Verdini et al. (2025). How to Connect Speech Foundation Models and Large Language Models? What Matters and What Does Not. *Interspeech 2025*

Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

How to Connect Speech Foundation Models and Large Language Models? What Matters and What Does Not

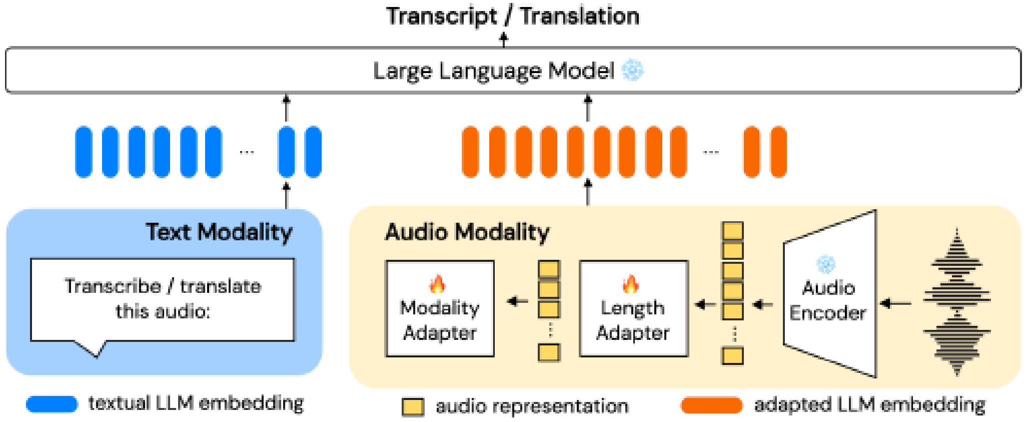
Francesco Verdini[1,3,*], *Pierfrancesco Melucci*[1,2,3,*], *Stefano Perna*[2,3,*], *Francesco Cariaggi*[3,4,*], *Marco Gaido*[5], *Sara Papi*[5], *Szymon Mazurek*[6,7], *Marek Kasztelnik*[7], *Luisa Bentivogli*[5], *Sébastien Bratières*[3,4], *Paolo Merialdo*[2], *Simone Scardapane*[1]

[1]Sapienza University of Rome, Italy; [2]Roma Tre University, Italy; [3]Translated, Italy; [4]Pi School, Italy; [5]Fondazione Bruno Kessler, Italy; [6]AGH University of Krakow, Poland; [7]ACC Cyfronet AGH, Poland

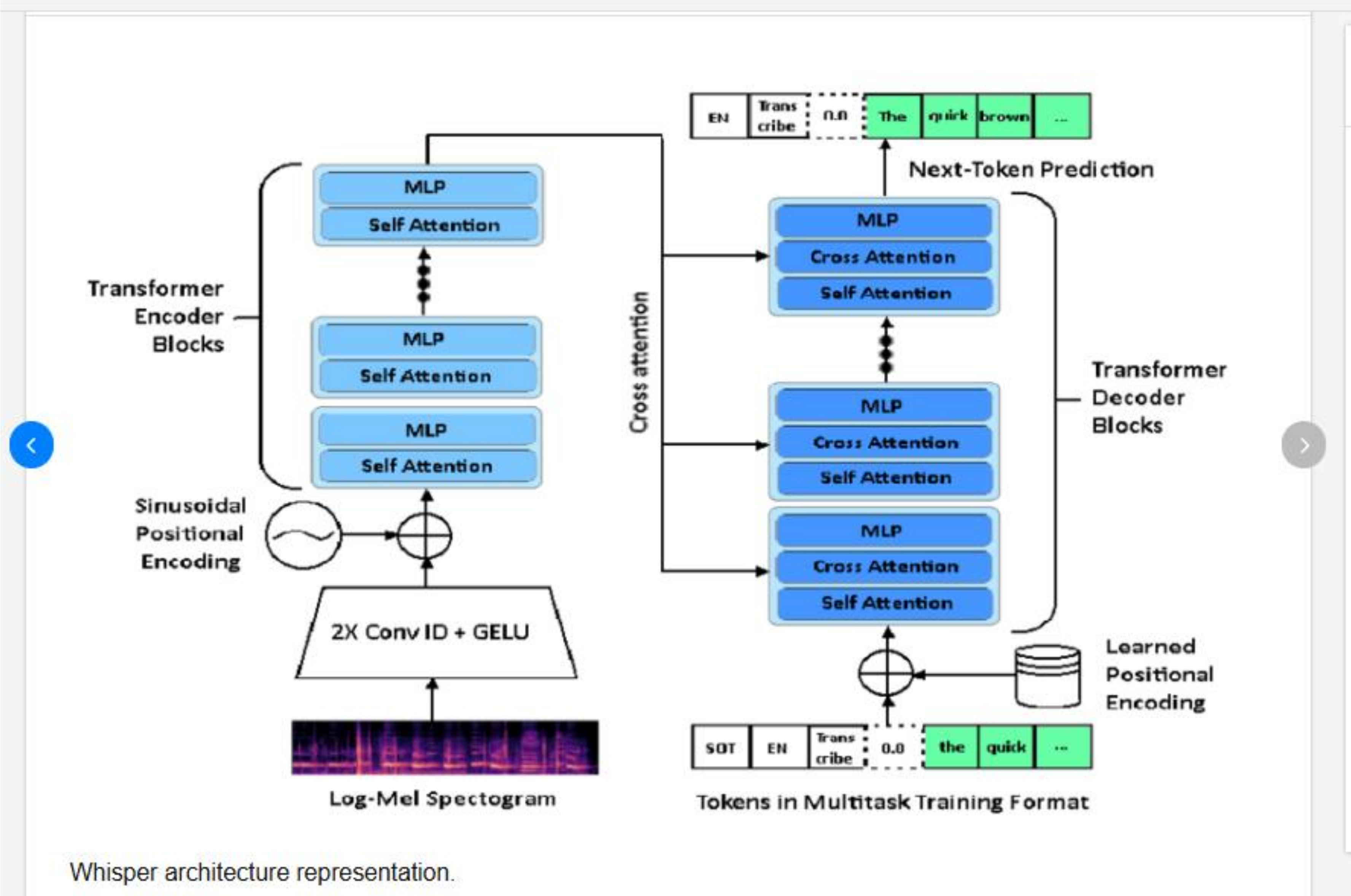
francesco.verdini@uniroma1.it, pierfrancesco.melucci@uniroma1.it, stefano.perna@uniroma3.it, francesco.cariaggi@picampus-school.com

Abstract

The remarkable performance achieved by Large Language Models (LLM) has driven research efforts to leverage them for a wide range of tasks and input modalities. In speech-to-text (S2T) tasks, the emerging solution consists of projecting the output of the encoder of a Speech Foundational Model (SFM) into the LLM embedding space through an adapter module. However, no work has yet investigated how much

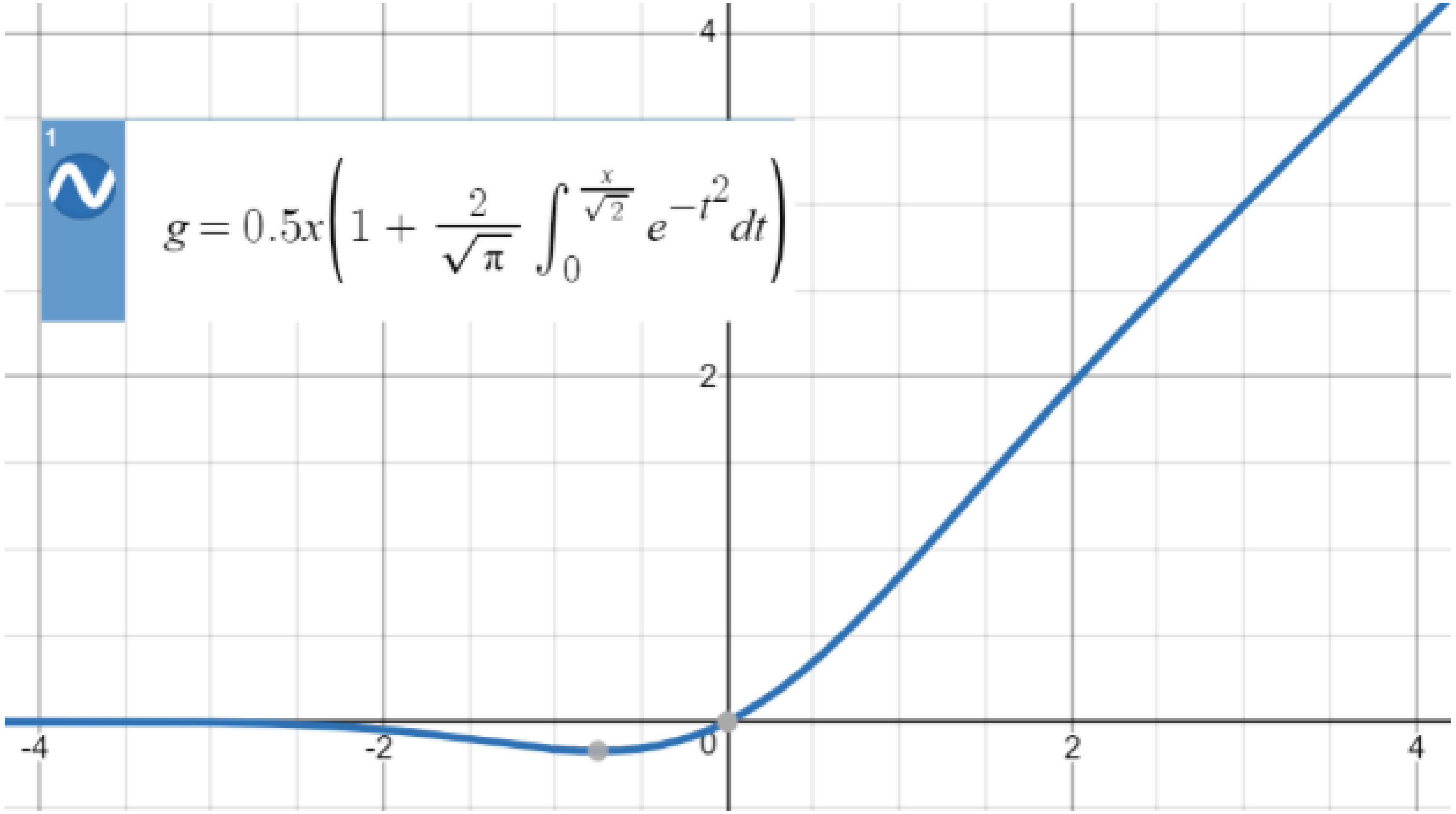


Whisper architecture

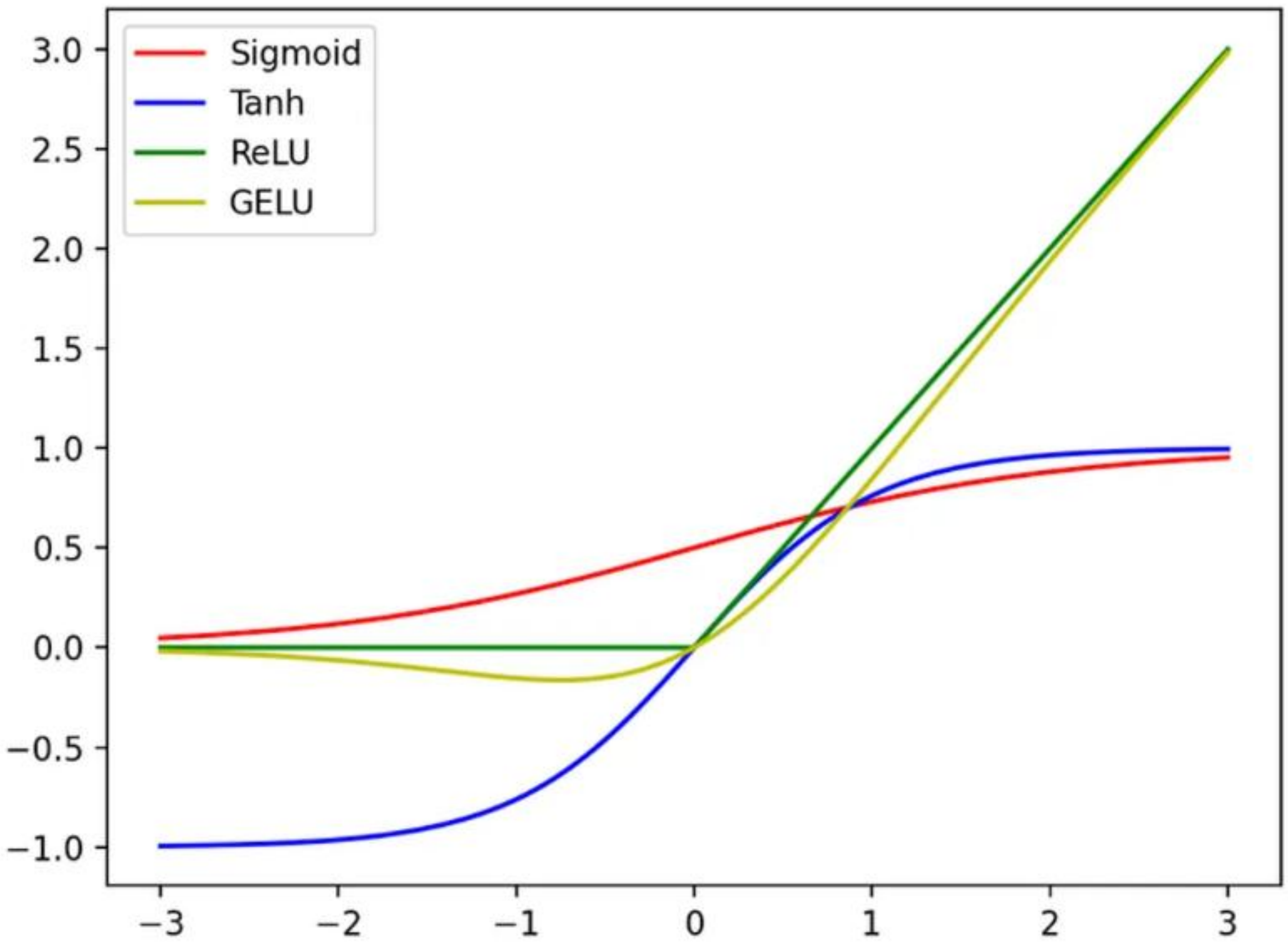


Whisper architecture representation.

GELU (Gaussian error linear unit)



$$g = 0.5x\left(1 + \frac{2}{\sqrt{\pi}} \int_0^{\frac{x}{\sqrt{2}}} e^{-t^2} dt\right)$$



GeLU ~ A smoother ReLU

Verdini et al. (2025). How to Connect Speech Foundation Models and Large Language Models?

- […] In speech-to-text (S2T) tasks, the emerging solution consists of projecting the output of the encoder of a Speech Foundational Model (SFM) into the LLM embedding space through an adapter module.
- However, no work has yet investigated how much the downstream-task performance depends on each component (SFM, adapter, LLM) nor whether the best design of the adapter depends on the chosen SFM and LLM.
- To fill this gap, we evaluate the combination of 5 adapter modules, 2 LLMs (Mistral and Llama), and 2 SFMs (Whisper and SeamlessM4T) on two widespread S2T tasks, namely Automatic Speech Recognition and Speech Translation.

Focus on the “adapter”. How they argue:

- Many architectural solutions have been proposed for the adapter
  - often employed to both reduce the LLM computational costs and the modality mismatch with the textual sequences.
- These methods span from fixed downsampling, obtained either with a stack of strided convolutions [9] or with window-level Q-Former [3], to modules with variable compression rates that reduce the input sequence based on its semantic content, such as Continuous Integrate-and-Fire (CIF) [10] and CTC compression [11].
- Nonetheless, a comprehensive study on the adapter choice is missing

Procedure

- explore impact on ASR and ST performance via systematic comparison of 20 different combinations:

- 5 adapters (proposed in the literature)
- 2 SFMs (Whisper-large-v3 [13] for ASR
  - and SeamlessM4T v2-large[14] for ST
- 2 LLMs (Llama [1] and Mistral [15])

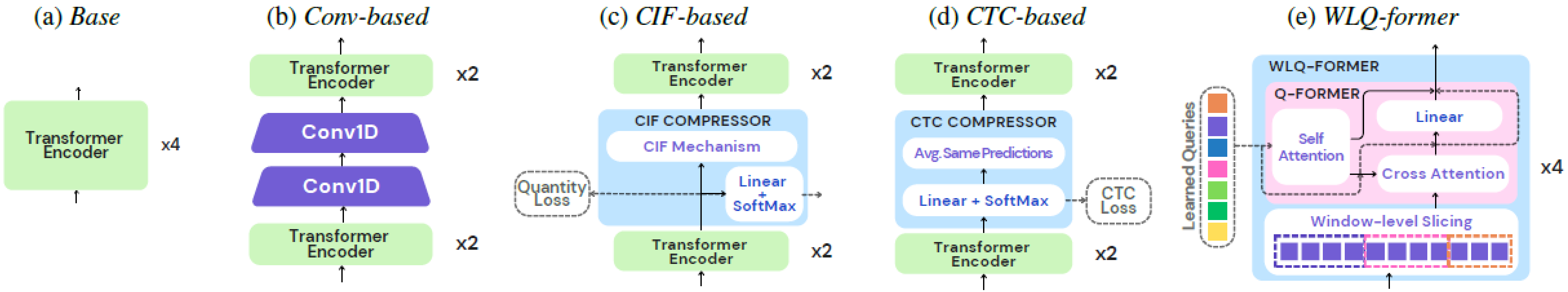


Figure 2: *Representation of the adapters analyzed in the paper.*

CIF = continuous integrate-and-fire
CTC = connectionist temporal classification
WL = window length
Q-former = transformer-based NN for cross-modal learning

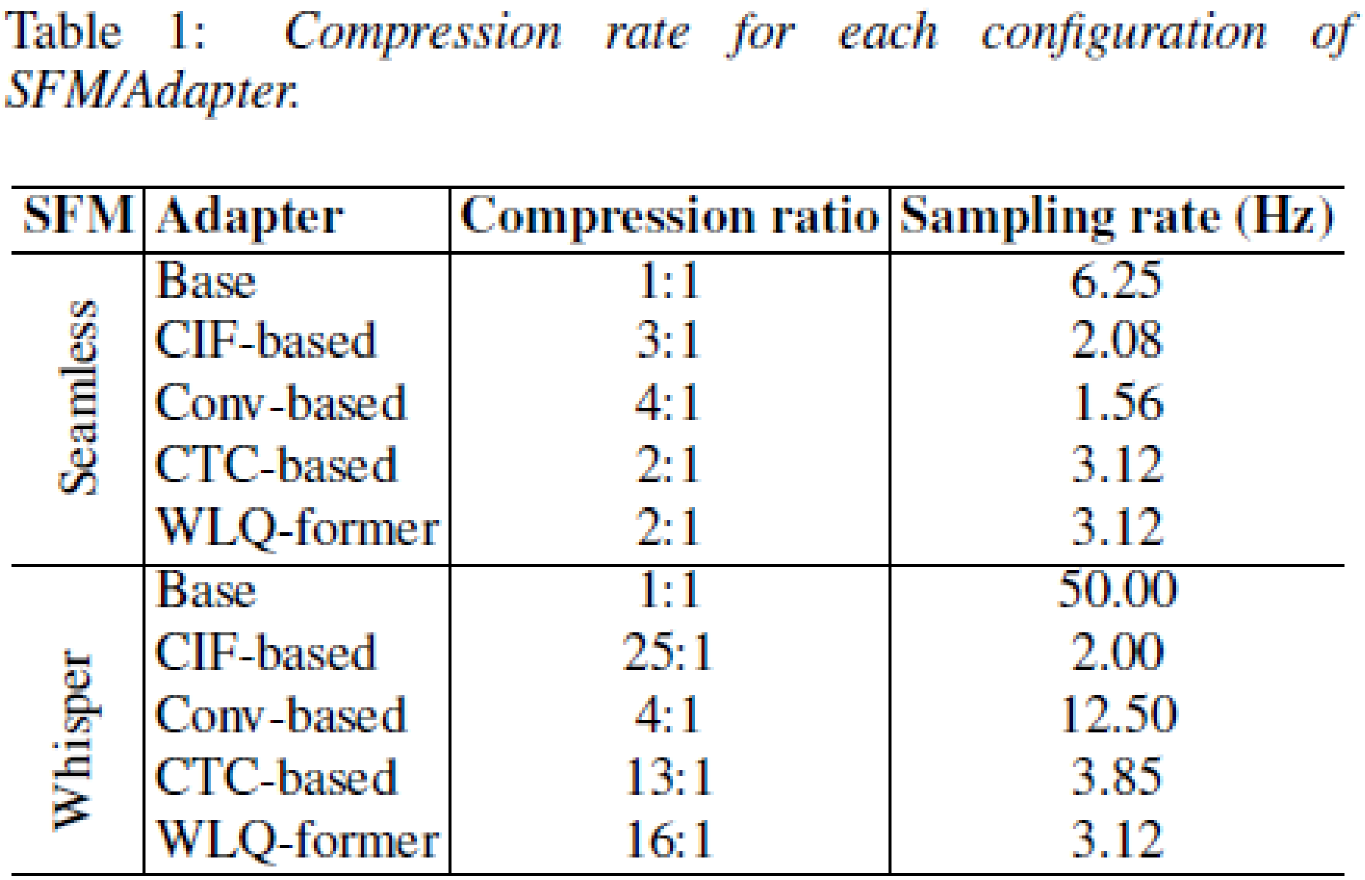


Table 1: *Compression rate for each configuration of SFM/Adapter.*

SFM	Adapter	Compression ratio	Sampling rate (Hz)
Seamless	Base	1:1	6.25
	CIF-based	3:1	2.08
	Conv-based	4:1	1.56
	CTC-based	2:1	3.12
	WLQ-former	2:1	3.12
Whisper	Base	1:1	50.00
	CIF-based	25:1	2.00
	Conv-based	4:1	12.50
	CTC-based	13:1	3.85
	WLQ-former	16:1	3.12

Product is 6.25 (1/16 sec)

Product is 50 (20ms frames)

Many other technical details (see paper)

Table 3: *ASR and ST results on CoVoST test sets. The best result for each (SFM, LLM) configuration is underlined, while the overall best is* ***bolded****. The difference with Base is statistically significant ($p < 0.05$) unless for scores marked with* *.

SFM	LLM	Adapter	ST - COMET (↑)						ASR - WER (↓)					
			en-de	de-en	es-en	fr-en	it-en	avg	en	es	fr	it	de	avg
SeamlessM4T	Mistral	Base	84.94	84.75	86.65	**84.71**	**85.42**	85.29	6.48	6.56	**9.69**	**7.8**	**8.36**	**7.78**
		CIF-based	84.31	84.33	86.31	84.32	85.07	84.87	7.10	6.92	10.23	8.60	9.38	8.45
		Conv-based	84.33	84.15	86.20	84.11	84.98	84.75	7.53	7.83	11.38	10.07	11.44	9.65
		CTC-based	82.95	82.48	85.20	82.85	83.57	83.41	7.94	7.90	12.51	10.31	12.29	10.19
		WLQ-former	84.67	84.71*	86.60*	84.59	85.29	85.17	**6.38***	6.80	9.83*	8.05	8.48*	7.91
	Llama 3.1	Base	85.12	84.15	86.17	84.08	84.78	84.86	7.15	7.46	10.67	9.20	9.96	8.89
		CIF-based	84.65	83.87	85.98	83.86	84.65	84.60	7.66	7.47*	12.36	10.18	10.50	9.63
		Conv-based	85.42	84.42	86.43	84.31	85.17	85.15	7.16*	7.08	10.79*	8.99	9.83*	8.77
		CTC-based	83.78	82.49	85.21	82.83	83.60	83.58	7.95	8.04	12.17	9.94	11.22	9.90
		WLQ-former	**85.65**	**84.84**	**86.66**	84.68	85.39	**85.44**	6.62	6.69	9.96	7.97	8.71	7.99
Whisper	Mistral	Base	78.98	81.38	84.79	81.63	82.69	81.89	11.37	7.57	12.81	10.14	10.88	10.55
		CIF-based	77.79	80.35	84.11	80.79	81.83	80.99	12.57	8.45	14.24	12.32	13.09	12.13
		Conv-based	78.73	81.26	84.72*	81.52*	82.58	81.76	11.78	7.60	13.23	10.67	11.52	10.96
		CTC-based	75.56	76.53	81.75	78.33	78.55	78.14	14.69	10.63	17.15	15.09	16.50	14.81
		WLQ-former	79.07*	81.44*	84.92	81.68*	82.92	82.00	11.82	8.21	13.60	15.77	12.55	12.39
	Llama 3.1	Base	80.43	82.15	85.21	82.33	83.06	82.64	9.90	**6.33**	11.27	8.52	9.09	9.02
		CIF-based	78.32	78.94	82.51	80.09	80.27	80.02	12.82	8.53	14.31	12.80	13.53	12.40
		Conv-based	80.84	82.57	85.49	82.60	83.51	83.00	9.90*	6.46	11.49*	8.75*	9.00*	9.12
		CTC-based	76.47	73.80	80.16	77.19	76.59	76.84	14.02	10.98	17.55	16.29	17.21	15.21
		WLQ-former	79.95	81.56	84.88	81.56	82.89	82.17	11.98	7.90	14.52	11.10	12.84	11.67

Conclusion

- experiments covering two tasks (ASR and ST) and 5 languages
- results demonstrate that the choice of the SFM is the most critical factor influencing downstream performance
- there is no one-size-fits-all solution for the adapter
  - the optimal choice varies depending on the specific combination of SFM and LLM.
- the Base and WLQformer adapters, which feature very different compression factors, demonstrate strong performance across tasks
  - suggesting that reducing sequence length mismatch between speech and text is less crucial than previously assumed.

Continuous Integrate-and-Fire (CIF)

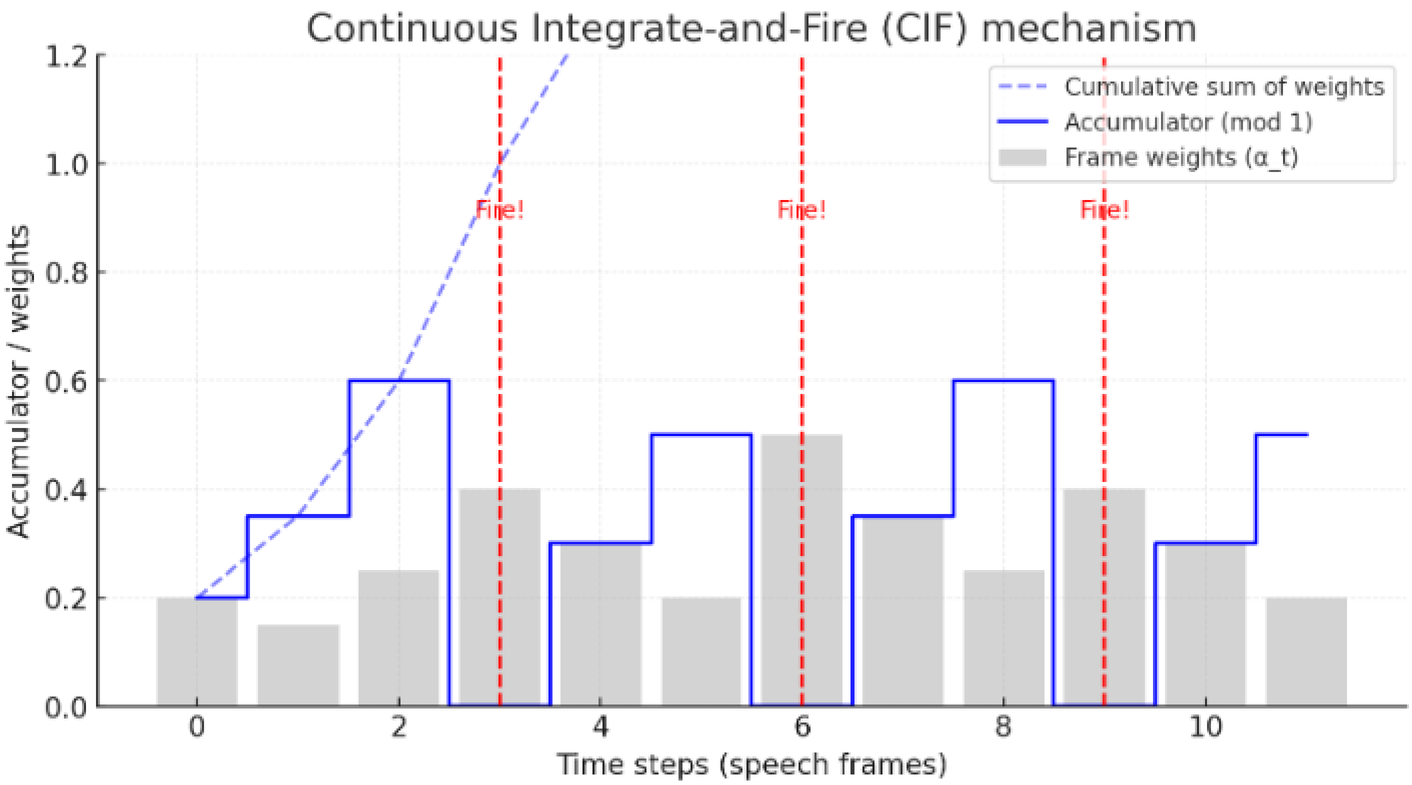
Problem

Alignment between speech (a **continuous-time sequence** with many acoustic frames, like 100 frames per second) with text (**discrete sequence**).

CIF

- neuroscience-inspired
- is an **alignment method** for bridging speech encoders with text decoders like when combining a speech foundation model with a large language model. Once you have token-level embeddings via CIF, you can feed them directly into a large language model as if they were text embeddings.

- **Alignment without supervision**: CIF automatically learns how many tokens to output and where to place them.


Continuous Integrate-and-Fire (CIF) mechanism
Accumulator / weights
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
Time steps (speech frames)
Fire!
Fire!
Fire!
Cumulative sum of weights
Accumulator (mod 1)
Frame weights (α_t)