Interspeech2025 Tutorial: *Interpretability Techniques for Speech Models*

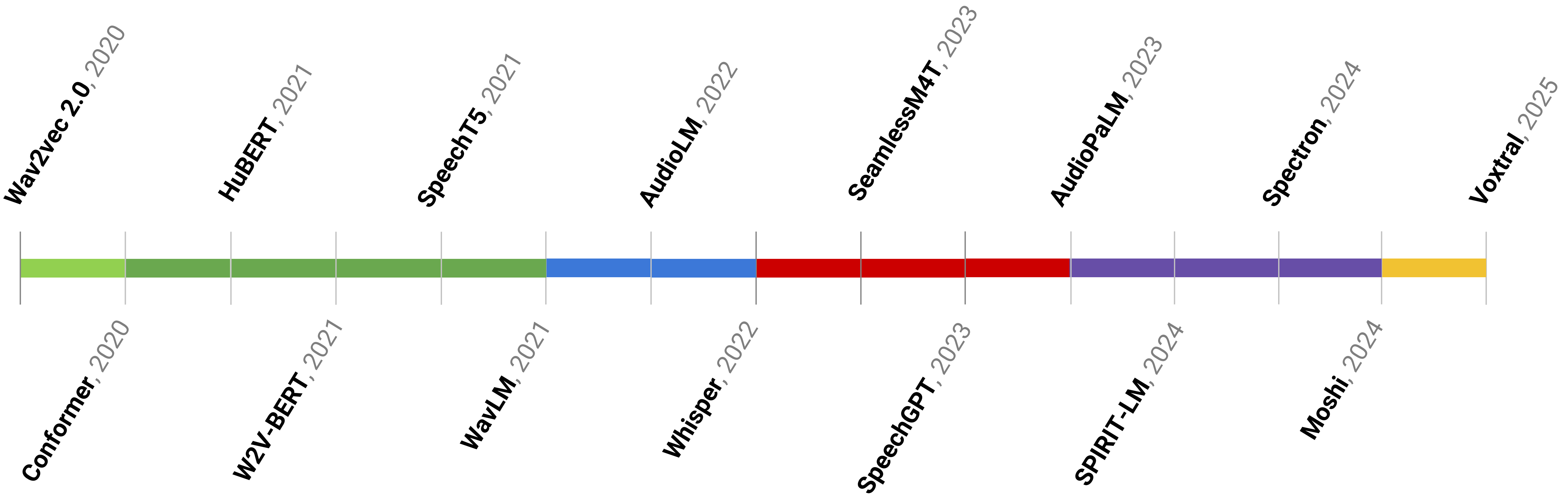
Context-Mixing in Speech Transformers

Hosein Mohebbi

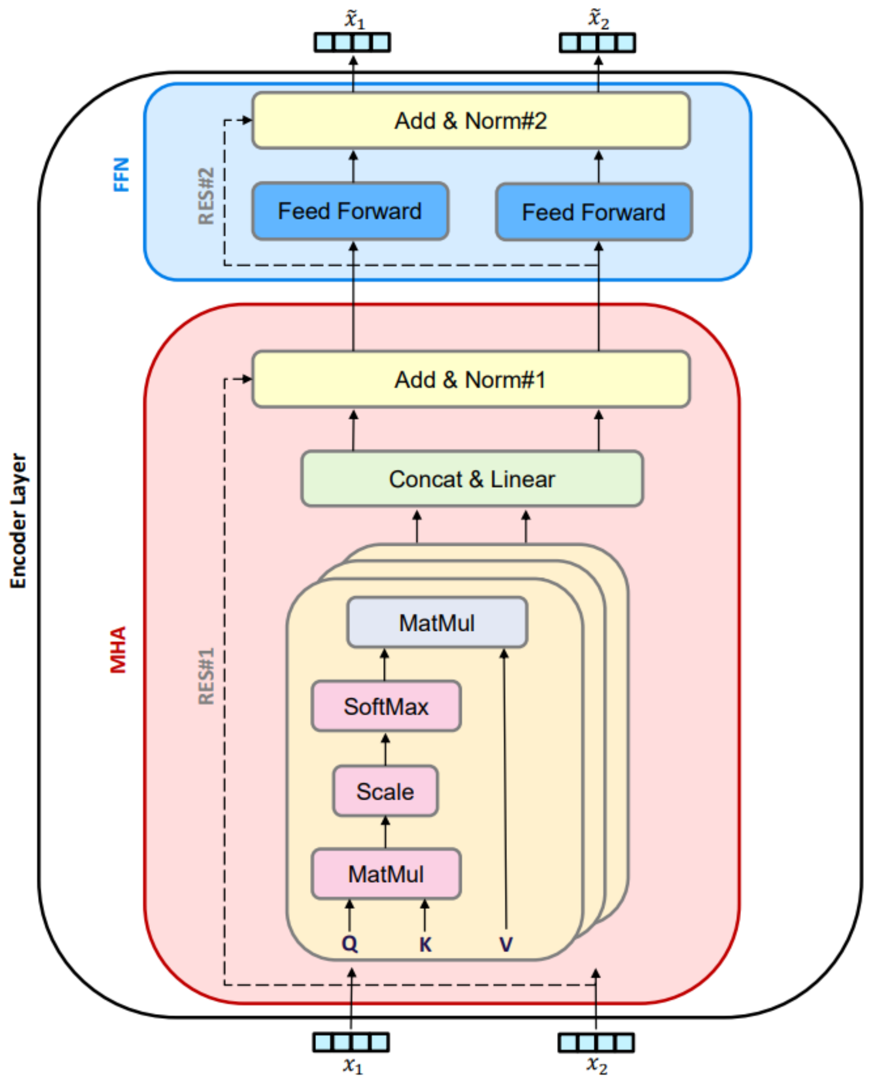
August 17, 2025

Rotterdam

Transformers for speech processing

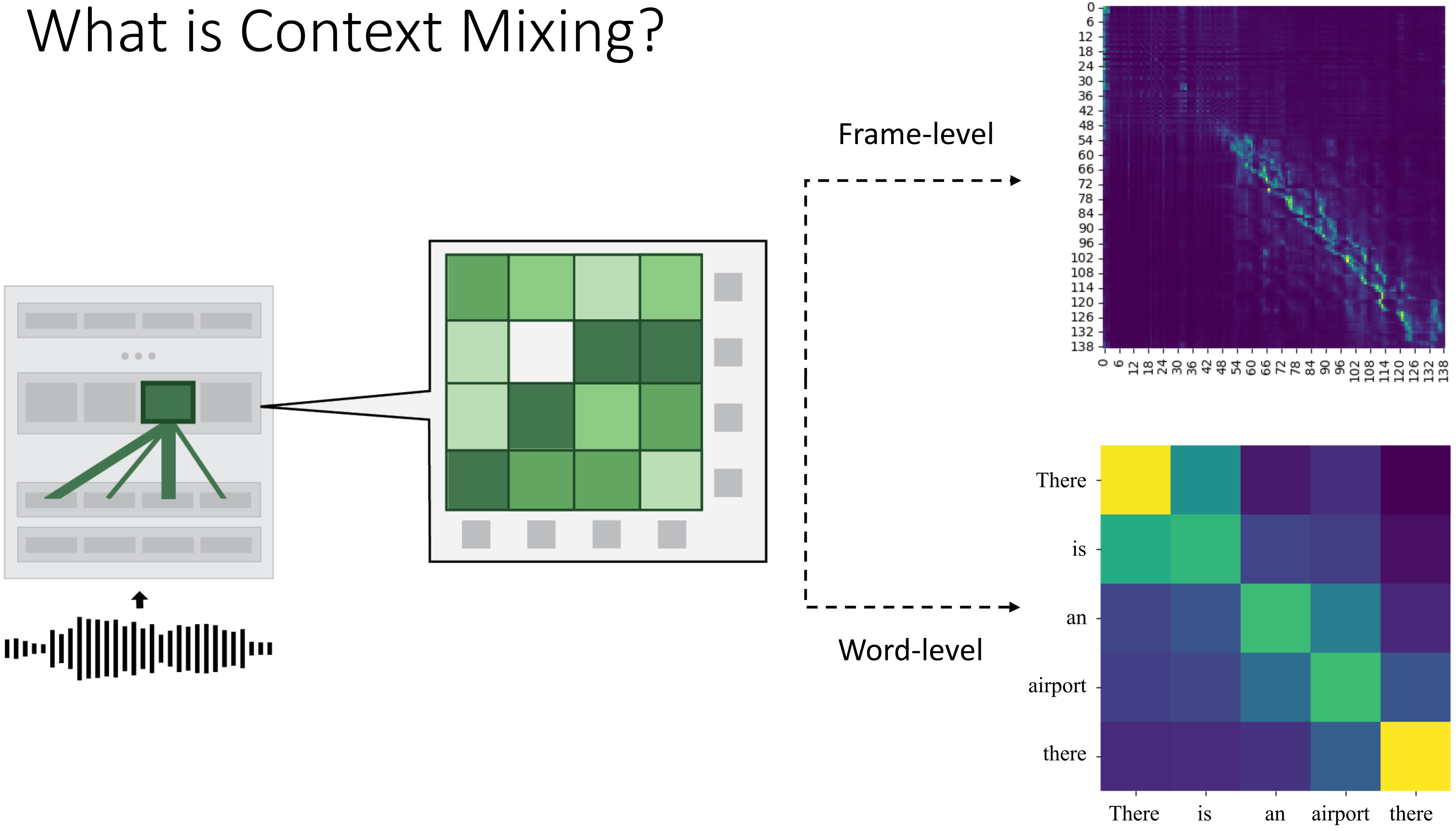


Transformer



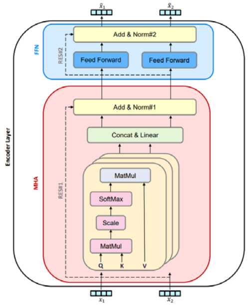
(Vaswani et al., 2017)

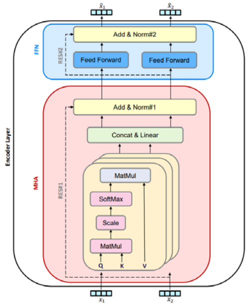
What is Context Mixing?

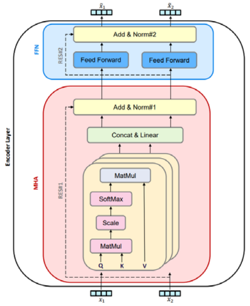


Measures of *Context-mixing*

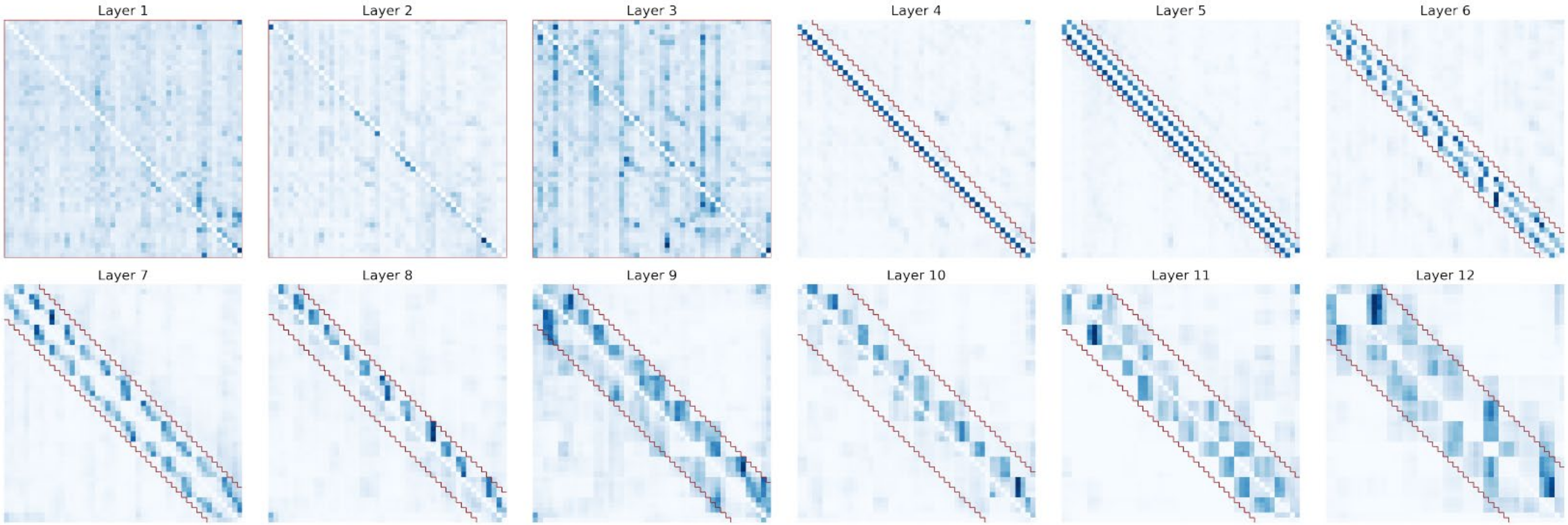
- Attention
- Attention-Norm
- Value Zeroing







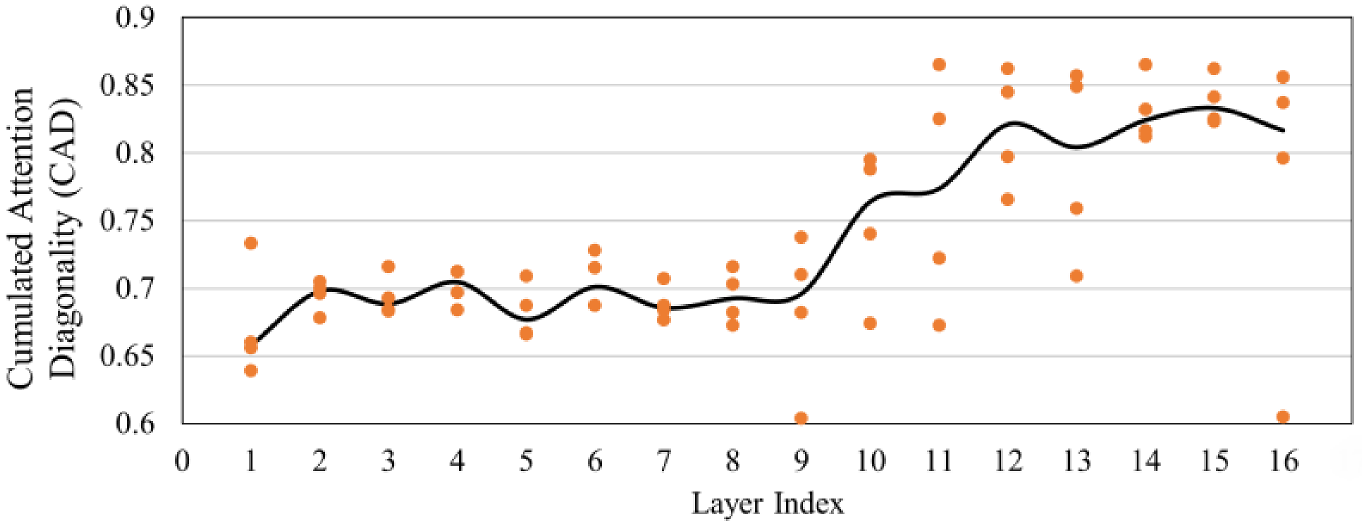
Diagonality

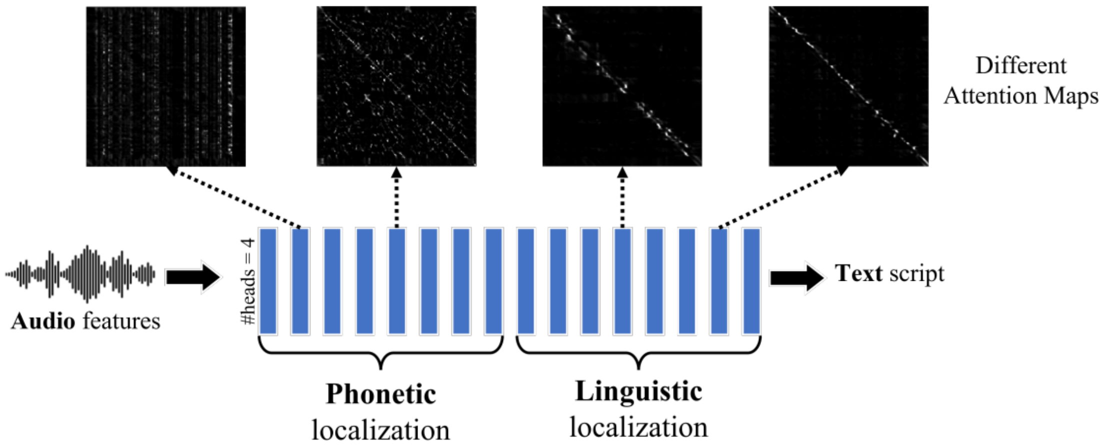


(Alastruey et al., 2022)

Two distinct roles

$$\mathrm{CAD}_h = \int_{r=0}^{1} \frac{1}{T} \sum_{i=1}^{T} \left(\sum_{\substack{j=\max(1,\\ i-r(T-1))}}^{\substack{\min(T,\\ i+r(T-1))}} A_h[i,j] \right) \mathrm{d}r = \int_{r=0}^{1} D(r)\mathrm{d}r$$





(Shim et al., 2022)

Diversity Loss

Diversity loss	dev	dev-other	test	test-other
$d^{\mathrm{A}}(m,n)$	6.37	6.02	6.31	6.10
$d^{\mathrm{Q}}(m,n)$	0.53	0.59	0.54	0.55
$d^{\mathrm{K}}(m,n)$	0.57	0.61	0.61	0.58
$d^{\mathrm{V}}(m,n)$	0.13	0.14	0.13	0.14

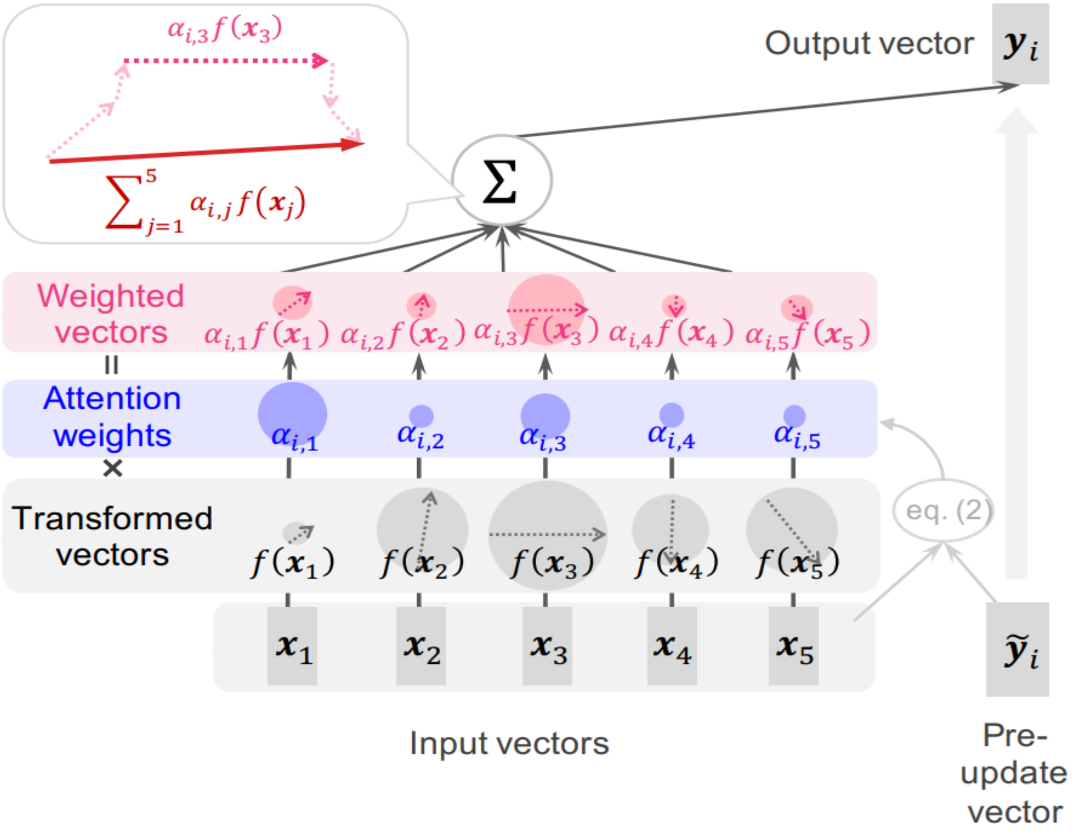
Attention → Highest

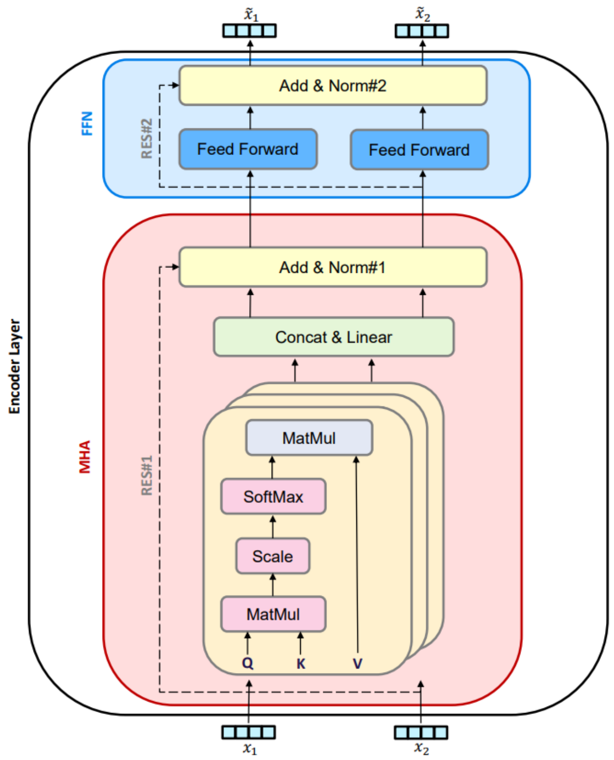
Value vectors → Lowest

Table 3: *Attention diversity losses summed over all layers of the Conformer acoustic encoder for the baseline full-context Librispeech model.*

(Audhkhasi et al., 2022)

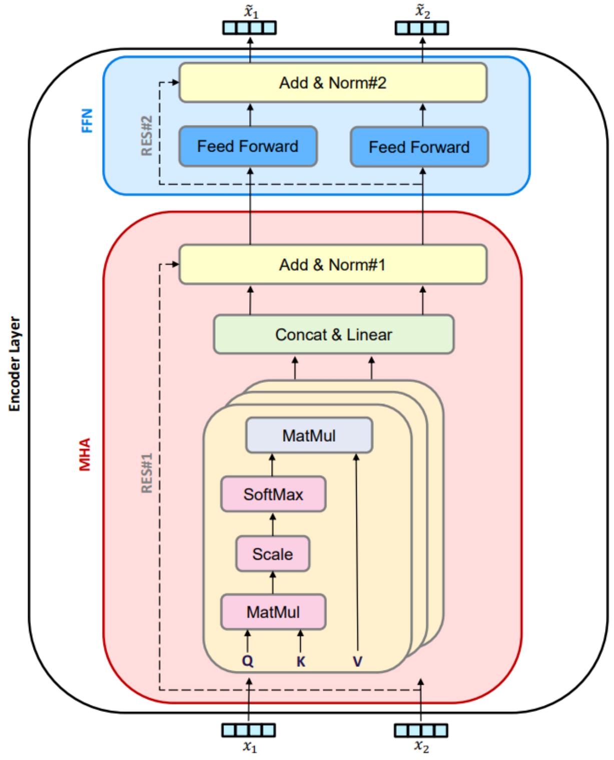
Attention-Norm





([Kobayashi et al., 2020](#))

Value Zeroing



(Mohebbi et al., 2023)

Value Zeroing

$$c_{i,j} = ?$$

(Mohebbi et al., 2023)

Value Zeroing

$$(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$$

$$\left.\begin{aligned} \boldsymbol{q}_i^h &= \boldsymbol{x}_i \boldsymbol{W}_{\boldsymbol{Q}}^{\boldsymbol{h}} + \boldsymbol{b}_Q^h \\ \boldsymbol{k}_i^h &= \boldsymbol{x}_i \boldsymbol{W}_{\boldsymbol{K}}^{\boldsymbol{h}} + \boldsymbol{b}_K^h \end{aligned}\right\} \quad \alpha_{i,j} = \underset{\boldsymbol{x}_j \in \mathcal{X}}{\mathrm{softmax}} \left(\frac{\boldsymbol{q}_i \boldsymbol{k}_j^\top}{\sqrt{d}} \right) \in \mathbb{R}$$

$$\boldsymbol{v}_i^h = \boldsymbol{x}_i \boldsymbol{W}_{\boldsymbol{V}}^{\boldsymbol{h}} + \boldsymbol{b}_V^h$$

$$\boldsymbol{z}_i^h = \sum_{j=1}^{n} \alpha_{i,j}^h \boldsymbol{v}_j^h$$

$$\boldsymbol{z}_i = \textsc{Concat}(\boldsymbol{z}_i^1, \ldots, \boldsymbol{z}_i^H)\boldsymbol{W}_O$$

$$\boldsymbol{z}_i = \mathrm{LN}_{\mathrm{MHA}}(\boldsymbol{z}_i + \boldsymbol{x}_i)$$

$$\tilde{\boldsymbol{x}}_i = \max(0, \boldsymbol{z}_i \boldsymbol{W}_1 + \boldsymbol{b}_1)\boldsymbol{W}_2 + \boldsymbol{b}_2$$

$$\tilde{\boldsymbol{x}}_i = \mathrm{LN}_{\mathrm{FFN}}(\tilde{\boldsymbol{x}}_i + \boldsymbol{z}_i)$$

$$C_{i,j} = ?$$

$$\boldsymbol{v}_j^h \leftarrow \mathbf{0}, \forall h \in H$$

$$C_{i,j} = \tilde{\boldsymbol{x}}_i^{\neg j} * \tilde{\boldsymbol{x}}_i$$

(Mohebbi et al., 2023)

Let’s Evaluate!

Evaluation

Controlled task: homophony in French

Target

livre (singular)

/livʁ/

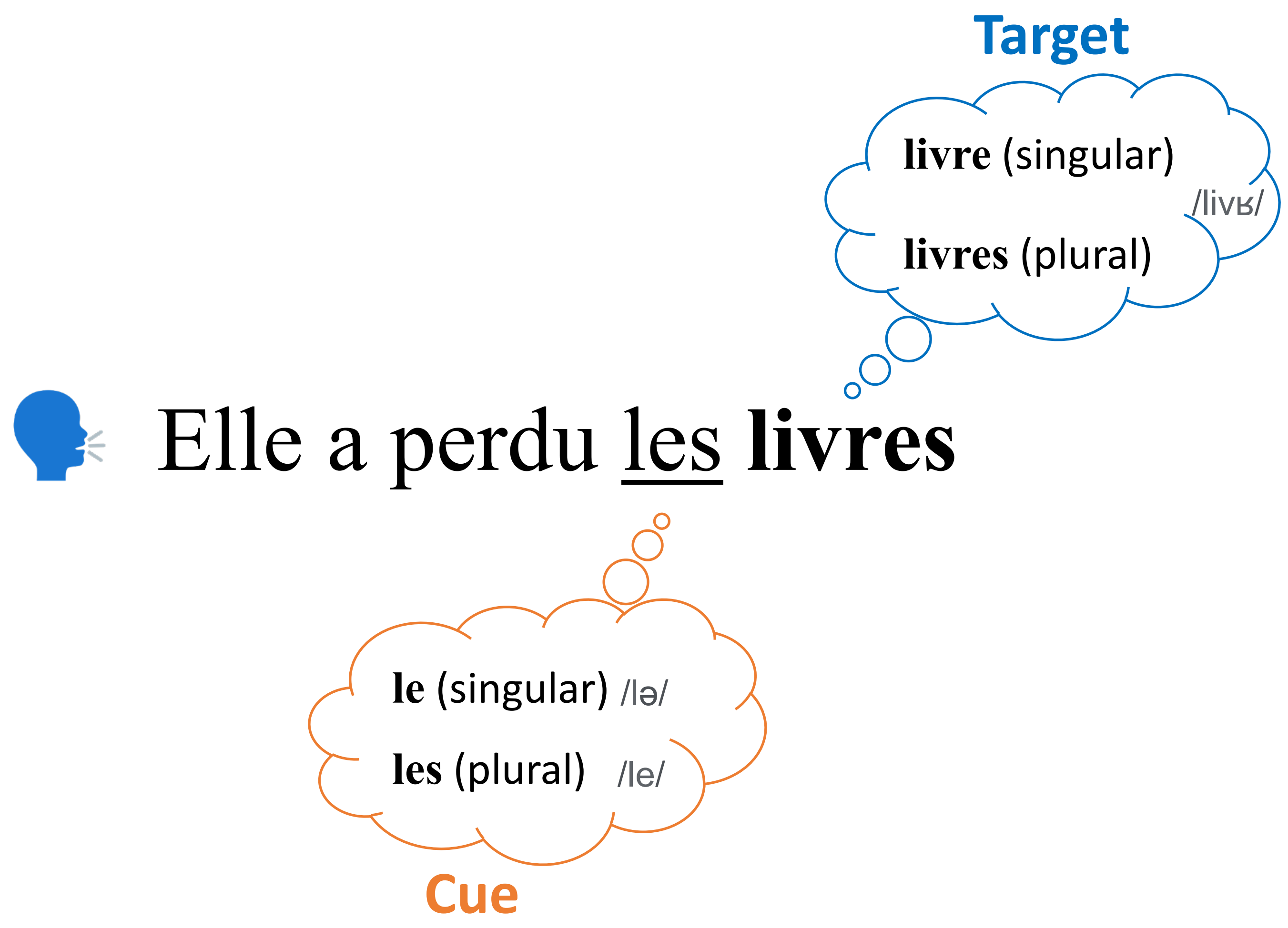
livres (plural)

Elle a perdu les **livres**

(She lost the books)

Evaluation

Controlled task: homophony in French

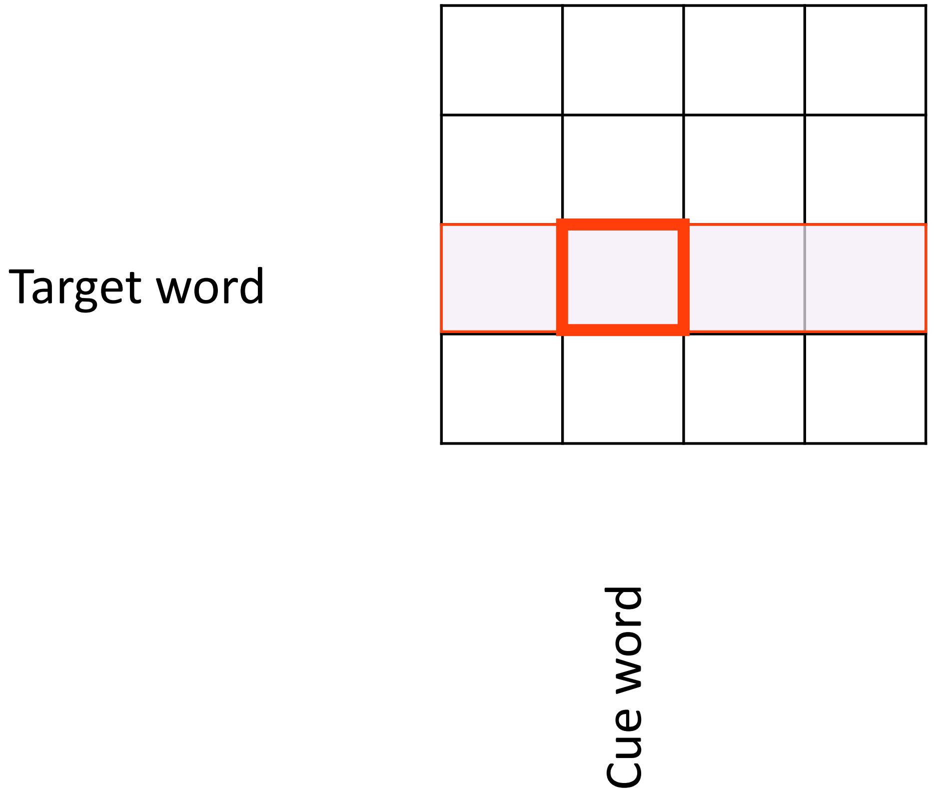


Defined Templates

Pattern	Examples of transcription	#
Det_Noun	C'est le septième **titre** de champion de Syrie de l'histoire du club Il y mène une **vie** d'études et de recherches	720
Pronoun_Verb	Chaque jour, leurs concurrents les voient sortir de pistes dont ils **ignorent** l'existence On y **trouve** une plage naturiste	257
Det_Noun_Verb	Peu après cette élimination, le **club** et Alexander se **séparent** à l'amiable À la fin, les **enfants** se **révoltent** et détruisent l'école.	23

Table 1: Examples of the extracted audios from the Common Voice corpus based on defined patterns. Last column shows the number of examples obtained. Cue and Target words are underlined and **bolded**, respectively.

Cue Contribution score



Cue Contribution

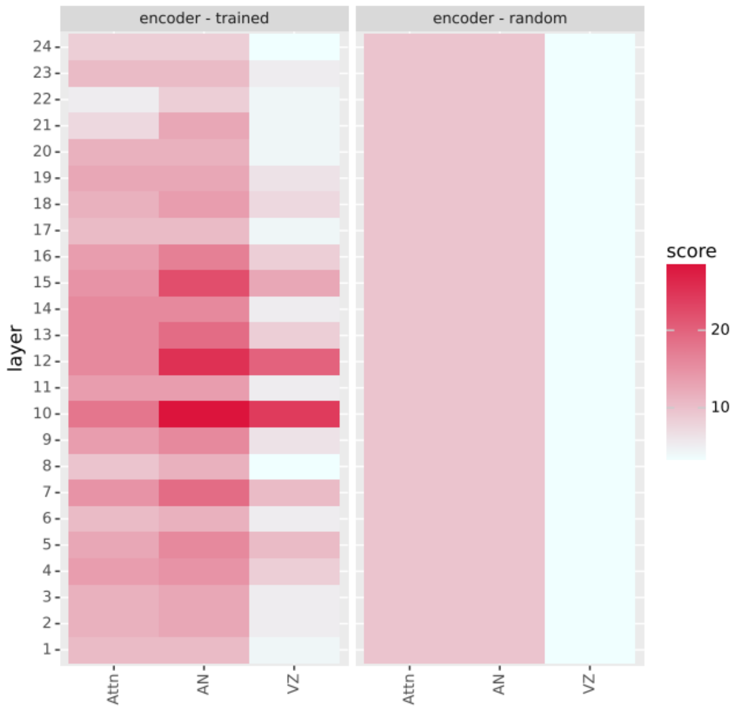


Figure 1: Layer-wise cue contribution according to different analysis methods averaged over all examples for XLSR-53, trained (left) vs. randomly initialized (right).

([Mohebbi et al., 2023](#))

Cue Contribution

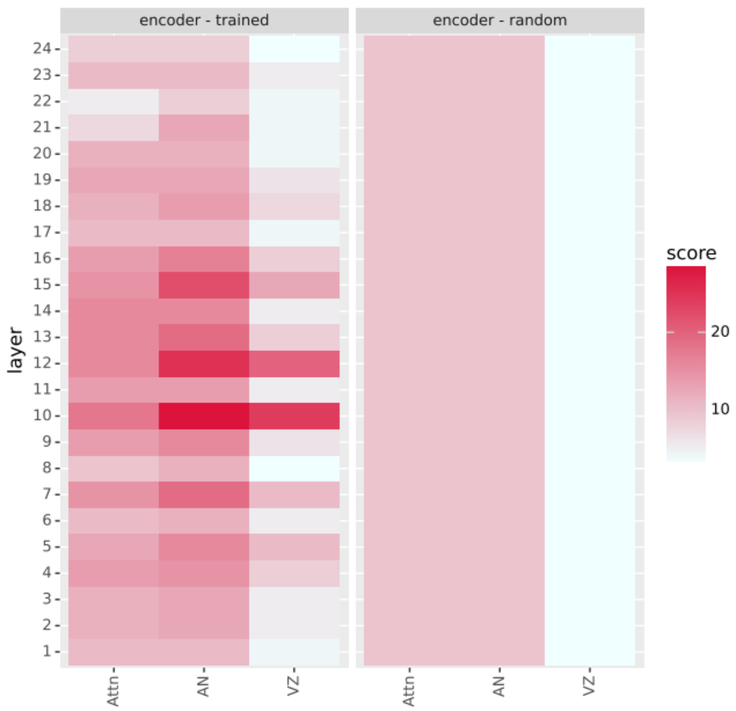


Figure 1: Layer-wise cue contribution according to different analysis methods averaged over all examples for XLSR-53, trained (left) vs. randomly initialized (right).

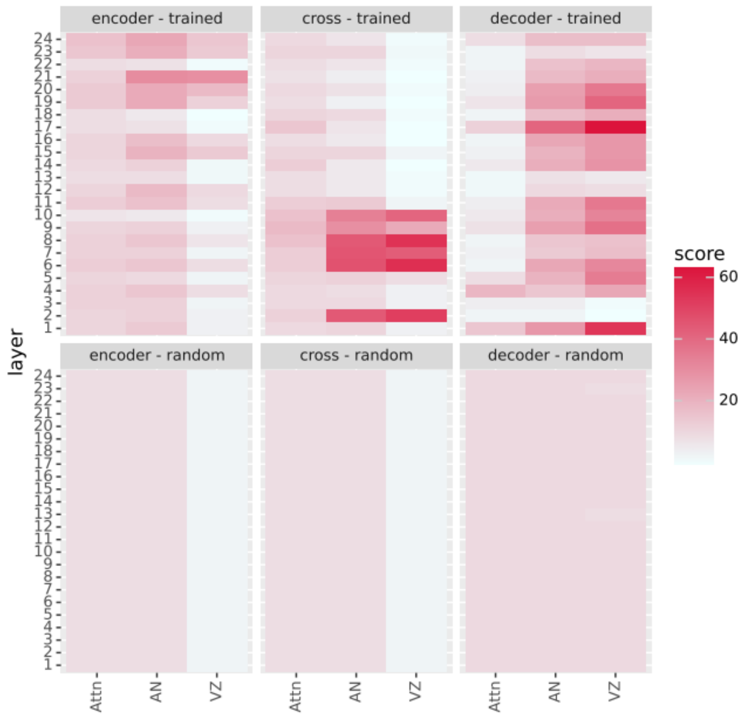


Figure 2: Layer-wise cue contribution according to different analysis methods averaged over all examples for Whisper-medium, trained (top) vs. randomly initialized (bottom).

(Mohebbi et al., 2023)

Cue Contribution v.s. ‘Number encoding’ probe

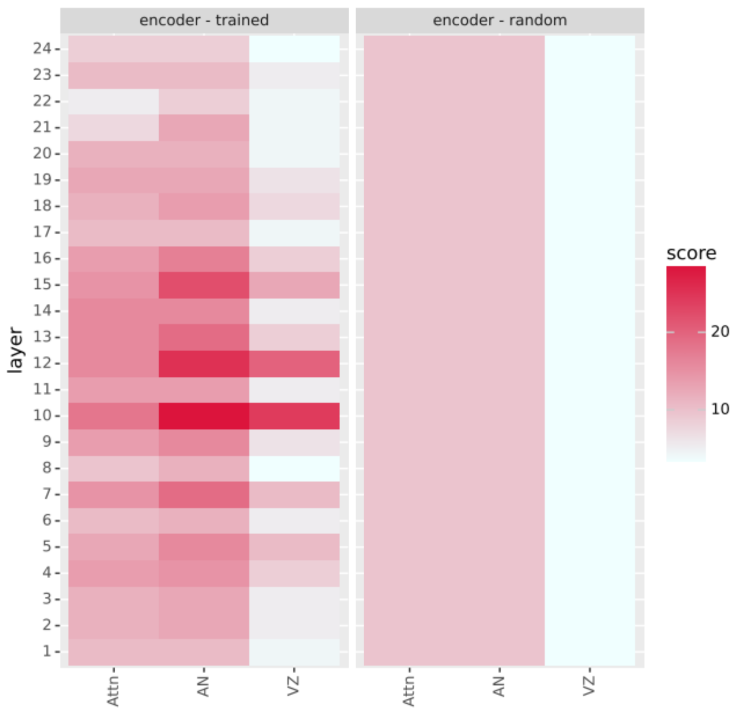


Figure 1: Layer-wise cue contribution according to different analysis methods averaged over all examples for XLSR-53, trained (left) vs. randomly initialized (right).

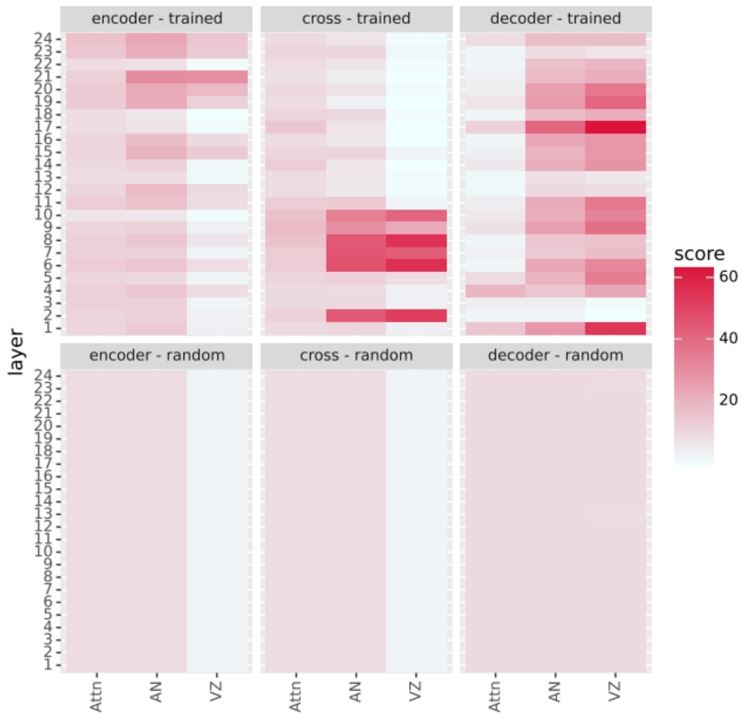


Figure 2: Layer-wise cue contribution according to different analysis methods averaged over all examples for Whisper-medium, trained (top) vs. randomly initialized (bottom).

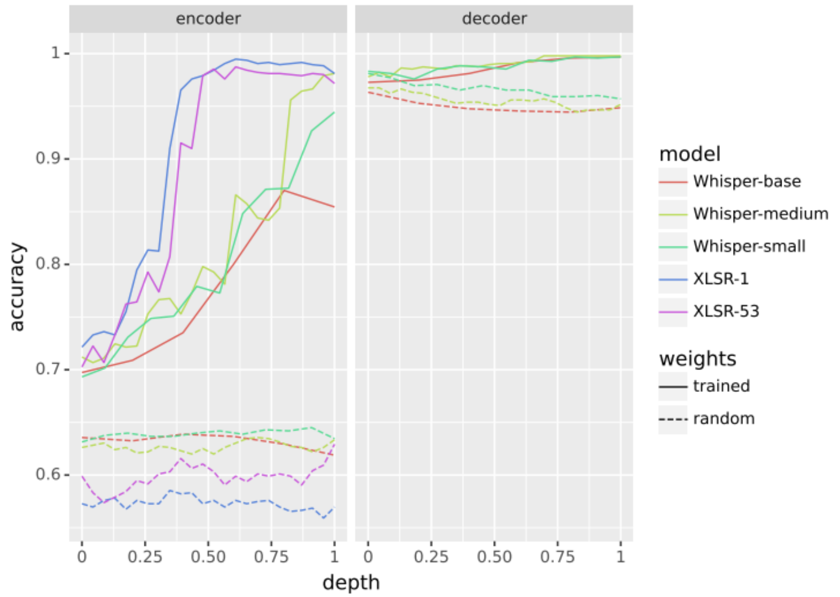


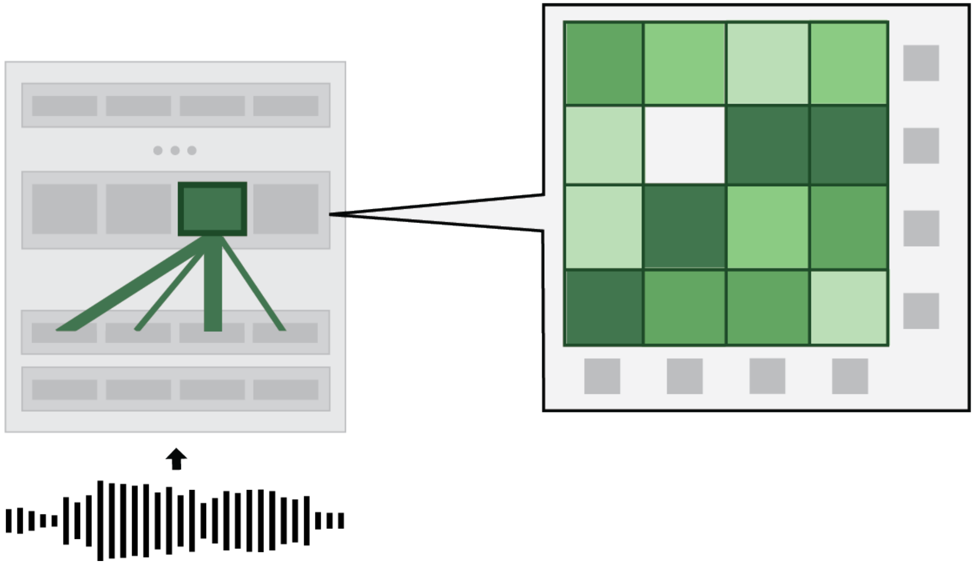
Figure 4: Accuracy of probing classifiers trained on frozen target representations obtained from various ASR models. The depth of Whisper-base (6) and Whisper-small (12), has been normalized to 1 to facilitate comparisons.

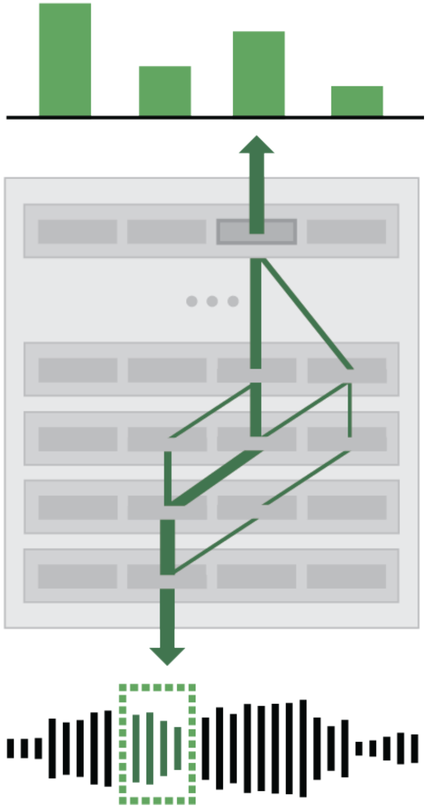
(Mohebbi et al., 2023)

Wrapping up

- Analyzed the pattern of attentions in speech Transformers (e.g., diagonality, diversity)
- Pointed out the limitations of attention as a measure of context-mixing
- Analyzed context-mixing beyond attention (using e.g., Attention-Norm, Value Zeroing)
- Context-mixing vs. Feature attribution

Context-Mixing vs. Feature Attribution





(illustration by Marianne)

Thanks!

Question?

Feel free to reach out for any questions:

Email: h.mohebbi@tilburguniversity.edu

X: @hmohebbi75

Thank you!

Website

Notebook