AAAI Workshop on Self-supervised Learning for Audio and Speech Processing
(AAAI SAS 2022)

Self-Adaptive Machine Speech Chain in Noisy Environment

Sakriani Sakti

Associate Professor of
Japan Advanced Institute of Science and Technology (JAIST), Japan
Adjunct Associate Professor of
Nara Institute of Science and Technology (NAIST), Japan
Visiting Research Scientist of RIKEN Center for
Advanced Intelligence Project AIP (RIKEN AIP), Japan

Co-Authors:

Andros Tjandra, Johanes Effendi, Sashi Novitasari, Satoshi Nakamura

(NAIST/RIKEN AIP, Japan)

Self-Supervised Learning

Self-Supervised Learning

- **A trend in the machine learning community:**
  - → Adopt self-supervised approaches to pre-train deep networks.
  - → Refer to specific techniques that learn general representations given a large amount of unlabeled data
  - → A portion of the input is used as a supervisory signal to predict the remaining portion of the input
  - → Utilized the learned representations to improve performance on a downstream task (i.e., speech recognition)
- **Some well-known approaches**
  - → CPC [Oord et al., 2018], APC [Chung et al., 2020]
  - → wav2vec [Schneider et al. 2019], wav2vec 2.0 [Baevski et al., 2020]
  - → HuBERT [Hsu et al., 2021], W2V BERT [Chung et al., 2021]

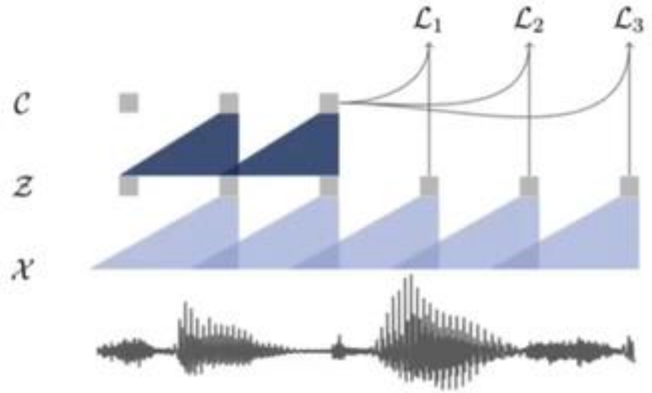


Figure from [Schneider et al. 2019]

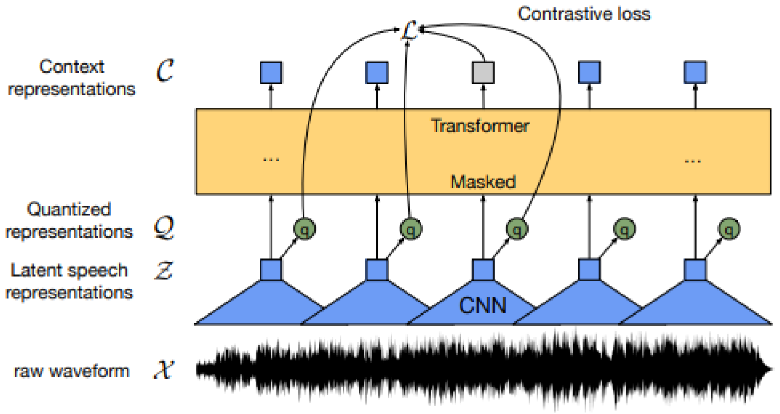


Figure from [Baevski et al., 2020]

Self-Supervised Learning

- **Objective of SSL:**

”Getting AI to Learn Like a Baby based on observation of its environment and interaction with people."

[Source: AITrends]

“Babies learn their first language through listening, talking, and interacting with adults. Can AI achieve the same goal without much low-level supervision?”

[Source: AAAI SAS 2022]

"Move the field of artificial intelligence beyond predictions and pattern-matching and toward machines that think like humans.“
“Need systems that can handle environment changes and do continual learning, lifelong learning"

[Bengio’s talk at Neurips 2019]

A self-adaptive machine that can handle environmental changes

Human Language Learning and Communication

From Baby Babble Into Language

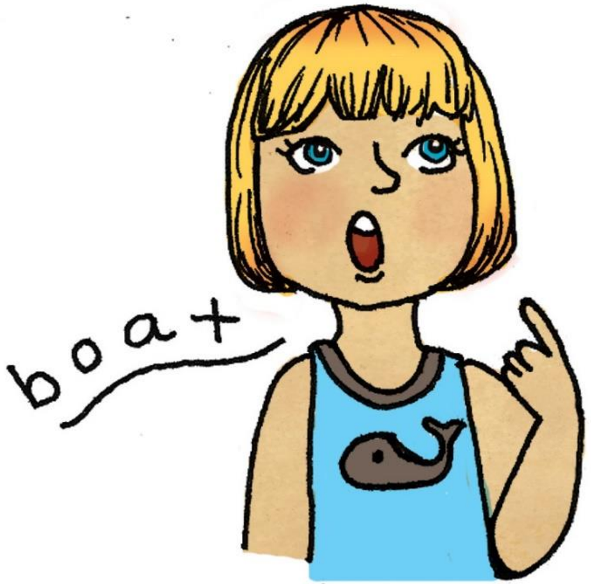
[Source: https://www.futurity.org/babies-babble-language-communication-1661482-2/
Image credit: Getty Images]

For infants to start producing their first words,
they must first begin matching
the sounds of babble and
the sounds of speech from the caregiver.

[Laing, et al. 2020]

Toddler Language Learning

- **Listening while Speaking**
  → Even when there is no parents or caregivers, the toddler can continue learning how to talk by constantly repeating their articulations & listening to sounds produced
  → A closed-loop speech chain has a critical auditory feedback mechanism



- **Hearing Loss = No auditory feedback**

[Source: https://www.cdc.gov/ncbddd/hearingloss/facts.html]

Children who lose their hearing often have difficulty to produce clear speech

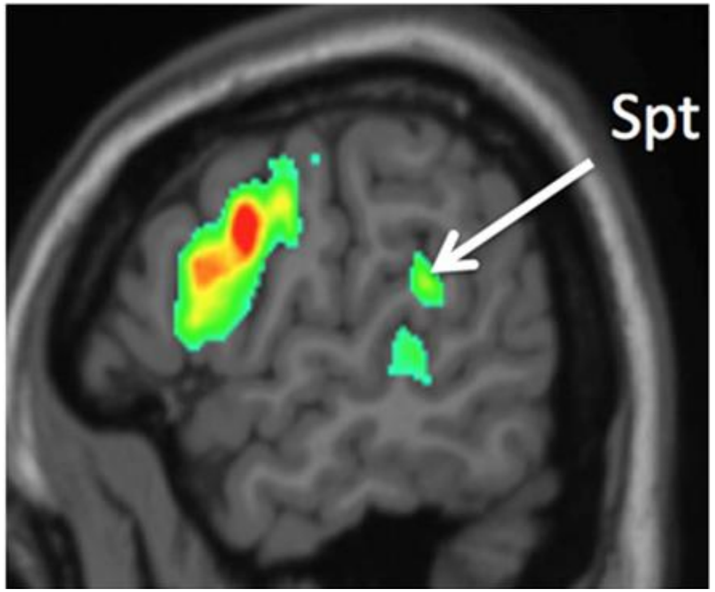
Adults who become deaf after becoming proficient with a language nonetheless suffer speech articulation declines as a result of the lack of auditory feedback

[Waldstein, 1990]

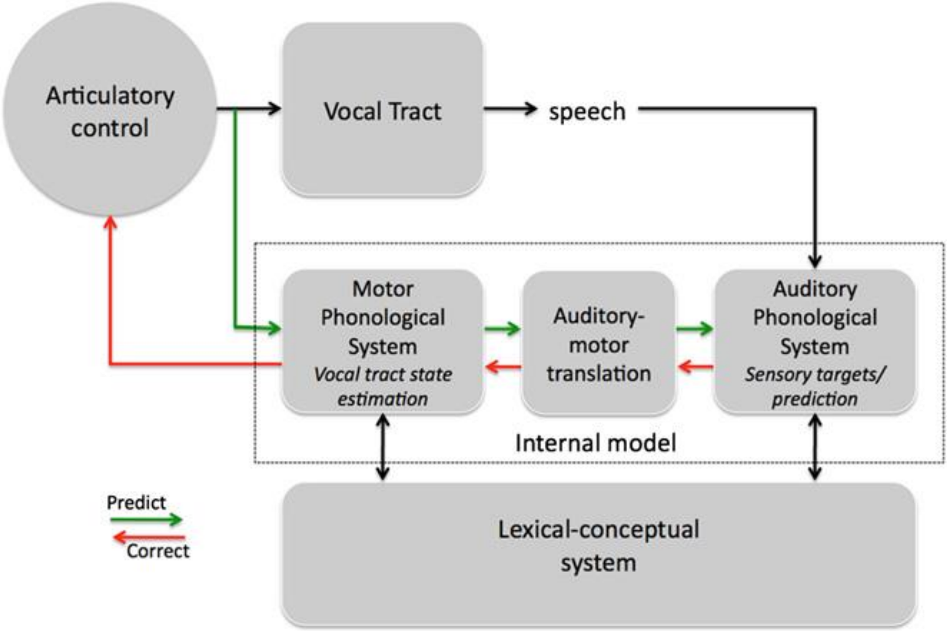
Sensorimotor response in the human brain

- **Sensorimotor Integration during Speech Processing**

  (1) the auditory system is critically involved in the production of speech
  (2) the motor system is critically involved in the perception of speech



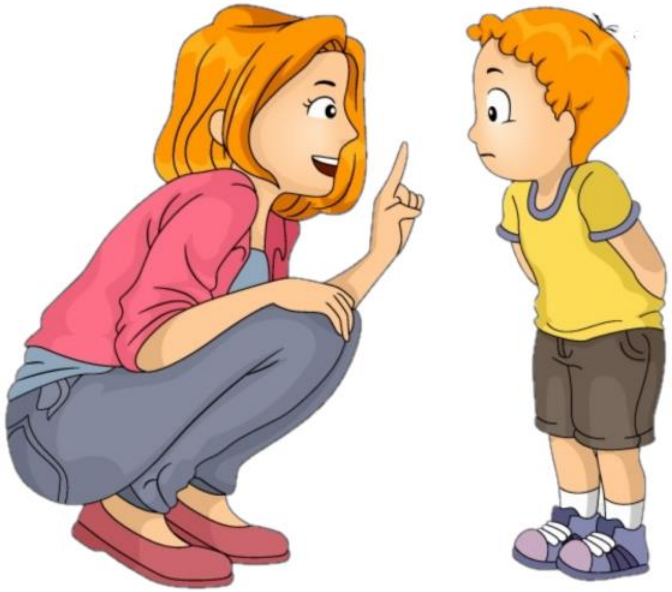
Spt exhibits sensorimotor response properties, activating both during the passive perception of speech and during covert (subvocal) speech articulation [Hickok et al, 2003]



An Integrated State Feedback Control (SFC) Model: Communication between auditory & motor systems is achieved by an auditory–motor translation system [Hickok et al. 2011]

Language Learning and Communication

- **During Language Learning**

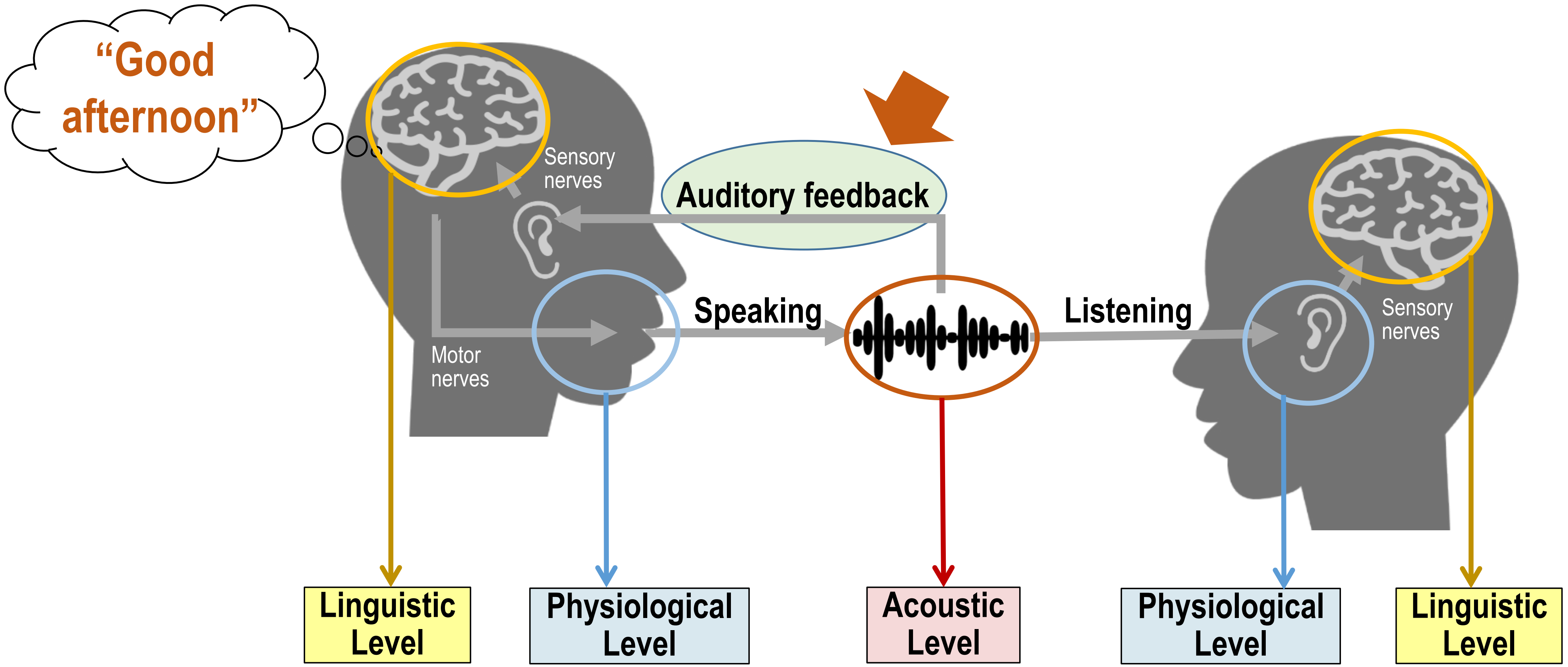


- **During Communication**

During speech production sensory feedback, such as auditory feedback, plays an important role in maintaining the fluidity of speech, as it allows speech motor movements to be monitored and production errors to be detected and corrected [Guenther, 2006].

Human Speech Chain

- **Speech Chain** [Denes & Pinson, 1993]



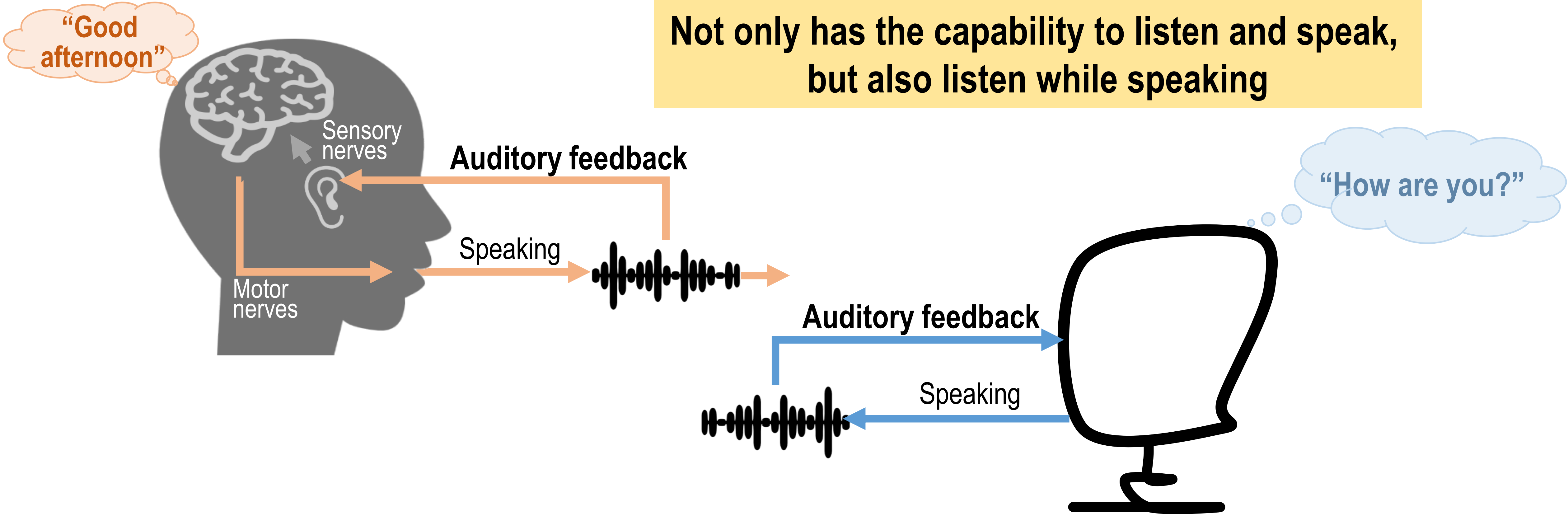
Machine Speech Chain:
Listening while Speaking by Deep Learning

Machine Speech Chain

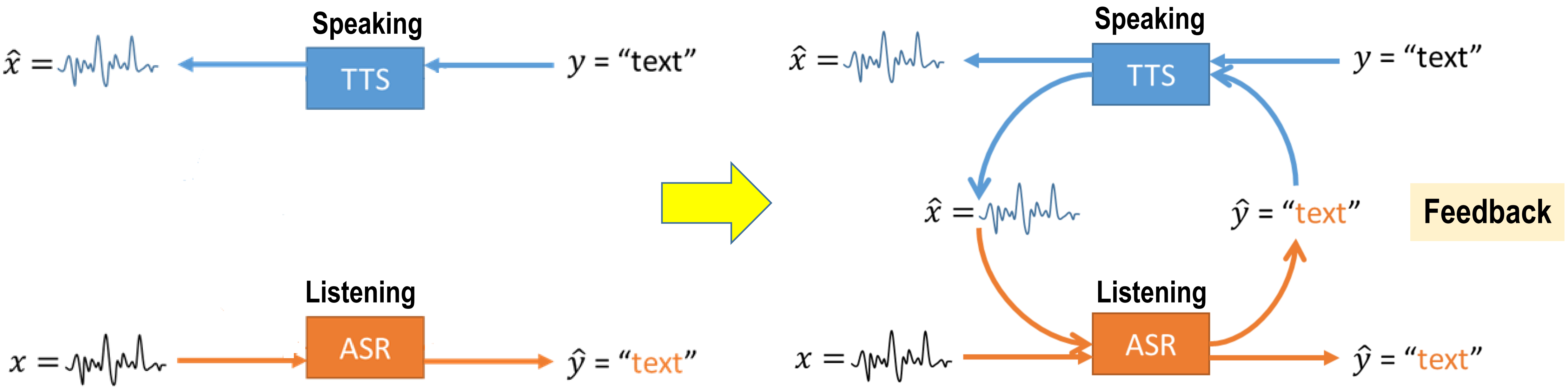
▪ Proposed Method

→ Develop a closed-loop speech chain model based on deep learning
→ The first deep learning model that integrates human speech perception & production behaviors

Not only has the capability to listen and speak, but also listen while speaking



Machine Speech Chain



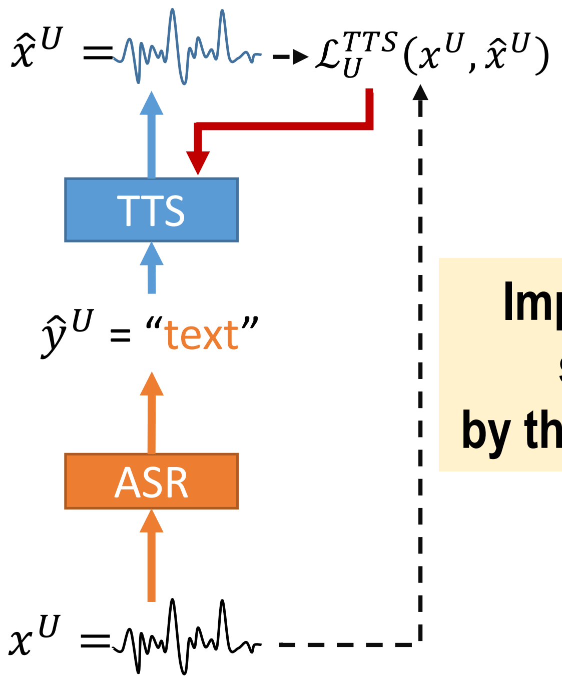
A closed-loop architecture:

→ **In training stage:**
- Allow to train with unlabeled data (low-level supervision)
- Allow ASR and TTS to teach each other using unlabeled data and generate useful feedback

→ **In Inference stage:** Possible to use ASR & TTS module independently

Learning with Unlabeled Data

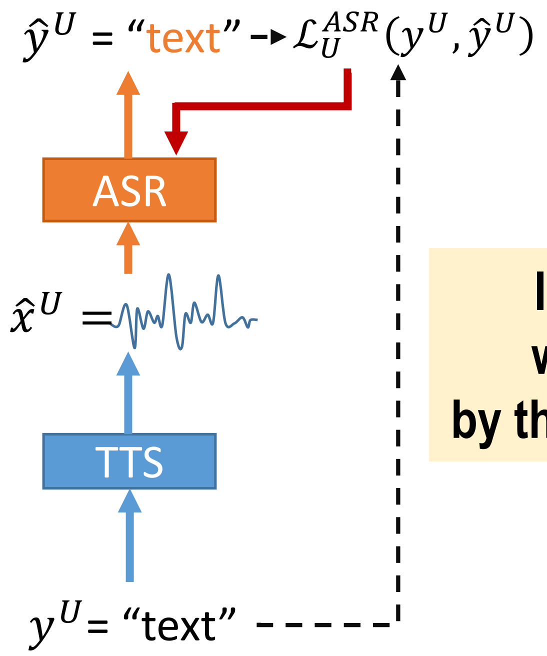
- **Learning with Speech only: ASR→TTS**



Improve TTS with speech only by the support of ASR

→ ASR predicts the transcription $\hat{y}^U$
→ Based on $\hat{y}^U$, TTS tries to reconstruct speech features $\hat{x}^U$
→ Calculate $\mathcal{L}_U^{TTS}(x^U, \hat{x}^U)$ between original speech features x^U and the predicted $\hat{x}^U$

- **Learning with Text only: TTS→ASR**



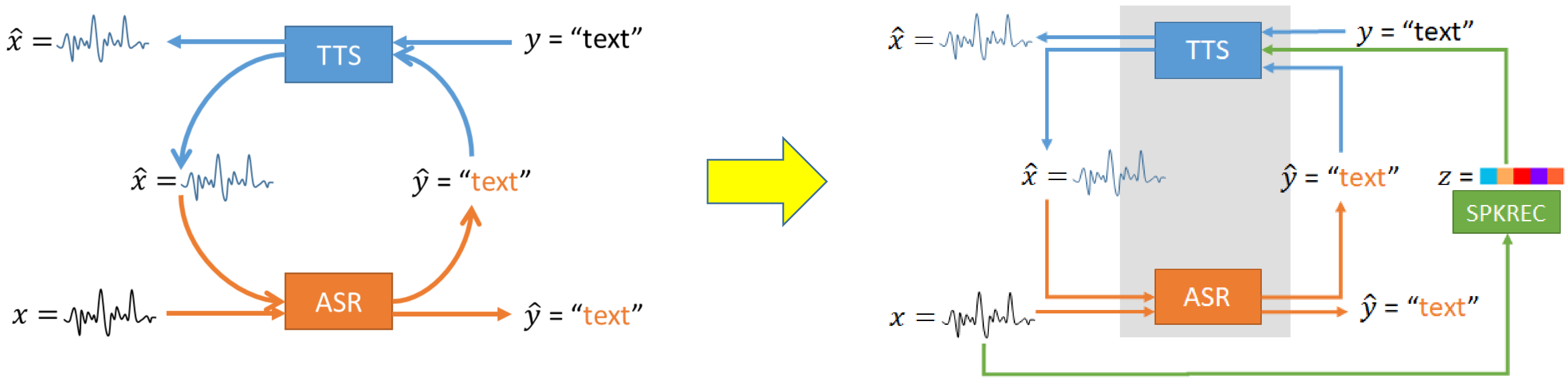
Improve ASR with text only by the support of TTS

→ TTS generates speech features $\hat{x}^U$
→ Based on $\hat{x}^U$, ASR tries to reconstruct text features $\hat{y}^U$
→ Calculate $\mathcal{L}_U^{ASR}(y^U, \hat{y}^U)$ between original text features y^U and the predicted $\hat{y}^U$

Multi-Speaker Machine Speech Chain

- **Handle Voice Characteristics from Unknown Speakers**

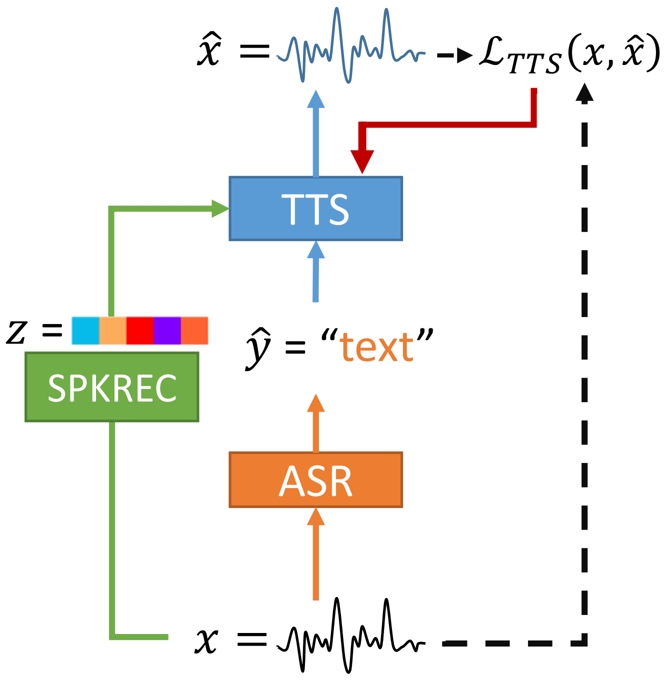
  → Basic Machine Speech Chain couldn't perform on unseen speaker
  → Integrate a speaker recognition system into the speech chain loop
  → Extend the capability of TTS to handle the unseen speaker using one-shot speaker adaptation



Utilizing [Deep speaker; Li et al., 2017]

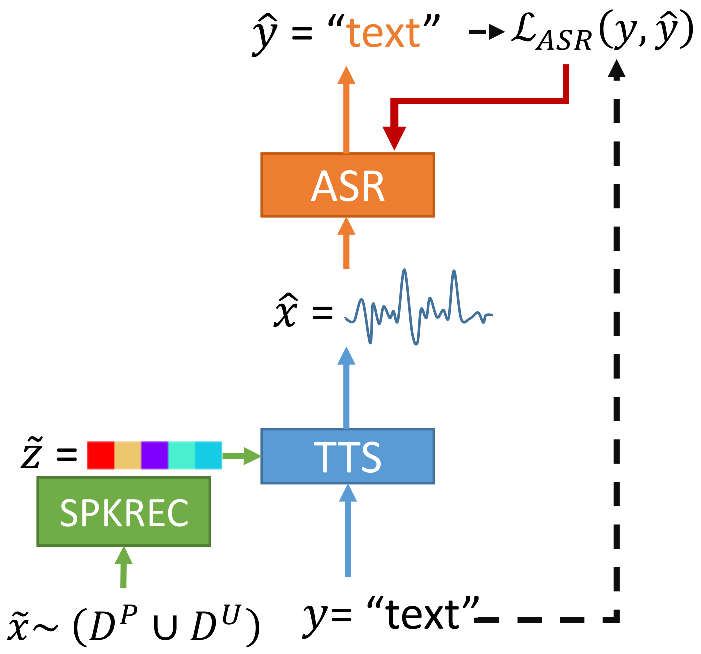
Learning in Multi-Speaker Speech Chain

- **Learning with Speech only: ASR⟶TTS**



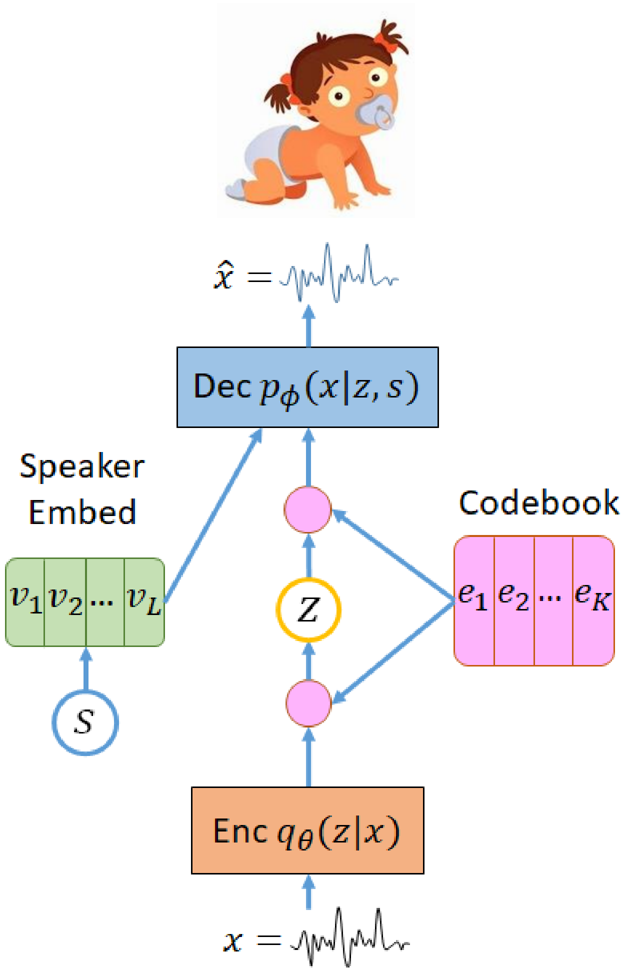
→ ASR predicts most possible transcription $\hat{y}$
→ SPKREC provides a speaker embedding z
→ Based on $[\hat{y}, \mathrm{z}]$, TTS tries to reconstruct speech $\hat{x}$

- **Learning with Text only: TTS⟶ASR**



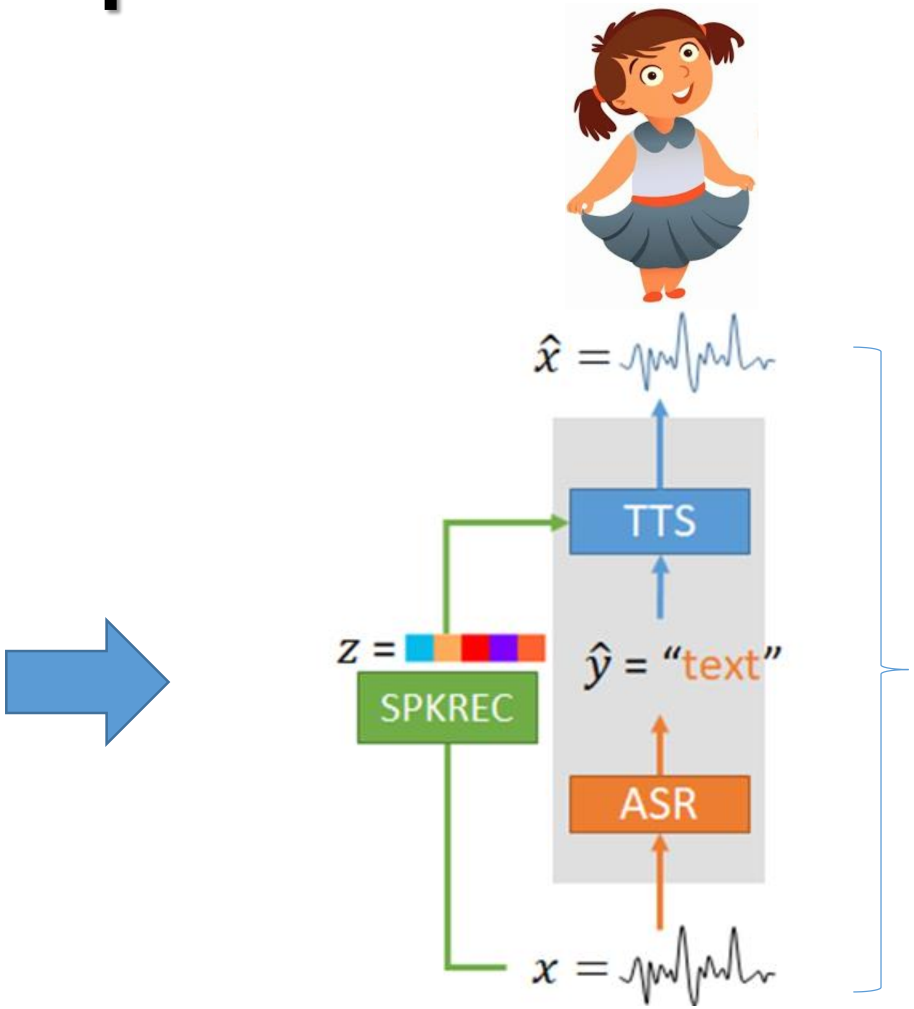
→ Sample a speaker vector $\tilde{z}$ from available speech
→ TTS generates speech features $\hat{x}$ based on $[y, \tilde{z}]$
→ Given $\hat{x}$, ASR tries to reconstruct text $\hat{y}$

Roadmap of Machine Speech Chain



Vector Quantized-Variational Autoencoder

[VQ-VAE; Oord et al, 2017]



Machine Speech Chain

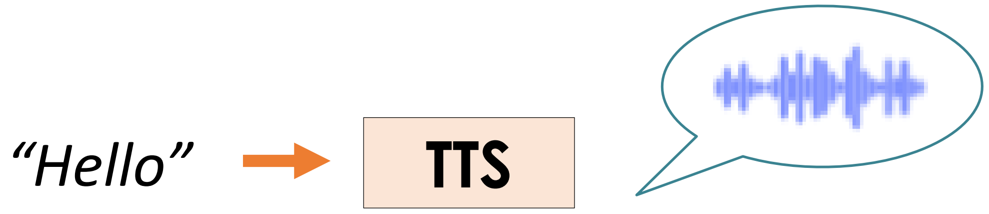
Do continual learning not only during training but also during inference

Self-Adaptive Mechanisms through Listening while Speaking

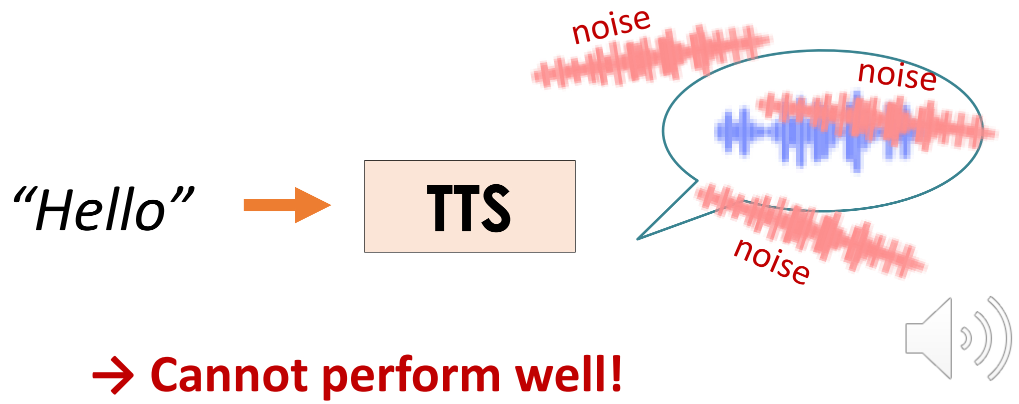
Speech Production

▪ State-of-the-art: Neural TTS

- Synthesizes a human-like speech in clean condition

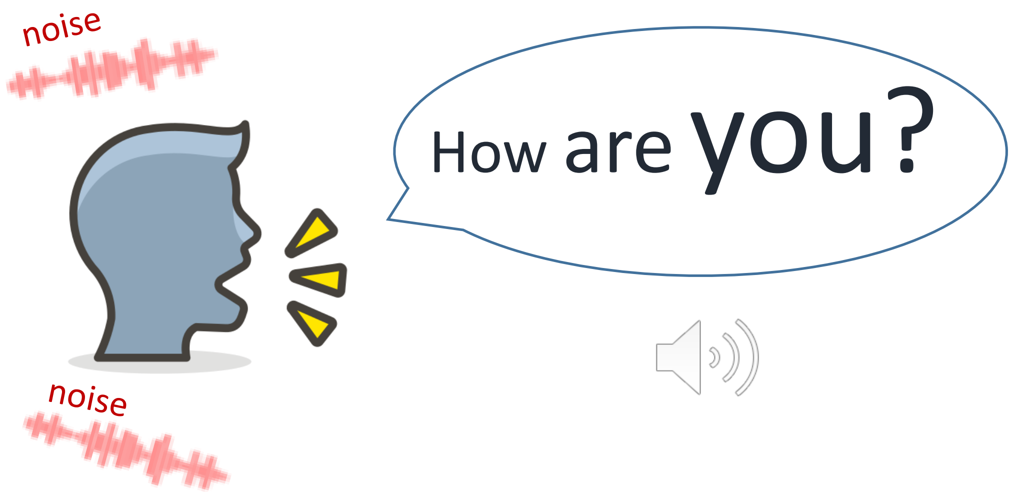


- Noisy condition?



→ **Cannot perform well!**

▪ How about Humans?



In noisy situation, human tend to speak louder (Lombard effect)

Existing Approaches

- **Parametric TTS in Noisy Condition**
  - HMM TTS speech modification to increase speech intelligibility in noise while keeping the speech energy fixed [Valentini-Botinhao et al., 2014; Schepker et al., 2015]
  - HMM TTS adapted to Lombard speech data [Raitio et al., 2014]

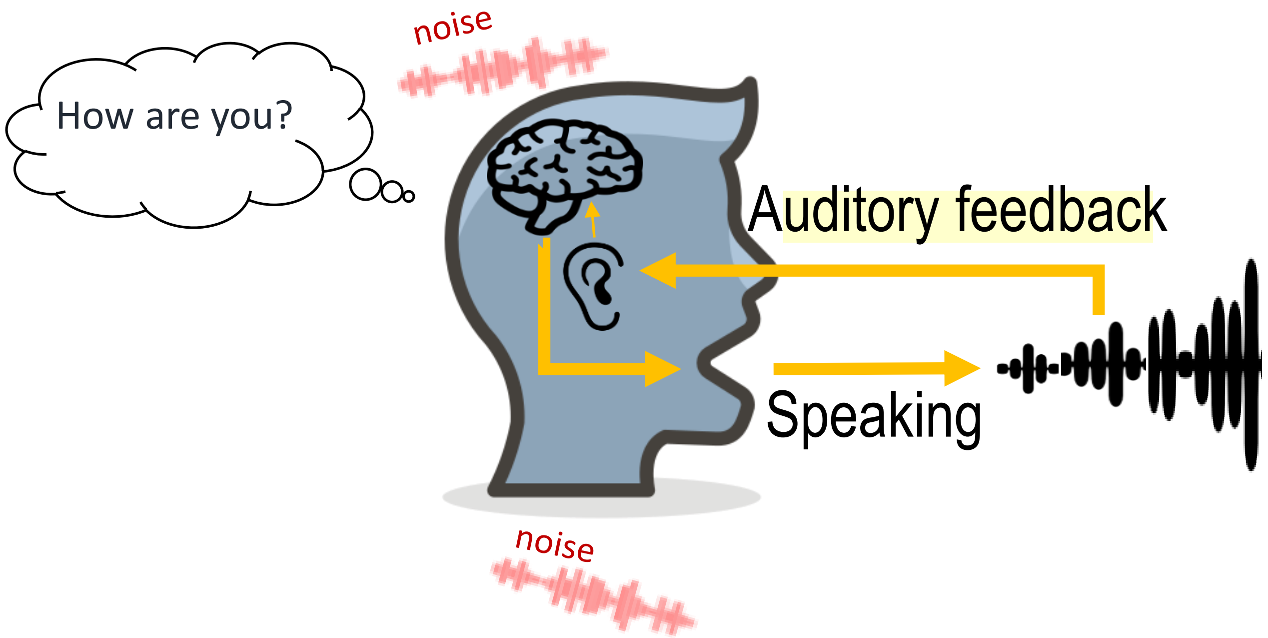
- **Neural TTS in Noisy Condition**
  - Transfer learning from a standard end-to-end TTS (clean) to an end-to-end Lombard TTS [Paul et al., 2020]
    → Lombard TTS is trained on a small Lombard dataset
  - End-to-end multi-style TTS [Hu et al., 2021]
    → Synthesizable speech styles: Normal speech, whispered speech, Lombard speech

Offline fine-tuning

Human:
No fine-tuning before speaking in noisy place

Human Speech Production in Noisy Speech

- **Dynamically Adaptation based on Auditory Feedback**

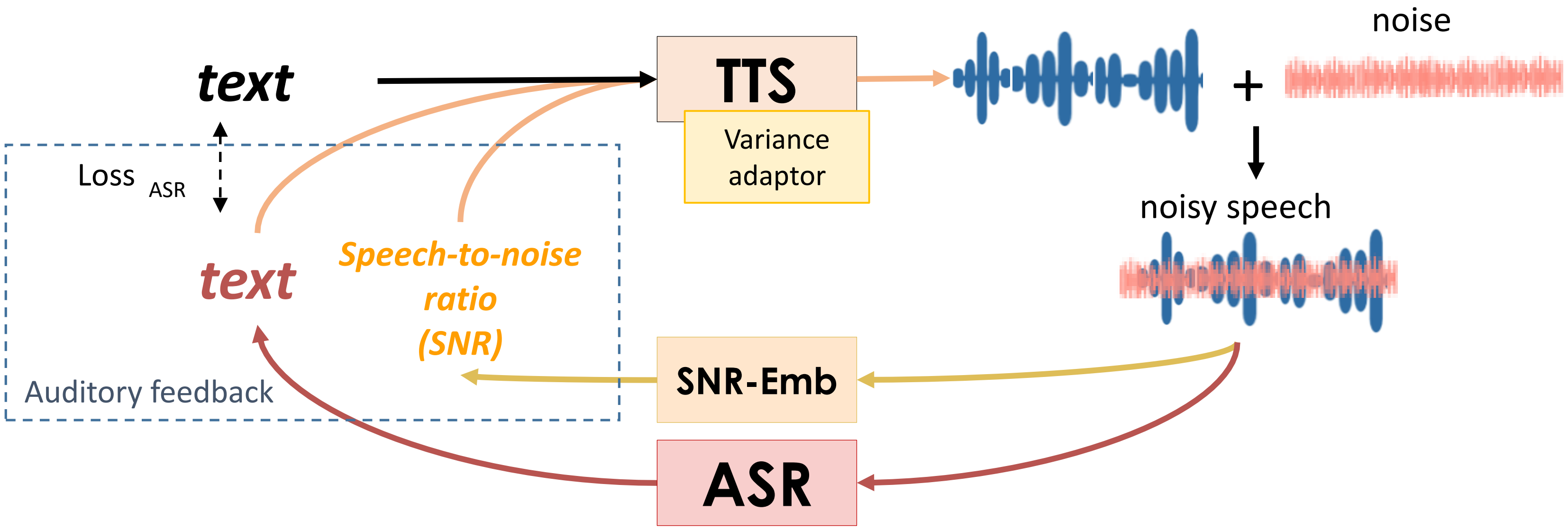


Humans speak while listen to their own speech (speech chain)

Dynamically adapt to the situation based on auditory feedback

Self-Adaptive Mechanism

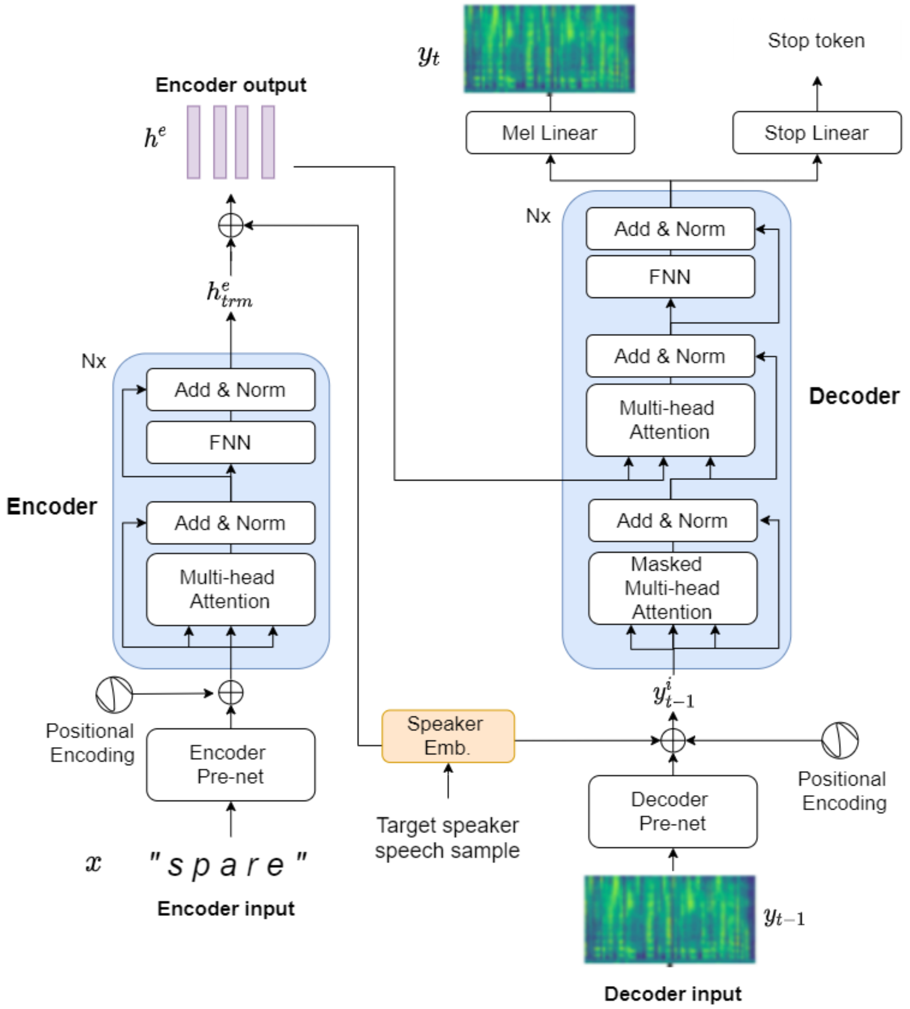
- **Machine Speech Chain Inference for TTS in Noisy Conditions**



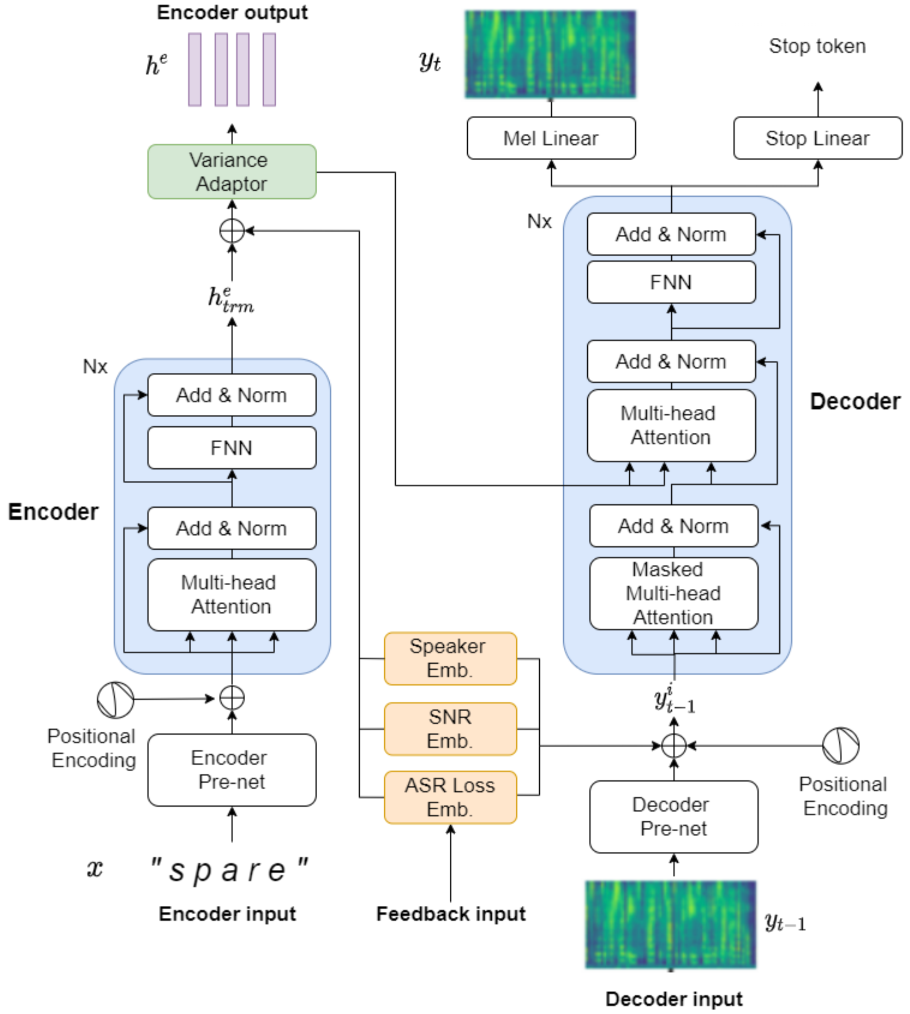
Aim: TTS dynamically adapt the situation by taking the auditory feedback and producing Lombard speech in noisy environments

Proposed Architecture

- **Transformer TTS**



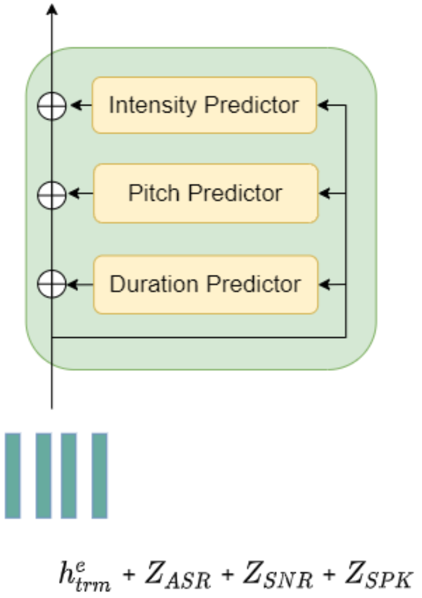
- **Transformer TTS with Auditory Feedback**



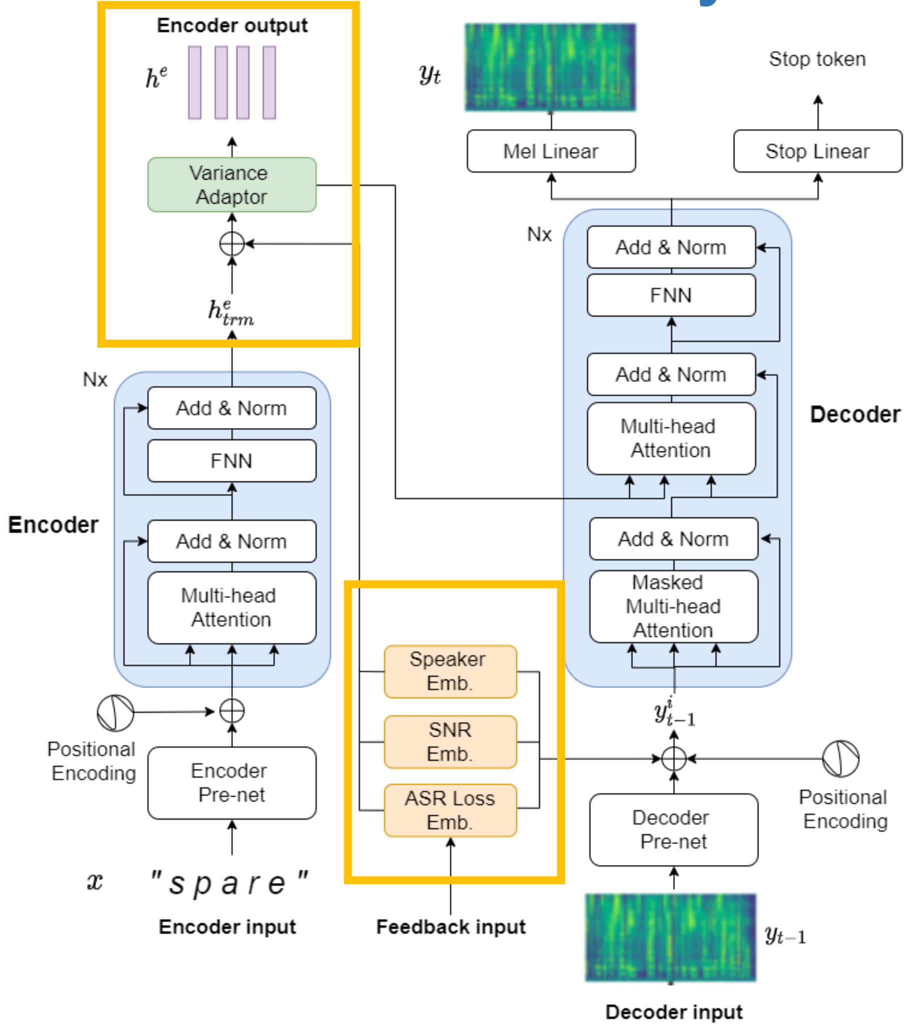
Proposed Architecture

- ## Auditory Feedback
  - **SNR Embedding**
  - **ASR-loss Embedding**

- ## Prosody Guide
  - **Variance adaptor**
    → Based on variance adaptor in Fast Speech [Ren et al., 2020]
    → Modified for autoregressive Transformer decoder



- ## Transformer TTS with Auditory Feedback



Experiment Setting: Data

A. Clean Wall Street Journal (WSJ) speech [Paul et al., 1992]

- Multi-speaker English speech, 81 hours of speech
- Training: *SI-284* set, dev: *dev92* set, test: *eval93* set

B. WSJ speech with additive noise

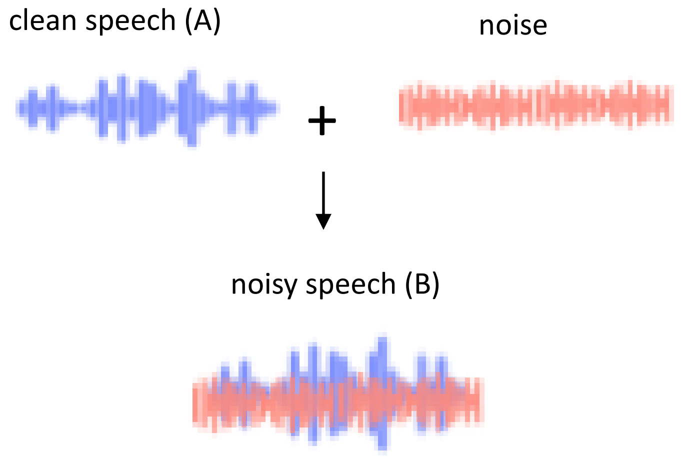
- Clean WSJ speech combined with noisy sound
  - Noise type : white noise and babble noise
  - SNR : SNR 0 and SNR -10

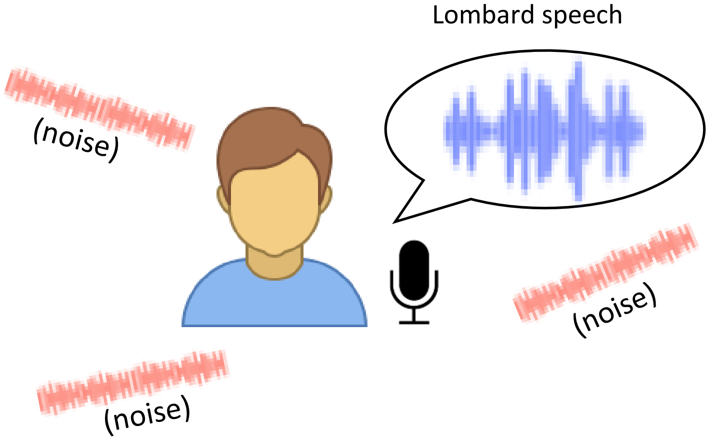
C. Natural Lombard speech

- Clean and noisy speech recorded from single male speaker
- Text: WSJ speech transcription (*dev92* + *eval93*)

D. Synthetic Lombard WSJ speech

- Clean WSJ speech with the intensity, pitch, and duration modified into Lombard speech





Experiment Setting: System Configuration

System	Structure	Training Data
TTS		
Baseline standard TTS	Transformer- 6 Enc, 6 Dec	Clean WSJ
Baseline standard TTS + Fine-tuning [Paul et al., 2020]		Clean WSJ + Synthetic Lombard WSJ
Proposed TTS		Clean WSJ + Synthetic Lombard WSJ
Feedback component		
ASR	Transformer- 12 Enc, 6 Dec (Speech-transformer [Dong et al., 2018])	Clean WSJ + Noisy WSJ
SNR recognition	4 convolutional + residual layers	Clean WSJ + Noisy WSJ (class: clean, SNR 0, SNR -10)

TTS Performance

Speech intelligibility measure (CER %) at different SNR levels using ASR trained on clean and noisy conditions

System	Clean	SNR 0	SNR -10
Baseline TTS			
Standard TTS	18.32	70.54	77.07
+ modification into Lombard speech	18.32	44.68	57.86
+ Fine-tuning with Lombard speech	13.40	28.12	46.13
Proposed TTS			
TTS + SNR emb.	<u>**11.58**</u>	22.82	42.00
TTS + SNR-ASR loss emb.	12.55	16.11	25.61
TTS + SNR-ASR loss emb. + var. adaptor	11.99	<u>**14.70**</u>	<u>**24.96**</u>
Topline (human natural speech)			
Natural speech	7.43	22.17	58.81
+ modification into Lombard speech	7.43	13.24	15.15
Natural Lombard speech	7.43	11.46	20.56

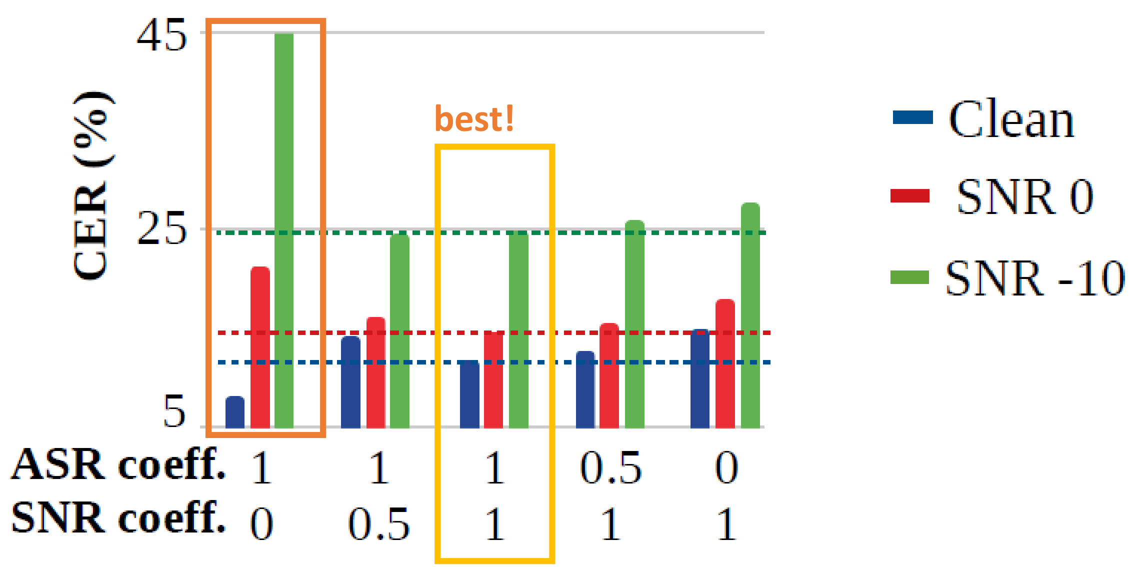
Best performance by TTS + SNR-ASR loss emb. + variance adaptor

- SNR and ASR feedback improved the speech intelligibility
- Variance adaptor guided the prosody change well by providing the target prosody information

How the auditory feedback affects TTS speech?

- Experiments by applying a coefficient to SNR embedding and ASR-loss embedding in encoder output and decoder input (default coefficient: 1)

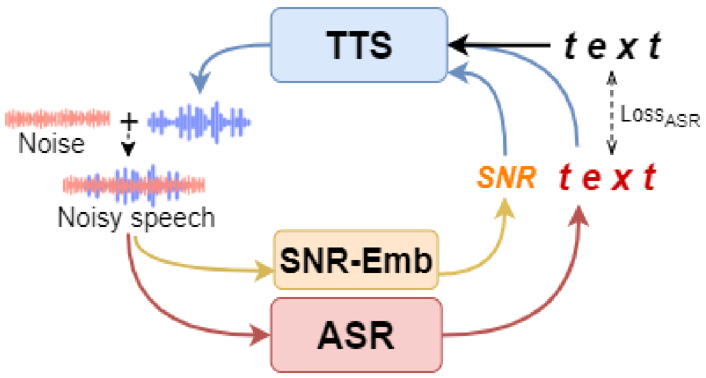
The effect of auditory feedback on speech intelligibility



- Clean condition: best performance with ASR feedback only (ASR coeff 1, SNR coeff 0)
- Noisy condition: best performance by equal amount of ASR + SNR feedback (coeff 1)

Both SNR and ASR-loss information are important to synthesize Lombard speech

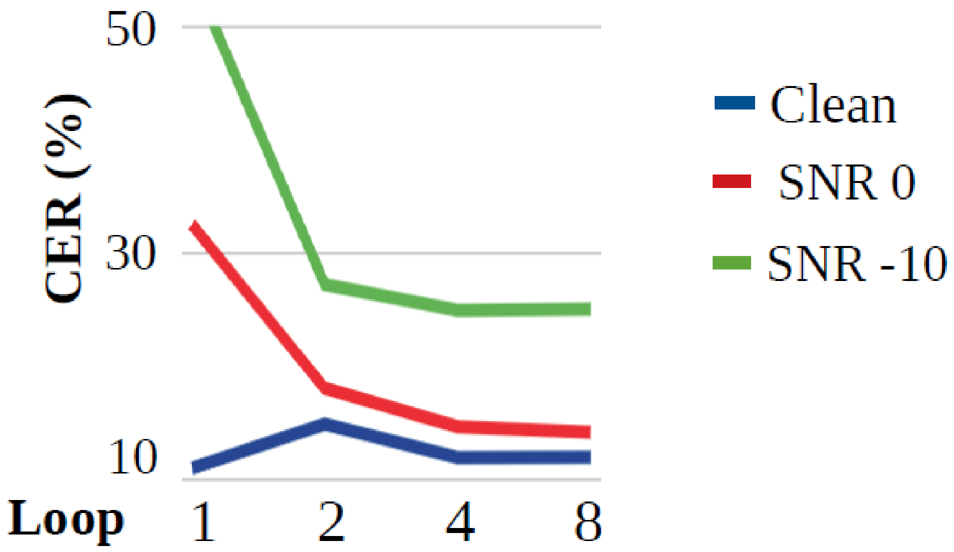
How the feedback loop affects TTS speech?



- Loop 1 : No feedback utilization
- Improvement significantly occurs after the 2nd loop

TTS performed dynamic adapt in several loops; listen to its voice in a noisy environment and then speak louder (similar to humans)

The effect of feedback loop on speech intelligibility



Summary

- **Machine Speech Chain:** **Inspired by the human speech chain, we proposed a machine speech chain to achieve low-level supervision**
  → Enables ASR & TTS to assist each other when they receive unpaired data
- **Multi-speaker Machine Speech Chain:** **Improved machine speech chain to handle voice characteristics from unknown speakers**
  → TTS can generate speech with similar voice characteristic only with one-shot speaker examples
- **Self-Adaptive Machine Speech Chain:** **Proposed TTS with auditory feedback**
  → Improved the TTS speech intelligibility in noisy condition
  → Dynamic adaptation with auditory feedback is critical not only for human but also in speech generation by machines

Machine Speech Chain Publications

General Machine Speech Chain Framework

- A. Tjandra, S. Sakti, S. Nakamura, "Listening while Speaking: Speech Chain by Deep Learning", in Proc. IEEE ASRU Workshop, 2017
- A. Tjandra, S. Sakti, S. Nakamura, "Machine Speech Chain with One-shot Speaker Adaptation", in Proc. INTERSPEECH, 2018
- A. Tjandra, S. Sakti, S. Nakamura, "End-to-end Feedback Loss in Speech Chain Framework via Straight-through Estimator", in Proc. IEEE ICASSP, 2019
- A. Tjandra, S. Sakti, S. Nakamura, "Machine Speech Chain," IEEE/ACM TASLP, Vol. 28, pp. 976-989, 2020

Multilingual Machine Speech Chain

- S. Nakayama, A. Tjandra, S. Sakti, S. Nakamura, "Speech Chain for Semi-supervised Learning of Japanese-English CS ASR & TTS", in Proc. IEEE SLT, 2018
- S. Nakayama, A. Tjandra, S. Sakti, S. Nakamura, "Zero-shot CS ASR and TTS with Multilingual Machine Speech Chain,“ in Proc. IEEE ASRU Workshop, 2019
- S. Nakayama, A. Tjandra, S. Sakti, S. Nakamura, "Code-Switching ASR and TTS using Semi-supervised Learning with Machine Speech Chain,“ IEICE Transactions on Information and Systems, Vol.E104-D, No.10, July. 7-8, 2021
- S. Novitasari, A. Tjandra, S. Sakti, S. Nakamura, "Cross-Lingual Machine Speech Chain for Javanese, Sundanese, Balinese, & Bataks Speech Recognition and Synthesis", in Proc. SLTU, 2020

Multimodal Machine Speech Chain

- J. Effendi, A. Tjandra, S. Sakti, S. Nakamura, “Listening while Speaking and Visualizing: Improving ASR through MC,” in Proc. IEEE ASRU Workshop, 2019
- J. Effendi, A. Tjandra, S. Sakti, S. Nakamura, "Augmenting Images for ASR & TTS through Single-loop & Dual-loop MC Framework," in Proc. INTERSPEECH, 2020
- J. Effendi, A. Tjandra, S. Sakti, Satoshi Nakamura, "Multimodal Chain:Cross-Modal Collaboration Through Listening, Speaking, and Visualizing,“ IEEE Access, No. 9, pp. 70286-70299, May. 6, 2021

Weakly Supervised Machine Speech Chain

- J. Effendi, S. Sakti, S. Nakamura, “Weakly-supervised Speech-to-text Mapping with Visually Connected Non-parallel Speech-text Data using Cyclic Partially-aligned Transformer,” Proc. of INTERSPEECH, Sep 2021

Incremental (Real-time) Machine Speech Chain

- S. Novitasari, A. Tjandra, S. Sakti, and S. Nakamura, “Seq-to-seq learning via attention transfer for incremental speech recognition,” INTERSPEECH, 2019
- T. Yanagita, SNeural iTTS: Toward Synthesizing Speech in Real-time with End-to-end Neural Text-to-Speech Framework,” Speech Synthesis Workshop, 2019
- S. Novitasari, A. Tjandra, T. Yanagita, S. Sakti, S. Nakamura, "Incremental Machine Speech Chain for Enabling Listening while Speaking in Real-time," in Proc. of INTERSPEECH, 2020

Dynamically Adaptive Machine Speech Chain

- S. Novitasari, S. Sakti, S. Nakamura, " Dynamically Adaptive Machine Speech Chain Inference for TTS in Noisy Environment: Listen and Speak Louder," in Proc. of INTERSPEECH, 2021

Citations

- **[Oord et al, 2018]** – A. Oord, Yazhe Li, and Oriol Vinyals. "Representation learning with contrastive predictive coding." arXiv preprint arXiv:1807.03748, 2018
- **[Chung et al, 2020]** – Y. Chung, J. Glass. "Generative pre-training for speech with autoregressive predictive coding." ICASSP 2020
- **[Schneider et al, 2019]** – S. Schneider, A. Baevski, R. Collobert, M. Auli, "Wav2vec: Unsupervised pre-training for speech recognition," https://arxiv.org/abs/1904.05862, 2019
- **[Baevski et al, 2020]** – A. Baevski, H. Zhou, A. Mohamed, M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations." NeurIPS 2020
- **[Hsu et al, 2020]** – W. Hsu, B. Bolte, Y. Tsai, K. Lakhotia, R. Salakhutdinov, A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," arXiv preprint arXiv:2106.07447, 2021
- **[Chung et al, 2020]** – Y. Chung, Y. Zhang, W. Han, C. Chiu, J. Qin, R. Pang, Y. Wu,"W2v-BERT: Combining Contrastive Learning and Masked Language Modeling for Self-Supervised Speech Pre-Training", arXiv:2108.06209, ASRU 2021
- **[Bengio et al, 2019]** – Y. Bengio, "From System 1 Deep Learning to System 2 Deep Learning," NeurIPS, 2019
- **[Laing et al, 2020]** – C. Laing, E. Bergelson, "From babble to words: Infants' early productions match words and objects in their environment," Cogn Psychol. 2020
- **[Waldstein, 1990]** – R.S. Waldstein, "Effects of postlingual deafness on speech production: Implications for the role of auditory feedback," J. Acoust. Soc. Am. 88, pp. 2099–2114, 1990
- **[Hickok, 2003]** – G. Hickok, B. Buchsbaum, "Temporal lobe speech perception systems are part of the verbal working memory circuit: Evidence from two recent fMRI studies," Behav. Brain Sci. 26, 2003
- **[Hickok, 2011]** – G. Hickok, J. Houde, F. Rong, "Sensorimotor Integration in Speech Processing: Computational Basis and Neural Organization", Neuron Perspective, Vol. 69, Issue 3, pp. 407-422, 2011
- **[Denes & Pinson, 1993]** -- P. Denes and E. Pinson, "The Speech Chain", ser. Anchor books. Worth Publishers, 1993. [Online]. Available: https://books.google.co.jp/books?id=ZMTm3nlDfroC
- **[Li et al. , 2017]** – C. Li, X. Ma, B. Jiang, X. Li, X. Zhang, X. Liu, Y. Cao, A. Kannan, and Z. Zhu, "Deep speaker: an end-to-end neural speaker embedding system," arXiv preprint arXiv:1705.02304, 2017
- **[Oord et al., 2017]** – A. Oord, O. Vinyals, K. Kavukcuoglu, "Neural Discrete Representation Learning," https://arxiv.org/abs/1711.00937, 2017
- **[Valentini-Botinhao et al., 2014]** – C. Valentini-Botinhao, J. Yamagishi, S. King, and R. Maia, "Intelligibility enhancement of HMM-generated speech in additive noise by modifying Mel cepstral coefficients to increase the glimpse proportion," Computer Speech and Language, vol. 28, pp. 665–686, 2014.
- **[Schepker et al., 2015]** – H. Schepker, J. Rennies, and S. Doclo, "Speech-in-noise enhancement using amplification and dynamic range compression controlled by the speech intelligibility index," The Journal of the Acoustical Society of America, vol. 138, no. 5, 2015.
- **[Raitio et al., 2014]** – T. Raitio, A. Suni, M. Vainio, and P. Alku, "Analysis of HMM based Lombard speech synthesis," in Proc. INTERSPEECH, 2011
- **[Paul et al., 2020]** -- D. Paul, M. P. Shifas, Y. Pantazis, and Y. Stylianou, "Enhancing speech intelligibility in text-to-speech synthesis using speaking style conversion," in Proc. INTERSPEECH, 2020
- **[Hu et al., 2021]** -- Q. Hu, T. Bleisch, P. Petkov, T. Raitio, E. Marchi, and V. Lakshminarasimhan, "Whispered and Lombard neural speech synthesis," in Proc. IEEE SLT, 2021
- **[Ren et al., 2020]** -- Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Liu, "FastSpeech 2: Fast and high-quality end-to-end text to speech," ArXiv, vol. abs/2006.04558, 2020
- **[Paul et al. , 1992]** – D. Paul and J. Baker, "The Design for the Wall Street Journal-based CSR Corpus," HLT, 1992
- **[Dong et al., 2018]** -- L. Dong, S. Xu, and B. Xu, "Speech-Transformer: A norecurrence sequence-to-sequence model for speech recognition," in Proc. ICASSP, 2018

Thank you