FACULTY OF ENGINEERING AND ARCHITECTURE

Generative Speech Enhancement with Self-Supervised Learning Models

[1] *Yanjue Song*, [2]Doyeon Kim, [2]Hong-Goo Kang, and [1]Nilesh Madhu

[1] IDLab, Ghent University - imec, Belgium
[2] Dept. of Electrical and Electronic Eng., Yonsei University, Seoul, Korea

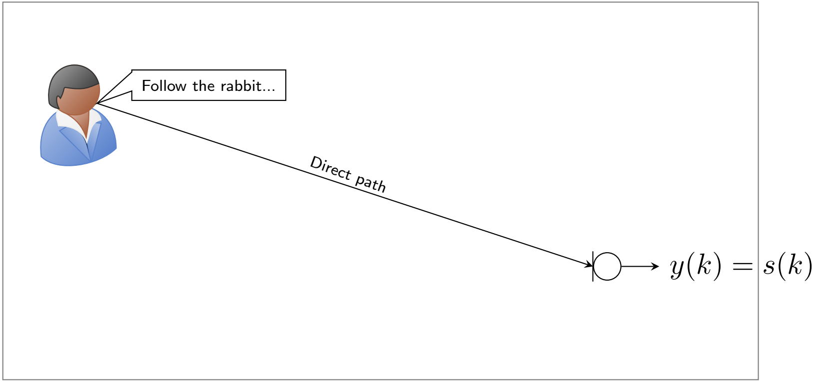
GHENT UNIVERSITY

imec

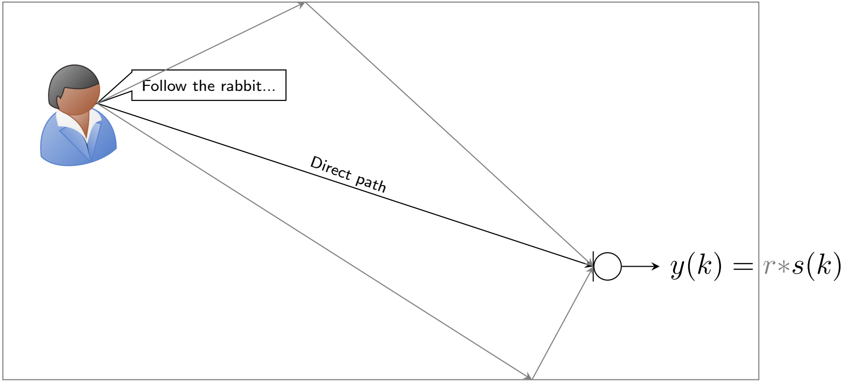
YONSEI UNIVERSITY

DSP&AI Lab. DIGITAL SIGNAL PROCESSING & ARTIFICIAL INTELLIGENCE LAB.

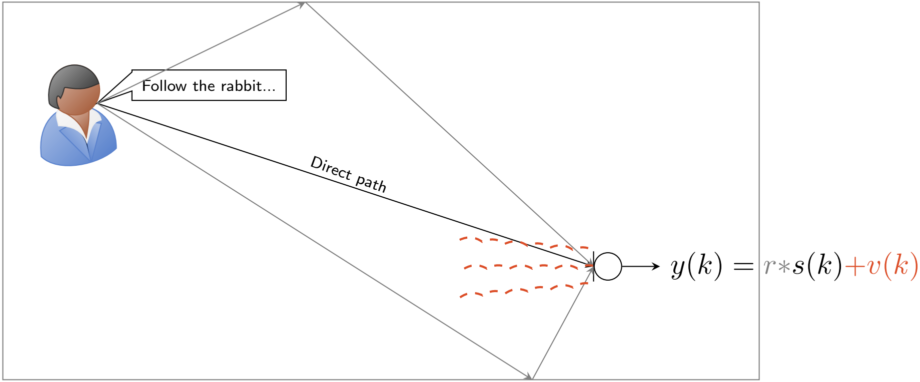
Speech Enhancement



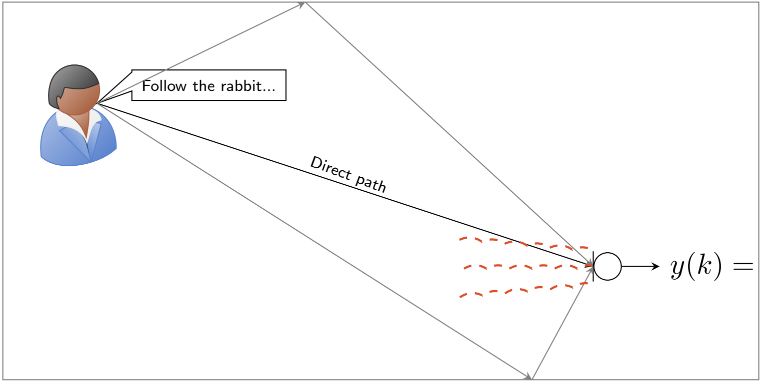
Speech Enhancement



Speech Enhancement

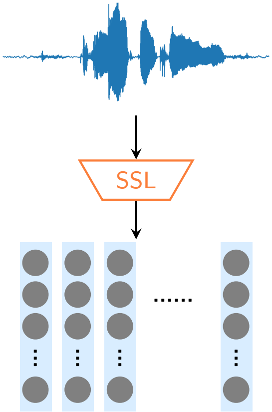


Speech Enhancement



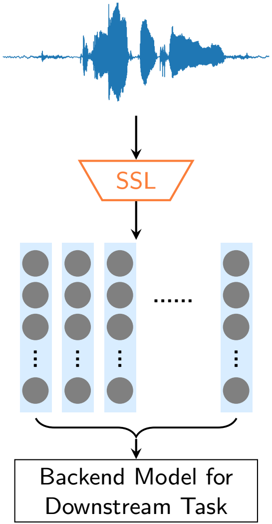
$$y(k) = r * \boxed{s(k)} + v(k)$$

Self-Supervised Learning (SSL) Models



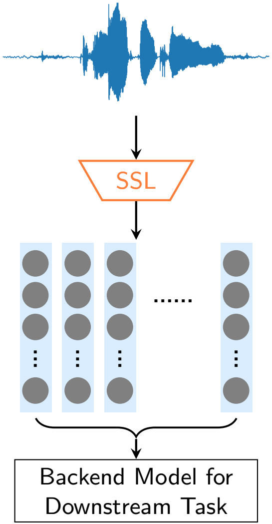
- Learning **representations** from data *without* human-labelled examples
- Extracted representations (embeddings) capable of various tasks
  - e.g., emotion recognition, speaker identification, ASR...
- Many different pre-trained models available

Self-Supervised Learning (SSL) Models



- Learning **representations** from data *without* human-labelled examples
- Extracted representations (embeddings) capable of various tasks
  - e.g., emotion recognition, speaker identification, ASR...
- Many different pre-trained models available

Self-Supervised Learning (SSL) Models



- Learning **representations** from data *without* human-labelled examples
- Extracted representations (embeddings) capable of various tasks
  - e.g., emotion recognition, speaker identification, ASR...
- Many different pre-trained models available

SSL Models in Speech Enhancement?

- **Which** pre-trained SSL model?
  $\rightarrow$ SSL model selection based on quantitative analysis of embeddings
  Song Y, Kim D, Madhu N, Kang H.-G. On the Disentanglement and Robustness of Self-Supervised Speech Representations. In *2024 International Conference on Electronics, Information, and Communication (ICEIC)* 2024 Jan 28 (pp. 662-665). IEEE.
- **How** to use it?
  $\rightarrow$ Improvement of the speech re-synthesis framework
  Song Y,, Kim D, Kang H.-G, and Madhu N. Spectrum-Aware Neural Vocoder Based on Self-Supervised Learning for Speech Enhancement. In *2024 32nd European Conference on Signal Processing (EUSIPCO)* 2024 Aug 26 (pp. 16-20). IEEE.

SSL Models in Speech Enhancement?

- **Which** pre-trained SSL model?
  $\rightarrow$ SSL model selection based on quantitative analysis of embeddings
  Song Y, Kim D, Madhu N, Kang H.-G. On the Disentanglement and Robustness of Self-Supervised Speech Representations. In *2024 International Conference on Electronics, Information, and Communication (ICEIC)* 2024 Jan 28 (pp. 662-665). IEEE.
- **How** to use it?
  $\rightarrow$ Improvement of the speech re-synthesis framework
  Song Y,, Kim D, Kang H.-G, and Madhu N. Spectrum-Aware Neural Vocoder Based on Self-Supervised Learning for Speech Enhancement. In *2024 32nd European Conference on Signal Processing (EUSIPCO)* 2024 Aug 26 (pp. 16-20). IEEE.

SSL Models in Speech Enhancement?

- **Which** pre-trained SSL model?
  $\rightarrow$ SSL model selection based on quantitative analysis of embeddings
  Song Y, Kim D, Madhu N, Kang H.-G. On the Disentanglement and Robustness of Self-Supervised Speech Representations. In *2024 International Conference on Electronics, Information, and Communication (ICEIC)* 2024 Jan 28 (pp. 662-665). IEEE.
- **How** to use it?
  $\rightarrow$ Improvement of the speech re-synthesis framework
  Song Y,, Kim D, Kang H.-G, and Madhu N. Spectrum-Aware Neural Vocoder Based on Self-Supervised Learning for Speech Enhancement. In *2024 32nd European Conference on Signal Processing (EUSIPCO)* 2024 Aug 26 (pp. 16-20). IEEE.

Selection Criteria

- How robust are these models in the real world?
  $\rightarrow$ Interference robustness

- What is extracted by the pre-trained models?
  $\rightarrow$ Preserved information

Materials

- Pretrained SSL models
  1. HuBERT: predicting clustering labels of masked frames
  2. wavLM: HuBERT with data augmentation (additive noise)
  3. wav2vec 2.0: contrastive learning of the quantised representations
  4. TERA: predicting masked spectrogram
- Data
  - Interference robustness
    - Valentini (speech) + DEMAND (noise) + MIT IR Survey (RIRs)
  - Preserved information: TIMIT with human annotation

Materials

- Pretrained SSL models
  1. HuBERT: predicting clustering labels of masked frames
  2. wavLM: HuBERT with data augmentation (additive noise)
  3. wav2vec 2.0: contrastive learning of the quantised representations
  4. TERA: predicting masked spectrogram
- Data
  - Interference robustness
    - Valentini (speech) + DEMAND (noise) + MIT IR Survey (RIRs)
  - Preserved information: TIMIT with human annotation

Robustness Metrics

- Measure the distance between the embeddings of the distorted speech ($\mathbf{e_x}$) and the clean reference ($\mathbf{e_s}$)
  1. Normalised Mean-square Error (MSE) $\downarrow$

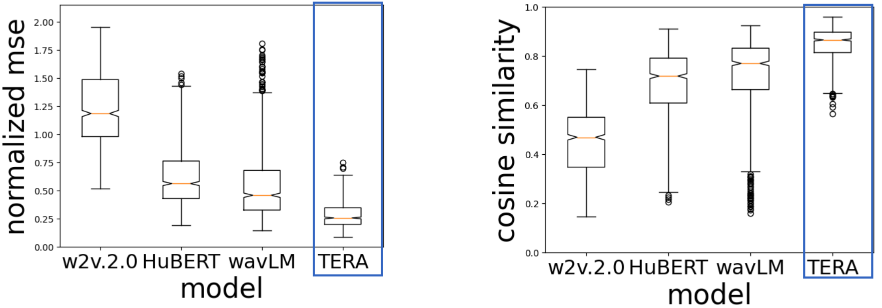
$$d(\mathbf{e_s}, \mathbf{e_x}) = \frac{1}{N}\left(\frac{\mathbf{e_s} - \mathbf{e_x}}{\sigma}\right)^T \cdot \left(\frac{\mathbf{e_s} - \mathbf{e_x}}{\sigma}\right)$$

  2. Cosine similarity (CS) $\uparrow$

$$c(\mathbf{e_s}, \mathbf{e_x}) = \frac{\mathbf{e_s}^T \mathbf{e_x}}{\|\mathbf{e_s}\| \|\mathbf{e_x}\|}$$

Robustness of SSL models

On the noisy and reverberant test set,



TERA shows highest robustness against interference.

Preserved Information

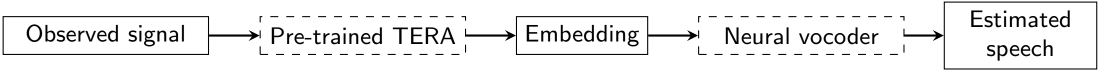
Logistic regression model (embeddings $\rightarrow$ labels) training accuracy

(Linearly separable)

Data source	**Target (total)**	**Accuracy (%)** HuBERT	TERA	wav2vec2.0	wavLM
sentence *sa1*	Phoneme (46)	93.2	**86.8**	89.1	92.7
	Word (12)	99.2	**94.5**	95.6	99.0
set *sx*	Sentence (330)	98.7	**73.8**	93.0	92.9
	Speaker (462)	90.0	**94.5**	94.7	53.0

- ▶ Contextual information (word prediction acc.) $>$ phonetic information (phoneme prediction acc.)
- $+$ Speaker information preservation
- $-$ Long-term contextual information

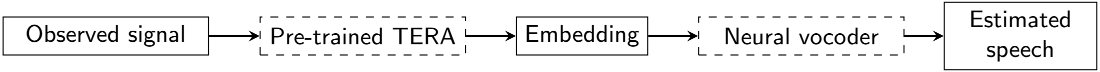
Leveraging Self-Supervised Learning for Speech Enhancement



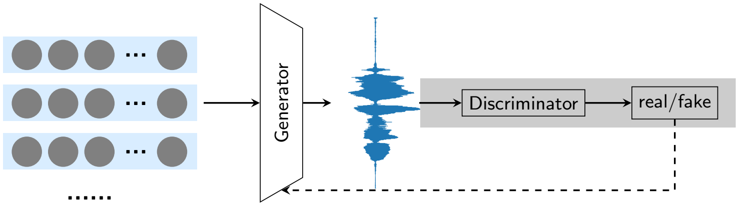
Baseline: denoising vocoder[1]

[1] Irvin B, Stamenovic M, Kegler M, Yang LC. Self-supervised learning for speech enhancement through synthesis. In ICASSP 2023.

Leveraging Self-Supervised Learning for Speech Enhancement



Baseline: denoising vocoder[1]

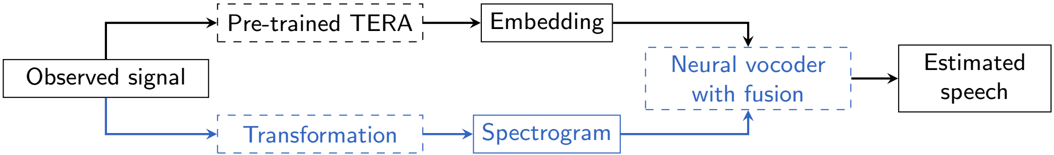


HiFi-GAN[2]-based neural vocoder

[1] Irvin B, Stamenovic M, Kegler M, Yang LC. Self-supervised learning for speech enhancement through synthesis. In ICASSP 2023.

[2] Kong J, Kim J, Bae J. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. In NeurIPS 2020.

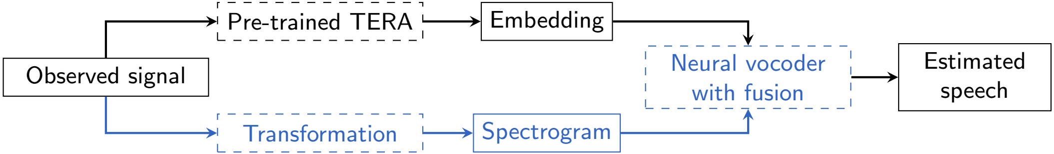
Proposed Framework



Proposed: **spectrum-aware** denoising vocoder

- Pre-trained SSL model == TERA
- Introduction noisy spectrogram for additional information
- Components to be optimised
  - What transformation/spectrogram?
  - How to fuse the two features?

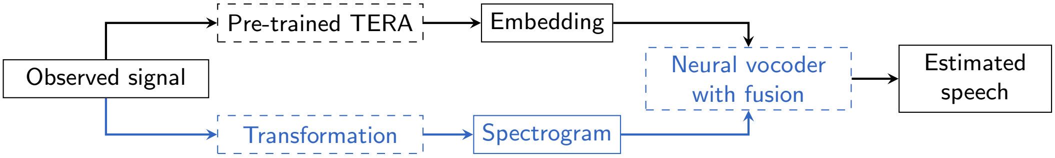
Proposed Framework



Proposed: **spectrum-aware** denoising vocoder

- Pre-trained SSL model == TERA
- Introduction noisy spectrogram for additional information
- Components to be optimised
  - What transformation/spectrogram?
  - How to fuse the two features?

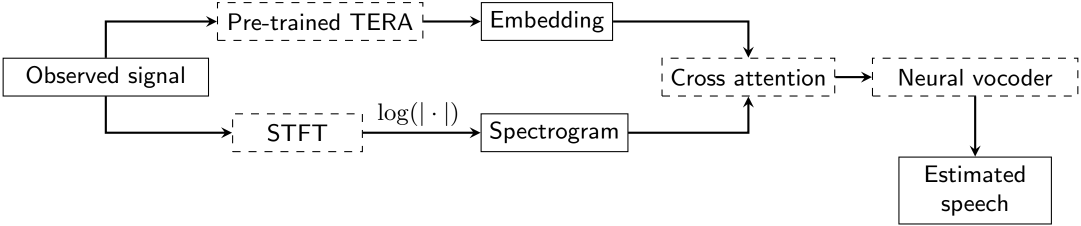
Proposed Framework



Proposed: **spectrum-aware** denoising vocoder

- Pre-trained SSL model == TERA
- Introduction noisy spectrogram for additional information
- Components to be optimised
  - What transformation/spectrogram?
  - How to fuse the two features?

Proposed System



Data and Evaluation Metrics

Training Dataset:

- DNS 2021 challenge dataset (RIR: SLR26 and SLR28)
- SNR $\in [-5, 20]$ dB
- T60s $\in [0.3, 1.3]$ sec

Test Dataset:

- CSTR VCTK dataset + NOISEX92 + MIT RIR
- SNR $\in \{-7, 0, 5, 10, 15\}$ dB
- T60s $\in [0.3, 1.3]$ sec

Evaluation metrics:

1. STOI
2. Speaker embedding (ECAPA-TDNN[3]) cosine similarity
3. DNSMOS
4. NISQAv2

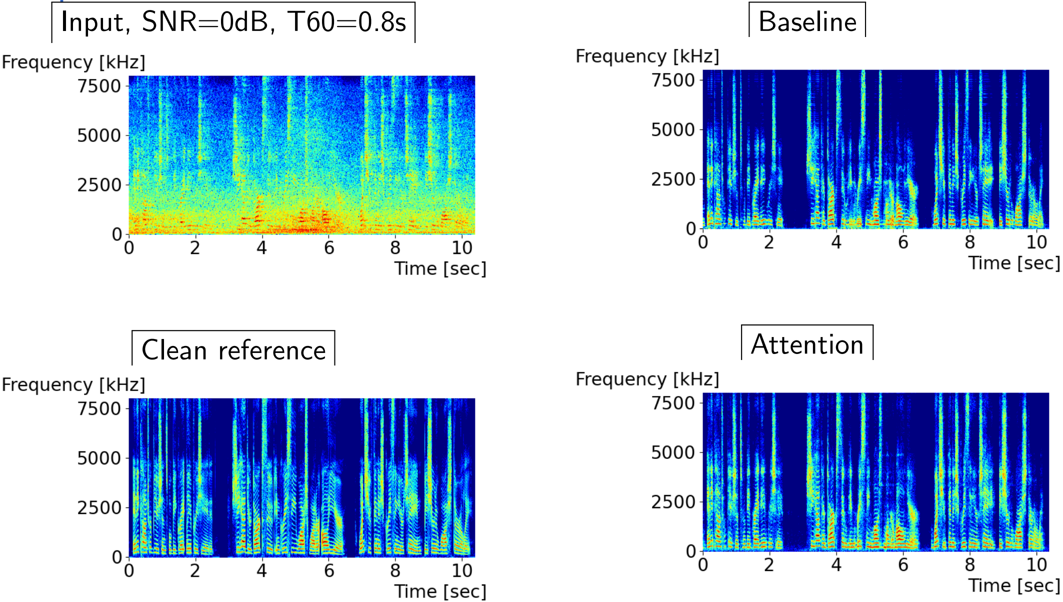
[3] Desplanques B, Thienpondt J, Demuynck K. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In Interspeech 2020.

Evaluation Results

- Improvement in the naturalness of synthesised audio

Model Description	STOI	DNSMOS			NISQAv2 (MOS)	Spk. embed. CS
		OVRL	**SIG**	**BAK**		
Distorted signals	0.770	1.814	2.458	2.005	1.659	0.600
Denoising vocoder (baseline)	0.808	**3.086**	3.379	**4.043**	3.097	0.551
Spectrum-aware vocoder (proposed)	0.811	3.054	**3.405**	3.892	3.691	0.529
+ Magnitude spectrum	**0.819**	2.999	3.374	3.835	3.566	0.552
+ Additive-fusion	0.814	3.017	3.306	3.997	**3.768**	**0.584**
Clean signals	1	3.668	3.951	4.209	4.550	1

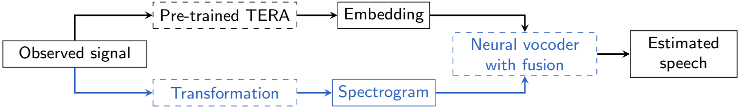
Samples


Input, SNR=0dB, T60=0.8s
Frequency [kHz]
7500
5000
2500
0
0
2
4
6
8
10
Time [sec]
Baseline
Frequency [kHz]
7500
5000
2500
0
0
2
4
6
8
10
Time [sec]
Clean reference
Frequency [kHz]
7500
5000
2500
0
0
2
4
6
8
10
Time [sec]
Attention
Frequency [kHz]
7500
5000
2500
0
0
2
4
6
8
10
Time [sec]

Conclusions

- TERA shows high robustness against interference
- Introduction of *noisy spectrum* improves the synthesis quality of the SSL-based neural vocoder
- Effective conditioning: *cross attention* block conditions noisy spectra by SSL embeddings

Spectrum-Aware Neural Vocoder Based on Self-Supervised Learning for Speech Enhancement

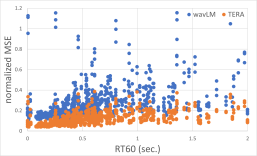


More samples:

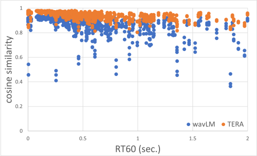
Robustness: Analysis by SNRs and T60s

Table 1: Analysis of pretrained SSL model according to the SNR (dB) of noise distortion.

Model		-7	0	5	10	15
wavLM	MSE ↓	0.967	0.593	0.430	0.352	0.295
	CS ↑	0.521	0.701	0.778	0.816	0.847
TERA	MSE ↓	0.452	0.334	0.265	0.212	0.166
	CS ↑	0.746	0.818	0.859	0.889	0.915



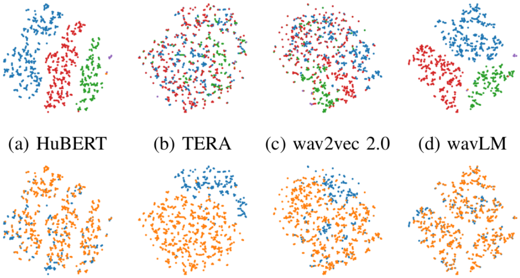
(a) Standardized MSE



(b) Cosine similarity

Figure 4: Analysis of pretrained SSL model according to the various RT60 (sec).

Preserved Information



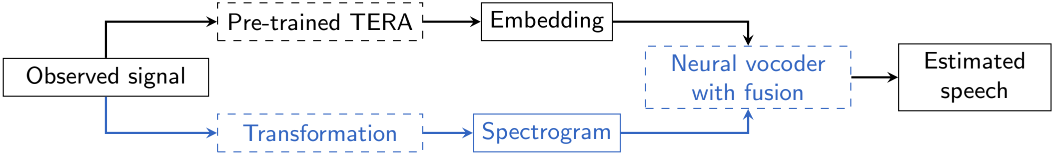
(a) HuBERT (b) TERA (c) wav2vec 2.0 (d) wavLM

(e) HuBERT (f) TERA (g) wav2vec 2.0 (h) wavLM

Figure 5: The t-SNE plot of embedding distributions of all ‘ao’ sounds in ’sa1’ from TIMIT training set, labeled by the words to which the phoneme belongs, or the speaker genders.

- How linearly-separable are the embeddings for one label?
  → **Training accuracy of multinomial logistic regression**

Research Questions



- **Q1**: How does the spectrum representation affect the system performance?
- **Q2**: How important is each hidden state of TERA?
- **Q3**: Which fusion method performs better (addition, cross-attention, FiLM)?
- **Q4**: For cross-attention and FiLM, which feature is best suited as the conditioning?

Evaluation Results: Preferred System

- **Q1**: How does the **spectrum representation** affect the system performance?
- Improvement in the naturalness of the synthetic audio
- Log-spectrogram works better

No.	Model Description	STOI	DNSMOS			NISQAv2					Spk. embed. CS
			OVRL	SIG	BAK	MOS	NOIS.	DIS.	COL.	LOUD.	
-	Distorted signals	0.770	1.814	2.458	2.005	1.659	1.697	3.073	2.328	2.505	0.600
-	Denoising vocoder (baseline)	0.808	**3.086**	3.379	**4.043**	3.097	**3.601**	3.325	2.953	3.726	0.551
1	Proposed reference	0.811	3.054	**3.405**	3.892	**3.691**	3.526	**3.998**	**3.494**	**3.992**	0.529
2	Magnitude spectrum feature	**0.819**	2.999	3.374	3.835	3.566	3.406	3.997	3.433	3.906	**0.552**
8	Clean embedding	0.949	3.121	3.429	3.977	4.161	3.876	4.330	3.979	4.238	0.899
-	Clean signals	1	3.668	3.951	4.209	4.550	4.251	4.596	4.301	4.476	-

Ablation Study: Embeddings as Input

- **Q4**: For cross-attention and FiLM, which feature is best suited as the conditioning?
- For attention fusion: spectrogram conditioned by embeddings

No.	Model Description	STOI	DNSMOS			NISQAv2					Spk. embed. CS
			OVRL	SIG	BAK	MOS	NOIS.	DIS.	COL.	LOUD.	
-	Distorted signals	0.770	1.814	2.458	2.005	1.659	1.697	3.073	2.328	2.505	0.600
-	Denoising vocoder (baseline)	0.808	**3.086**	3.379	**4.043**	3.097	3.601	3.325	2.953	3.726	**0.551**
1	Proposed reference	**0.811**	3.054	**3.405**	3.892	**3.691**	3.526	**3.998**	**3.494**	**3.992**	0.529
6	Attention conditioned by spectrum	**0.811**	2.966	3.261	**3.968**	3.522	**3.602**	3.862	3.276	3.876	0.524
8	Clean embedding	0.949	3.121	3.429	3.977	4.161	3.876	4.330	3.979	4.238	0.899
-	Clean signals	1	3.668	3.951	4.209	4.550	4.251	4.596	4.301	4.476	-

Ablation Study: Hidden Layers

- **Q2**: How important is each hidden state of TERA?
- The last layer contributes the most.
- Beneficial to include all layers

Combination weights for TERA hidden state layers

Variant	**Layer1**	**Layer2**	**Layer3**	**Layer4**
1	-0.002	-0.011	0.036	0.098
2	0.003	0.016	-0.105	-0.248
4	0.017	0.025	-0.479	-1.229
5	0.015	0.116	-0.586	-1.495
8	-0.068	-0.048	-0.041	0.115

No.	Model Description	**STOI**	**DNSMOS**			**NISQAv2**					**Spk. embed. CS**
			OVRL	**SIG**	**BAK**	**MOS**	**NOIS.**	**DIS.**	**COL.**	**LOUD.**	
-	Distorted signals	0.770	1.814	2.458	2.005	1.659	1.697	3.073	2.328	2.505	0.600
-	Denoising vocoder (baseline)	0.808	3.086	3.379	**4.043**	3.097	**3.601**	3.325	2.953	3.726	**0.551**
1	Proposed reference	**0.811**	**3.054**	**3.405**	**3.892**	**3.691**	3.526	**3.998**	**3.494**	**3.992**	0.529
3	TERA - last hidden state	0.798	2.955	3.303	3.876	3.605	3.609	3.955	3.351	3.870	0.524
8	Clean embedding	0.949	3.121	3.429	3.977	4.161	3.876	4.330	3.979	4.238	0.899
-	Clean signals	1	3.668	3.951	4.209	4.550	4.251	4.596	4.301	4.476	-

Ablation Study: Fusion Methods

- **Q3**: Which **fusion method** performs better?
- Attention/addition both boost the objective scores

No.	Model Description	STOI	DNSMOS			NISQAv2					Spk. embed. CS
			OVRL	SIG	BAK	MOS	NOIS.	DIS.	COL.	LOUD.	
-	Distorted signals	0.770	1.814	2.458	2.005	1.659	1.697	3.073	2.328	2.505	0.600
-	Denoising vocoder (baseline)	0.808	**3.086**	3.379	**4.043**	3.097	3.601	3.325	2.953	3.726	0.551
1	Proposed reference	0.811	3.054	**3.405**	3.892	3.691	3.526	3.998	**3.494**	**3.992**	0.529
4	Additive-fusion	**0.814**	3.017	3.306	3.997	**3.768**	**3.932**	**4.032**	3.465	3.984	**0.584**
5	FiLM	0.739	2.696	3.005	3.827	2.828	3.409	3.408	2.614	3.434	0.387
8	Clean embedding	0.949	3.121	3.429	3.977	4.161	3.876	4.330	3.979	4.238	0.899
-	Clean signals	1	3.668	3.951	4.209	4.550	4.251	4.596	4.301	4.476	-