UNIVERSITÀ
DI TRENTO

FONDAZIONE
BRUNO KESSLER

Over-generation cannot be rewarded:
Length-Adaptive Average Lagging

Sara Papi, Marco Gaido, Matteo Negri, Marco Turchi

{spapi,mgaido,negri,turchi}@fbk.eu

AutoSimTrans Worshop @ NAACL2022

Evaluation in SimulST

Simultaneous Speech Translation (SimulST) evaluation involves measuring:

- **Translation quality**: measures how good the translation is (e.g. with BLEU as in standard offline speech translation evaluation)
- **Latency**: measures the delay between the source speech and the generated translation

Latency Metrics

Many metrics have been proposed for simultaneous machine translation and adapted to SimulST:

➜ **Average Lagging** or AL (Ma et al., 2019) is the most popular and widely used

Latency Metrics

Many metrics have been proposed for simultaneous machine translation and adapted to SimulST:

➔ **Average Lagging** or AL (Ma et al., 2019) is the most popular and widely used

The AL goal is to quantify how much time the system is out of sync with the speaker

Average Lagging: formulation

$$AL = \frac{1}{\tau'(|\mathbf{X}|)} \sum_{i=1}^{\tau'(|\mathbf{X}|)} d_i - d_i^*$$

index of the target token when the end of audio is reached

$$d_i^* = (i-1) \cdot \frac{\sum_{j=1}^{|\mathbf{X}|} T_j}{|\mathbf{Y}|}$$

Average Lagging: formulation

system delay
to be evaluated

$$AL = \frac{1}{\tau'(|\mathbf{X}|)} \sum_{i=1}^{\tau'(|\mathbf{X}|)} d_i - d_i^*$$

index of the target token when the end of audio is reached

$$d_i^* = (i-1) \cdot \frac{\sum_{j=1}^{|\mathbf{X}|} T_j}{|\mathbf{Y}|}$$

Average Lagging: formulation

system delay
to be evaluated

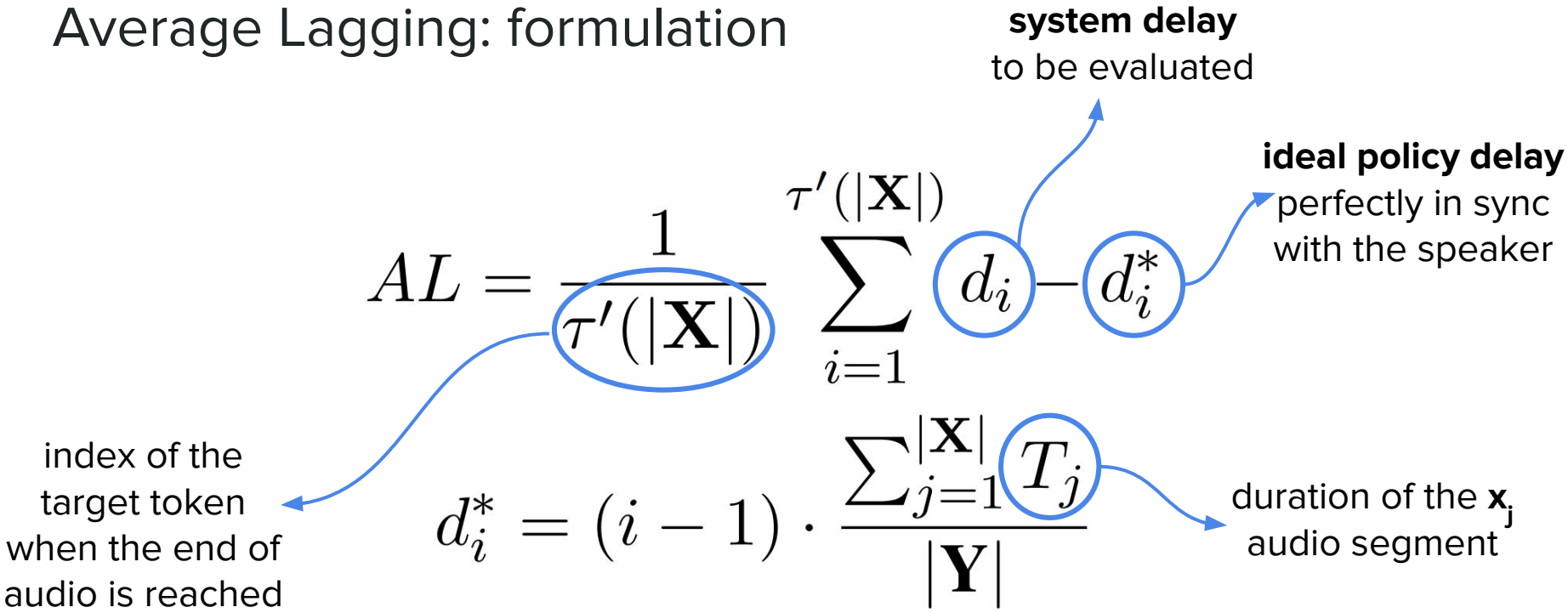
ideal policy delay
perfectly in sync
with the speaker

$$AL = \frac{1}{\tau'(|\mathbf{X}|)} \sum_{i=1}^{\tau'(|\mathbf{X}|)} d_i - d_i^*$$

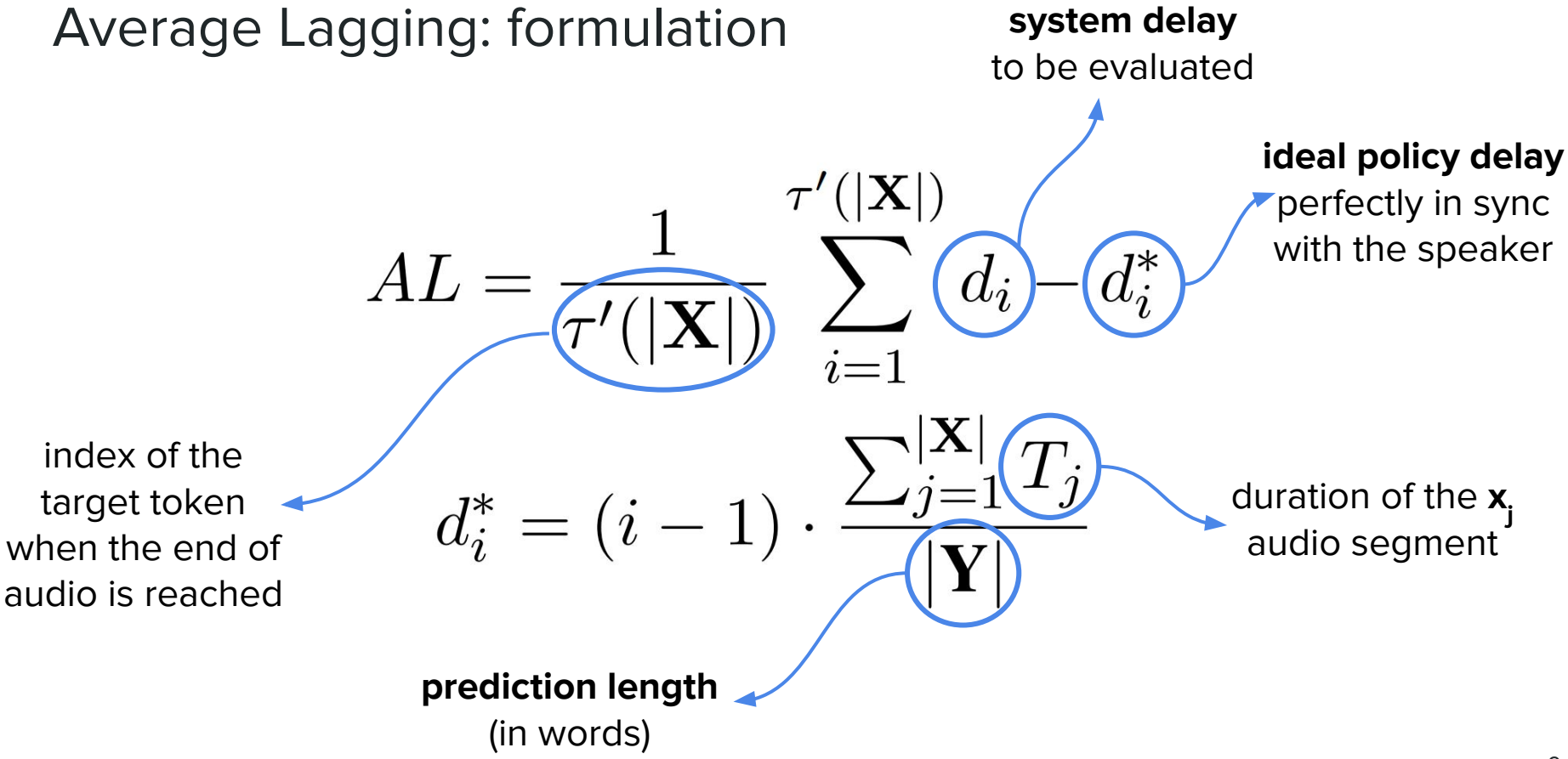
index of the target token when the end of audio is reached

$$d_i^* = (i-1) \cdot \frac{\sum_{j=1}^{|\mathbf{X}|} T_j}{|\mathbf{Y}|}$$

Average Lagging: formulation



Average Lagging: formulation



Average Lagging and under-generation:

In adapting the metric for speech, Ma et al. (2020) noticed that the metric was not robust to **under-generation**:

- The problem is more frequent in SimulST due to the presence of silences or long pauses

➜ The lagging behind the ideal policy becomes negative and this favors under-generative systems

Average Lagging and under-generation:

In adapting the metric for speech, Ma et al. (2020) noticed that the metric was not robust to **under-generation**:

- The problem is more frequent in SimulST due to the presence of silences or long pauses

➔ The lagging behind the ideal policy becomes negative and this favors under-generative systems

For this reason, they proposed to change the ideal policy calculation

Average Lagging: ideal policy for under-generation

$$AL = \frac{1}{\tau'(|\mathbf{X}|)} \sum_{i=1}^{\tau'(|\mathbf{X}|)} d_i - d_i^*$$

$$d_i^* = (i-1) \cdot \frac{\sum_{j=1}^{|\mathbf{X}|} T_j}{|\mathbf{Y}^*|}$$

Reference length instead of prediction length

The Problem of Over-generation

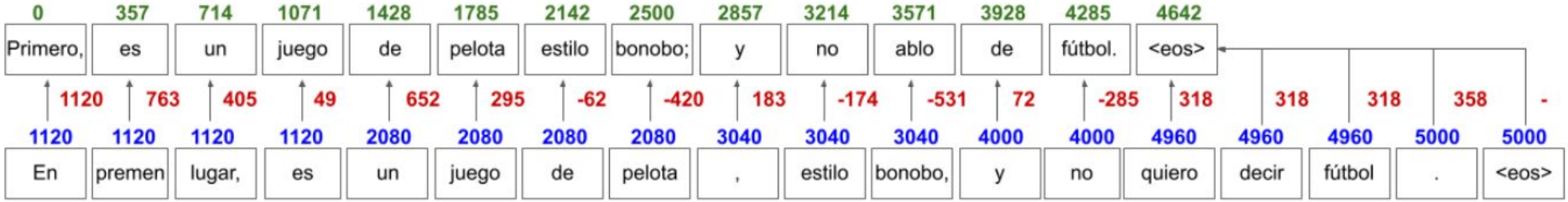
Still, Ma et al. (2020) did not consider the **over**-generation:

➜ Older systems were more affected by under-generation

BUT

➜ Newer systems can generate more than one word at a time which sometimes results in over-generation

The Problem of Over-generation: an example



AL (automatic): 198ms

Real delay: 846ms

How frequent is over-generation?

We investigate if it represents a problem considering:

- **CAAT (Liu et al., 2021)**: state-of-the-art SimulST system with adaptive policy
- **Wait-k (Ma et al, 2020)**: SimulST system adopting the most popular (fixed) decision policy in simultaneous
- **Offline with wait-k (Papi et al, 2022)**: offline system used in simultaneous by adopting wait-k decision policy

Average Word Length Difference (AWLD)

We measure AWLD between reference and predictions:

$$\text{AWLD} = \frac{1}{N} \sum_{s=1}^{N} |\mathbf{Y}| - |\mathbf{Y}^*|$$

number of samples

- Positive values ➔ the system tends to over-generate
- Negative values ➔ the system tends to under-generate

AWLD Statistics

Model	k=3	k=5	k=7	k=9	k=11
wait-k	-5.57	-3.82	-2.30	-1.13	-0.74
offline wait-k	0.48	0.49	0.53	0.74	0.80
CAAT	1.57	0.96	0.61	0.35	0.18

AWLD Statistics

Model	k=3	k=5	k=7	k=9	k=11
wait-k	**-5.57**	**-3.82**	**-2.30**	**-1.13**	**-0.74**
offline wait-k	0.48	0.49	0.53	0.74	0.80
CAAT	1.57	0.96	0.61	0.35	0.18

wait-k strongly under-generates

AWLD Statistics

Model	k=3	k=5	k=7	k=9	k=11
wait-k	-5.57	-3.82	-2.30	-1.13	-0.74
offline wait-k	**0.48**	**0.49**	**0.53**	**0.74**	**0.80**
CAAT	1.57	0.96	0.61	0.35	0.18

offline wait-k slightly over-generates

AWLD Statistics

Model	k=3	k=5	k=7	k=9	k=11
wait-k	-5.57	-3.82	-2.30	-1.13	-0.74
offline wait-k	0.48	0.49	0.53	0.74	0.80
CAAT	**1.57**	**0.96**	**0.61**	**0.35**	**0.18**

CAAT tends to over-generate, especially at low latency

Our solution: Length-Adaptive Average Lagging (LAAL)

Our metric accounts also for over-generation phenomena by considering in the ideal policy computation:

- reference length when the prediction is shorter
- prediction length when the prediction is longer

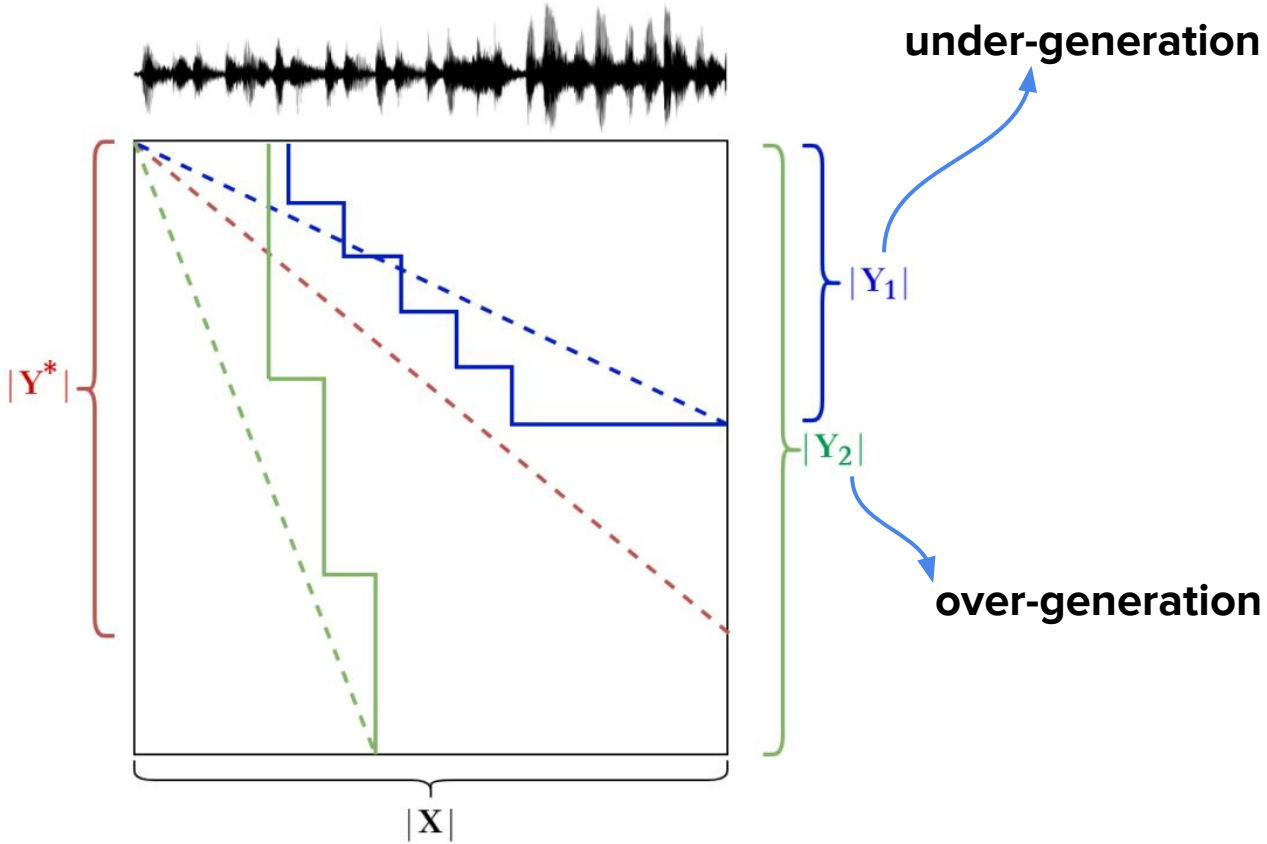
➜ the correction is made at sentence-level and can be applied both to over- and under-generative SimulST systems

LAAL: formulation

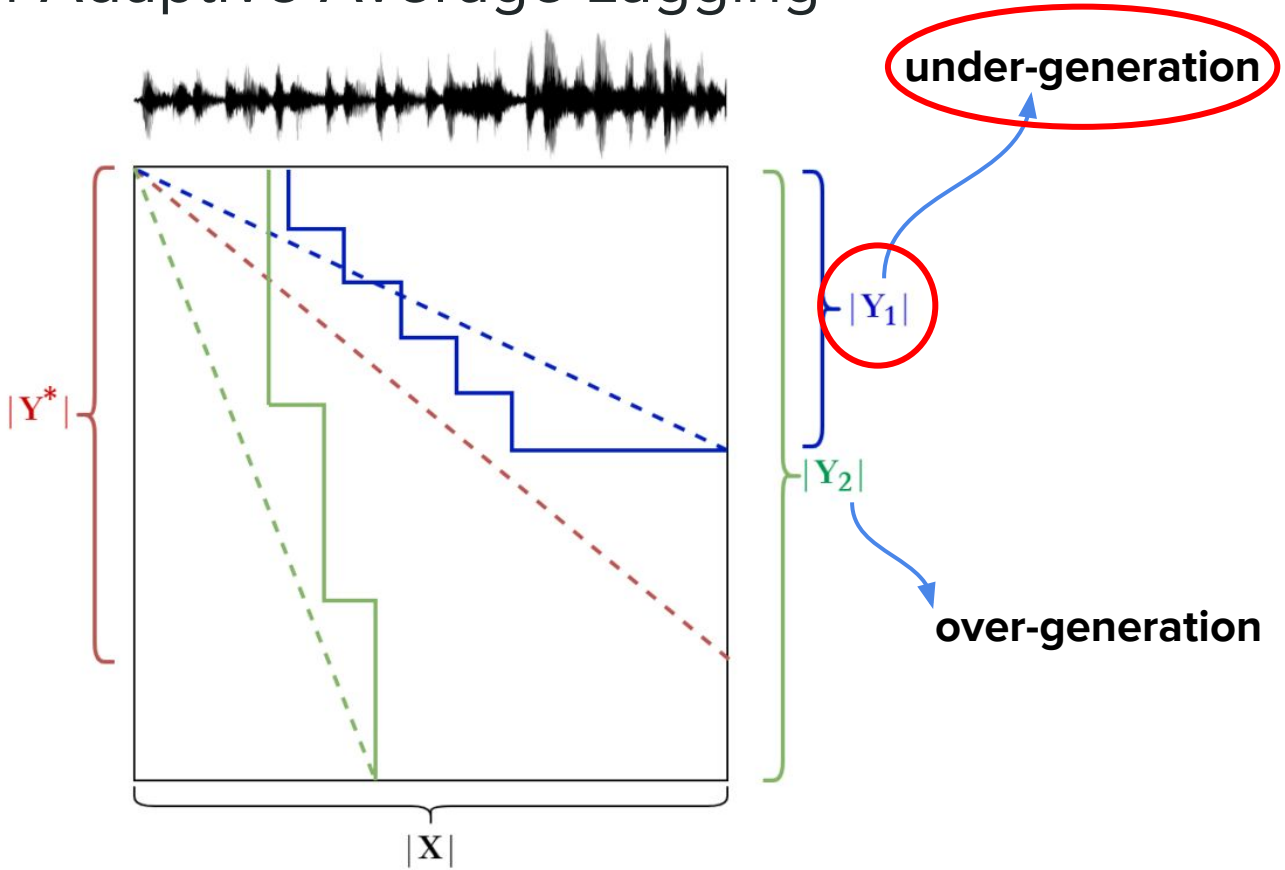
For each sentence, we take the maximum between prediction and reference lengths in the ideal policy:

$$d_i^* = (i-1) \cdot \frac{\sum_{j=1}^{|\mathbf{X}|} T_j}{\max\{|\mathbf{Y}|, |\mathbf{Y}^*|\}}$$

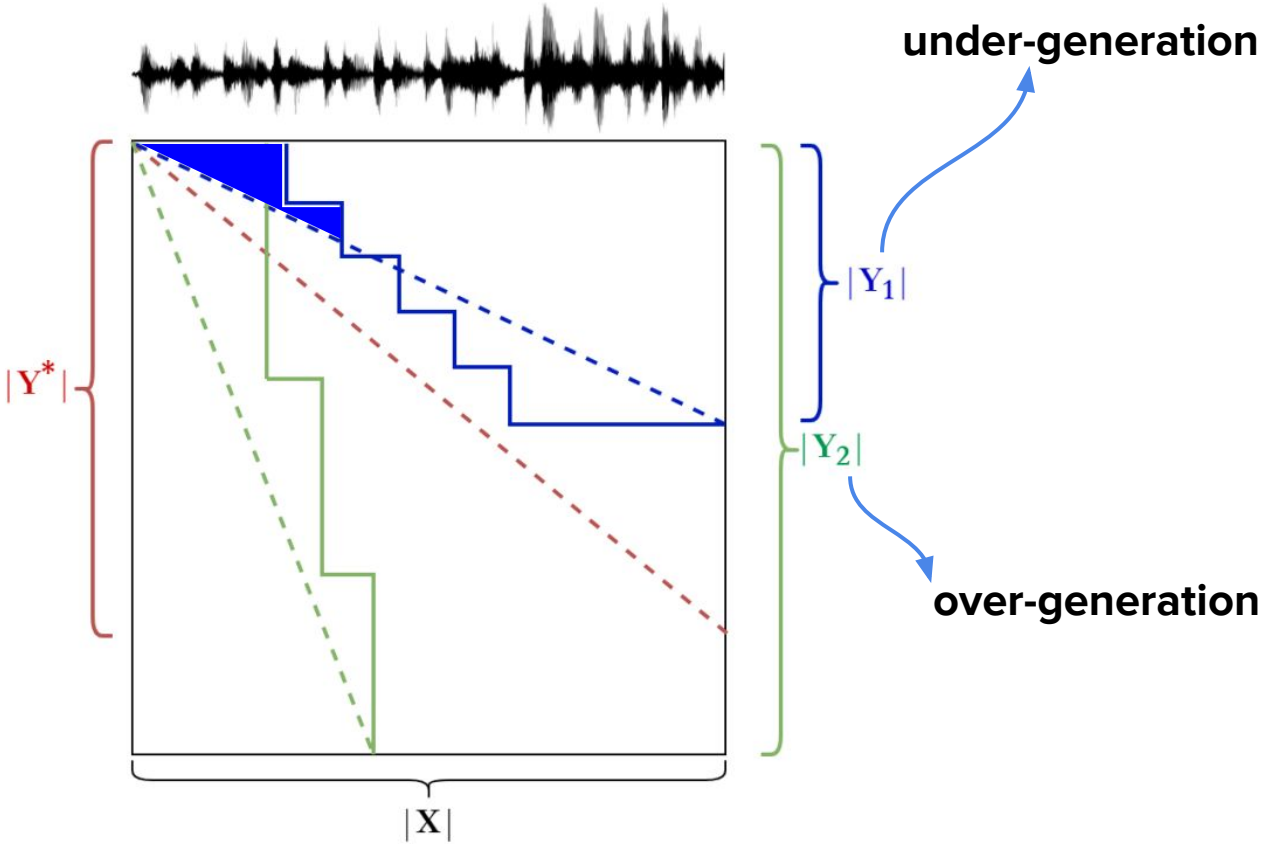
Length-Adaptive Average Lagging



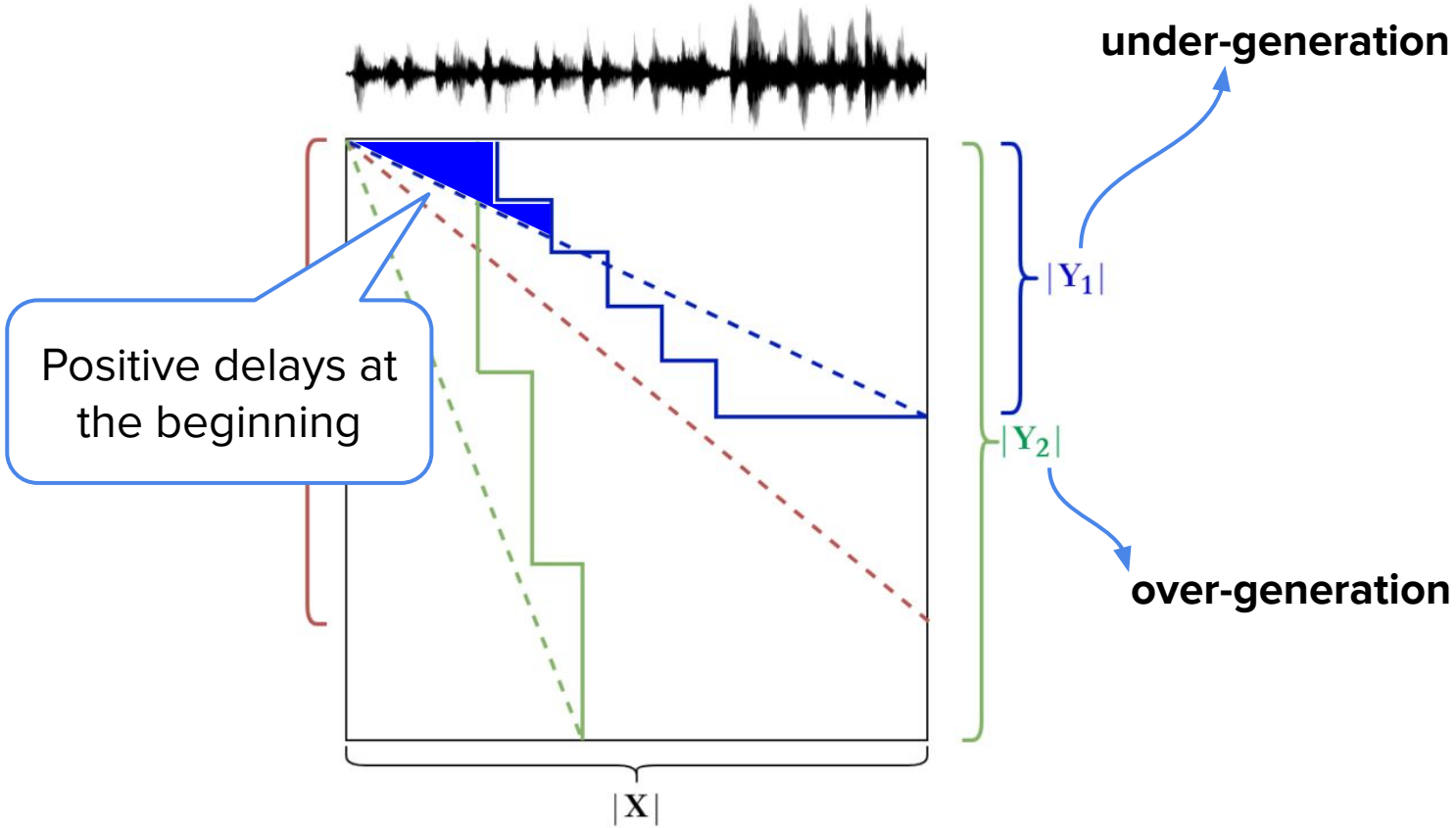
Length-Adaptive Average Lagging



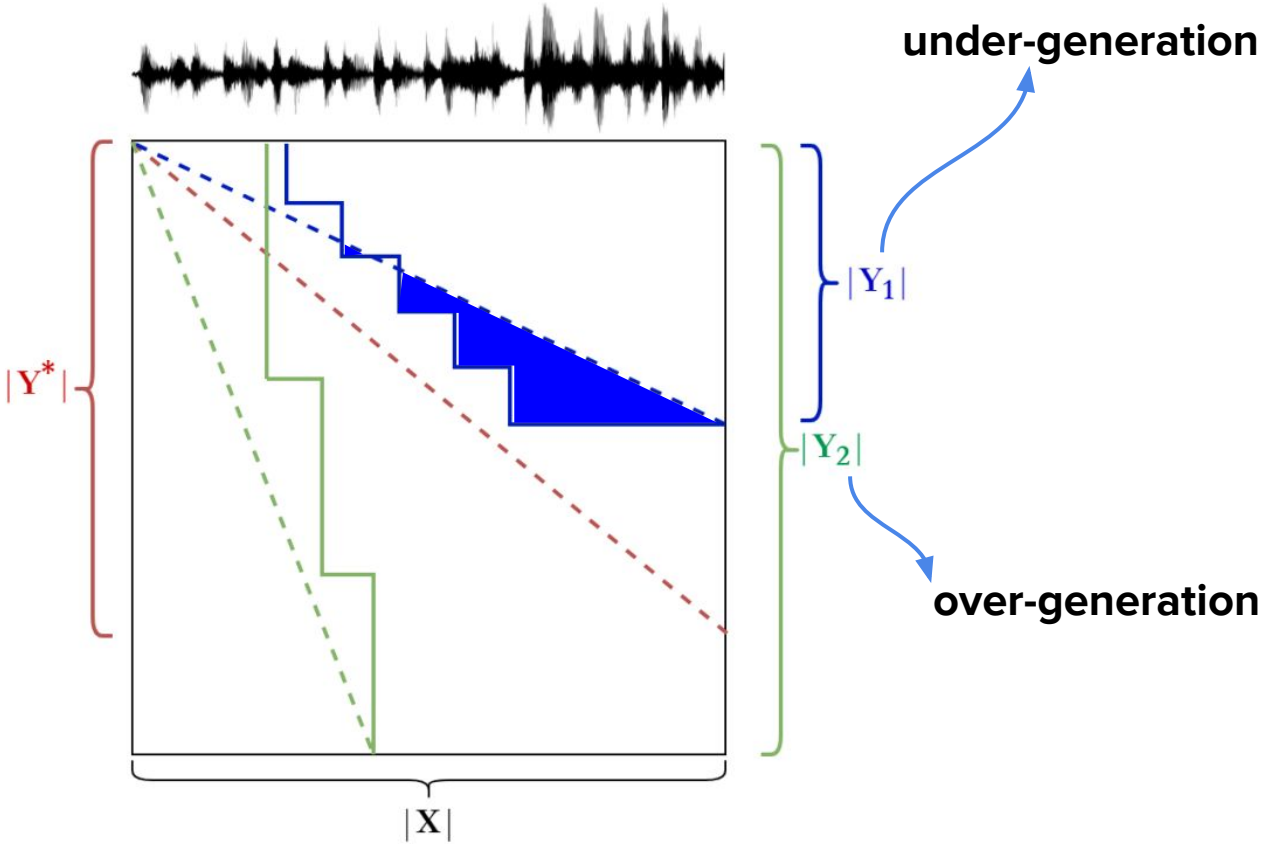
Length-Adaptive Average Lagging



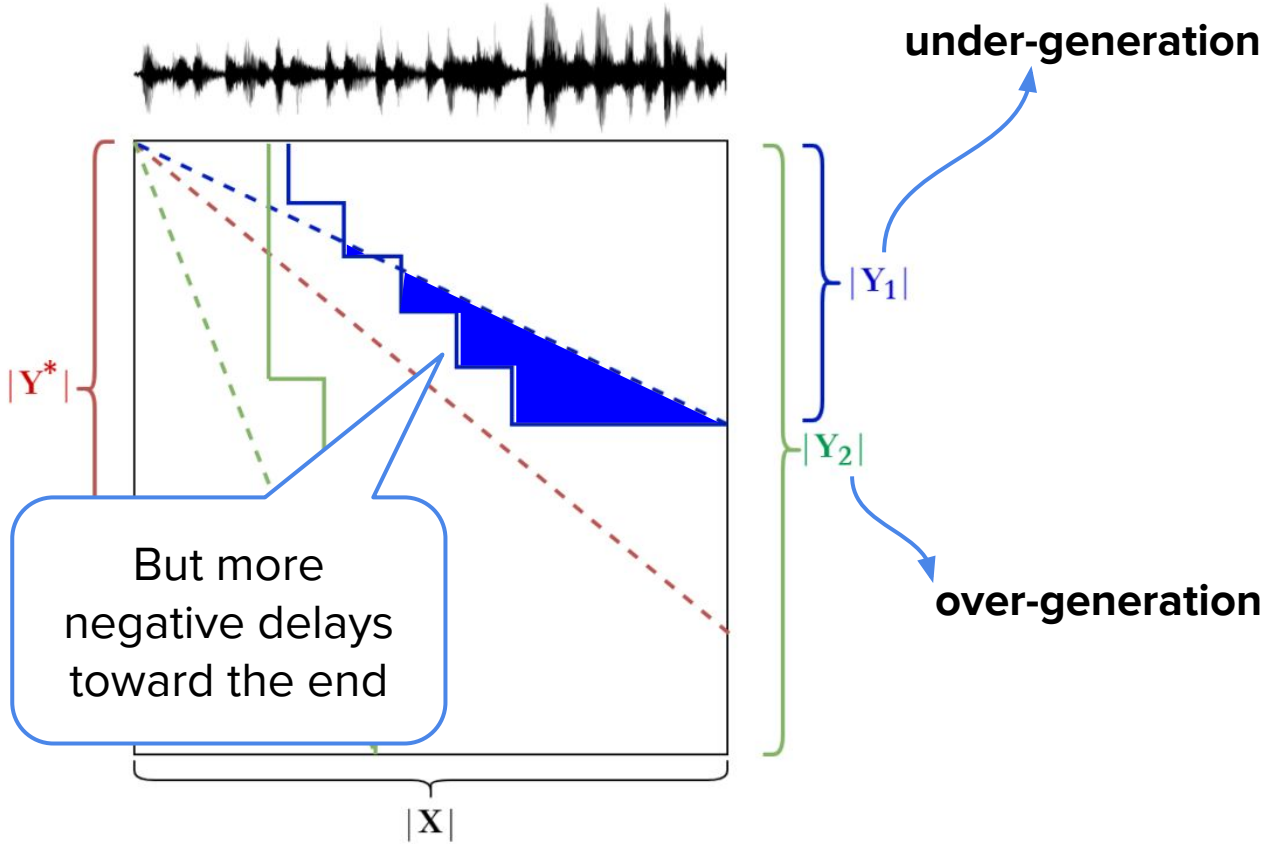
Length-Adaptive Average Lagging



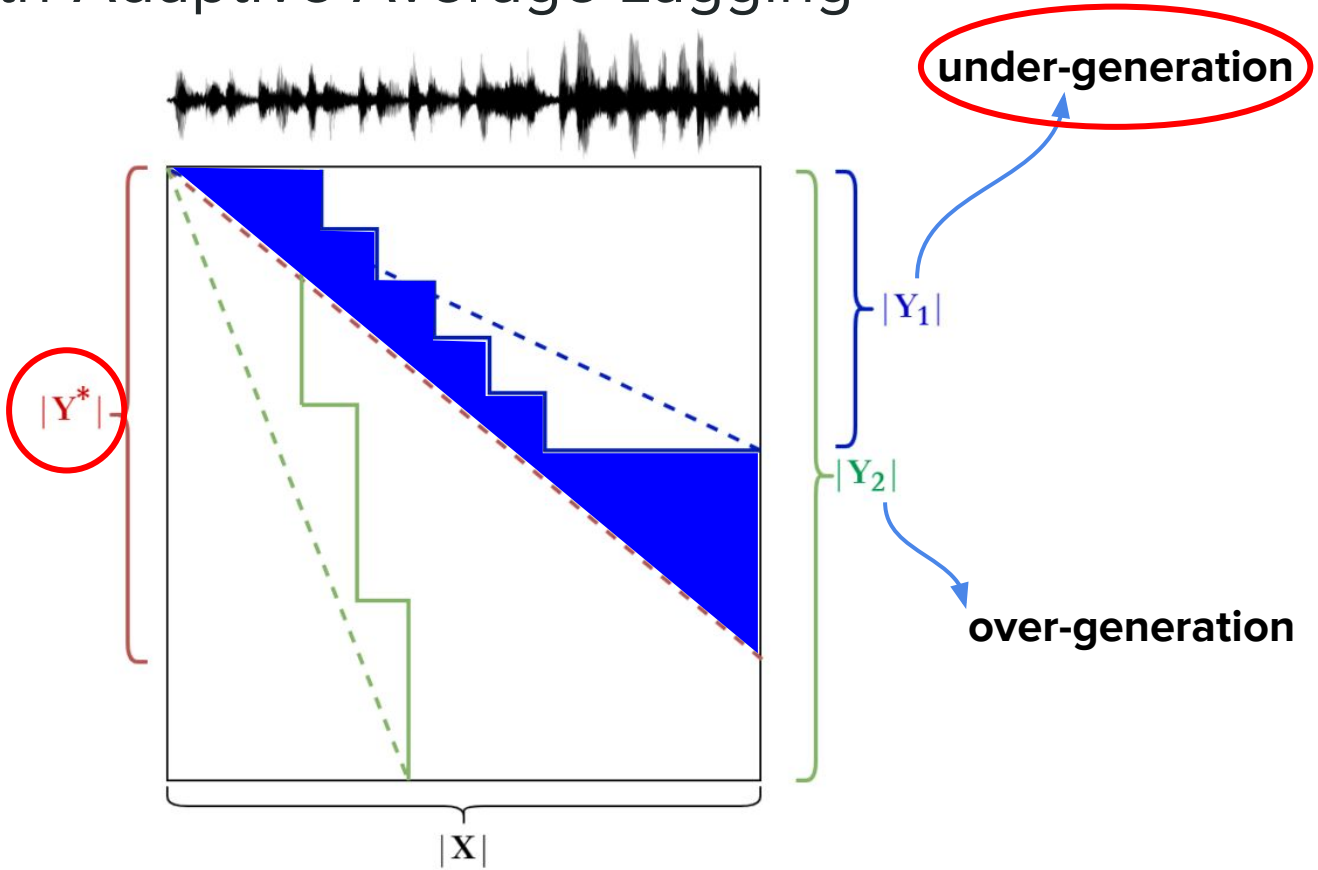
Length-Adaptive Average Lagging



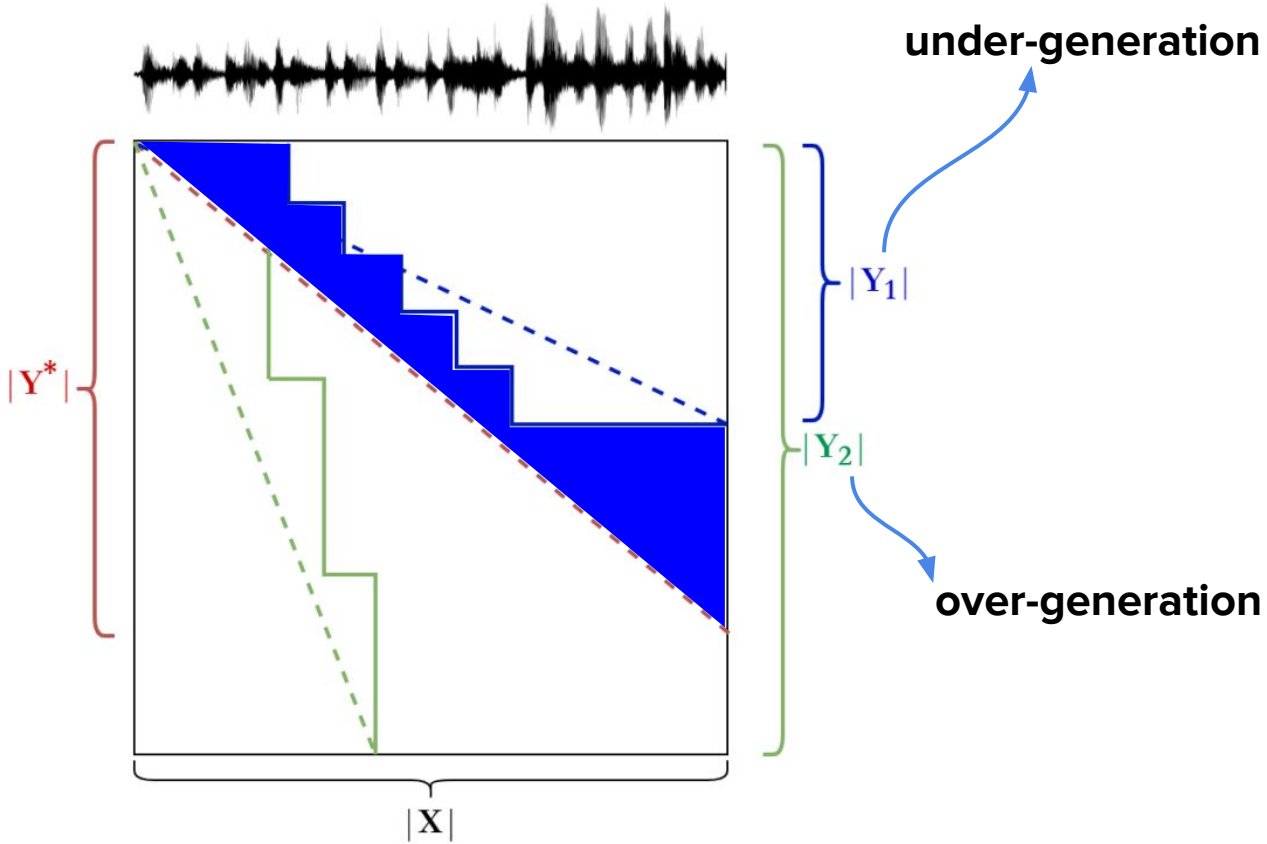
Length-Adaptive Average Lagging



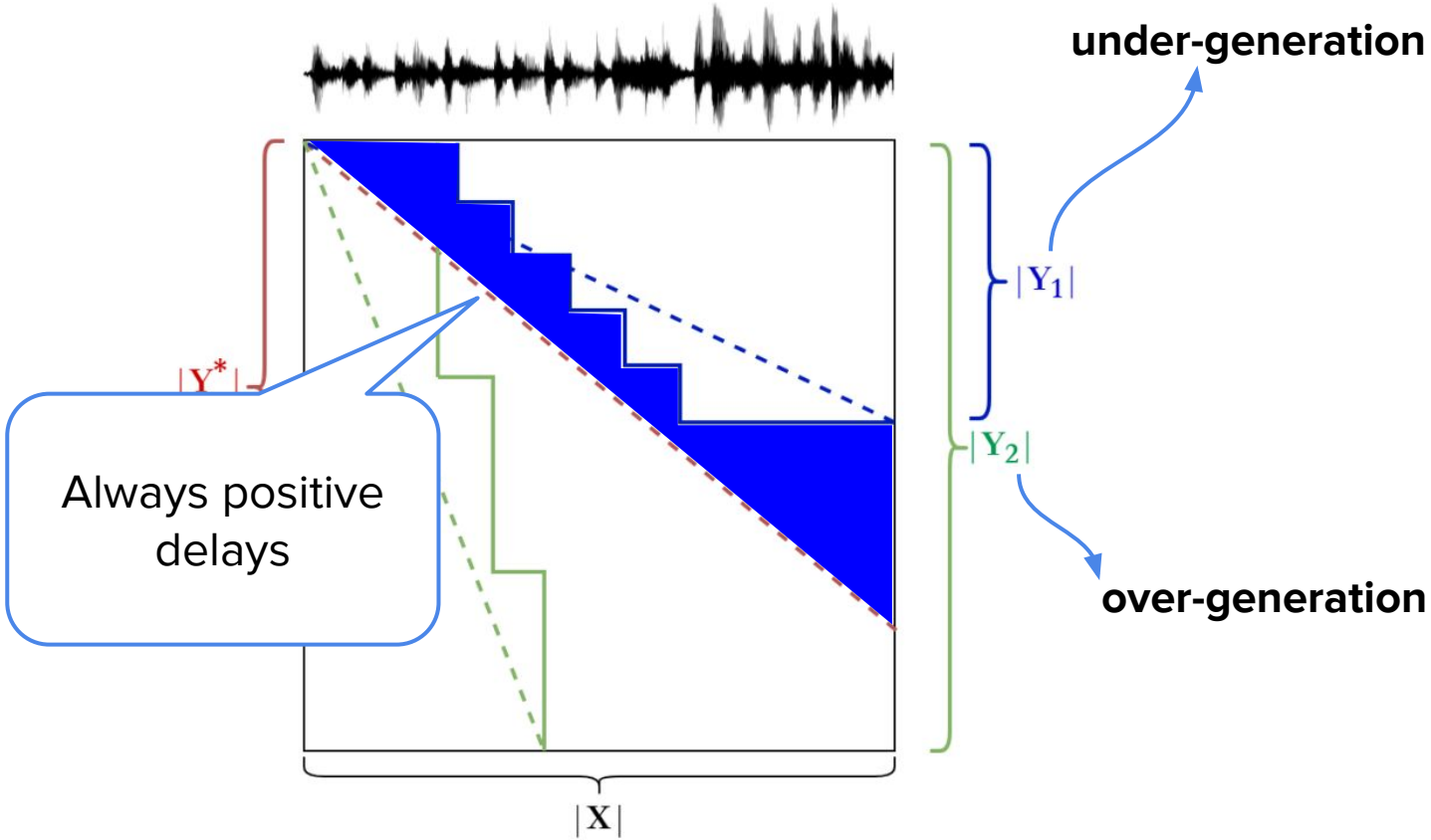
Length-Adaptive Average Lagging



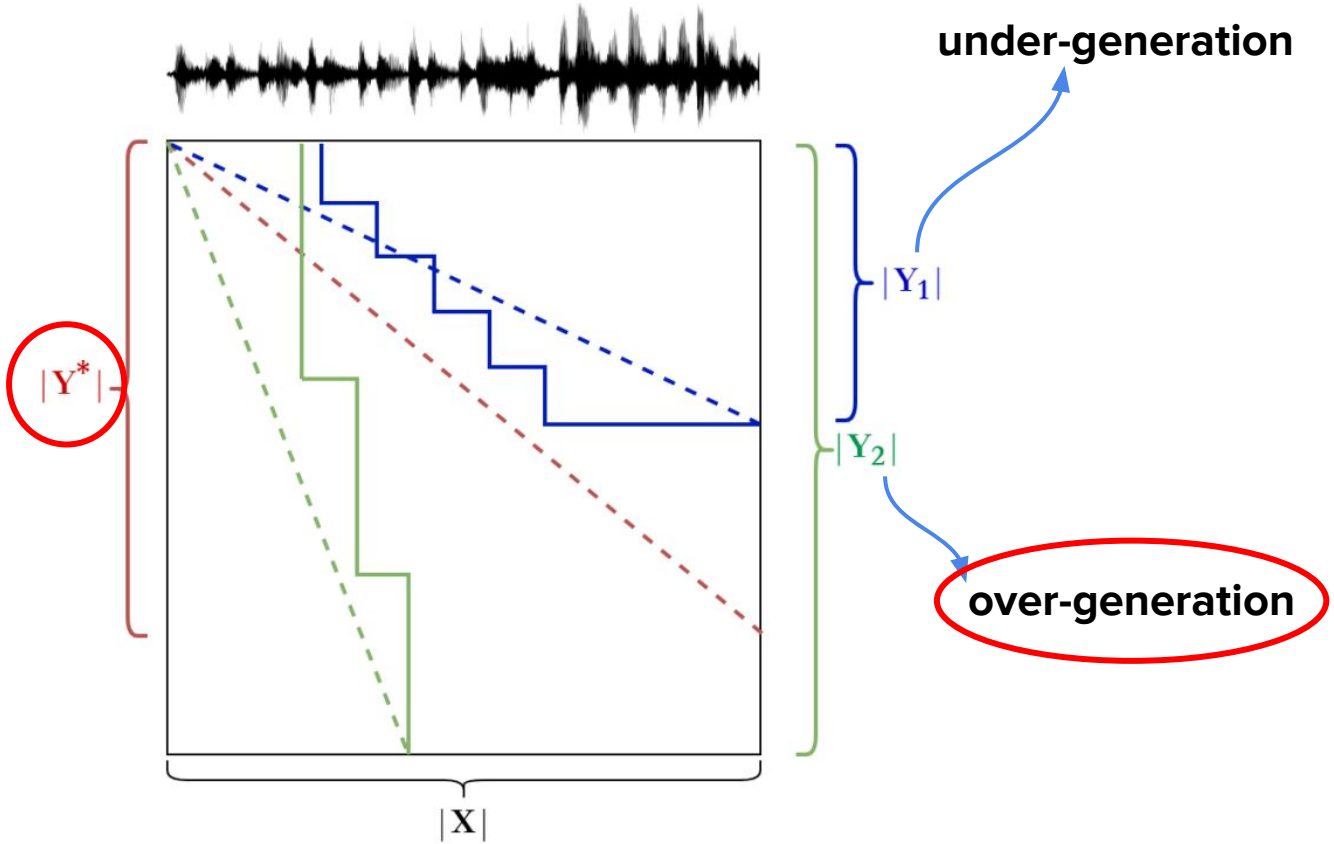
Length-Adaptive Average Lagging



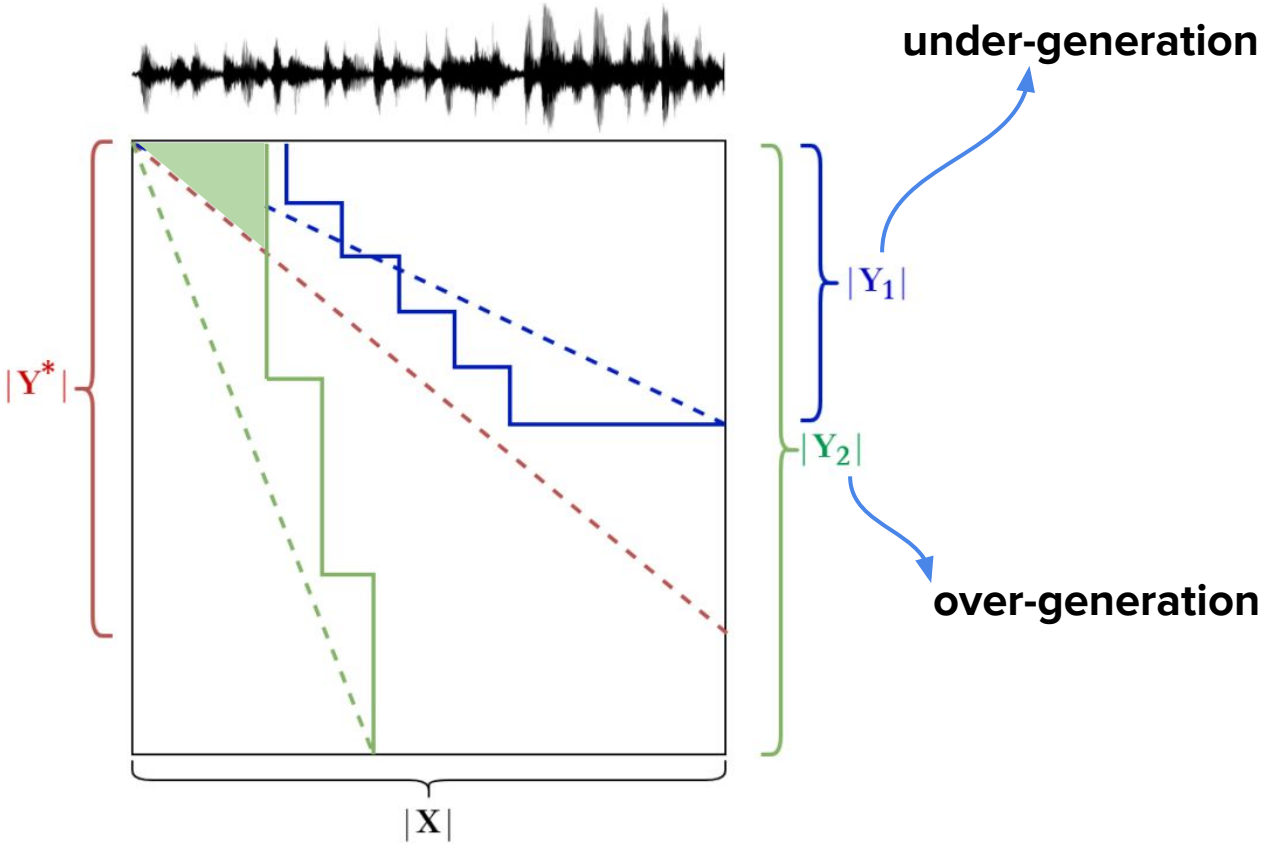
Length-Adaptive Average Lagging



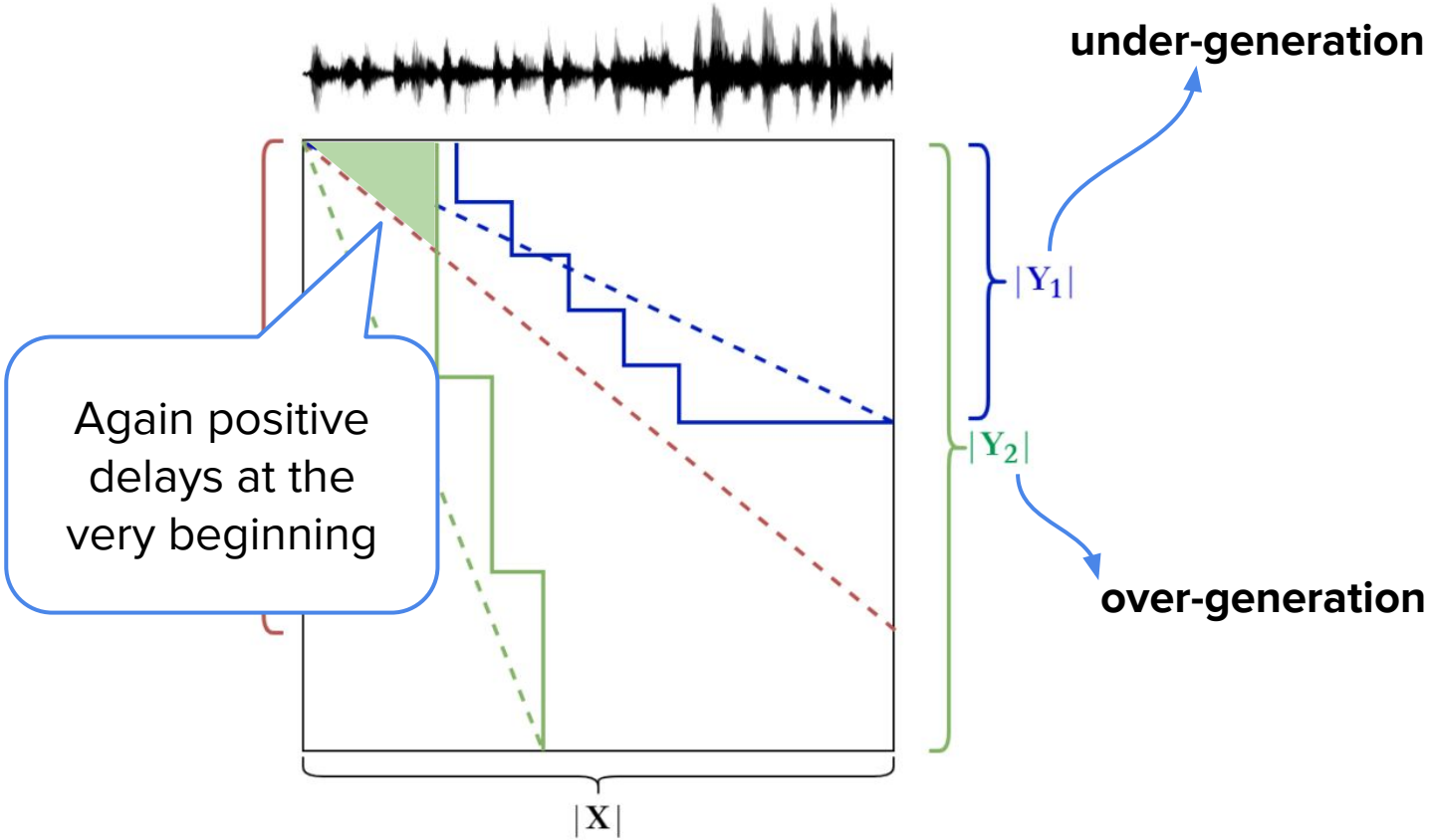
Length-Adaptive Average Lagging



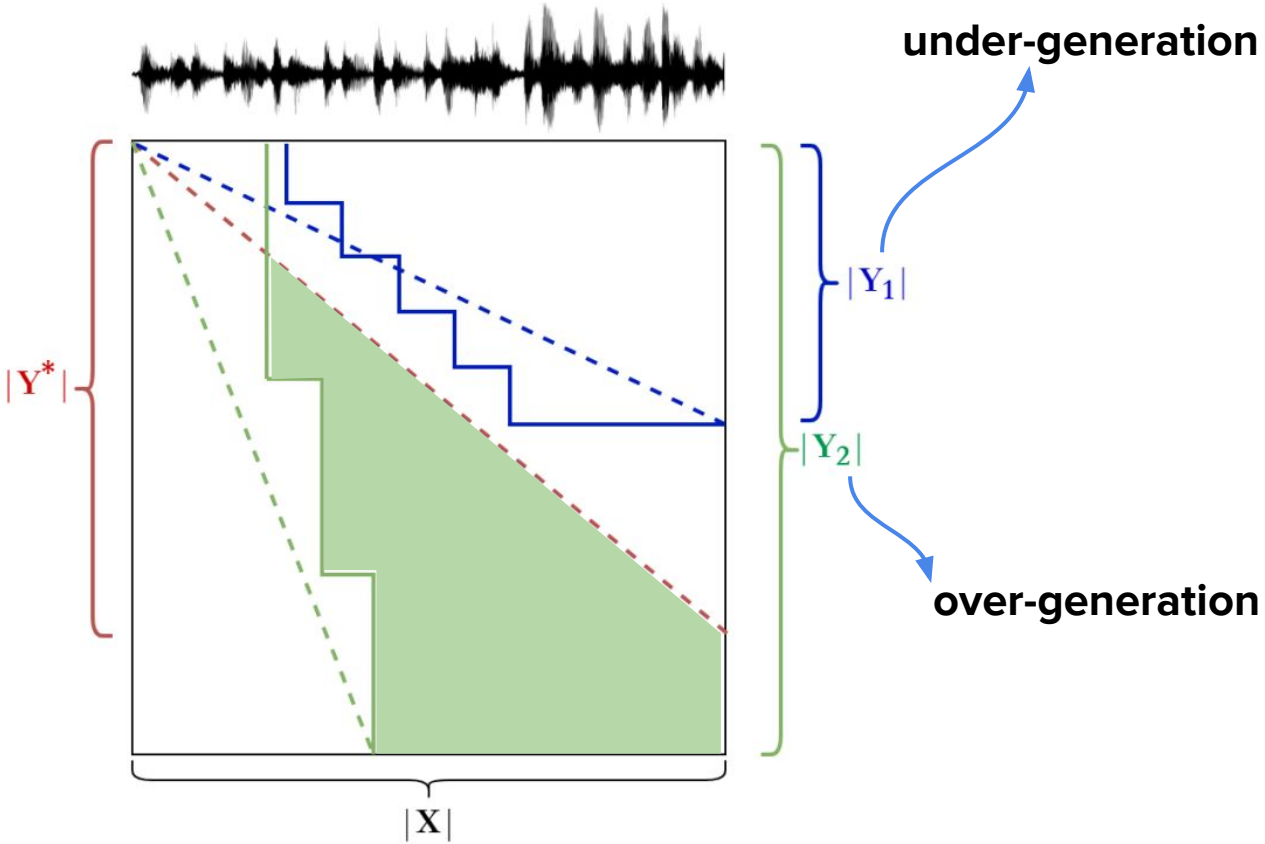
Length-Adaptive Average Lagging



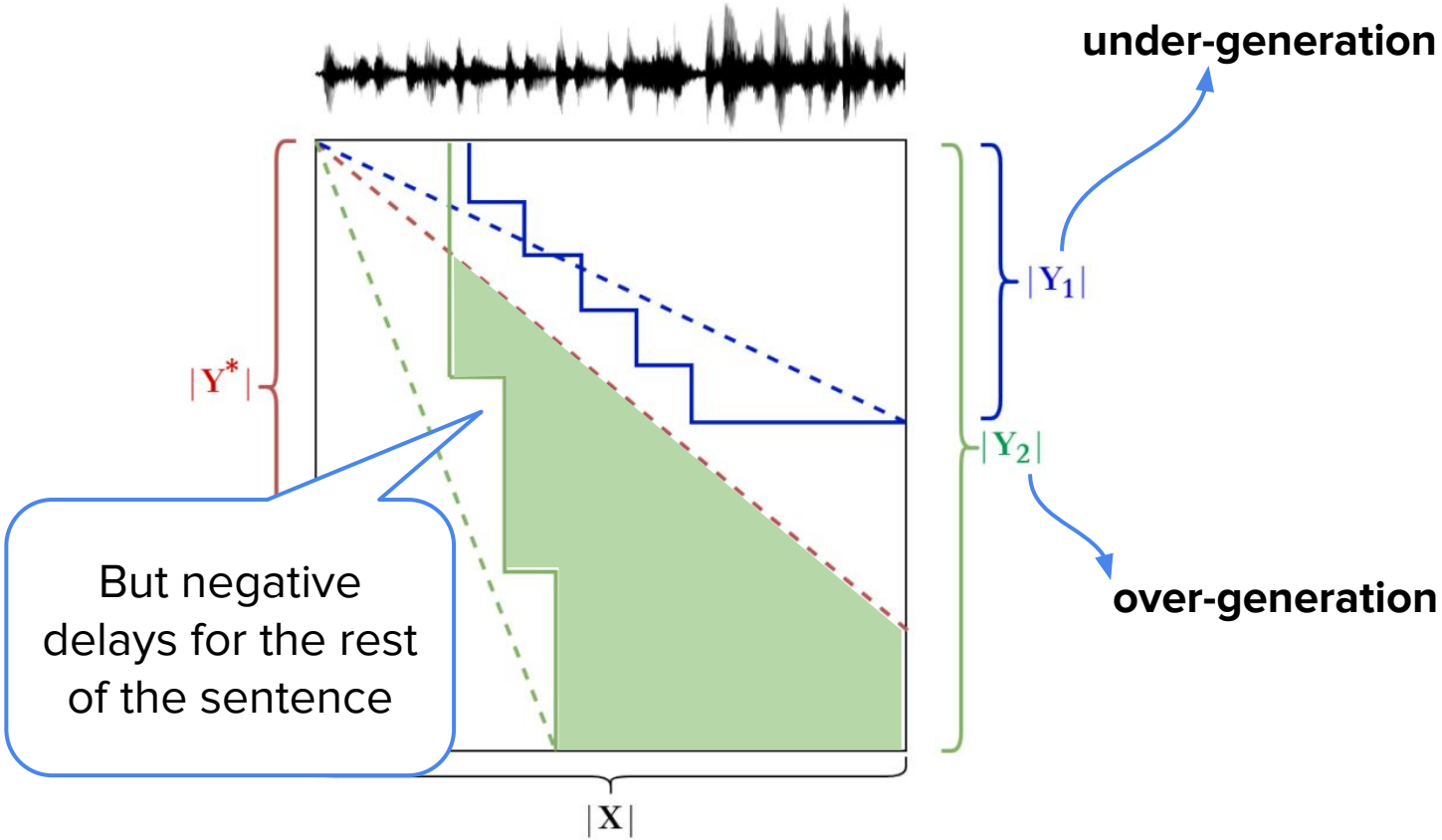
Length-Adaptive Average Lagging



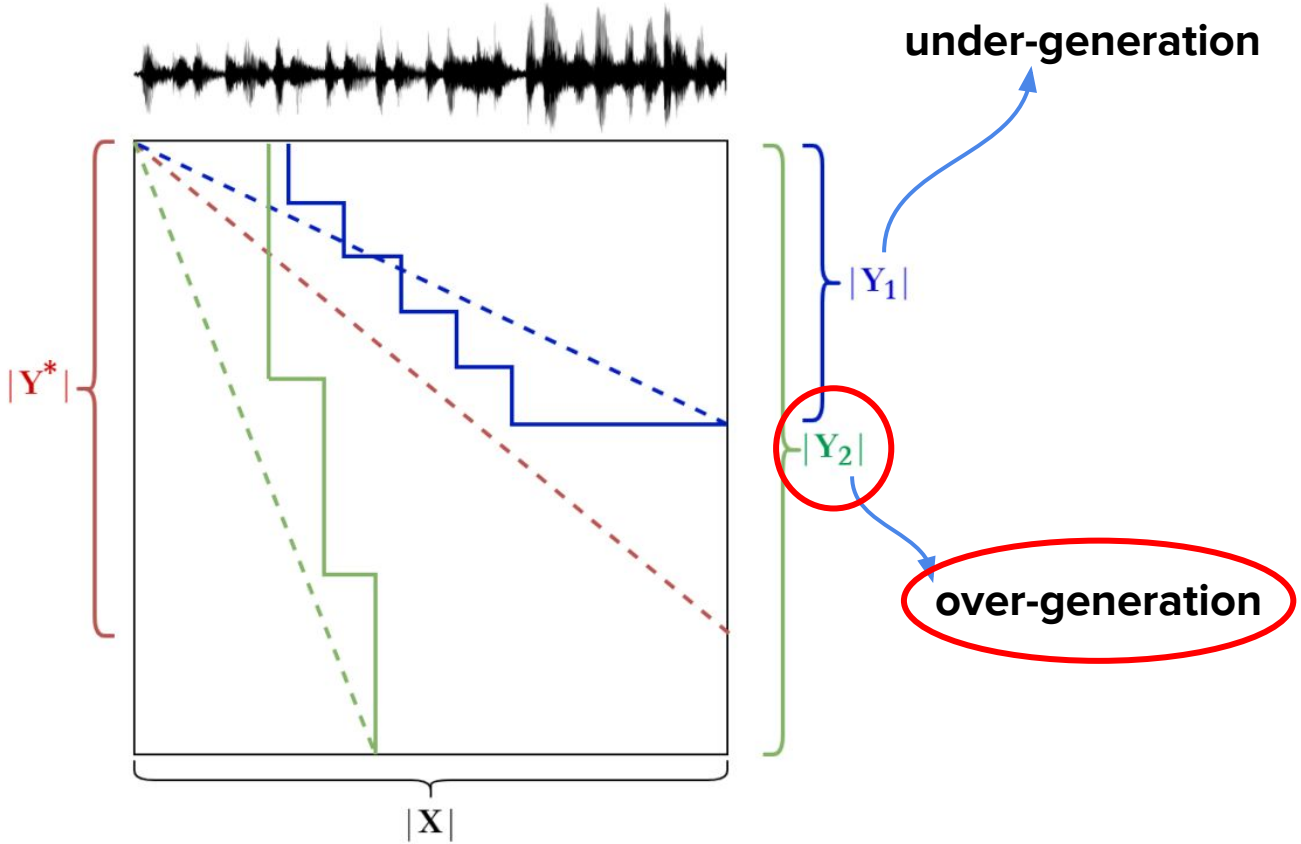
Length-Adaptive Average Lagging



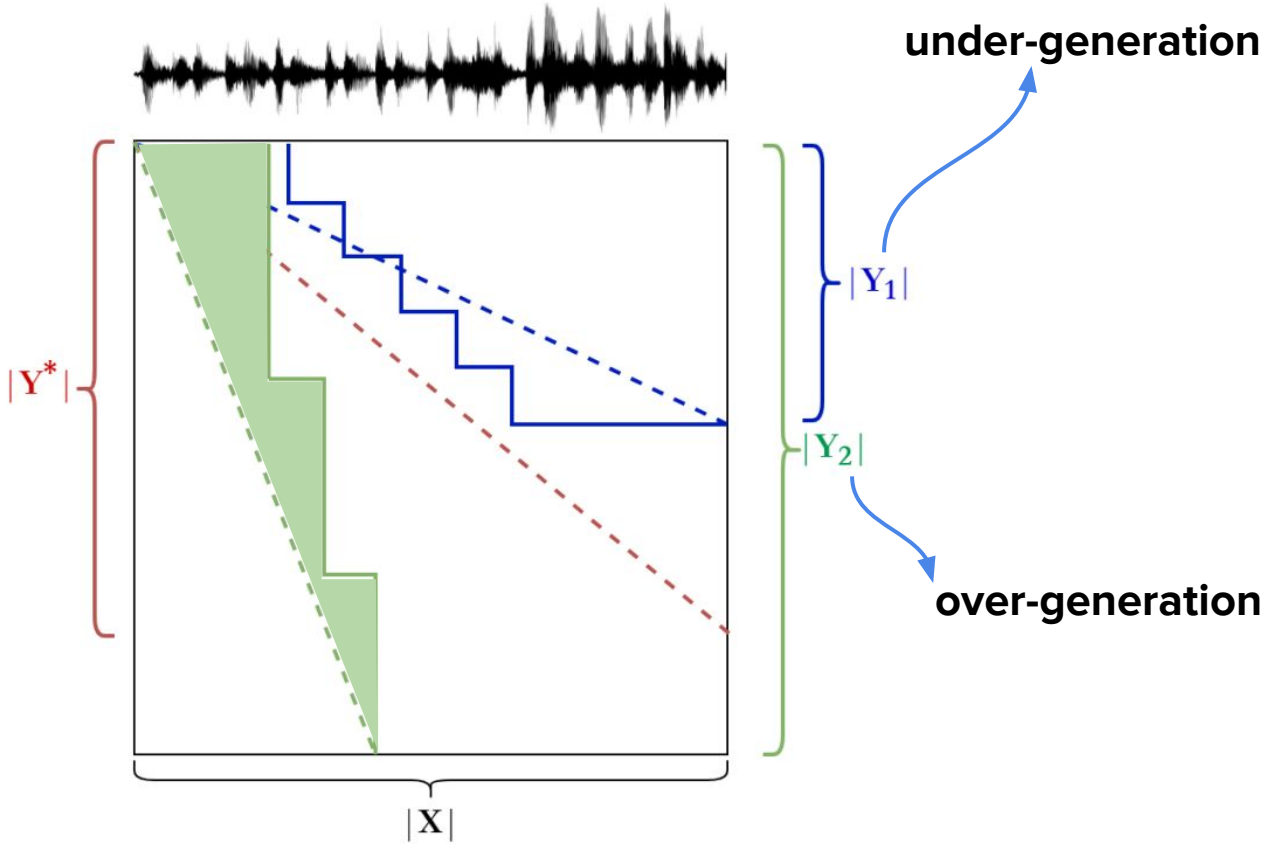
Length-Adaptive Average Lagging



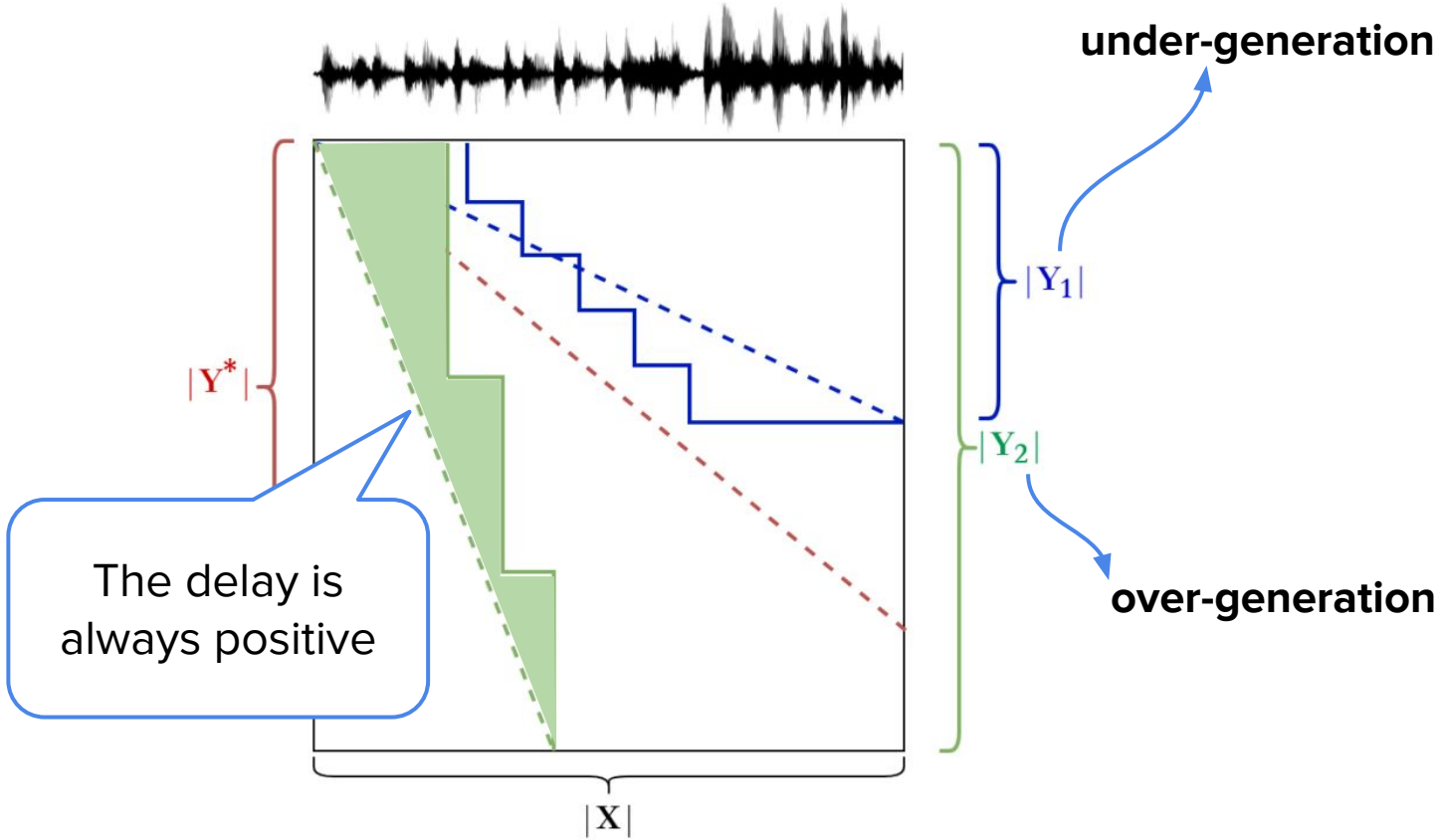
Length-Adaptive Average Lagging



Length-Adaptive Average Lagging



Length-Adaptive Average Lagging



LAAL: results

Model	Metric	k=3	k=5	k=7	k=9	k=11
wait-k	AL	1761	1970	2272	2582	2931
	LAAL	1778	2001	2332	2655	3003
offline wait-k	AL	1522	1959	2463	2926	3350
	LAAL	1682	2093	2588	3043	3457
CAAT	AL	735	1149	1533	1905	2265
	LAAL	1018	1365	1708	2046	2382

LAAL: results

Model	Metric	k=3	k=5	k=7	k=9	k=11
wait-k	AL	1761	1970	2272	2582	2931
	LAAL	1778	2001	2332	2655	3003
offline wait-k	AL	1522	19[illegible]	2463	2926	3350
	LAAL	1682				3457
CAAT	AL	735				2265
	LAAL	1018	1365	1708	2046	2382

Slight differences between AL and LAAL

LAAL: results

Model	Metric	k=3	k=5	k=7	k=9	k=11
wait-k	AL	1761	1970	2272	2582	2931
	LAAL	1778	2001	2332	2655	3003
offline wait-k	AL	1522	1959	2463	2926	3350
	LAAL	1682	2093	2588	3043	3457
CAAT	AL	735				2265
	LAAL	1018				2382

LAAL increases by ~100ms compared to AL

LAAL: results

Model	Metric	k=3	k=5	k=7	k=9	k=11
wait-k	AL	1761	1970	2272	2582	2931
	LAAL	1778	2001			
offline wait-k	AL	1522	1959			
	LAAL	1682	2093			
CAAT	AL	735	1149	1533	1905	2265
	LAAL	1018	1365	1708	2046	2382

LAAL is at least 100ms higher compared to AL

LAAL: results

Model	Metric	k=3	k=5	k=7	k=9	k=11
wait-k	AL	1761	1970	2272	2582	2931
	LAAL	1778	2001	2332	2655	3003
offline wait-k	AL	1522				3350
	LAAL	1682				3457
CAAT	AL	735				2265
	LAAL	1018	1365	1708	2046	2382

LAAL latency increases up to ~300ms at very low latency

In Conclusion

- Current SimulST systems evaluation metrics do not take into account over-generation
- The problem is quite frequent in the output of current SimulST systems
- Our proposed metric, Length-Adaptive Average Lagging, takes into account for both under- and over-generation phenomena at sentence level
- Our experiments show that LAAL gives a more precise measure of latency compared to AL
- **Use LAAL for more reliable reliable system evaluations!**

What’s next?

The ideal policy formulation of AL/LAAL assumes that:

- only one word at a time is emitted
- the words are equally distributed in the speech

➔ These strong assumptions are not valid in general and imply that the obtained evaluation could be unreliable

➔ This should be considered in future studies for even better latency estimates

UNIVERSITÀ DI TRENTO

FONDAZIONE BRUNO KESSLER

Thanks for your attention!

Over-generation cannot be rewarded: Length-Adaptive Average Lagging

Sara Papi, *Marco Gaido, Matteo Negri, Marco Turchi*

{spapi,mgaido,negri,turchi}@fbk.eu