GOING BEYOND YOUR EXPECTATIONS IN LATENCY METRICS FOR SIMULTANEOUS SPEECH TRANSLATION

Jorge Iranzo-Sánchez Javier Iranzo-Sánchez Adrià Giménez Jorge Civera

www.mllp.upv.es MLLP Machine Learning and Language Processing UNIVERSITAT POLITÈCNICA DE VALÈNCIA

INTRODUCTION TO SIMULST EVALUATION

- Task: Translate unbounded input stream in real-time
- **Translation quality** (BLEU, COMET) and **latency** (LAAL, ATD).
- Independent segments of 2-10 seconds, take the **mean** of this.
- This hides current **problematic behaviour of SimulST systems.**
- Evaluation of past IWSLT shared tasks results to show this.
- We propose methods and give recomendations to compare SimulST system latencies.

THE PROBLEM

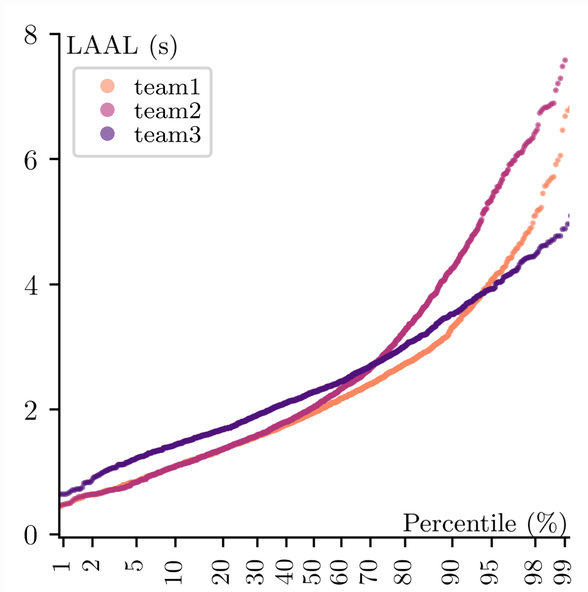
IWSLT 2024 en→ja

		LAAL				
team	BLEU	M	mdn	p90	p95	p99	max
...							
Team 2	19.3	2.4	2.0	4.2	5.3	7.5	31.3
Team 3	17.9	2.4	2.2	3.5	3.9	4.8	10.4

- Determining latency only on **metric mean** is highly unrealistic.
- False conclusions about model performance and stability.
- Example:
  - Translation quality: Team 2 > Team 3.
  - If looking only latency mean, Team 2 > Team 3.
  - **However**, Team 2 has higher latency spikes!
  - Could we really affirm that Team 2 is better than Team 3?

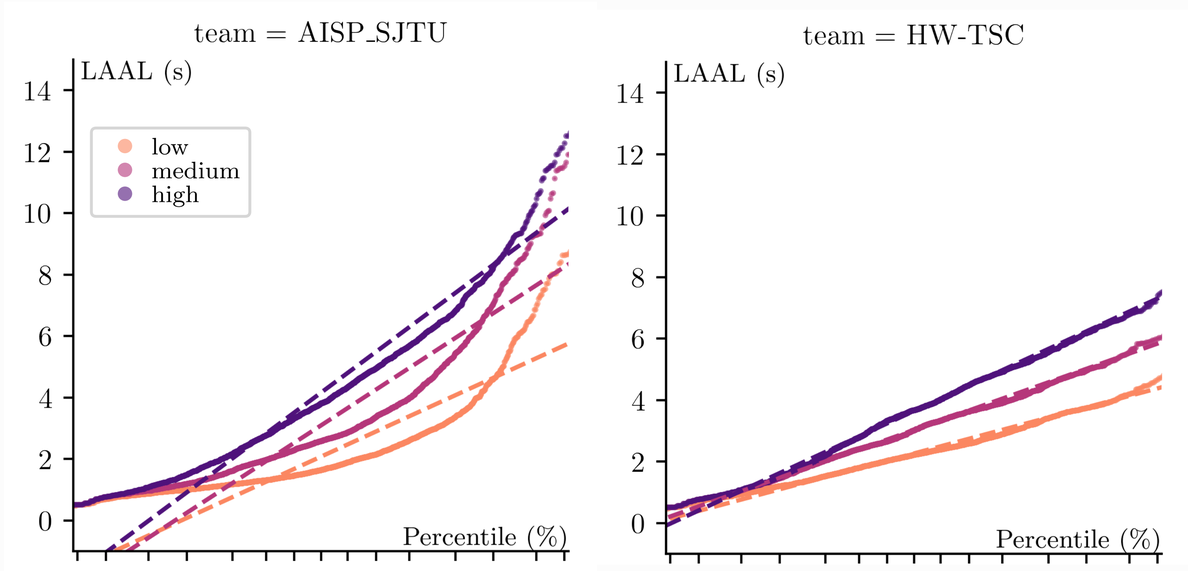
NORMALITY

IWSLT 2024 en→ja



- Graphical representations can give us a better look at system behaviour.
- Percentiles are a good way to compare latency between systems.

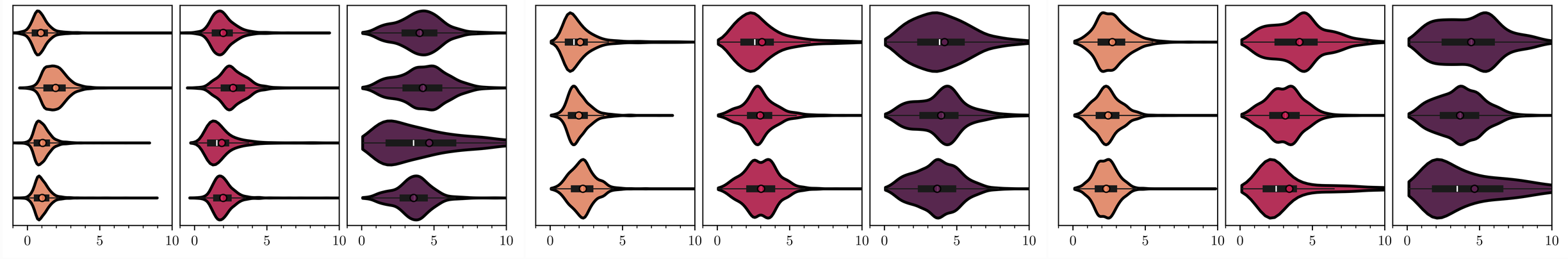
IWSLT 2022 en→zh



- Normality of distribution as a way to measure representativeness of the mean of the distribution.
- Normal probability plots → Normality and percentiles combined.

LATENCY DISTRIBUTION

- IWSLT 2022 LAAL distributions, some participations clear have unusual behaviour compared to the rest.



PHENOMENA HIDDEN BY THE MEAN: OVER-WAIT

- **Over-wait** OW_s^r: % of samples with duration $> t$ with ratio between latency scores and input length $> r$.
- Over-wait can be easily be used to detect degeneration to offline behaviours.

$\text{OW}_r^{0.75}$, AISP_SJTU - IWSLT 2022

	r			
Lat. band	0.75	0.85	0.95	1.00
low	6.5	6.5	6.2	6.2
medium	17.0	16.0	15.7	15.7
high	48.3	38.7	33.1	32.6