Temporal-hierarchical features from noise-robust speech foundation models for non-intrusive intelligibility prediction

Santiago Cuervo, Ricard Marxer

Université de Toulon, Aix-Marseille Université, CNRS, LIS

The 4th Clarity Workshop on Machine Learning Challenges for Hearing Aids (Clarity-2023)

Outline

(1) Motivation: getting the non-intrusive setup closer to the intrusive setup

Speech foundation models.

Emergent signal-noise disentanglement in noise-robust foundation models.

(2) Our model: extracting features across time and layers

Time-wise and layer-wise transformers with binaural blocks.

(3) Preliminary experiments and results

Peek into the benefits of noise-robust backbones, multi-scale modeling and binaural blocks.

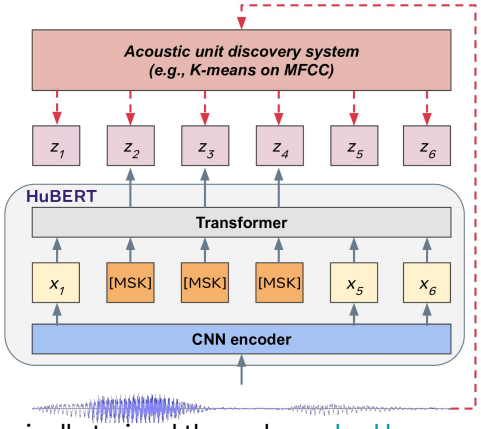
Our submission E011.

Speech foundation models

Foundation model def.

Very **large** deep learning models **trained on large diverse datasets** that **can be applied to many unseen tasks without (or with little) extra training**.

Very large transformers trained through self-supervision or weak supervision on very large speech corpuses.



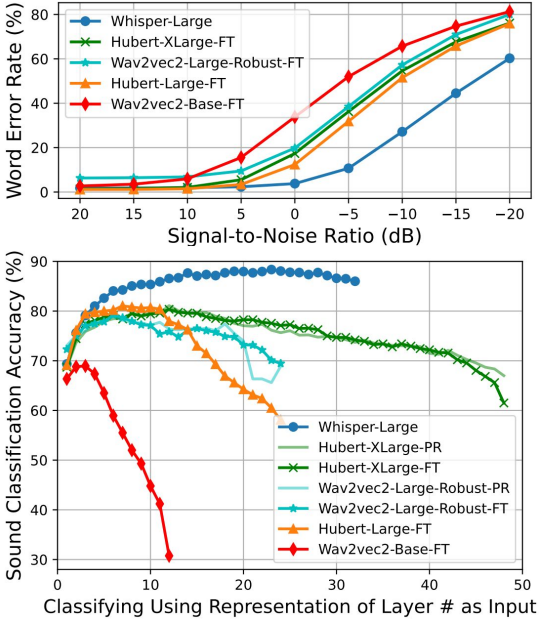
Typically trained through masked language modeling, as HuBERT here depicted.

	Model size* (# parameters)	Training data (# hours of speech)	Robust
Wav2vec 2.0 (2020)	316M	60K	No
HuBERT (2021)	1B	60K	No
WavLM (2022)	316M	94K	Yes
Whisper (2022)	1.5B	680k	Yes

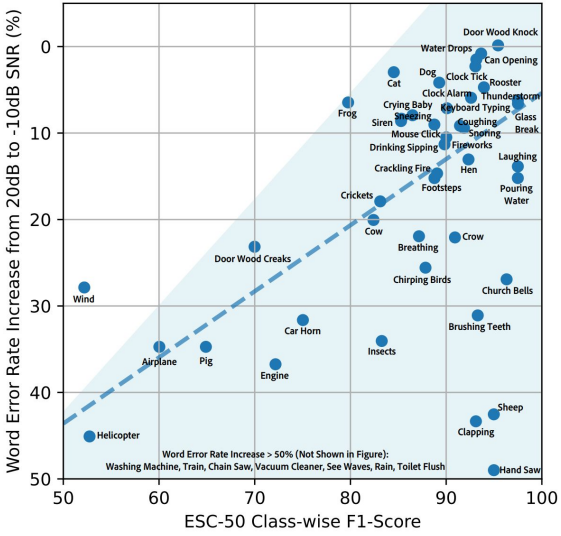
* In all cases we consider only the `largest` models

Motivation: Whisper exhibits some degree of noise-signal disentanglement

(1) Noise-robust models are strong general audio-event taggers



(2) ASR performance improves for background noises for which the model has good classification accuracy



All plots from Gong et al. (2023)

Gong et al. conclusion: *Whisper's ASR robustness stems NOT from being noise-invariant, but from internally conditioning predictions on the noise-type.*

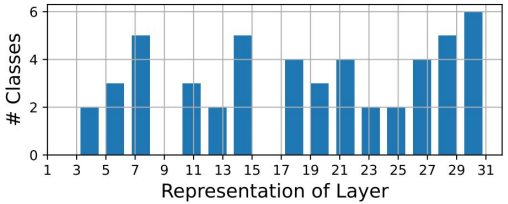
This suggests that there is some disentanglement of signal and noise in its representations.

Hypothesis 1: *this is a general feature of robust speech foundation models.*

Hypothesis 2: *such representations are better for non-intrusive intelligibility prediction as they bring it closer to the intrusive setup, where noise-signal disentanglement is a given.*

Model: Time-wise and layer-wise transformers with binaural blocks

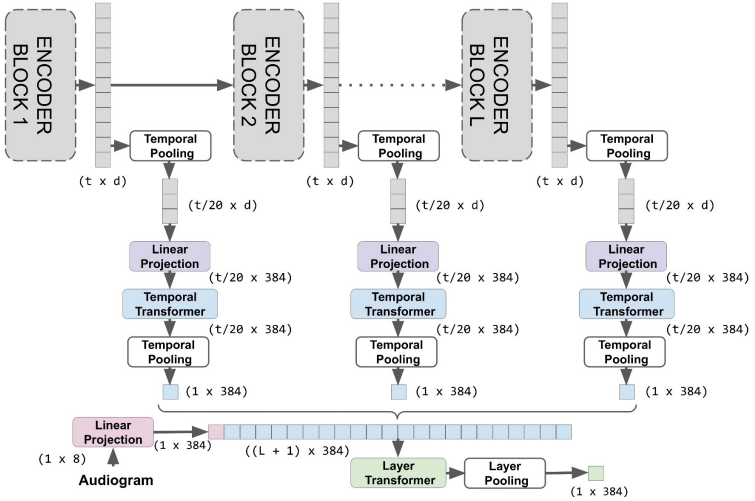
(1) Background noise information is distributed across layers.



Number of noise classes (out of 50) in which the corresponding Whisper's encoder layer is the best predictor. Taken from Gong et al. (2023).

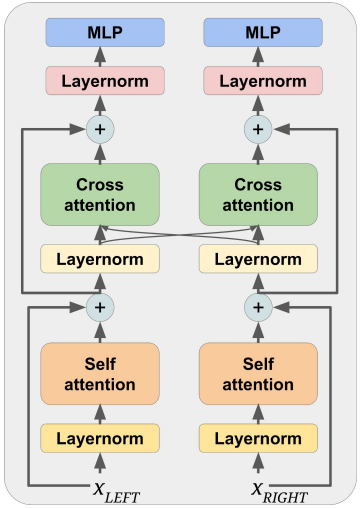
(2) We use a transformer not only across time, but also across layers to adaptively extract temporal and multi-layer features (based on the `T1-Tr` model from Gong et al. (2023)).

We inject audiogram information as an additional layer.



(3) We use binaural transformer blocks

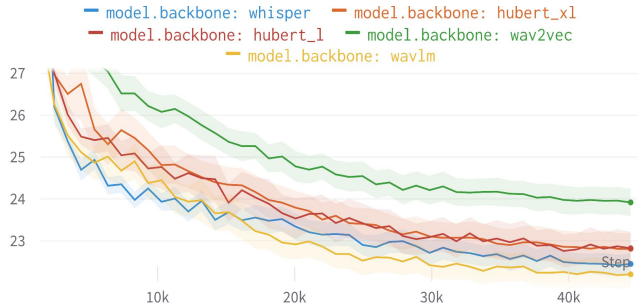
It allows binaural interactions by using cross-attention between audio channels.



Preliminary experiments and results

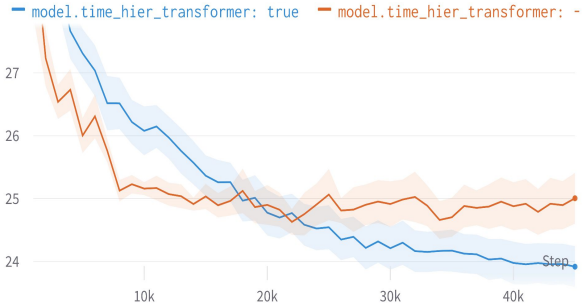
Mean validation RMSE across training with a confidence interval of 90%

(1) Noise-robustness matters



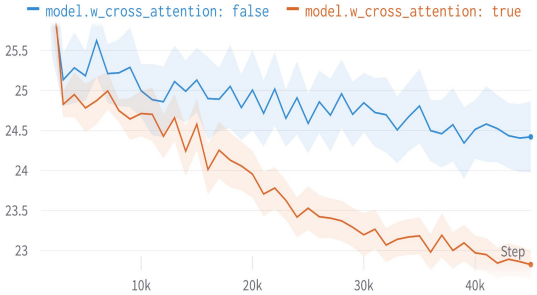
6 independent runs for each model and each train split (total 90 runs)

(2) Hierarchical modeling matters



6 independent runs using wav2vec 2.0 as backbone on `train.1` and using or not temporal-hierarchical transformers (total 12 runs).

(3) Binaural cross-attention matters



3 independent runs using WavLM and Whisper as backbones on `train.1` and using or not binaural cross-attention (total 6 runs).

Our submission E011

For each backbone (out of Whisper and WavLM), and each split, we took the best model according to RMSE on the validation set**. We made an ensemble by fitting a weighted average of the predictions from each backbone to minimize validation error:

```
0.41 * whisper_preds + 0.59 * wavlm_preds
```

Results on the test set

	RMSE	NCC
test.1	27.81	0.73
test.2	24.56	0.79
test.3	22.66	0.83
Average	25.12	0.78

Table 1: *Results on the validation set for each train split in terms of RMSE and normalized cross correlation (NCC).*

Model	Split	RMSE ↓	NCC ↑
Baseline	train.1	29.819	0.663
	train.2	30.060	0.677
	train.3	30.350	0.665
Whisper	train.1	24.249 ± 1.106	0.812 ± 0.017
	train.2	$\mathbf{23.164 \pm 0.144}$	0.823 ± 0.010
	train.3	22.457 ± 0.496	0.838 ± 0.011
WavLM	train.1	$\mathbf{23.796 \pm 2.447}$	$\mathbf{0.818 \pm 0.037}$
	train.2	23.466 ± 0.281	$\mathbf{0.827 \pm 0.004}$
	train.3	$\mathbf{21.588 \pm 0.648}$	$\mathbf{0.848 \pm 0.007}$
Ensemble	train.1	**21.069**	**0.857**
	train.2	**20.836**	**0.866**
	train.3	**19.284**	**0.877**

** For **train.1** we used as validation set the samples from the CEC2 challenge from **train.2**. Similarly, for **train.2** we used the ones from **train.3**, and for **train.3** the ones from **train.1**.

Thank you. Questions?

Email: santiago.cuervo@lis-lab.fr