

Neural Speech Completion

Kazuki Tsunematsu¹, Johanes Effendi^{1,2},
Sakriani Sakti^{1,2}, Satoshi Nakamura^{1,2}

¹Nara Institute of Science and Technology (NAIST), Japan

²RIKEN Center for Advanced Intelligence Project (AIP), Japan



Outline

- Motivation Background
- Related Works
- Proposed Framework
 - Text-to-text Completion System
 - Speech-to-text Completion System
 - Speech-to-speech Completion System
- Experiments
- Conclusions

Motivation Background and Related Works

Human-Human Communication

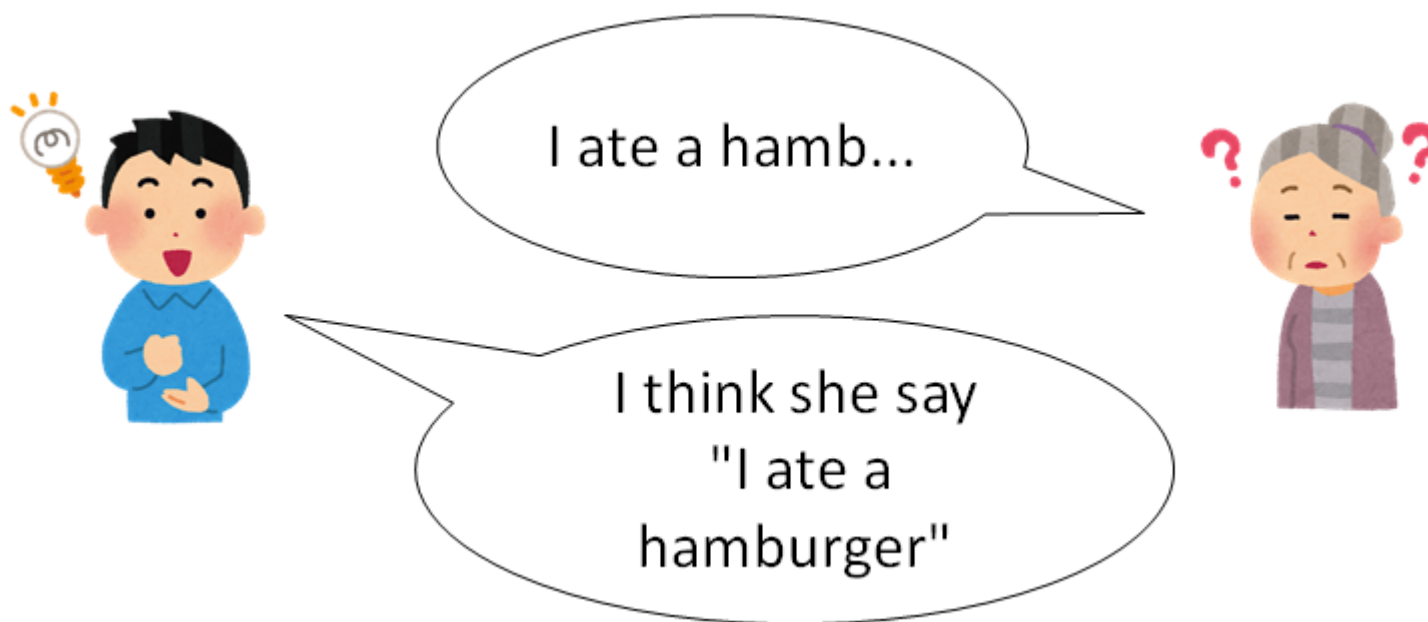
■ Speech Communication

■ Transformation:

- ✓ Intention in the speaker's mind → Understanding in the listener's mind
- ✓ Auditory system and brain play a decisively proactive role

■ Anticipation:

Often predict the end of a sentence even when the other person has not finished it

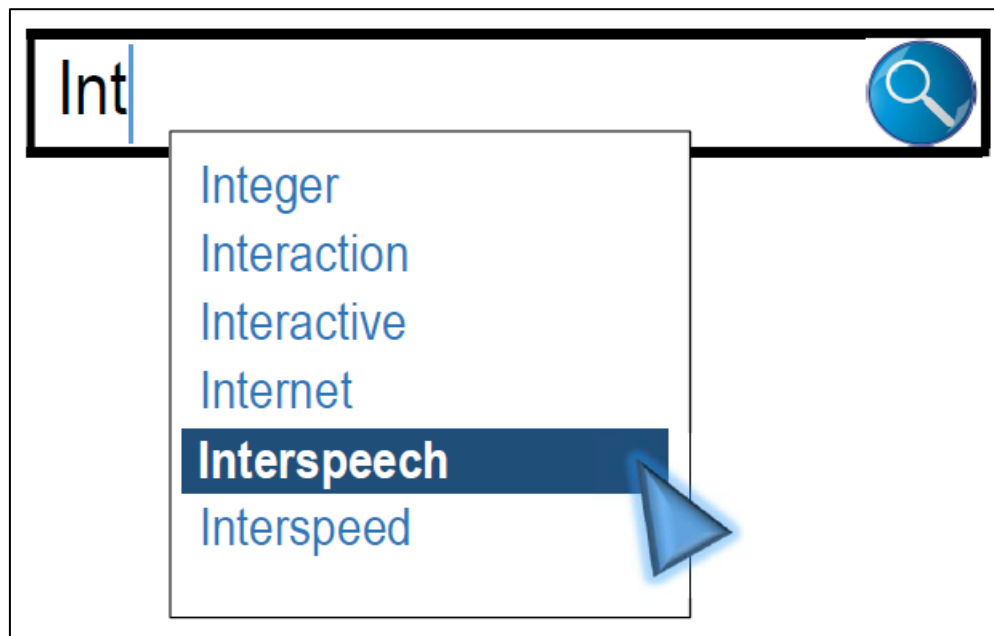


Human-Machine Communication

- **Anticipation in Human-Machine Communication**

- **Text Autocomplete**

Predicts the next word a user intends to enter after only a few characters have been typed



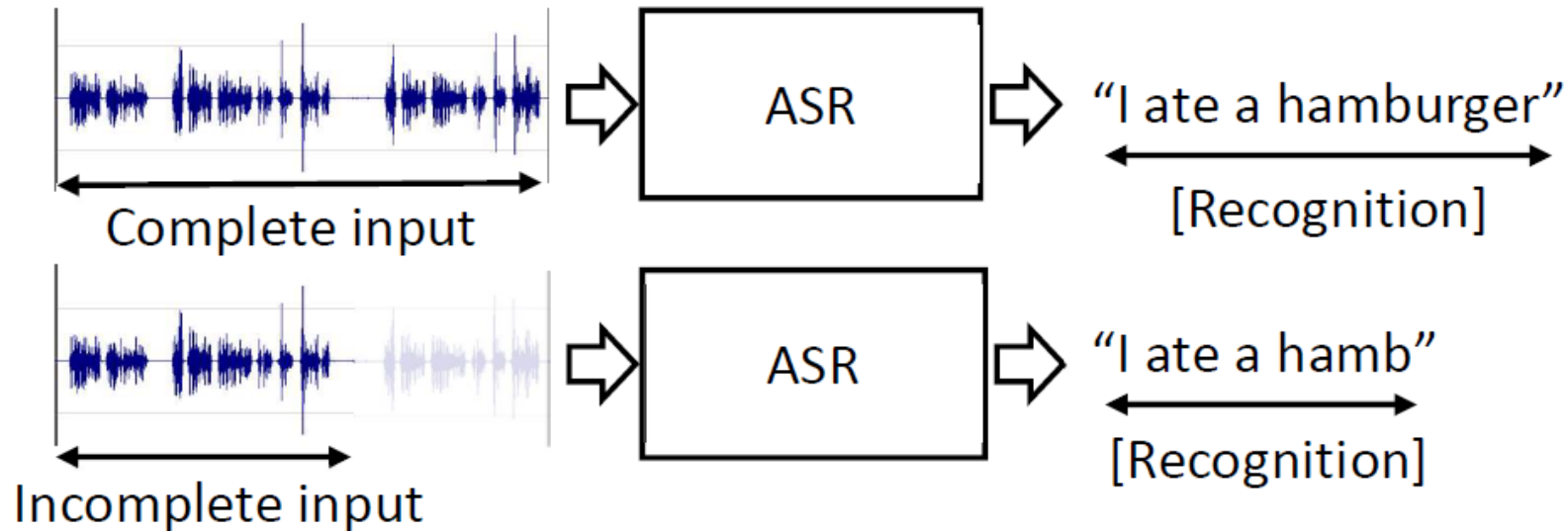
Remains limited to
text-based
human-machine
interaction

- **Widely used in many application**

Search engines, text editors, and command-line interpreters [Hollingsworth, 2018]

Related Works

- Speech-based Interaction
 - Automatic Speech Recognition



Passively recognizes what is being said

Related Works

■ Speech-based Interaction

■ Limited studies address the speech completion task

[Goto et al., 2002] proposed a speech recognizer that was extended with a vocabulary tree to provide candidates with the complete text

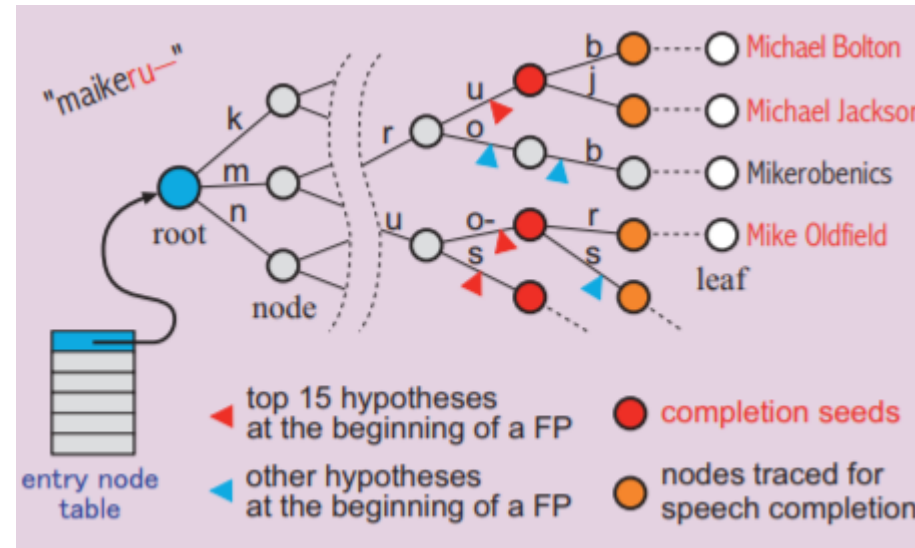
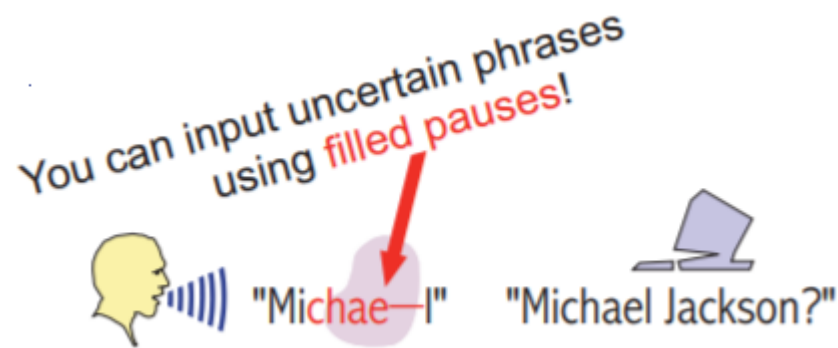


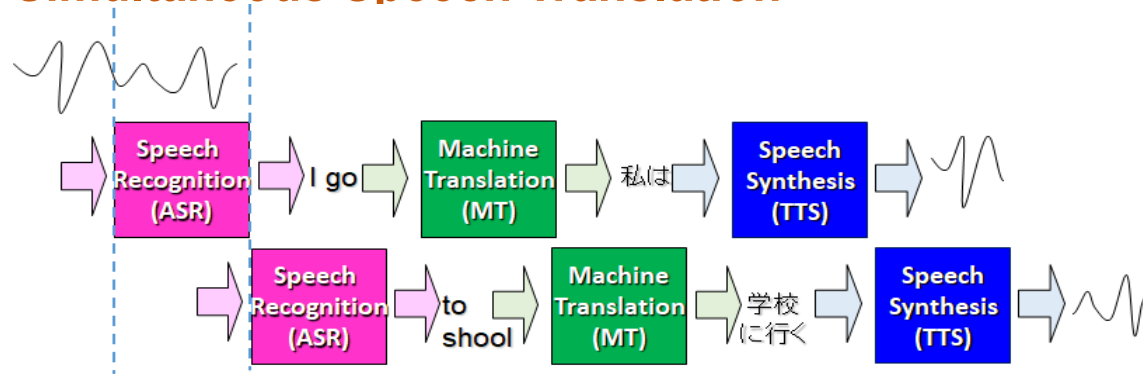
Figure source: <https://staff.aist.go.jp/m.goto/PAPER/ICSLP2002POSTERGoto.pdf> [Goto et al., 2002]

Still based on the traditional HMM speech-to-text system

Related Works

■ Speech-based Interaction

■ Simultaneous Speech Translation



- **[Niehues et al., 2018]** discussed the challenge of MT system:
Translate partial sentences of the source language
into complete sentences of the target language
- **The ability to predict is a prerequisite for being a successful simultaneous interpreter**
- **Incorporating the prediction within incremental ASR may help the performance of partial recognition**

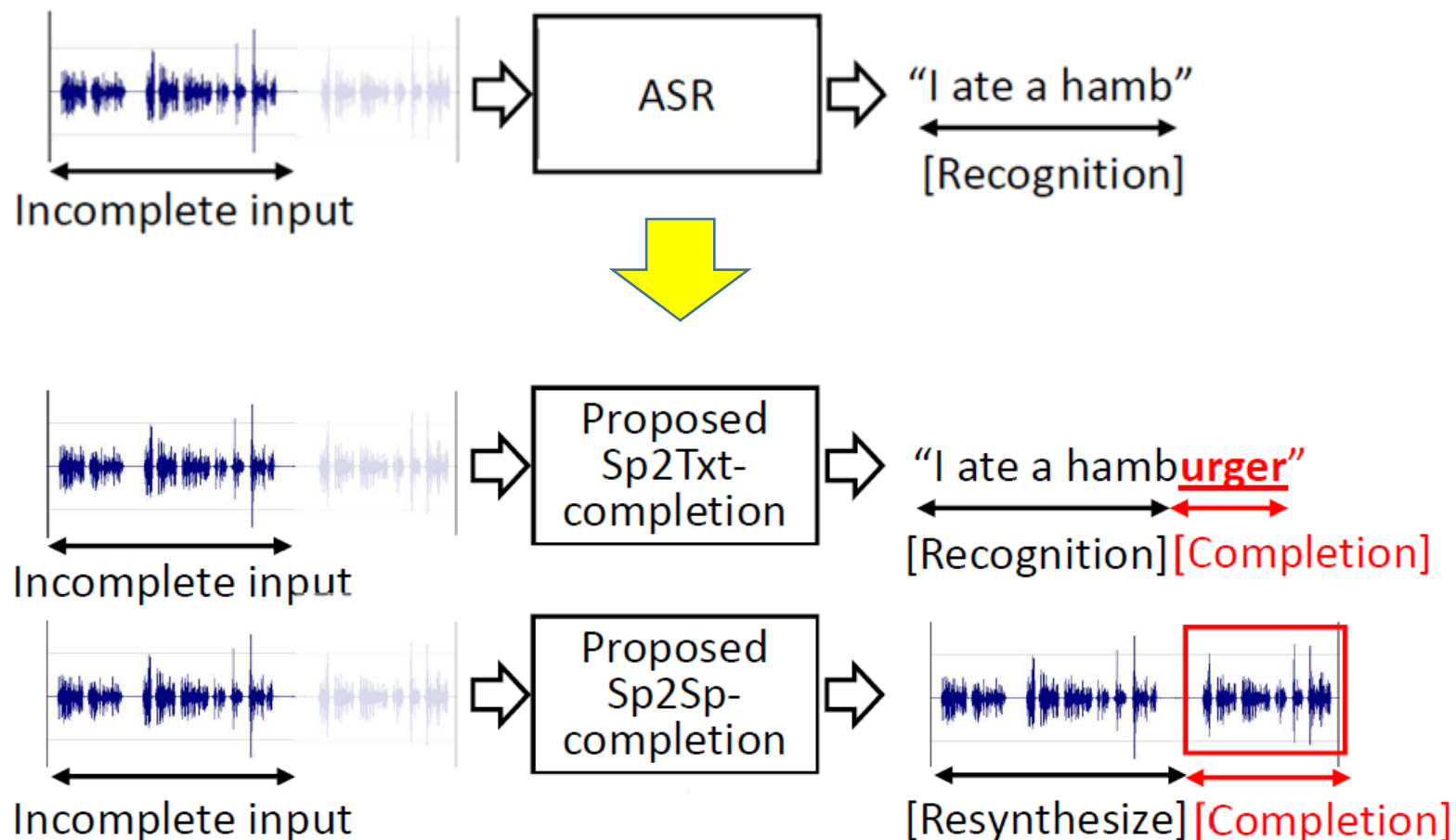
**Voice search, speech translation,
dialog system
may require a system:**

- ✓ Not only recognizes what has been said
- ✓ But also predicts what will be said

Proposed Framework

Proposed: Neural Speech Completion

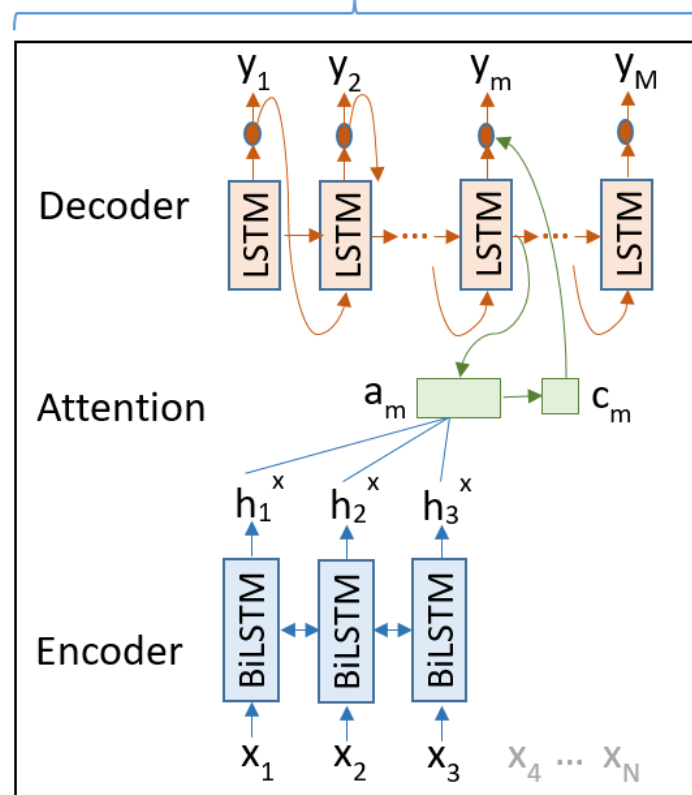
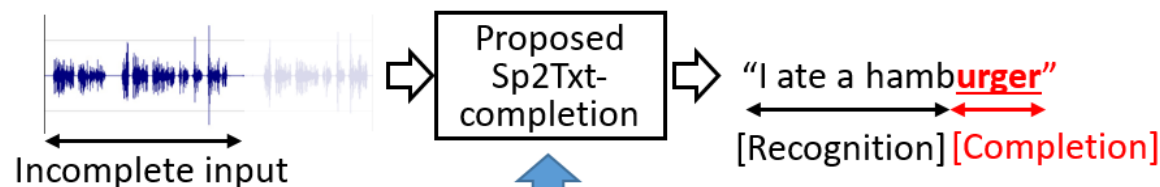
■ Overview



Investigate neural text-to-text,
speech-to-text and
speech-to-speech
completion framework

Proposed: Neural Speech Completion

Sequence-to-Sequence Attention-based Neural Networks



Standard ASR system

Input: $x = [x_1, \dots, x_N]$ with length N

Output: $\hat{y} = [\hat{y}_1, \dots, \hat{y}_M]$ with length M

Input: $x = [x_1, \dots, x_P]$ with length $P < N$

Output: $\hat{y} = [\hat{y}_1, \dots, \hat{y}_Q]$ with length $Q < M$

Text-to-text completion system

Input: $y = [y_1, \dots, y_Q]$ with length $Q < M$

Output: $\hat{y} = [\hat{y}_1, \dots, \hat{y}_M]$ with length M

Speech-to-text completion system

Input: $x = [x_1, \dots, x_P]$ with length $P < N$

Output: $\hat{y} = [\hat{y}_1, \dots, \hat{y}_M]$ with length M

Speech-to-speech completion system

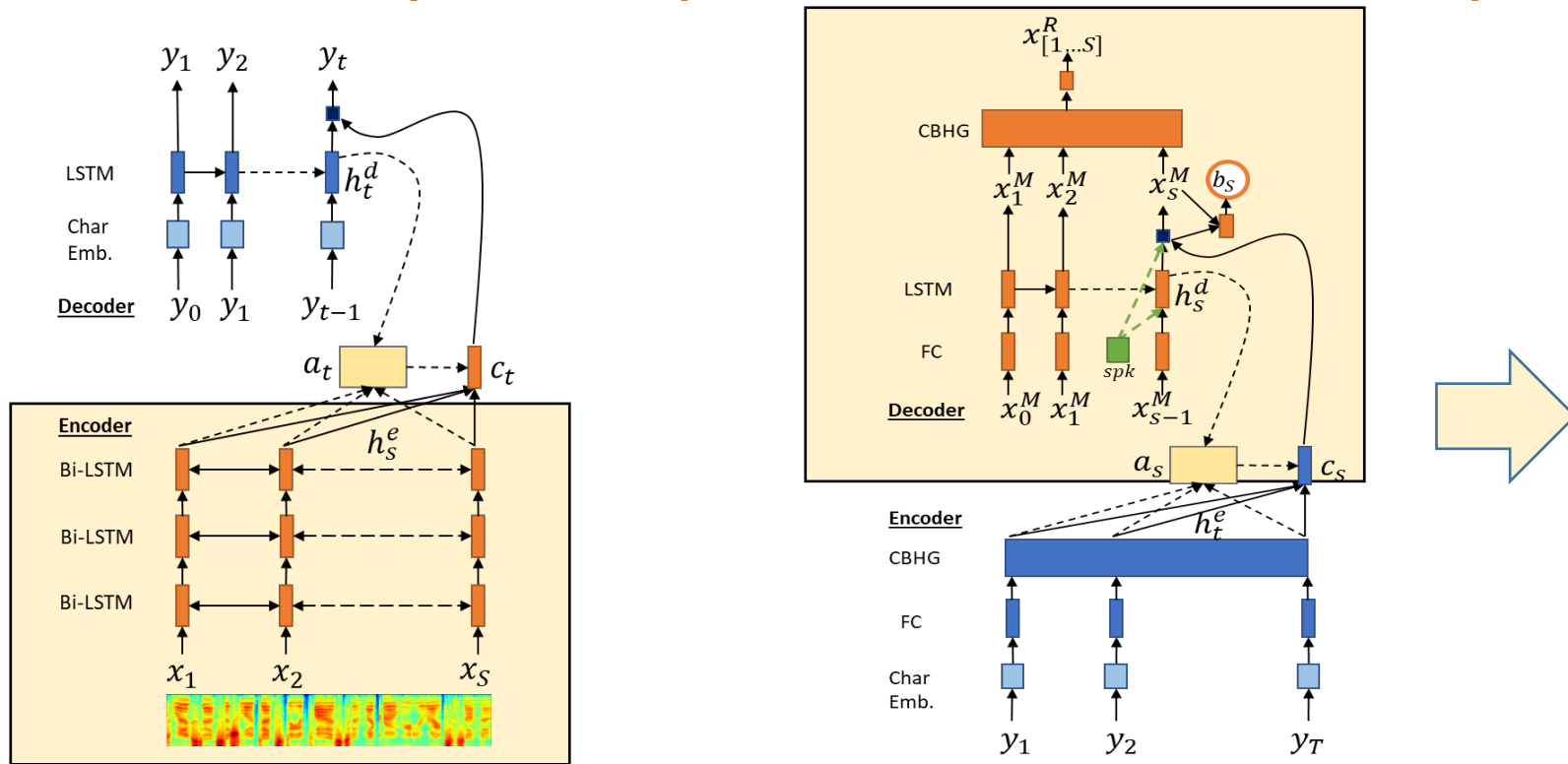
Input: $x = [x_1, \dots, x_P]$ with length $P < N$

Output: $\hat{x} = [\hat{x}_1, \dots, \hat{x}_N]$ with length N

Proposed Framework

■ Speech-to-speech Completion System

- Utilized the pretrained speech-to-text encoder and text-to-speech decoder



Pretrained speech-to-text system
Similar to [LAS, Chan et al. 2015]

Pretrained text-to-speech system
Similar to [Tacotron: Wang et al., 2017]

Experiments

Experimental Set-Up

■ Data

- Domain-specific sentences with synthesized speech utterances
- English Basic Travel Expression Corpus (BTEC) [Kikui et al., 2006]

✓ Text Data

Corpus	Number of Sentences		
	Train	Val	Test
BTEC	157448	4870	510

- ✓ Speech was generated with Google TTS
- ✓ Incomplete data: 25%, 50%, and 75% partial lengths

■ Features

- ✓ Text: Character-based (26 letters (a-z) and three special tags [<s>, </s>, <spc>])
- ✓ Speech: 80-dim log Mel Spectrogram (speech reconstruction with Griffin-Lim)

Baseline Systems and Completion Task

■ Baseline Systems

- **RNN-LM** [Kombrink et al., 2011]
 - ✓ Language model - Predictive model for the next token, given the previous one
 - ✓ Performed repeatedly on the incomplete part until [EOS]
- **BERT** [Devlin et al., 2014]
 - ✓ Language understanding – Bidirectional Encoder Representations from Transformer (BERT)
 - ✓ Replaced the incomplete part with [MASK] for prediction

■ Completion Task

Input	I ate a hamb
Word Completion	I ate hamb <u>urger</u>
Sentence Completion	I ate hamb <u>urger at a restaurant</u>

Performance: Proposed Completion System

■ Word Completion Task

■ Comparison:

- ✓ Human (15 subjects; TOEIC score > 730)
- ✓ Proposed Text-to-text Completion System

■ Results: Character Error Rate (CER)

	CER (%)
Proposed System	2.70
Human	7.21
Human (best)	5.50

**Proposed text-to-text completion system
outperformed human completion**

Performance: Proposed Completion System

■ Sentence Completion Task

■ Objective Evaluation:

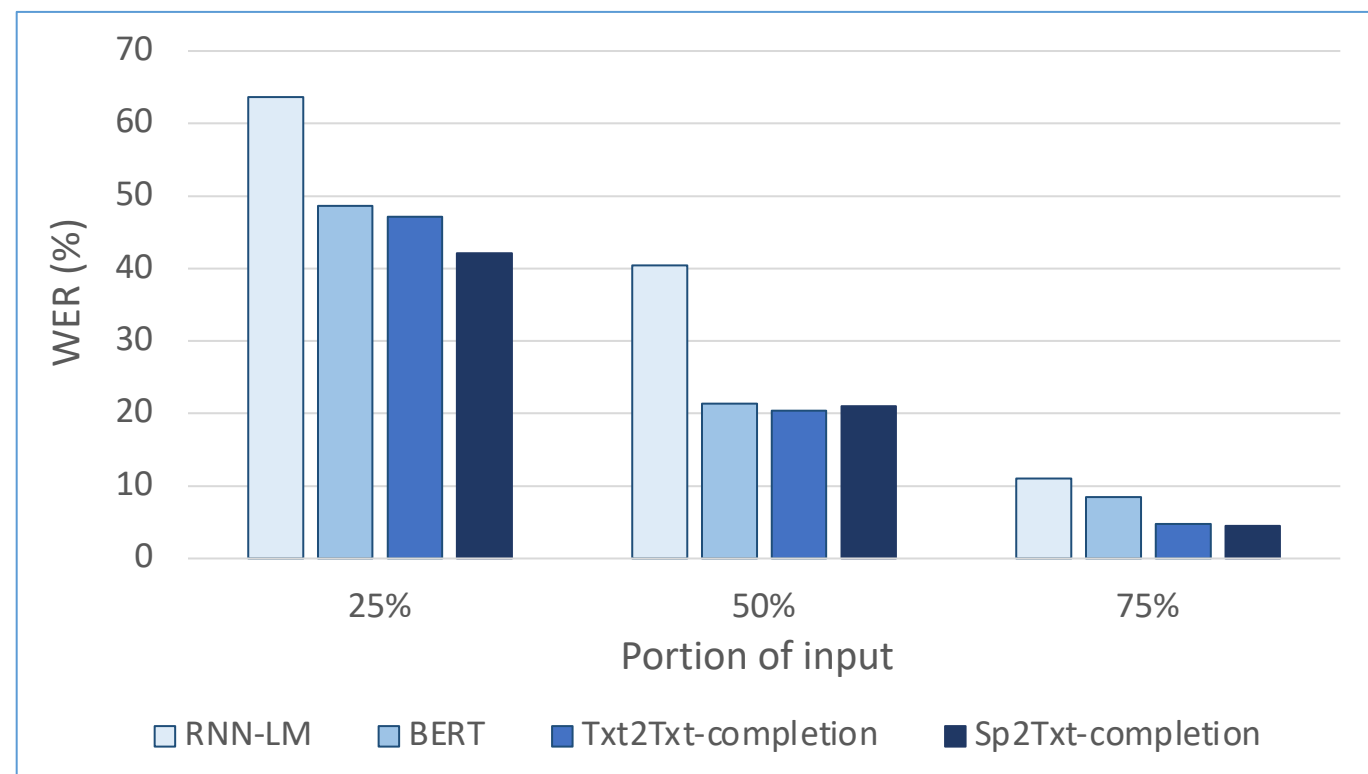
- ✓ Word error rate

■ Comparison:

- ✓ RNN-LM
- ✓ BERT
- ✓ Proposed Text-to-text Completion System
- ✓ Proposed Speech-to-text Completion System

Proposed text-to-text and speech-to-text completion system outperformed the baseline RNN-LM and BERT

■ Results: Word Error Rate (WER)



Performance: Proposed Completion System

■ Sentence Completion Task

■ Subjective Evaluation:

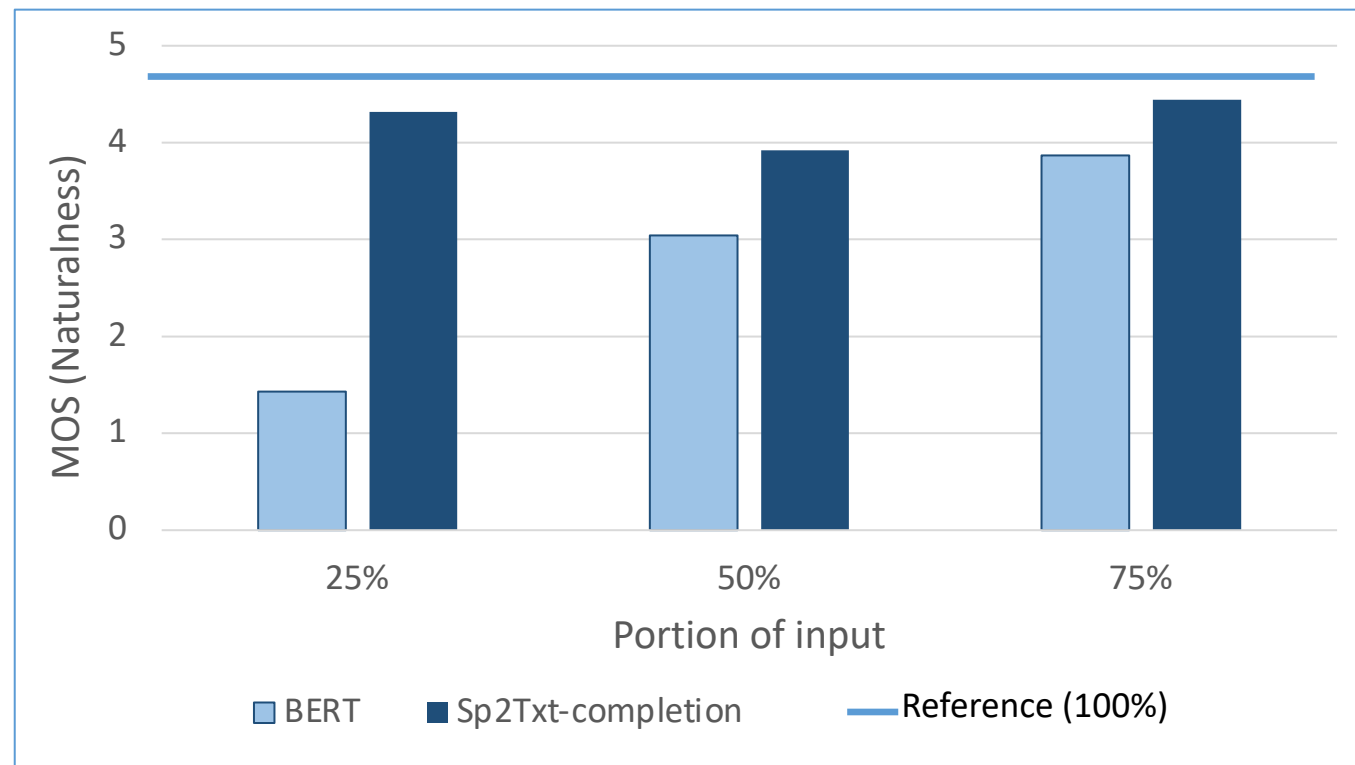
- ✓ Naturalness
- ✓ 12 subjects (TOEIC score > 730)

■ Comparison:

- ✓ BERT
- ✓ Proposed Speech-to-text Completion System
- ✓ References

Proposed speech-to-text completion system provided more natural suggestions than BERT

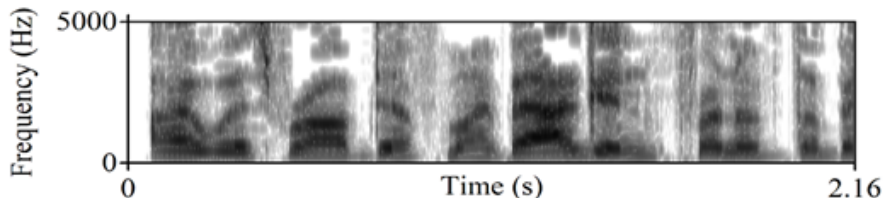
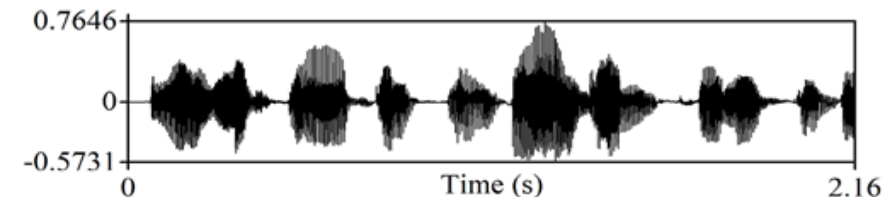
■ Results: Naturalness (MOS)



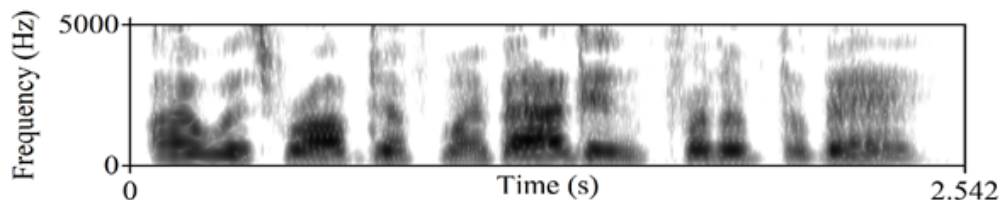
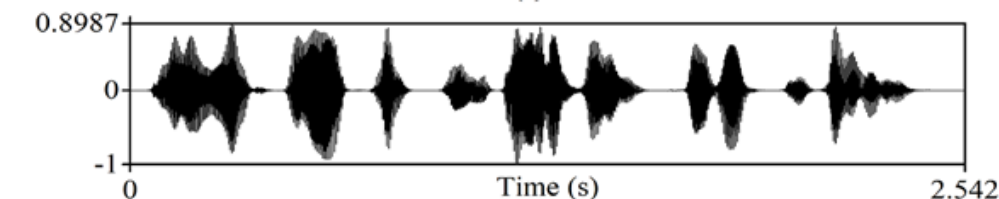
Performance: Proposed Completion System

■ Speech-to-speech Completion System

■ Input:



■ Output:



■ Reference:



Proposed speech-to-speech completion system

- Resynthesized the speech input
- Completed the remaining part
- While retaining the characteristic of speech style

Conclusions

Conclusions

- **Utilizing seq2seq deep learning framework for new task:**
Word- and sentence-based speech completion
- **Our proposed system provided**
 - ✓ Complete full-length results closer to the reference transcriptions
 - ✓ A lower error rate & more natural suggestions than the baseline
- **The speech-to-speech completion system**
 - ✓ Resynthesized the speech input and completed the remaining part
 - ✓ While retaining the characteristics of the speaker's speech style
- **A simple yet efficient approach enables people to easily reproduce the works**
- **Future works**
 - ✓ Refine our framework and incorporate it within an incremental ASR
 - ✓ Investigate the capability using multi-speaker natural speech data

Citations

- **[Hollingsworth, 2018]** – S. Hollingsworth, “Google autocomplete: A complete SEO guide,” Search Engine Journal, 2018.
- **[Goto et al., 2002]** – M. Goto, K. Itou, and S. Hayamizu, “c,” in Proc. ICSLP, 2002
- **[Niehues et al., 2018]** – J. Niehues, N. Pham, T. Ha, M. Sperber, and A. Waibel, “Low-latency neural speech translation,” arXiv preprint arXiv:1808.00491, 2018.
- **[Chan et al., 2016]** – W. Chan, N. Jaitly, Q. Le, and O. Vinyals, et al., “Listen, attend and spell: A neural network for large vocabulary conversational speech recognition,” in Proc. ICASSP, 2016
- **[Wang et al., 2017]** – Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio et al., “Tacotron: A fully end-to-end text-to-speech synthesis model,” arXiv preprint, arXiv:1703.10135, 2017
- **[Kikui et al., 2006]** – G. Kikui, S. Yamamoto, T. Takezawa, and E. Sumita, “Comparative study on corpora for speech translation,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 5, pp.1674–1682, 2006
- **[Kombrink et al., 2011]** – S. Kombrink, T. Mikolov, M. Karafiat, and L. Burget, “Recurrent neural network based language modeling in meeting recognition,” in Proc. INTESPEECH, 2011
- **[Devlin et al., 2014]** – J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pretraining of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810.04805, 2014

Speech Samples

<https://sites.google.com/ahclab.naist.jp/neuralspeechcompletion/home>

Thank you

