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, France

santiago.cuervo@lis-lab.fr, ricard.marxer@lis-lab.fr

Abstract

We present a method for non-intrusive speech intelligibility prediction leveraging temporal and hierarchical features from noise-robust speech foundation models. First, a temporal transformer is applied along the time axis to sequences of representations obtained at each layer of the foundation model. The resultant features are then averaged along the time axis, yielding one embedding per layer. Next, a layer-wise transformer is applied to the set of layers' representations to derive multi-level features. The resulting sequence is averaged along the layer axis. This pipeline is applied to each channel of the binaural signal, obtaining one embedding per channel. To account for non-linear binaural interactions, each transformer block has a cross-attention layer between the two channels. Finally, the embeddings from both channels are averaged, yielding the final representation used to predict the percentage of correctly recognized words in the utterance through a linear projection. Predictions are conditioned on the listener's audiogram, treated as an additional layer before the layer-wise transformer. We performed experiments using the CPC2 dataset with Whisper and WavLM as backbones. Our results show significant improvements over the baseline model.

Index Terms: non-intrusive intelligibility prediction, speech foundation models, hierarchical model

1. Introduction

Recently it was shown that Whisper [1], an ASR model trained with a 680k hour labeled speech corpus recorded in diverse conditions, captures in its representations rich linearly-accessible information from the background noise present in speech utterances [2]. The authors suggested that recognizing speech conditioned on the noise type is the mechanism behind Whisper's widely noted noise-robust ASR performance.

Motivated by this finding, we hypothesized that by leveraging such representations, which effectively exhibit a disentanglement of signal and noise, we could build models with improved performance on non-intrusive intelligibility prediction. This would be achieved by bringing the non-intrusive setup closer to the intrusive setup, in which separation of signal and noise is assumed to be given.

In this work we propose a model inspired by two key insights from [2]:

a) Noise-robust foundation models seem to exhibit disentanglement of signal and noise in its representations.

b) At least in Whisper, noise information is distributed across multiple layers of the model.

Based on a) we used Whisper and WavLM [3] to extract speech features. WavLM was not among the models studied in [2], however we included it because its training involves a denoising task with diverse noise sources, which we speculate could also promote noise-signal disentanglement. Its strong performance on diarization and separation tasks [3] could be evidence of it. To tackle b) we used a model with an architecture similar to the `Tl-Tr` model proposed in [2], in which transformers [4] across time and layers are used for feature extraction.

2. Model

Our model is illustrated in Figure 1. Each audio channel from the binaural signal is processed by a noise-robust foundation model with L layers, yielding a sequence of representations at each layer for each channel. Sequences are shortened along the time axis by applying average pooling with a kernel width of 20 and stride of 20. Next, embeddings are linearly projected to a 384-dimensional space. Shortening and dimensionality reduction are performed to reduce computational costs. After temporal pooling, a single-head temporal transformer with internal dimension 384 is applied at each layer, producing sequences of contextual features. Global average pooling is then applied along the time axis to each layer's contextual sequence, resulting in a single 384-dimensional embedding per layer. Layer representations are concatenated, forming an $L \times 384$-dimensional tensor. At this point, we inject the listener's information by linearly projecting the audiogram to a 384-dimensional space and concatenating it with the layers' representations across the layer axis. The sequence is used as input for a single-head layer transformer, yielding an $(L+1) \times 384$-dimensional tensor of multi-level features. The sequence is compressed again by global average pooling, resulting in a final 384-dimensional representation per channel. The two channels' representations are averaged and linearly projected to produce the correctness prediction. We use a sigmoid layer at the output to bound the predictions between 0% and 100%.

To allow for non-linear binaural interactions, each transformer block has a cross-attention layer between the output of its self-attention layer and the output of the self-attention layer at the same level in the other channel's transformer block (Figure 1, right).

The parameters of the foundation model are frozen during training. All the trainable parameters are shared between audio channels and the parameters of the temporal transformer and the linear projection for dimensionality reduction are also shared across layers.

3. Experimental setup

We performed experiments with two foundation models: Whisper LARGE and WavLM LARGE. For each model we ran 3 independent experiments with different random seeds on each of the training data sets (`train.1`, `train.2` and `train.3`). For `train.1` we used as validation set the samples from the

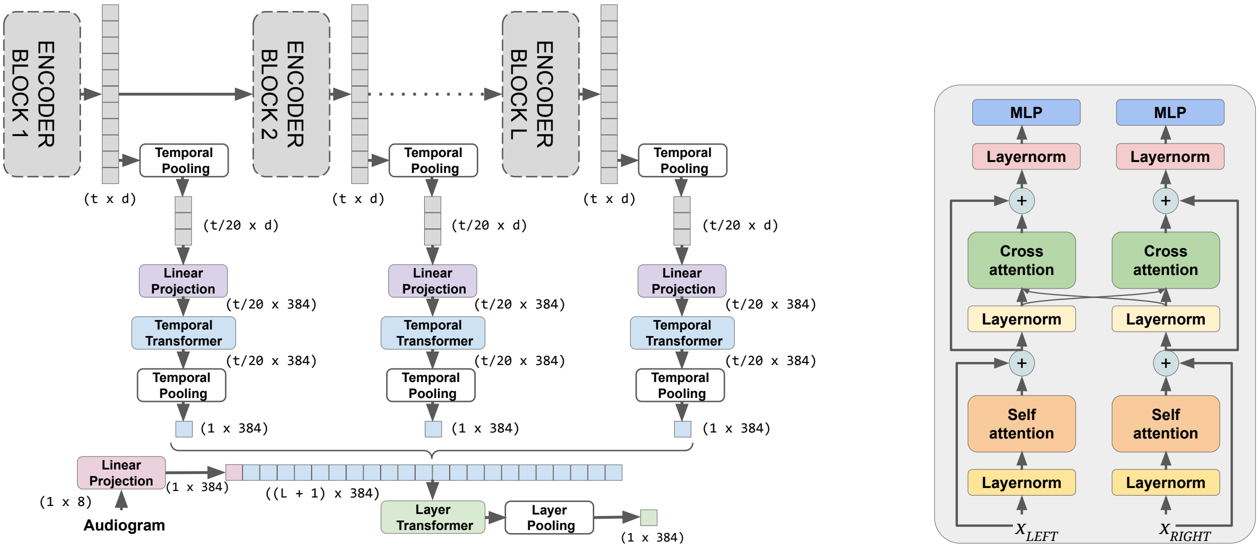

Figure 1: *Model architecture. With the exception of the blocks in grey, blocks with the same color indicate shared parameters.* ***Left:*** *Pipeline applied to each channel of the binaural signal to obtain one representation per channel. The representations from both channels are then averaged and linearly projected to predict the intelligibility score.* ***Right:*** *Transformer binaural block used in the temporal and layer transformers. The cross-attention layer enables modeling of non-linear binaural interactions.*

CEC2 challenge from `train.2`. Similarly, for `train.2` we used the CEC2 samples from `train.3`, and for `train.3` the ones from `train.1`.

All models were trained to minimize a Huber loss for 80,000 steps using the Adam optimizer [5] with a learning rate of 3e−5, $\beta_1 = 0.9$, $\beta_2 = 0.98$, and a batch size of 160. We used a cosine learning rate schedule with a linear warm-up of 2000 steps. To all transformer layers we apply dropout [6] with $p = 0.1$. A training run takes roughly 9.3 hours on a single node with an NVIDIA A100-80 GB GPU and it requires about 18.4 GB of GPU memory when using Whisper features (inner dimension $d = 1280$), and 14.4 GB when using WavLM features (inner dimension $d = 1024$).

4. Results and analysis

Table 1 shows the results on the validation set for each of the training splits. We compare our results with the CPC2 baseline, a logistic regression model that maps HASPI [7] scores onto correctness values. Note that the results given for the baseline are from training it on what would normally be the validation set. For example, the results reported for `train.1` are for a baseline system trained on the CEC2 samples from `train.2`, which is the validation set for `train.1` on the other models. Therefore, the out-of-training scores for the baseline are likely worse, and the ones displayed are likely overly optimistic. We also report results for an ensemble model in which the predictions from the best-performing WavLM and Whisper based models are combined using a weighted average to compute the final prediction. The weights are optimized on each training set.

Overall, the results show a significant improvement using our model compared to the baseline. Models with the WavLM backbone outperform those with Whisper in most cases, providing further evidence that WavLM has comparable or better signal-noise disentanglement capabilities compared to Whisper. However, it should be noted that models with WavLM exhibit higher variance in performance. The ensemble shows the best performance, possibly indicating that Whisper's and WavLM's representations differ in a way that allows them to compensate for each other's biases to some degree.

5. Conclusion

We proposed a model based on extracting temporal and hierarchical features from noise-robust speech foundation models exhibiting signal-noise disentanglement. Results show that our model consistently and significantly outperforms the baseline. Furthermore, WavLM outperforms Whisper as a backbone, suggesting it may have better signal-noise disentanglement. This should encourage further research into using WavLM as backbone for tasks like audio event tagging, where currently Whisper is state-of-the-art. An ensemble of models using both backbones performed best, and was used for our final `E011` submission. Preliminary studies suggest the proposed multi-level feature extraction and binaural cross-attention meaningfully impact performance. We leave ablation studies and detailed analyses and benchmarking for future work.

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`	**23.164 ± 0.144**	0.823 ± 0.010
	`train.3`	22.457 ± 0.496	0.838 ± 0.011
WavLM	`train.1`	**23.796 ± 2.447**	**0.818 ± 0.037**
	`train.2`	23.466 ± 0.281	**0.827 ± 0.004**
	`train.3`	**21.588 ± 0.648**	**0.848 ± 0.007**
Ensemble	`train.1`	**21.069**	**0.857**
	`train.2`	**20.836**	**0.866**
	`train.3`	**19.284**	**0.877**

6. Acknowledgements

We are grateful to the French National Research Agency for their support through the ANR-20-CE23-0012-01 (MIM) grant.

7. References

[1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022.

[2] Y. Gong, S. Khurana, L. Karlinsky, and J. Glass, "Whisper-at: Noise-robust automatic speech recognizers are also strong general audio event taggers," 2023.

[3] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, July 2022. [Online]. Available: https://www.microsoft.com/en-us/research/publication/wavlm-large-scale-self-supervised-pre-training-for-full-stack-speech-processing/

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[5] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, San Diego, CA, USA, 2015.

[6] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014.

[7] J. M. Kates and K. H. Arehart, "The hearing-aid speech perception index (haspi)," *Speech Communication*, vol. 65, pp. 75–93, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639314000545