# Self-adaptive Incremental Machine Speech Chain for Lombard TTS with High-granularity ASR Feedback in Dynamic Noise Condition

Sashi Novitasari[1], Sakriani Sakti[2,1], and Satoshi Nakamura[1]

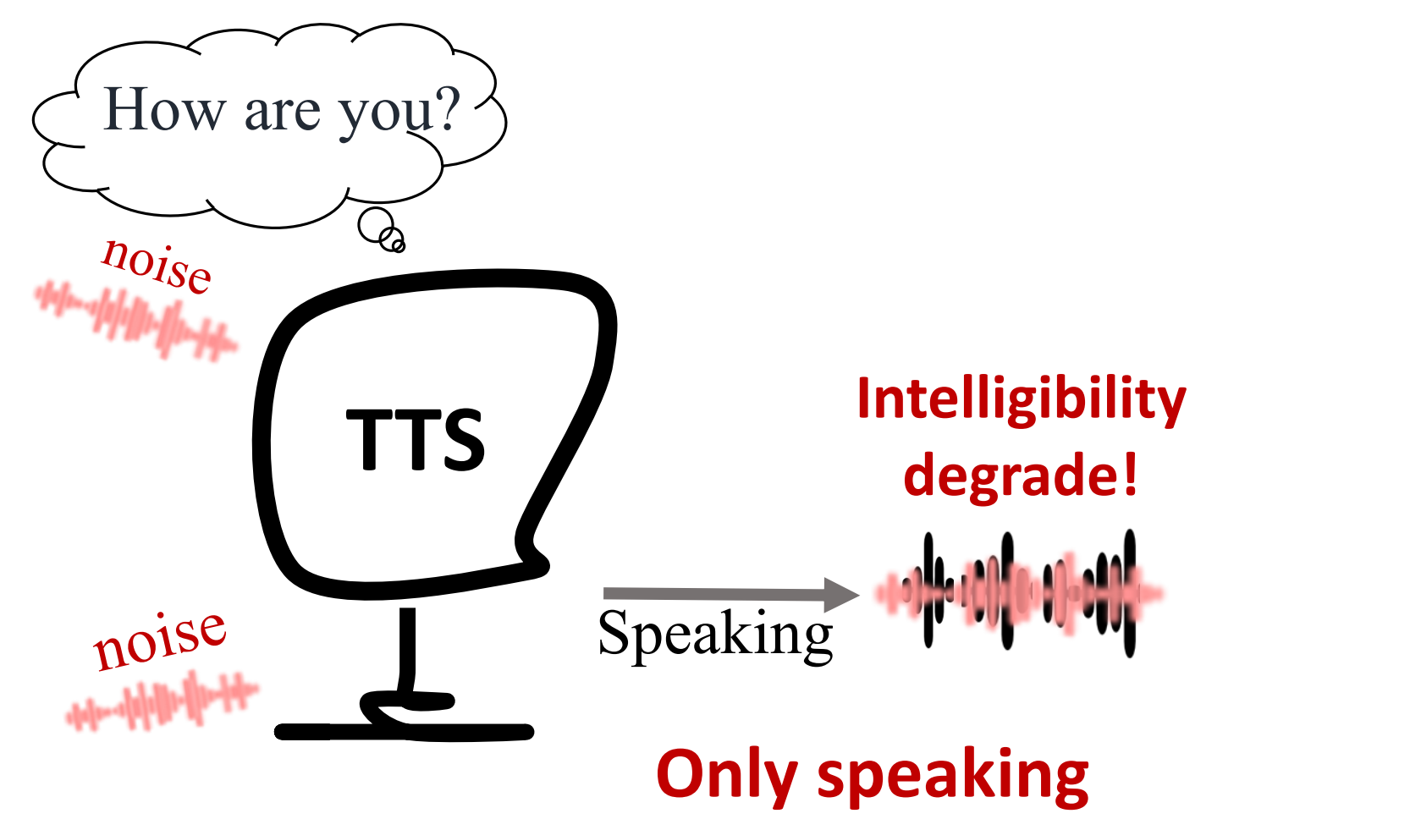
[1]Nara Institute of Science and Technology, Japan; [2]Japan Advanced Institute of Science and Technology, Japan

ICASSP 2023
4 - 10 JUNE, RHODES ISLAND, GREECE

## I. BACKGROUND

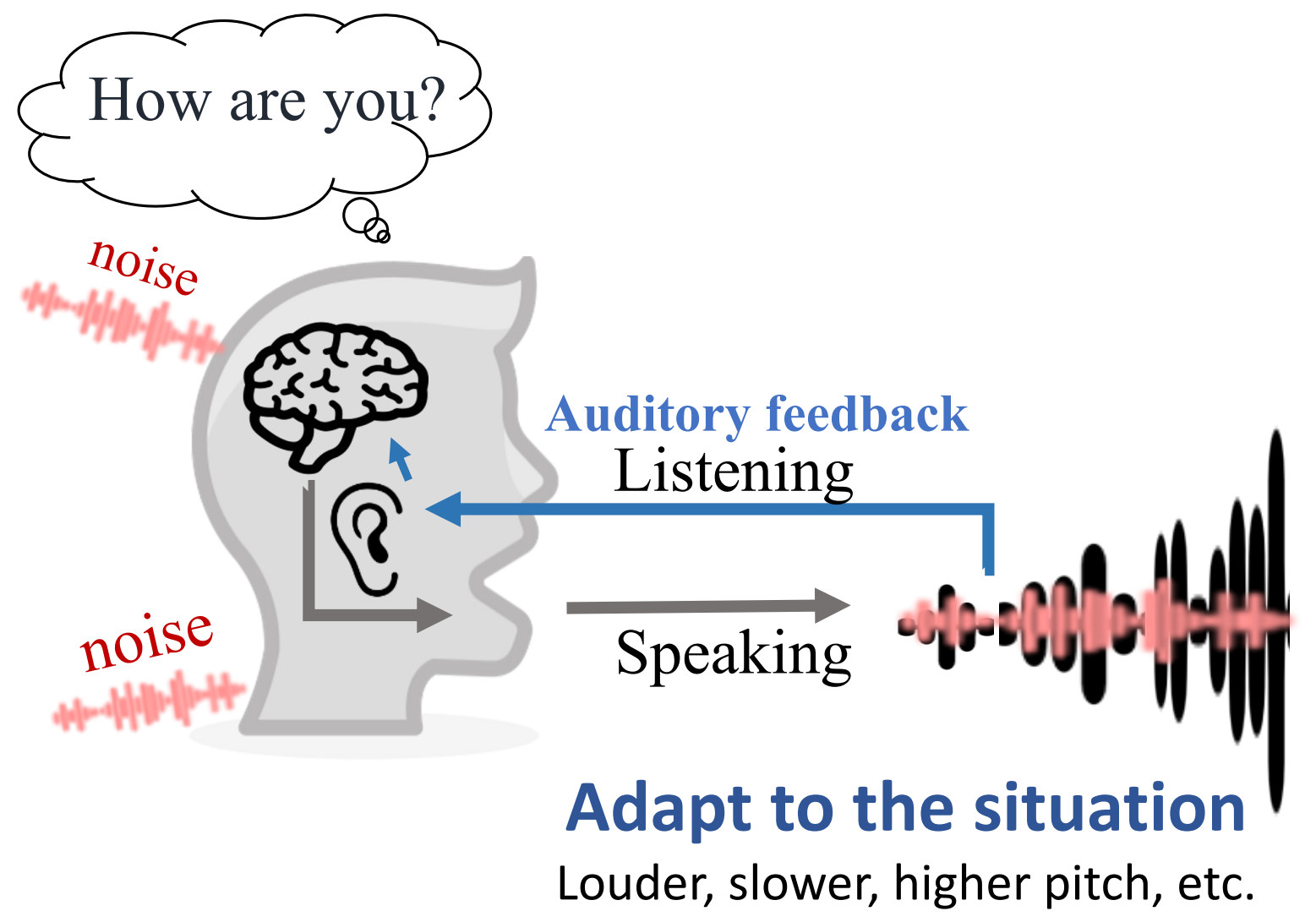
### A. TTS in noisy place

**Standard TTS**



- Noise degrades the TTS speech intelligibility
- TTS is commonly trained using speech data from the clean environment
- Common solution: Fine-tune the TTS using Lombard speech data (static noise) [Raitio et al., 2011; Paul et al., 2020]

**Human**



- **Speech chain**. Speaking while listening [Denes and Pinson, 1993]
- **Lombard effect**. Speaking effort adjustment in noisy place to improve the speech intelligibility

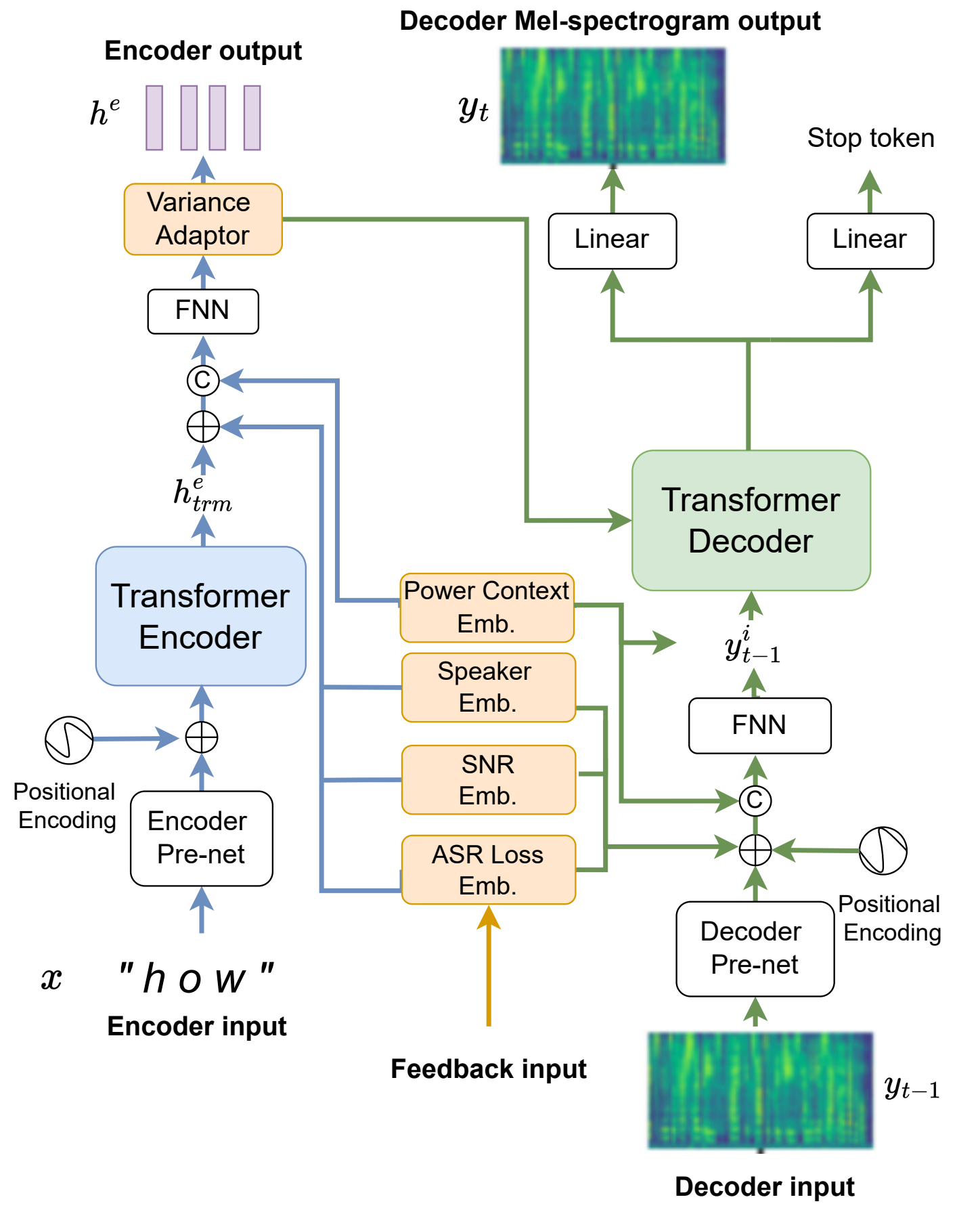
**TTS limitation:**
- No auditory feedback mechanism
- Cannot self-adapt to noisy situation

### B. Adapt-ITTS: Self-adaptive incremental TTS with machine speech chain mechanism [Novitasari et al., 2022]

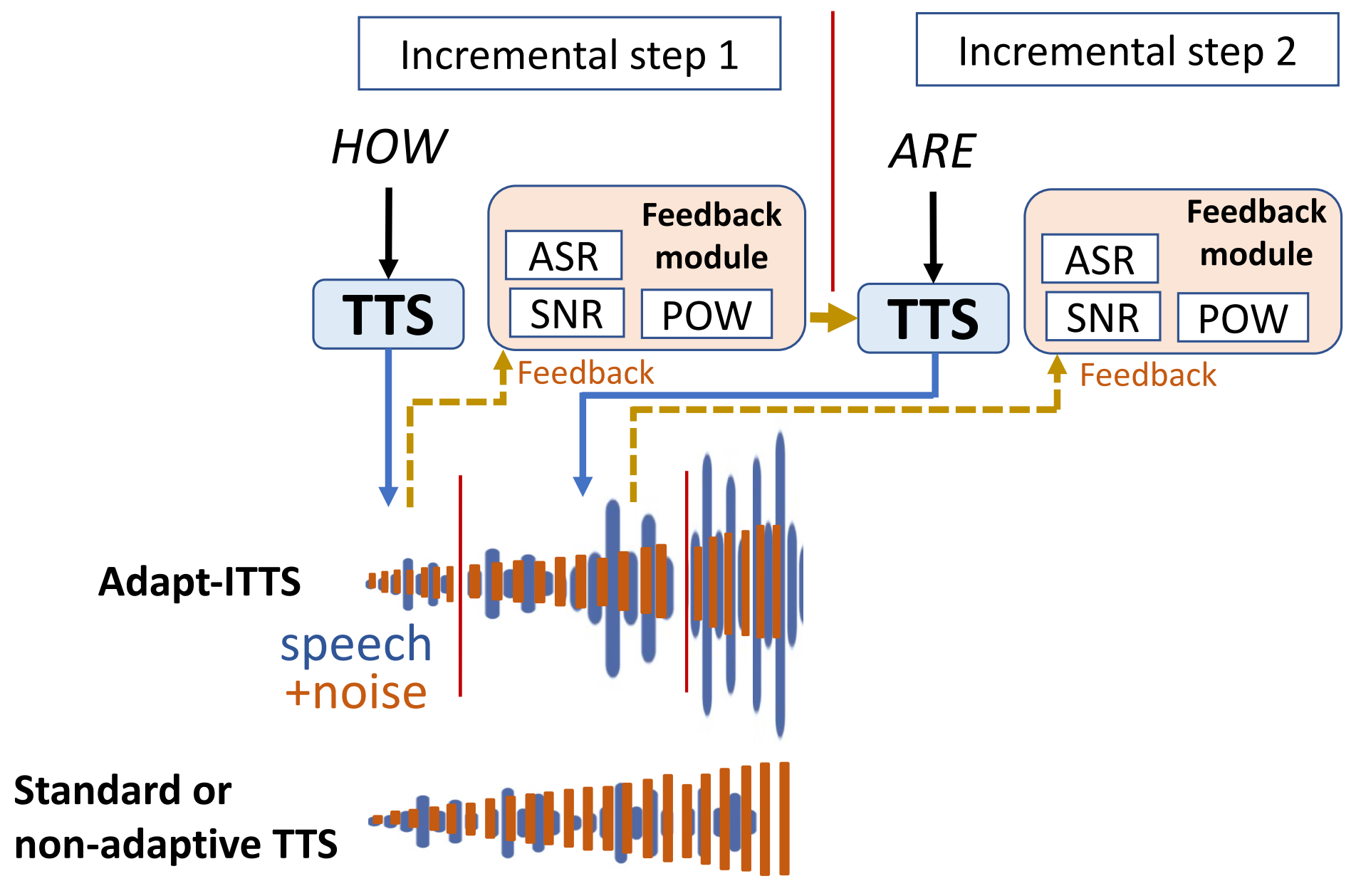
**End-to-end incremental TTS (ITTS)** that adapts the speaking style using the auditory feedback based on the prev. synthesized speech + noise

Autoregressive Transformer-based TTS with variance adaptor and feedback modules:
- ASR loss, based on the noisy synth. speech
- SNR, speech-to-noise ratio
- POW, synth. speech power

**Architecture**



**Mechanism** (incremental)



**Limitation:**
- Low-granularity ASR loss feedback only based on positive character class (limited information)
- Training based on synthetic Lombard data

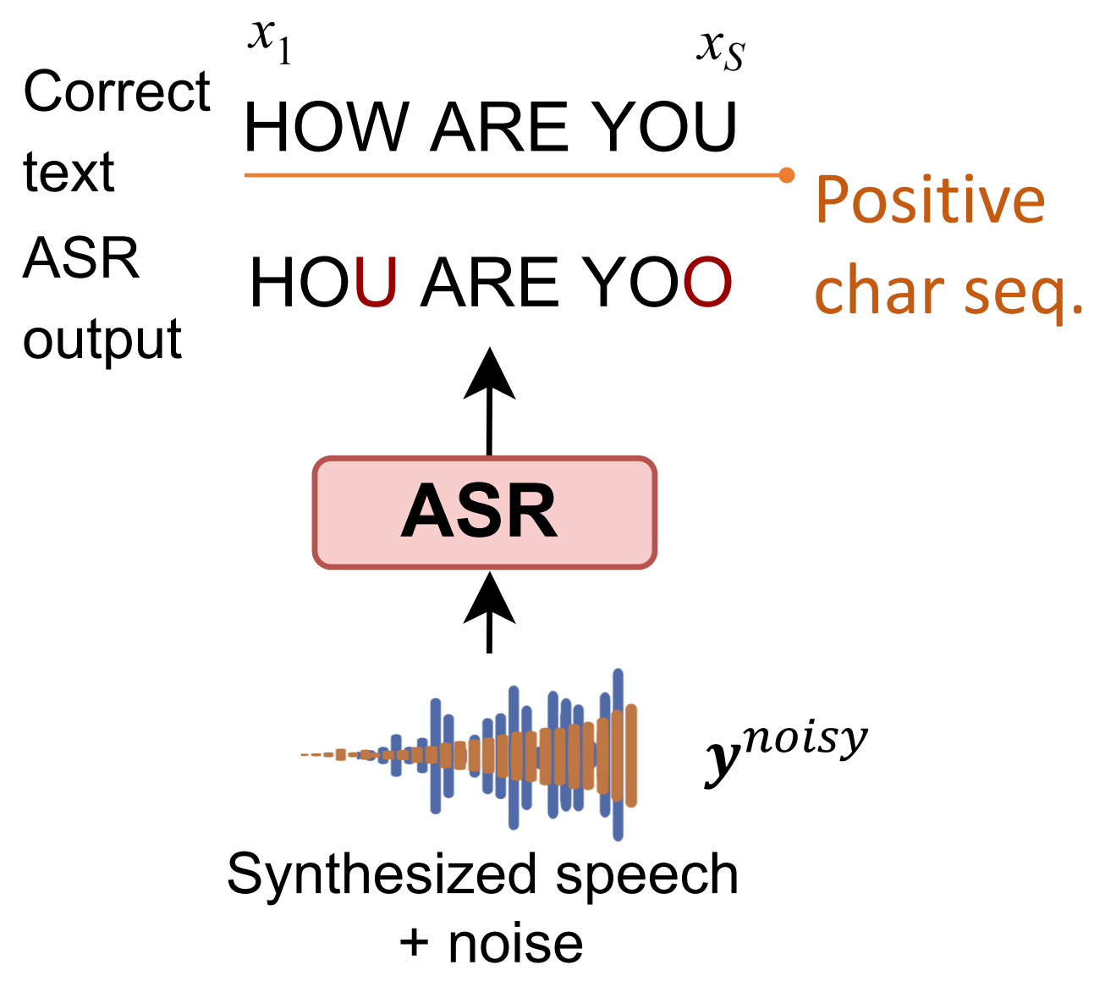
## II. PROPOSED METHOD

### Adapt-ITTS with high-granularity ASR feedback

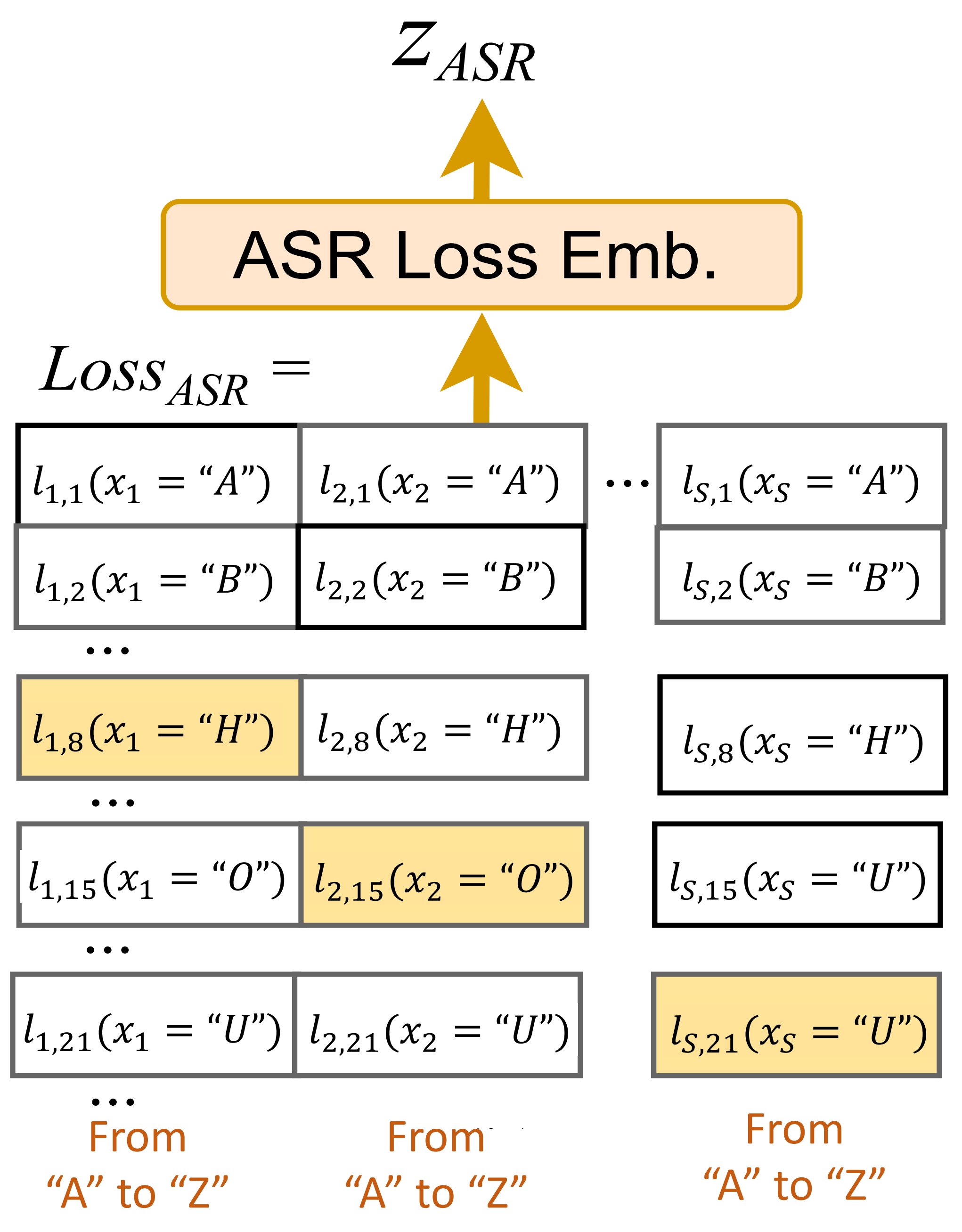
**Improve the Adapt-TTS by enriching the ASR auditory feedback information**

- For each incremental step, use the character-vocabulary level ASR feedback based on the losses of the positive and negative classes
- ASR feedback is an ASR loss embedding ($z_{ASR}$)
- Character-level ASR loss
- Generated by transcribing noisy TTS speech using an ASR

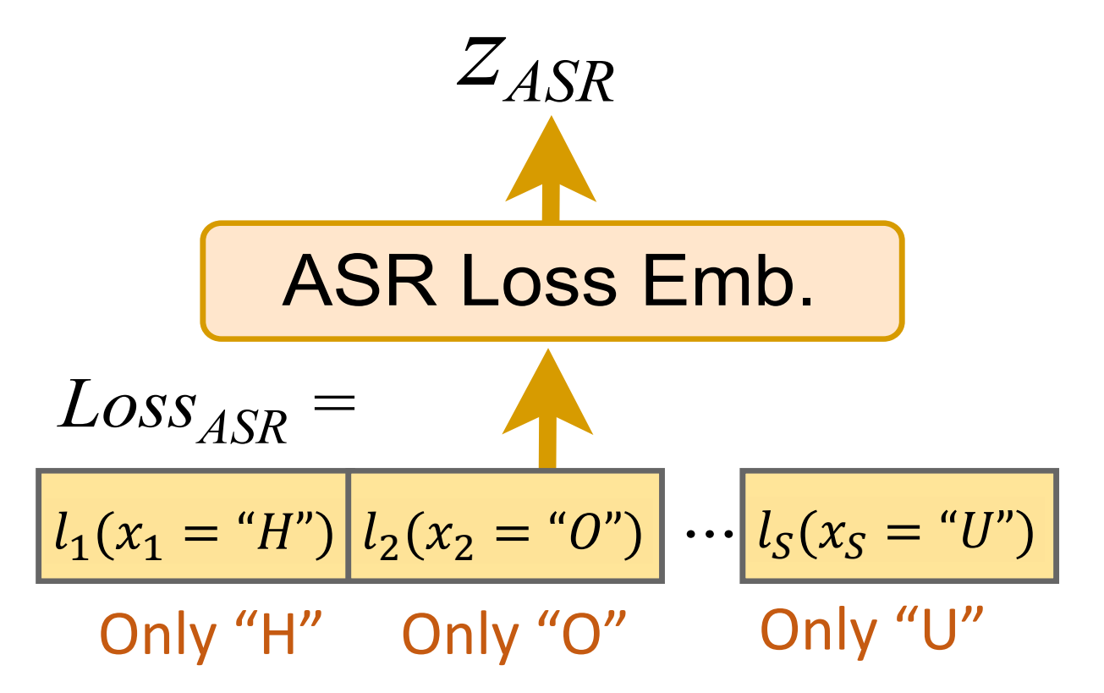
$$z_{ASR} = ASR\ Loss\ Embedding\ (Loss_{ASR}(\boldsymbol{x}, \boldsymbol{p_x}))$$
$$\boldsymbol{p_x} = p_{ASR}(\boldsymbol{x}|\boldsymbol{y}^{noisy})$$



### Proposed ASR feedback generation method



Low-granularity ASR feedback [Previous work]



- MCE loss for each character timestep ($s$)
- For positive character class only

$$l_s(x_s, p_{x_s}) = -\sum_{c=1}^{C} \mathbb{1}(x_s = c) * \log p_{x_s}[c]$$

🟨 Positive character class ($x_S^c = 1$)
□ Negative character class ($x_S^c = 0$)

- BCE loss for each character ($c$) in dictionary for each character timestep($s$)

$$l_{s,c}(x_S^c, p_{x_s}) = -(x_S^c * \log p_{x_s}[c] + (1 - x_S^c) * \log(1 - p_{x_s}[c]))$$

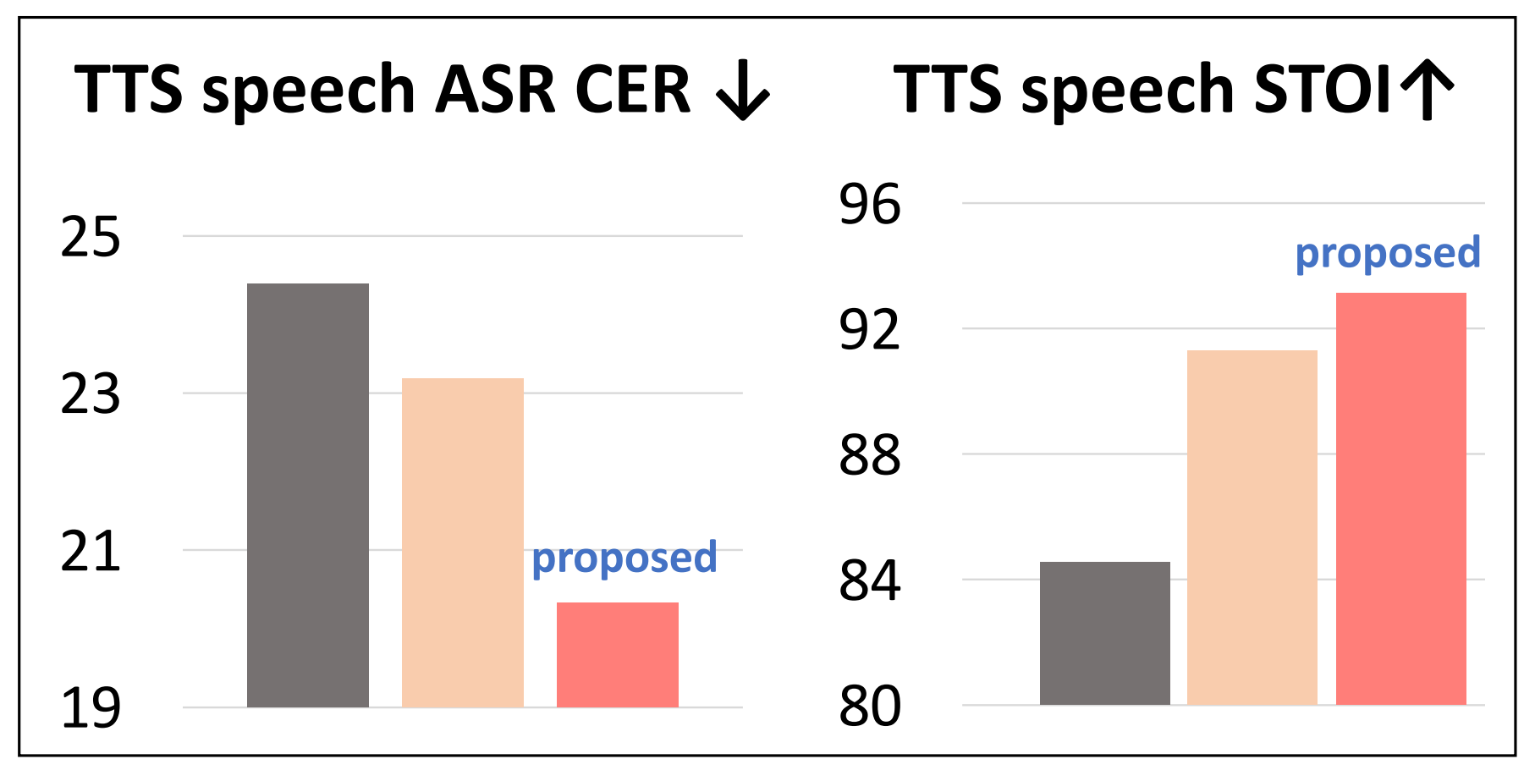
## III. EXPERIMENT

### A. Setting

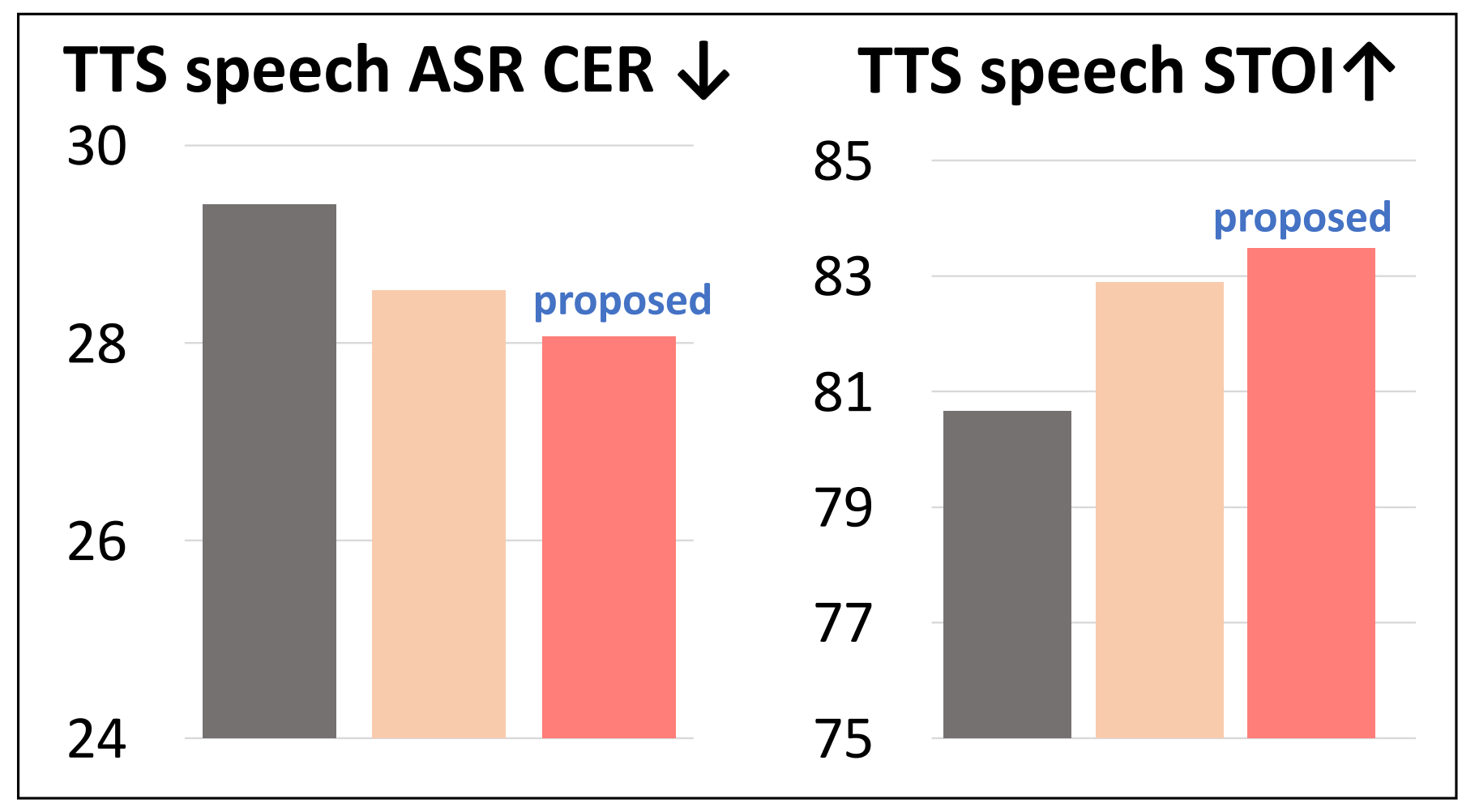
- Training data:
  - WSJ [Paul & Baker, 1992]: Natural normal speech + synthetic Lombard speech (multi-speaker)
  - Hurricane [Cooke et al., 2013]: Natural normal speech + natural Lombard speech (single speaker)
- Architecture: Autoregressive transformer + variance adaptor + feedback modules

### B. Result: TTS speech intelligibility in noisy situation

**WSJ**



**Hurricane**



**TTS speech intelligibility measure:**
- ASR CER (%) : Character error rate.
- STOI (%) : Short-term objective intelligibility measure.

**Noise:**
- Babble noise and white noise
- SNR: -10 dB, 0 dB, clean (no noise)

**Model:**
**Baseline**
▨ Finetune-TTS

**Adapt-ITTS**
🟧 With ASR feedback: low granularity
🟥 **With ASR feedback: high granularity (proposed)**

**The proposed high-granularity ASR feedback improved the incremental TTS speech intelligibility**

## IV. CONCLUSION

**Adapt-ITTS with the high-granulated ASR feedback for the self-adaptive speech synthesis in noisy conditions**

- Adapt-ITTS adapts the speech style based on noise conditions
- Short-term feedback in an incremental mechanism
- The proposed ASR feedback improved Adapt-ITTS intelligibility in noisy conditions

**Scan for speech samples**

or

https://sites.google.com/view/adapt-lombard-tts/home