S., Hershey, J. R., Wilson, K., Thorpe, J., Chinen, M., Patton, B., & Saurous, R. A., *Differentiable consistency constraints for improved deep speech enhancement, ICASSP 2019.*

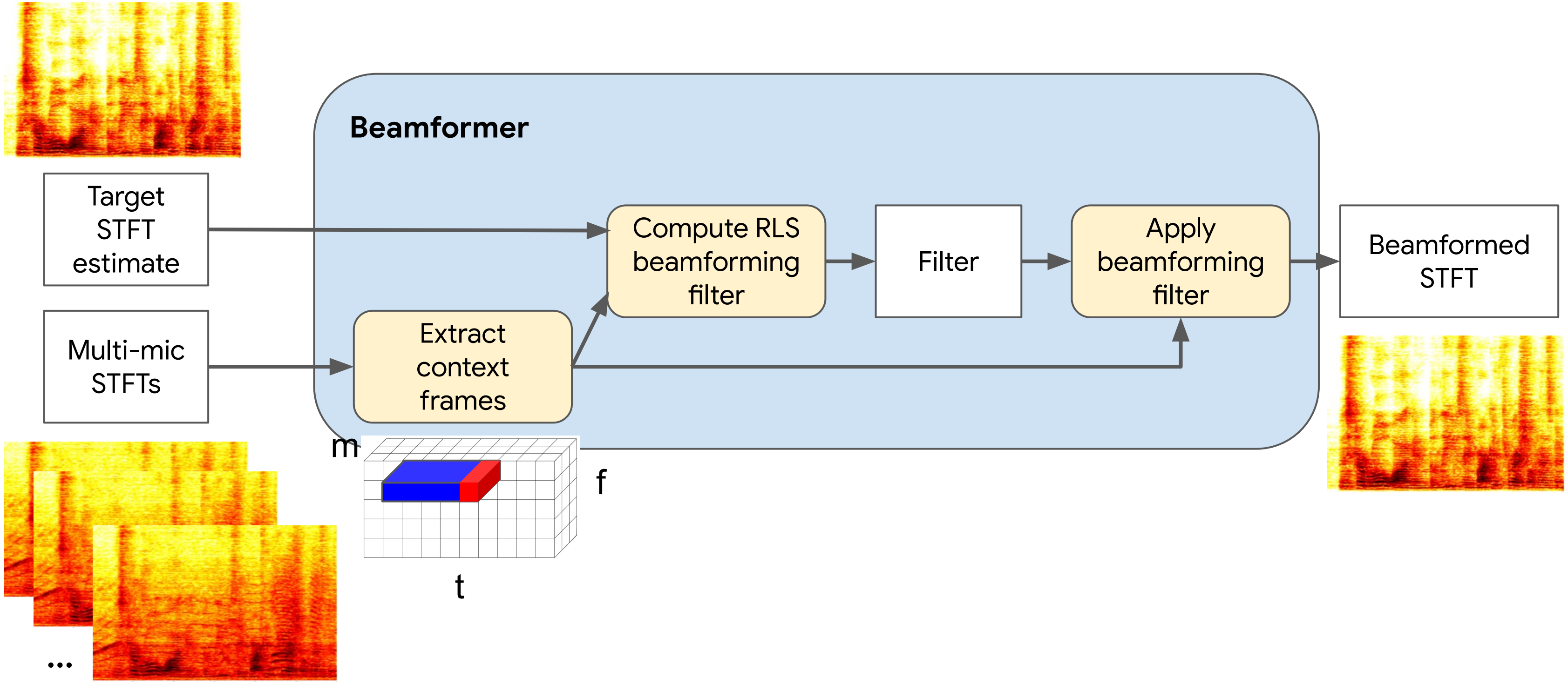
[3] Wilson, K., Chinen, M., Thorpe, J., Patton, B., Hershey, J., Saurous, R. A., Lyon, R. F., *Exploring tradeoffs in models for low-latency speech enhancement, IWAENC 2018.*

Training for enhancement

- Trained with consistent multi-resolution compressed STFT loss on target and interferer.



Causal multi-frame RLS beamforming



Google Research

Wang, Z. Q., Erdogan, H., Wisdom, S., Wilson, K., Raj, D., Watanabe, S., ... & Hershey, J. R., Sequential multi-frame neural beamforming for speech separation and enhancement. *IEEE Spoken Language Technology Workshop (SLT)* 2021.

Causal multi-frame RLS beamforming

- Optimization problem for filter **W** to predict target **x** from input **y**:

$$\hat{\mathbf{W}}_t = \min_{\mathbf{W}_t} \quad L_t(\mathbf{W}_t) = \sum_{\tau=0}^{t} \lambda_{t,\tau} \|\mathbf{x}_\tau - \mathbf{W}_t^T \mathbf{y}_\tau\|^2$$

Note that the classic unweighted RLS uses $\lambda_{t,\tau} = \lambda^{t-\tau}$, where λ is an exponential averaging weight usually chosen with value between 0.98 and 1.0.

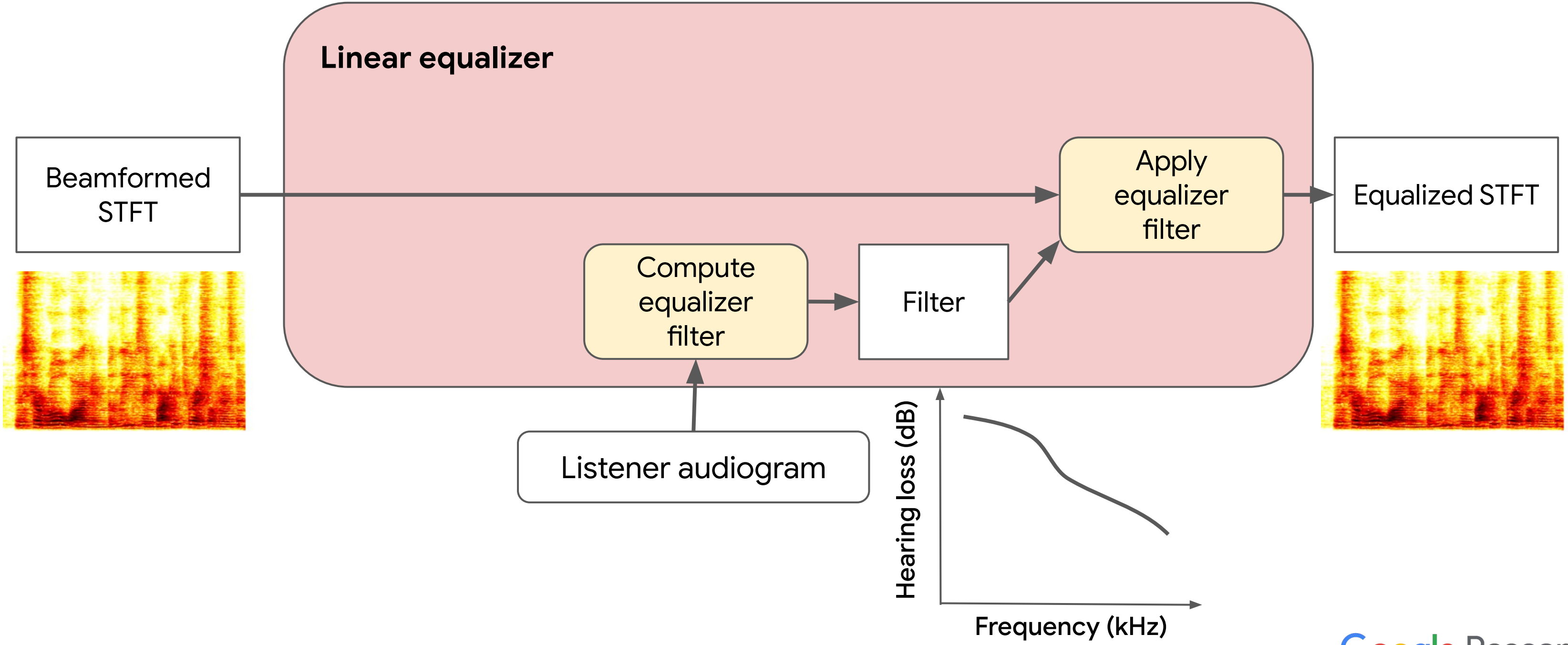
- Non-causal solution:

$$\mathbf{W}_t = \mathbf{R}_{yy,t}^{-1} \mathbf{R}_{xy,t}^T \qquad \mathbf{R}_{yy,t} = \sum_{\tau=0}^{t} \lambda_{t,\tau} \mathbf{y}_\tau \mathbf{y}_\tau^T \qquad \mathbf{R}_{xy,t} = \sum_{\tau=0}^{t} \lambda_{t,\tau} \mathbf{x}_\tau \mathbf{y}_\tau^T.$$

- Canonical causal recursive solution (**no matrix inverses!**):

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

Linear equalizer

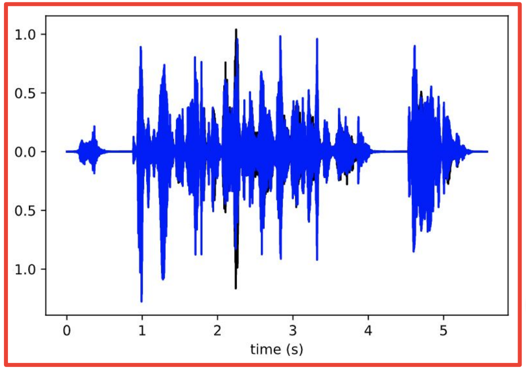


Google Research

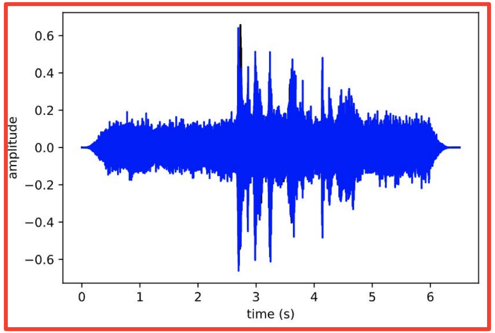
Audio demos

Description: male voice target with female voice interferer (Scene S07458)

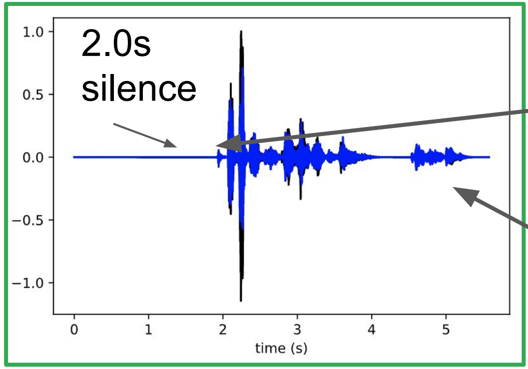
Description: male voice target with noise interferer (Scene S08143)



Baseline

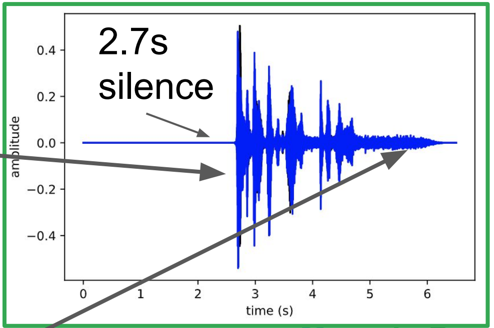


Enhancement output before beamformer



Target speech onset

Our submission (enhancement + beamformer + linear equalizer)



Attenuated but non-zero interferer

Google Research

Audio demos

Noise interferer example:
(i.e. hairdryer, dishwasher, kettle, fan)

baseline

our submission

Description: male voice target with noise interferer

Speech interferer example:
(i.e. another male or female voice**)

**interferer begins speaking immediately; the target starts speaking after 2 seconds

baseline

our submission

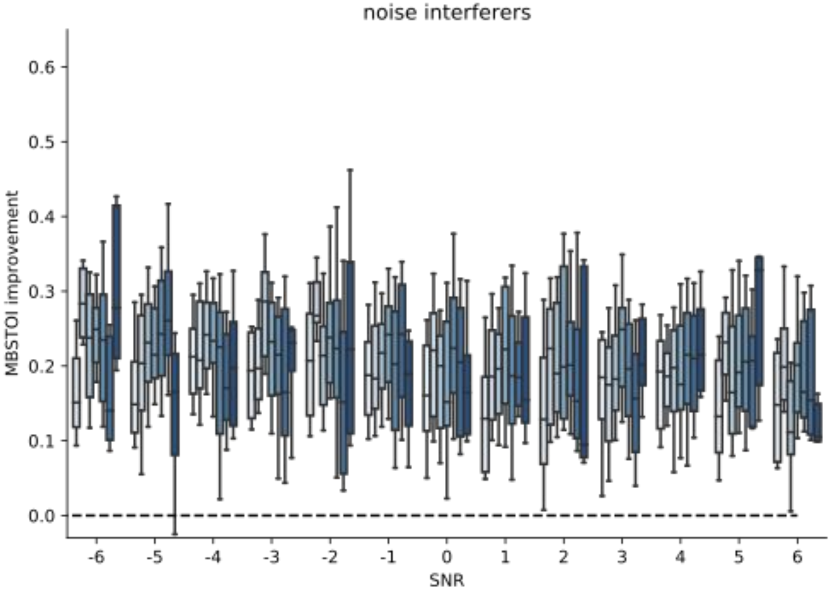
Description: male voice target with female voice interferer

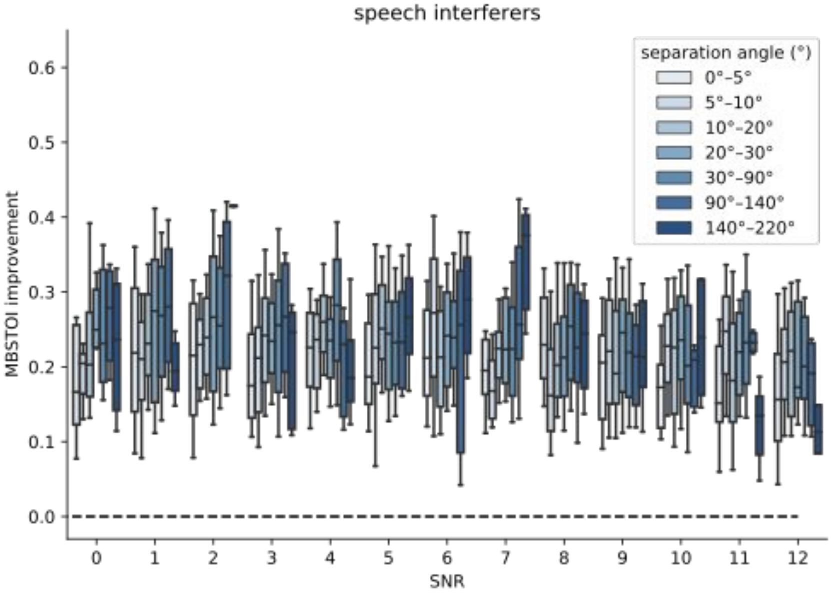
Google Research

MBSTOI results

Dev baseline: 0.41 mean, 0.41 median
Dev proposed: 0.632 mean, 0.642 median

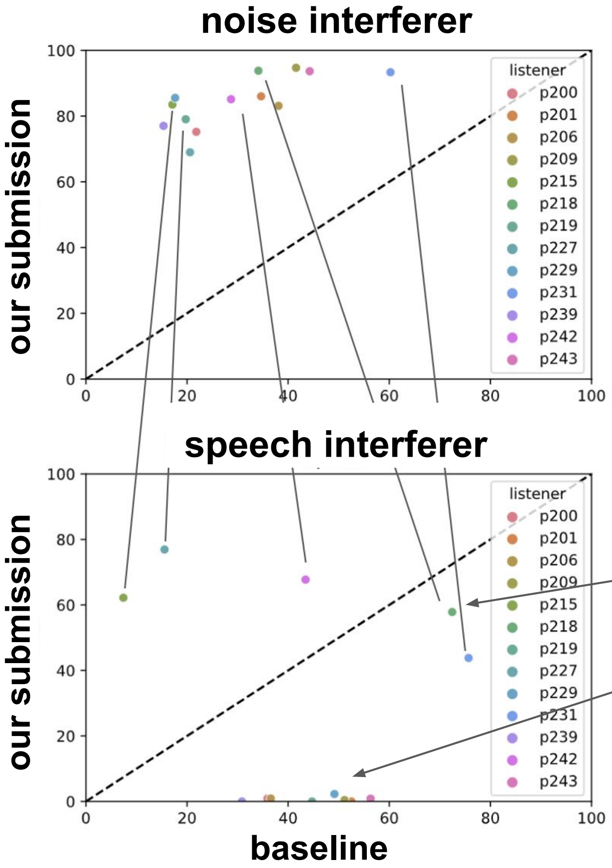
Eval baseline: 0.310 mean, 0.314 median
Eval proposed: 0.644 mean, 0.6652 median





Google Research

Listening test results (preliminary)



- For noise interferers, +~40% boost in correctness.
  - Direction of improvement consistent with MBSTOI.
- For speech interferers, highly mixed results.
  - Next slide: investigate 2 listeners responses (**p218**, **p219**).

Google Research

p219

correctness	hypothesis	prompt	SNR
0.0	confusion	There was a pause broken by the girl	0
0.0	send a message	I wish you could have seen them	0
0.0		I just didn't want to risk it	0
0.0	consumer confusion	For once her mileage was beginning to show	0
0.0		But a very nice little bear all the same	2
0.0	£50	I'm going up myself to have a look round	2
0.0		He's gone said Sue drama in her voice	2
0.0	duration	I must take advice said Sir George stubborn an...	2
0.0		She wasn't going to let him intimidate her any...	4
0.0		George is quite an interesting character in th...	4
0.0	round Arch	I must confess I was deeply depressed he said ...	4
0.0	English language is similar	Of course they would say that wouldn't they	4
0.0		I pray that it is you reading this my darling	6
0.0	surface while talking	We can attack an orange for not being an apple	6
0.0	finance	Well not quite for nothing he said	6
0.0	fire	You just walk out stand and smile	6
0.0	ice	For the first time he was proud of them	6
0.0	Happy New Years Eve	The building was a great show box of concrete	8
0.0	something	Well after all she is your mother	8
0.0		We shan't have to get involved with all that	8
0.0	High Barnet	I suspect that may not be possible Edward told...	8
0.0	what is a points of interest	He might even buy us a drink now	10
0.0	spoils you again	He hasn't got the bottle for anything else	10
0.0		So we return to the original crux	10
0.0		But there was no general media outcry	10
0.0		You know the one with the broken leg	12
0.0		He would like to have the baby	12
0.0		We seem to have made a little camp	12
0.0		Right now I am off the international scene	12
0.0		War wasn't going to roll through our village	12
0.0		Like the telephone system or the electrical wi...	12

Lister p219 had 0% correct and had no response to highest SNR examples - possibly only heard one speaker.

Listener p218 seems to randomly alternate between correct and incorrect - possibly confusing which speaker is target

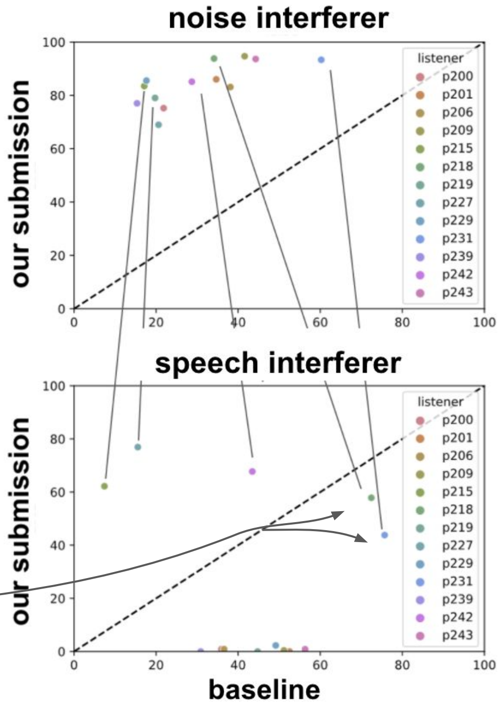
? denotes examples where listener transcript differs significantly from actual target

p218

correctness		hypothesis	prompt	SNR
0.000000	?	on the next book	She was trying to frighten me off of course	0
100.000000		beyond this point there will be no further dev...	Beyond this point there will be no further dev...	0
87.500000		my faith in Justice was 0 that day	My faith in justice was zero that day	0
0.000000	?	an easy route	She took a deep breath to steady herself	0
66.666667		I have some games sufficient confidence to loo...	I had now gained sufficient confidence to look...	0
66.666667		on paper news on Sunday was brilliant idea	On paper 'News on Sunday' was a brilliant idea	2
77.777778		we need to discover how we live with	We need to discover how to live with them	2
87.500000		Tactics play a big part in the cycle race	Tactics play a big part in cycle racing	2
12.500000	?	the middle page shadow	The wobbly singing of the little choir stopped	2
14.285714	?	all thought for the	Could you not consider leaving the room	2
0.000000	?	firearms in	Did the seaside do this to people	4
100.000000		I was about to telephone the police	I was about to telephone the police	4
12.500000		what are the man took they a joke	While they work the men talk and joke	4
100.000000		they had a good evening together all the same	They had a good evening together all the same	4
0.000000	?	comparison and one other	He needs to leap into the next league	6
0.000000	?	in various ways	But then how could he have seen through it	6
87.500000		he was the only one apart from me	She was the only one apart from me	6
87.500000		was it Jean who told you all this	Was it she who told you all this	6
50.000000		campsite it's a reflection of his excitable fo...	Perhaps that is a reflection of his enthusiast...	8
0.000000	?	depends considerably upon	We had to stop selling the turf then he said	8
0.000000	?	wonderful	Reference has been made to the complexity of p...	8
88.888889		it's not the end of the world	It's not like the end of the world	8
100.000000		you can be what you like he said	You can be what you like he said	10
100.000000		you may speak to them if you wish	You may speak to them if you wish	10
75.000000		I don't want to know anymore	I don't want to know any more	10
0.000000	?	press collection	You look like a little Dutch girl	10
100.000000		this morning she could barely taste anything	This morning she could barely taste anything	12
0.000000	?	try this	Have you been up to the house yet	12
100.000000		it must be the most beautiful house	It must be the most beautiful house	12
100.000000		I said nothing but my mother didn't seem to no...	I said nothing but my mother didn't seem to no...	12

Listening test results (preliminary)

- Methodology: for each utterance, I reviewed the transcript and ground truth and made binary decision of correct or incorrect.
- 8 listeners had total scores near zero
  - 4 gave no responses for the highest SNR utterances, suggesting they were listening for the intereferer and got confused when they only heard one speaker
  - 2 consistently incorrect, except for one (mid level SNR) utterance where they got it correct.
  - 2 consistently got incorrect for all examples, but appeared confident in noting many words in each utterance
- 7 listeners had non-zero total scores
  - 2 seem to alternate between incorrect and correct utterance transcripts (see p218 and p231)
  - 5 listeners appear to have completely valid responses
- Conclusion: 5 of 15 listeners appear to have completely valid responses.



Google Research

Future work

- Ablations
  - Training augmentation
  - Enhancement-only
- Explore if allowing some noise in the first 2 seconds helps avoid target/interferer confusion; more generally, explore if allowing some noise allows listeners to adapt and actually enhance intelligibility.
- Real-world target identification methods (not relying on first 2 seconds of interferer)
  - Visual
  - Spatial (e.g. direction)
  - Speaker ID
- Should target/interferer speakers be from same dataset?

Thank You

Samuel J. Yang and Scott Wisdom
Research Scientists

Google Research