Recent Advances in Speech Translation

Dr. Satoshi Nakamura*

with Katsuhito Sudo, Sakriani Sakti[#] ,

and Ryo Fukuda, Sashi Novitasari, Tomoya Yanagita,

Kosuke Doi, Yasumasa Kano, Yuka Ko, Yuki Yano, Hirotaka Tokuyama, Yui Oka

*Director, Data Science Center,

Professor, Graduate School of Science and Technology,

Nara Institute of Science and Technology

([#]currently with JAIST)

Talk Outline

- Recent advances
  - Machine translation
- Speech translation
  - Speech translation research history
  - Simultaneous speech translation
- Summary and future directions

Recent MT progress

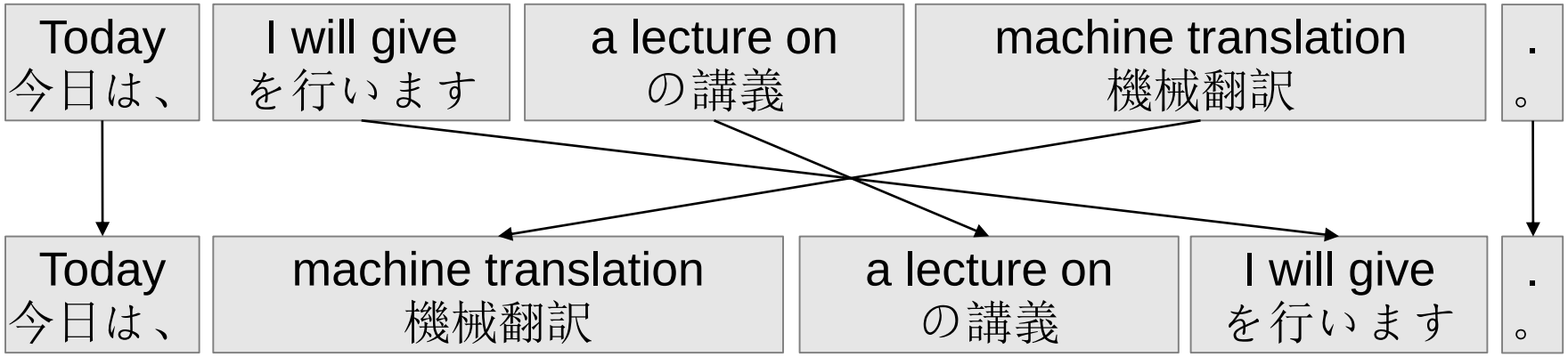
- Rule-based MT：
  - Linguists generate translation rules
- Corpus-based MT:
  - Example-Based: Automatic rule extraction from corpus
    [M.Nagao84, Sato et.al.,89, Sumita et. al., 91]
  - Statistical MT: Statistical Modeling of MT based on Noisy Channel Model [P.F.Brown, et.al. 93]
  - Phrase-base SMT
    - P.Koehn, et al., “Statistical Phrase-based Translation”, UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY INFORMATION SCIENCES INST 2003
- Tree-to-string
  - Statistical MT based on Tree Structure
    - Y.Liu, et al,” Adjoining Tree-to-String Translation”, Proc. ACL 2011
- Neural Machine Translation
  - Combination of Encoder and Decoder by LSTM
    - I.Sutskever, “Sequence to Sequence Learning with Neural Networks”, Proc. NIPS 2014
- Attention NMT
  - Add Attention to encoder and decoder
    - D.Bardanau, et al., “Neural Machine Translation by Jointly Learning to Align and Translate”, Proc. of ICLR 2014
- Transformer NMT
  - Self attention by multiple heads. Transformer.
    - Ashish Vaswani et al., Attention Is All You Need, Proc. of NIPS 2017.

Phrase Based Statistical Machine Translation

- Divide the sentence into small phrases and translate

SVO: Subject Verb Object

Today I will give a lecture on machine translation .



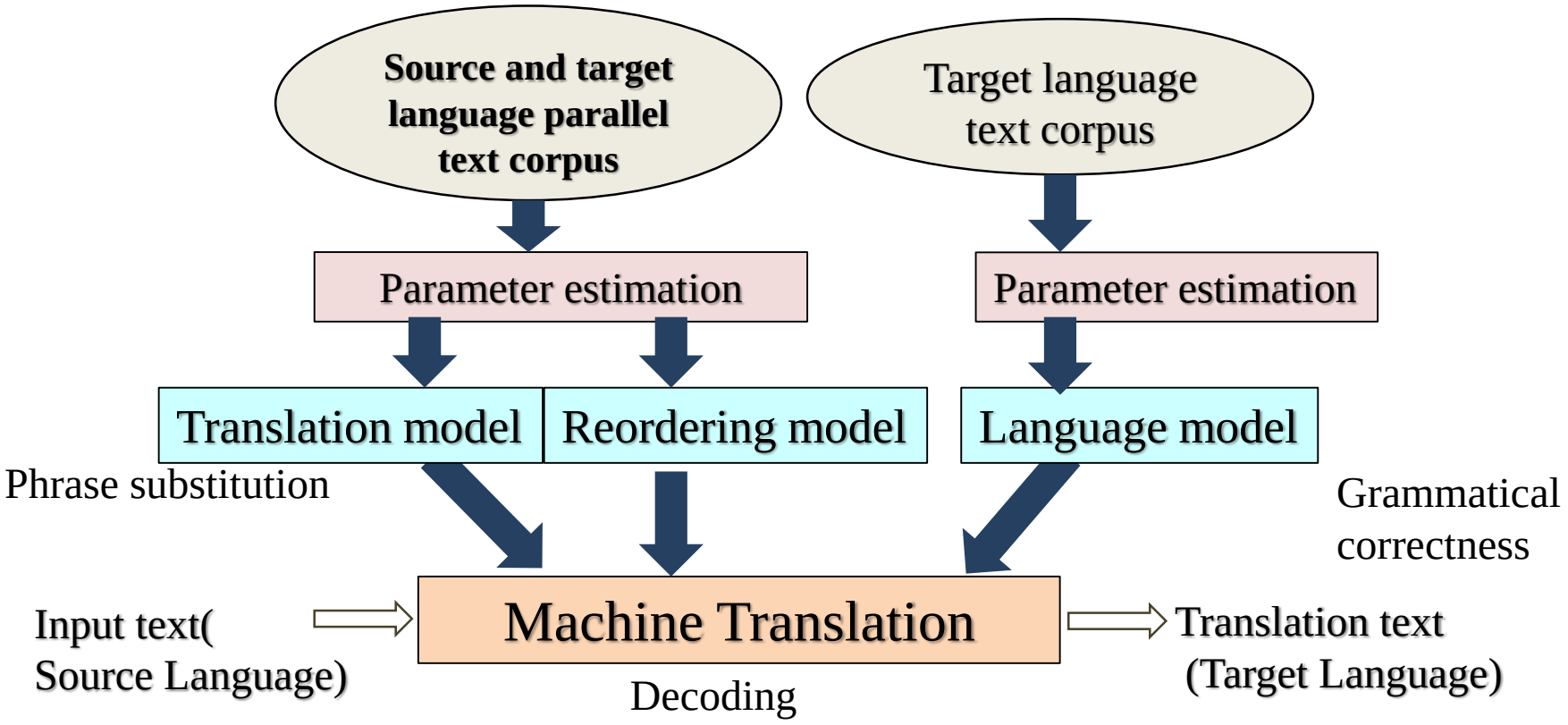
今日は、機械翻訳の講義を行います。
kyowa kikaihonyaku no kogi wo okonaimasu

SOV: Subject Object Verb

- Score translations with translation model (TM), reordering model (RM), and language model (LM)

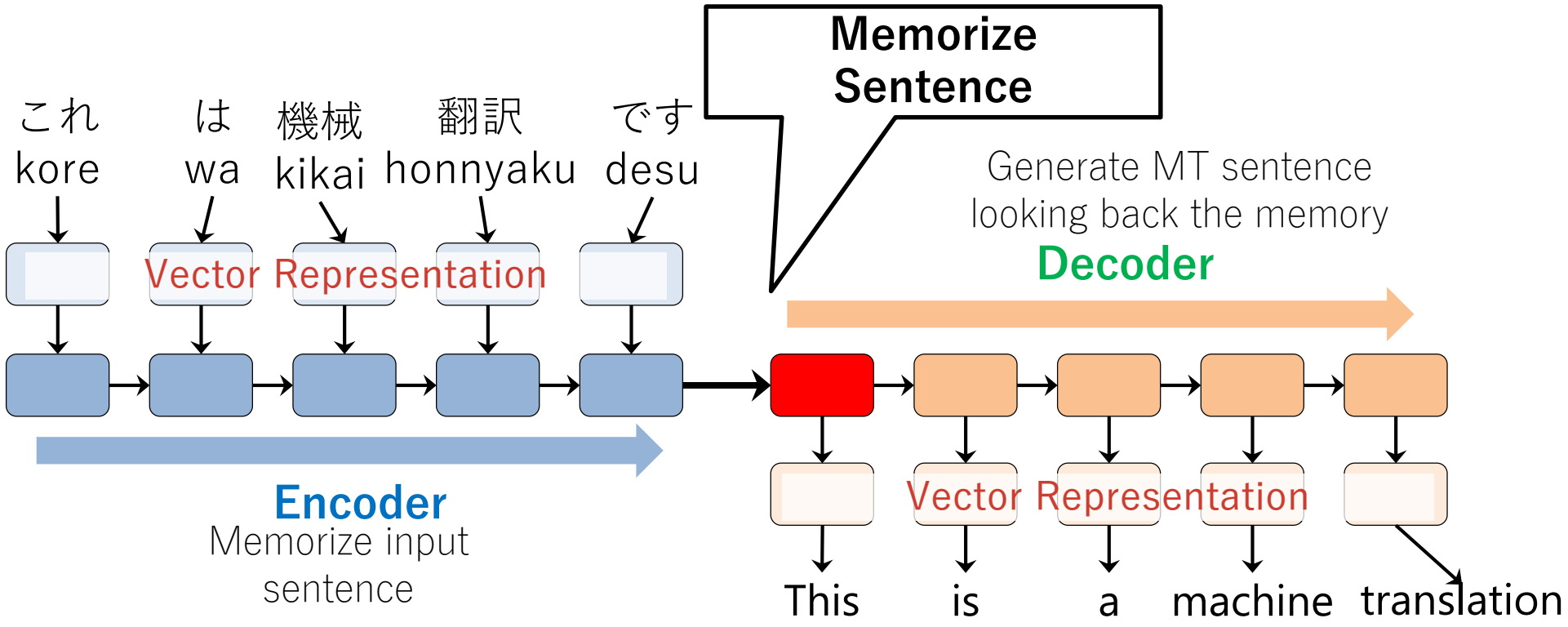
Statistical MT

- Translation model, Language Model, Distortion Model



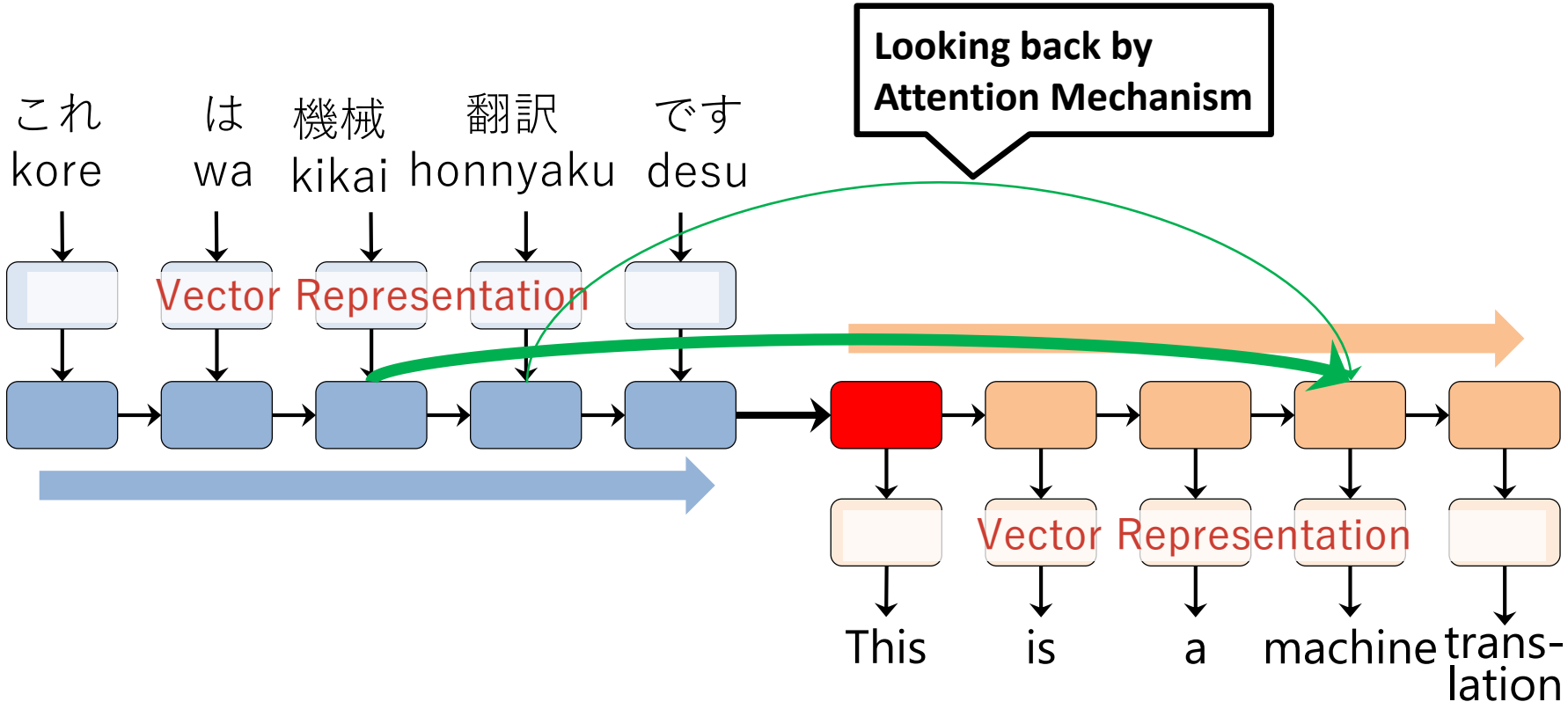
Encoder-decoder Model [Sutskever+ 14]

- Memorize input sentence by LSTM recurrent neural network
- Generate output sentence by LSTM recurrent neural network



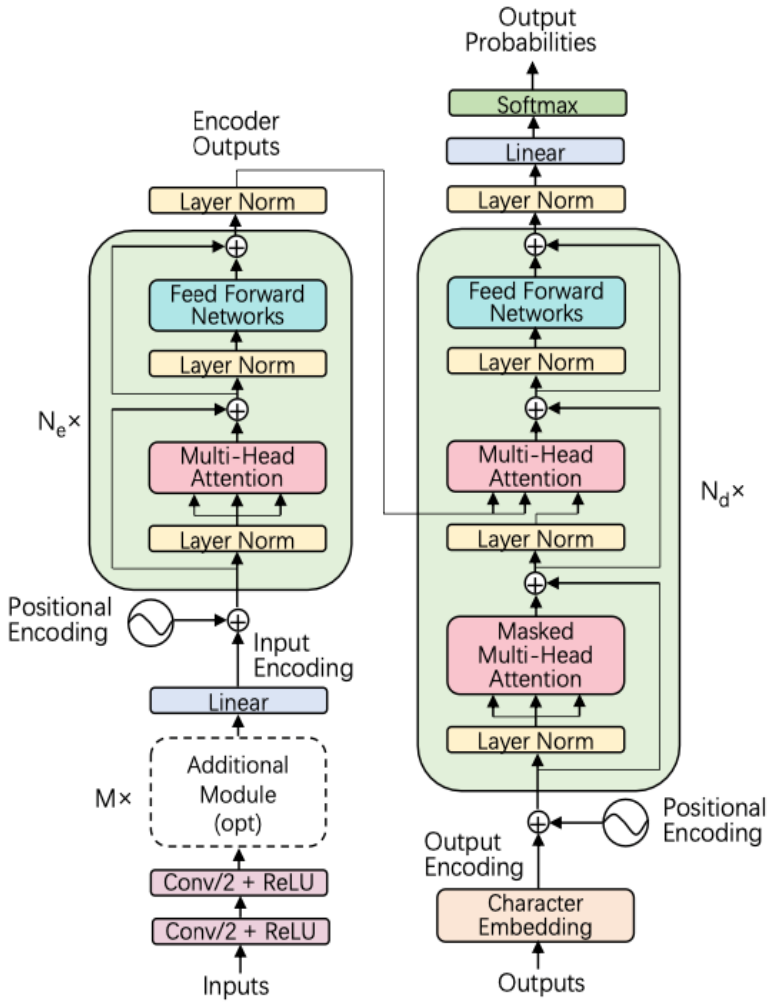
Attention Mechanism [Bahdanau+ 14]

- Better Memorization of Sentence and Looking-back Mechanism
  - Weighted-sum by the attention



Transformer: Fully Attention-based NN [Vaswani+ 2017]

- No RNNs, stacked NNs *Self-attention* to extract context dependent information
- *Positional encoding* instead of recurrent steps
- *Multi-head attention* to utilize various aspects



History of Speech Translation Research in Japan

1986 — 1992 NAIST — 1999 — 2006 — 2008 — 2010

ATR — ATR — ATR — NICT

VoiceTra

Read Speech

- Syntactically correct
- Clear utterance
- Limited domain
  Ex. “Conference Registration”

Daily Conversation

- Standard expression
- Unclear utterance
- Limited domain
  Ex. “Hotel Reservation”

Wider and Real Domain

- Wider and real domain
  “International Travel”
- Realistic expressions
- Noisy speech
- J-E, J-C speech translation

+ More languages for translation

- 21 multilateral text translation
- Multilate
- Network

C-STAR

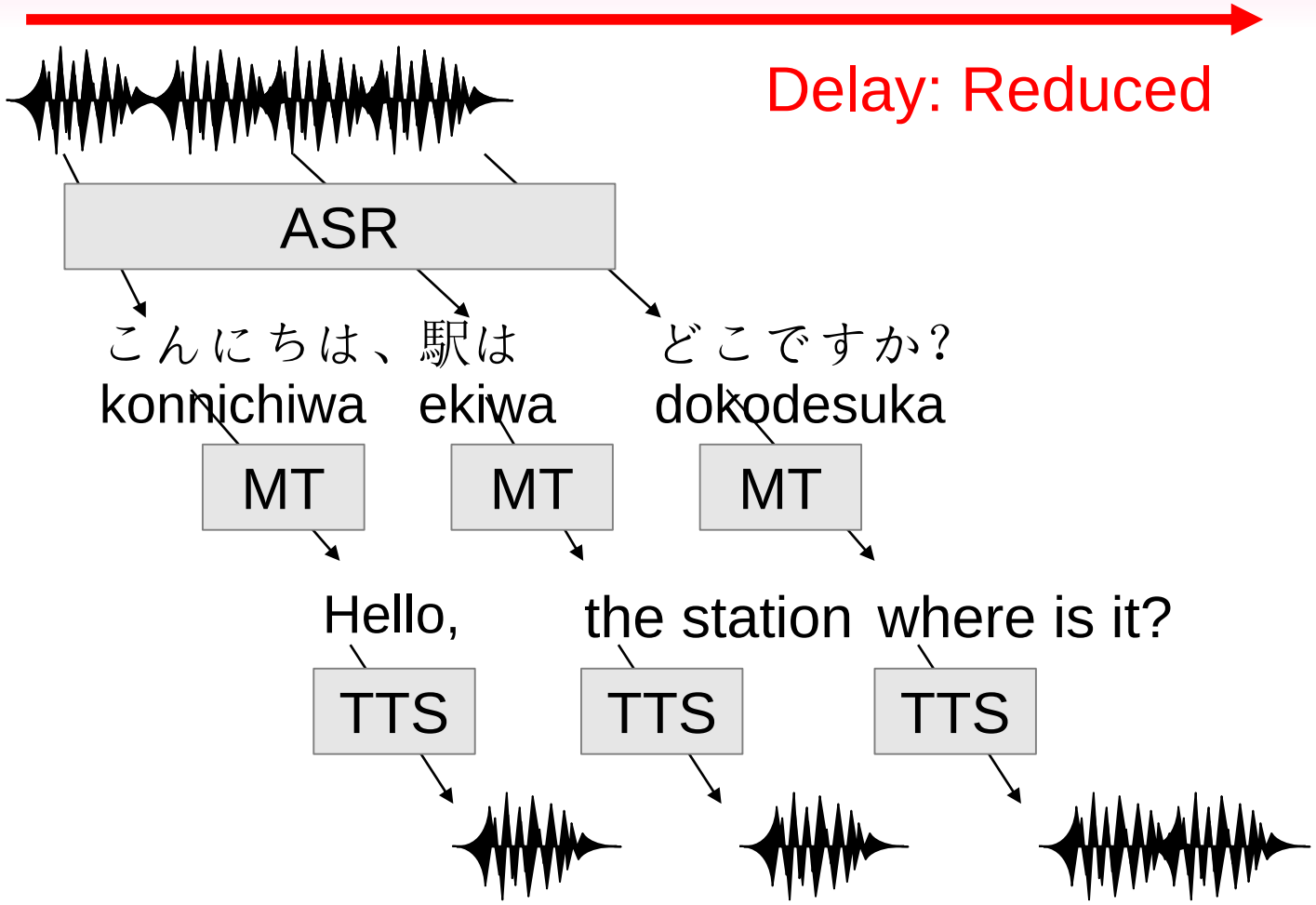
- Multilateral translation for 7 world languages
- Evaluat

Simultaneous Speech Translation

Fundamentals

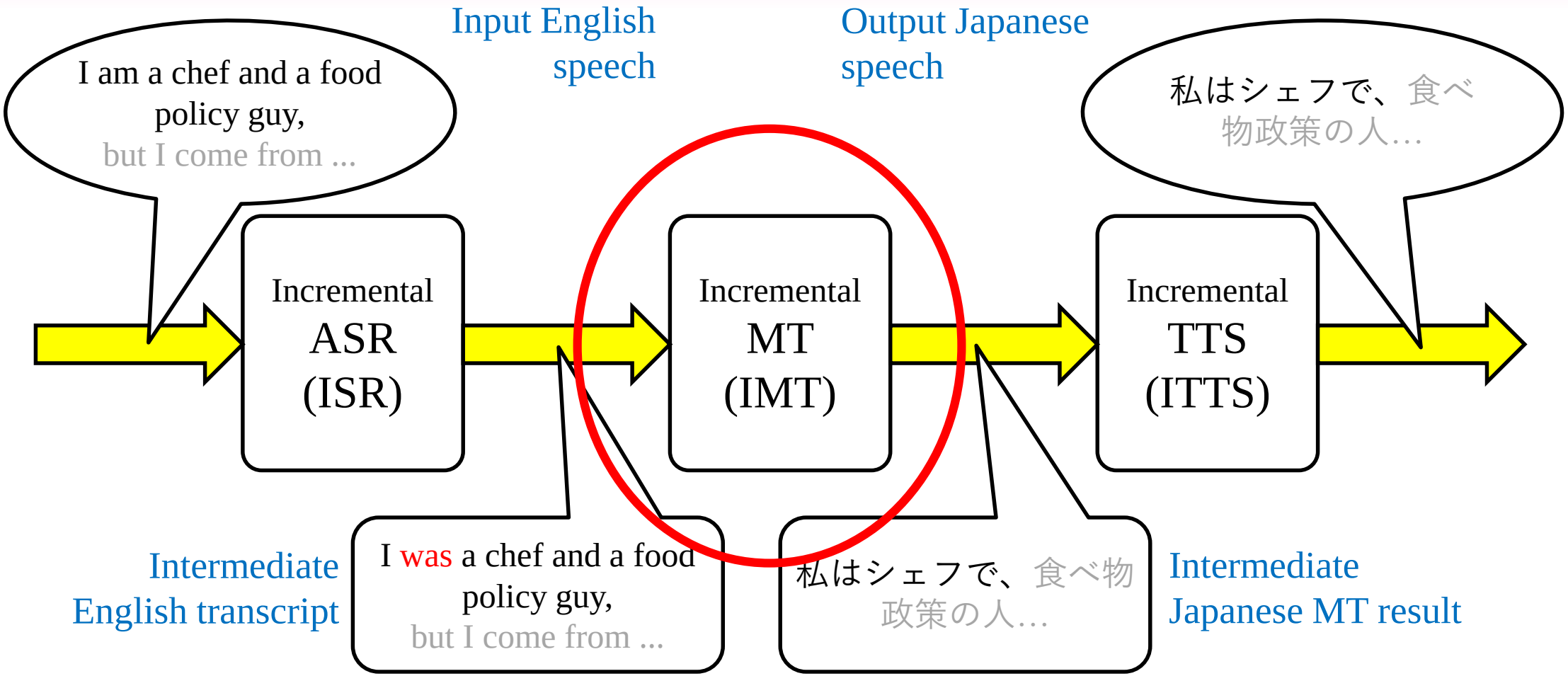
Rule-based Machine Translation → Statistical Machine Translation → Neural Machine Translation

Simultaneous Incremental Speech Translation

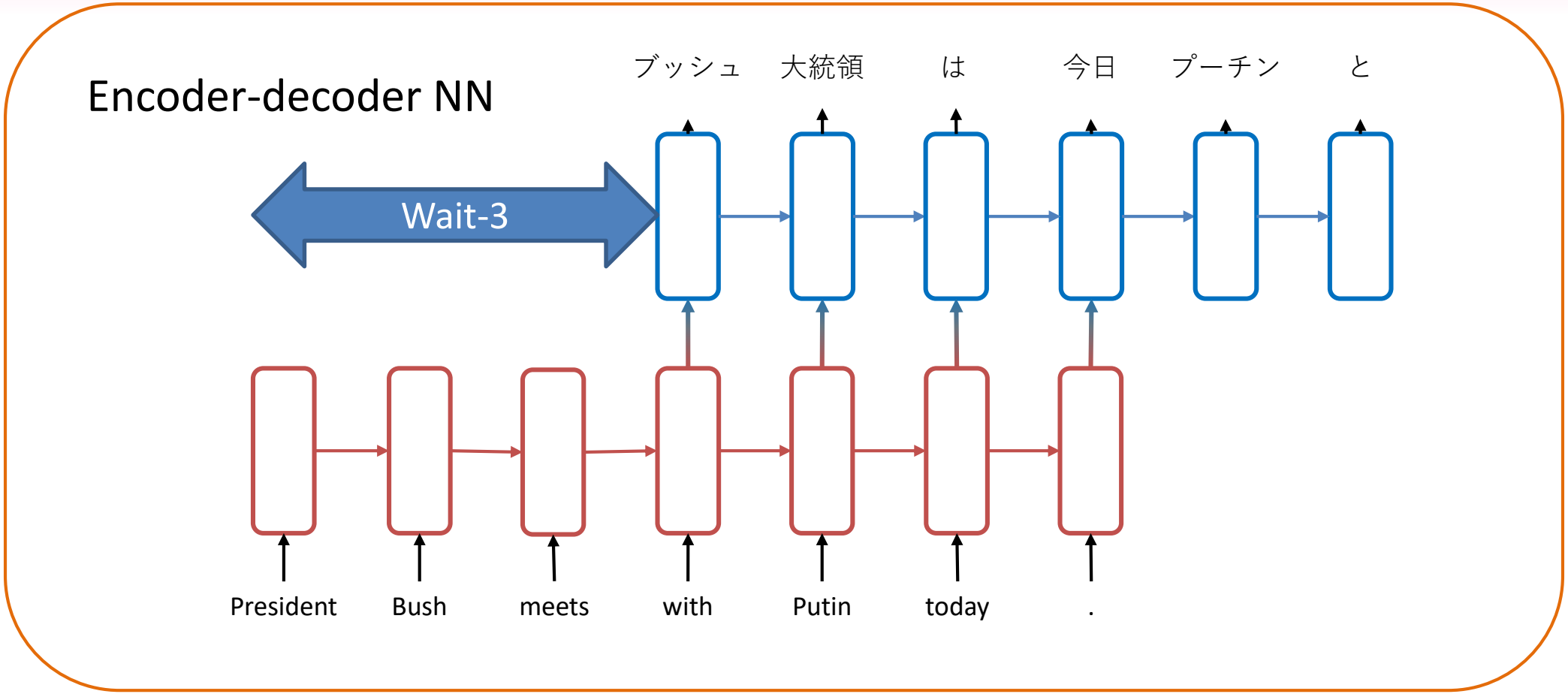


But, this is not easy!

Cascade Simultaneous S2S Translation System



STACL: Wait-k Algorithm [M.Ma, et al., 2018]

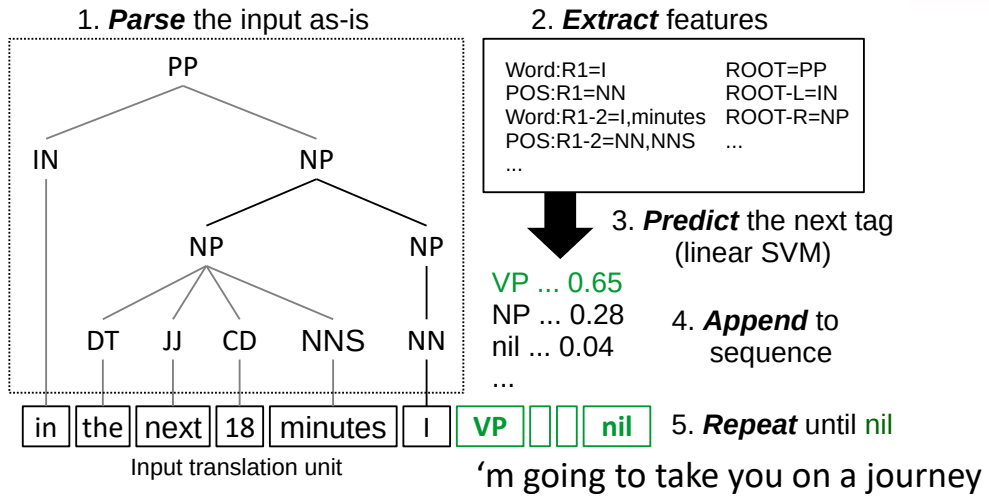


Mingbo Ma, et al., "STACL: Simultaneous Translation with Integrated Anticipation and Controllable Latency", arXiv:1810.08398v3 [cs.CL] 3 Nov 2018

Translation Timing Control by Syntactic Prediction in SMT (2015,2021)

- Syntactic Prediction [Oda, et al., 2015]
  - Incremental bottom up parsing
  - Feature extraction and syntactic prediction

Subject Verb Object : English
Subject Object Verb: Japanese



- Wait MT output when specific labels appear.
  - Control MT output timing according reordering

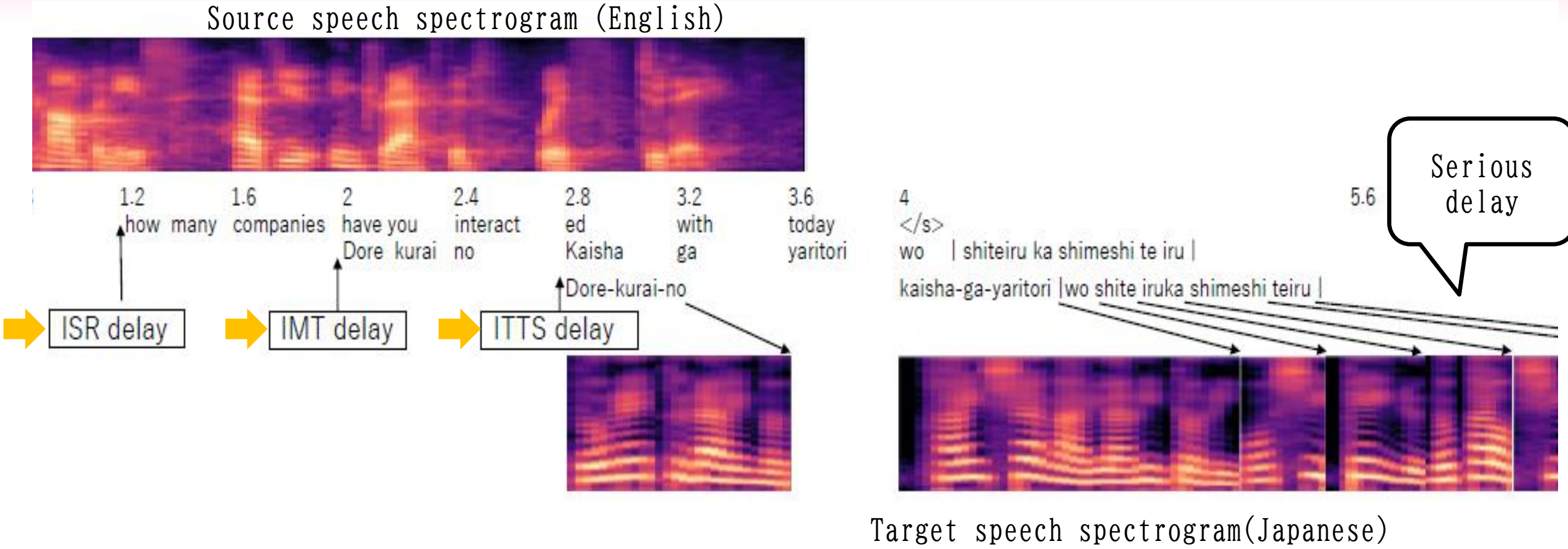
- Use LSTM and BERT to predict next tag. [Kano, et al., 2021]

Incremental parsing and syntactic prediction	in the next 18 minutes[NP] predict [VP] (wait) i 'm going to take **[keep]** i 'm going to take you on a journey [VP end]
MT results	18 分 で あ る[NP] [VP](wait) を 行 っ て い ま す[keep] 皆さん を 旅 に お連れ し ます [VP end]

Oda, Yusuke *et al.*, Syntax-based Simultaneous Translation through Prediction of Unseen Syntactic Constituents, Proc. of ACL-IJCNLP 2015.
Y.Kano, K.Sudo, S.Nakamura, "Simultaneous Neural Machine Translation with Constituent Label Prediction", Proc. of WMT 2021.

Video

Problem in the current Simultaneous Speech Translation System



Evaluation: Quality (ASR & MT for TED Talks)

ISR

	Model	WER (%)
Non-incremental	LSTM	25.46
	Transformer	20.74
Incremental	LSTM (low latency)	31.88
	LSTM (high latency)	32.43
	Transformer (low latency)	32.06
	Transformer (high latency)	25.01

ISR+IMT

	BLEU-4 (%)	Subjective evaluation	
		Adequacy	Fluency
Gold transcript+MT	15.7	3.41	3.93
Non-incremental ASR+MT	12.8	3.20	4.01
IMT (low latency EVS:8.81s)	5.1	2.80	3.03
IMT (medium latency EVS:11.87s)	8.4	2.98	3.54
IMT (high latency EVS:16.91s)	9.4	3.34	3.80

R.Fukuda, et al, "SIMULTANEOUS SPEECH-TO-SPEECH TRANSLATION SYSTEM WITH TRANSFORMER-BASED INCREMENTAL ASR, MT, AND TTS", Proc. Oriental COCOSDA 2021

Summary

- Remarkable progress
  - Statistical Models
  - Recurrent Deep Neural Network
  - Progress in Speech Translation

- Automatic Simultaneous Speech Translation
  - Data Collection
  - Automatic Simultaneous Speech Translation for distant language-pairs

- Further Research
  - Higher Quality with Shorter Latency
  - Evaluation of Simultaneous Speech Interpretation
  - Context/ Situation dependency
  - Semantics, Discourse Analysis
  - Para-linguistics/ Multi-modal

Thank you for listening

Human Interpreting[A.Mizuno 2016]

E-J Translation Example

(1) The relief workers (2) say (3) they don't have (4) enough food, water, shelter, and medical supplies (5) to deal with (6) the gigantic wave of refugees (7) who are ransacking the countryside (8) in search of the basics (9) to stay alive.

(1) 救援担当者は (9) 生きるための (8) 食料を求めて (7) 村を荒らし回っている (6) 大量の難民達の (5) 世話をするための (4) 十分な食料や水，宿泊施設，医療品が (3) 無いと (2) 言っています.

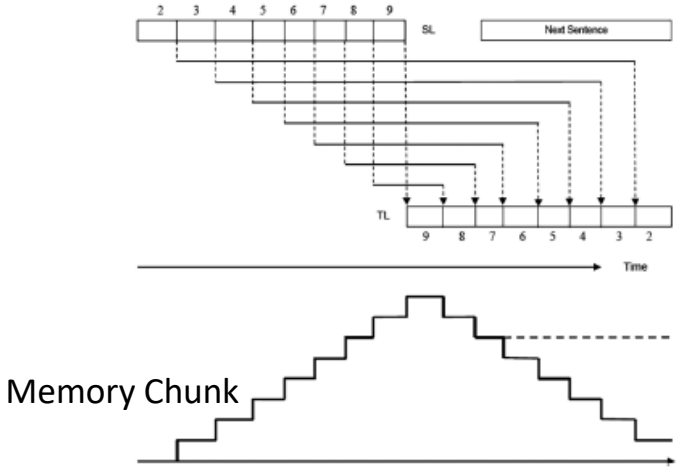
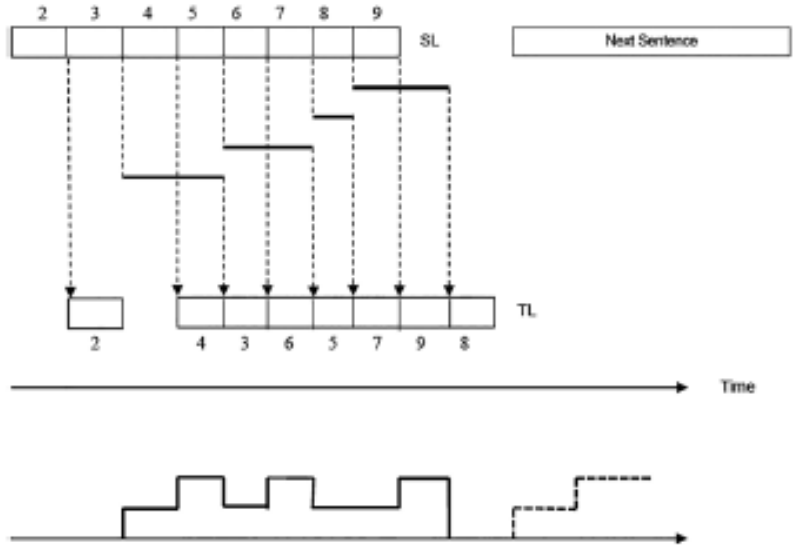


Fig.4 Translation to seek syntactic correspondence and its load
The dotted line of lower right indicates assumed load when next sentence comes in before the completion of translation of previous sentence.

Necessary #Chunk＞３！

(1) 救援担当者達の (2) **話では** (4)食料，水，宿泊施設，医薬品が，(3) 足りず (6) 大量の難民達の (5) 世話が出来ないとのことです. (7) **難民達は**今村々を荒らし回って，(9) 生きるための (8) 食料を求めているのです.



Necessary #Chunk＜３！

Recent Progress of ASR

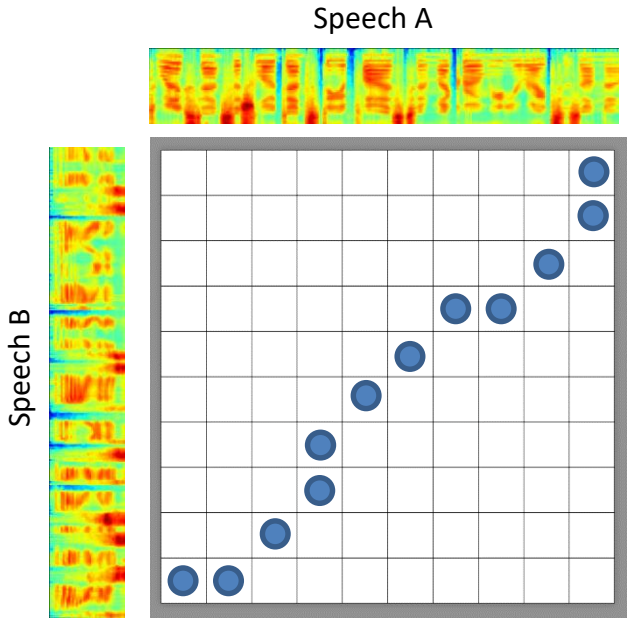
- Traditional Technologies
  - Template Matching, Dynamic Programing [Sakoe 71]
  - Hidden Markov Modeling, N-Gram Model [Mercer 83, etc]
  - Neural Network, TDNN[Waibel 89], LSTM [Hochreiter 97]
  - Weighted Finite State Transducer [Mohri 2006]
  - Big Training Data, Data Collection through Trial Service

- Deep Learning
  - DNN-HMM [Hinton 2012]
    - Estimate State Posterior Probability by DNN
  - Connectionist Temporal Classification [Graves 2013]
    - Predict Phoneme Label every frame
  - Listen, Attend, and Spell [Chan 2016]
    - CTC+Attention: End-to-end modeling
  - Transformer ASR
    - Faster calculation by Multiple-head attention

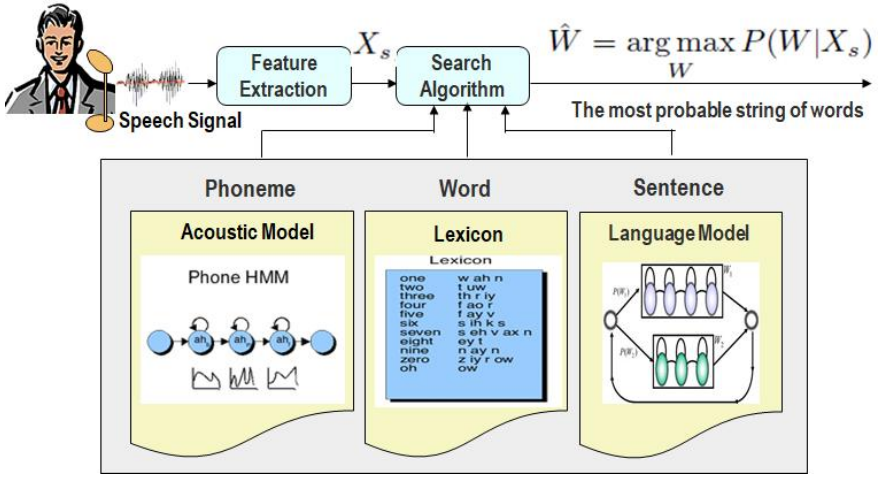
Speech Recognition

1960 → 1990 → 2014+

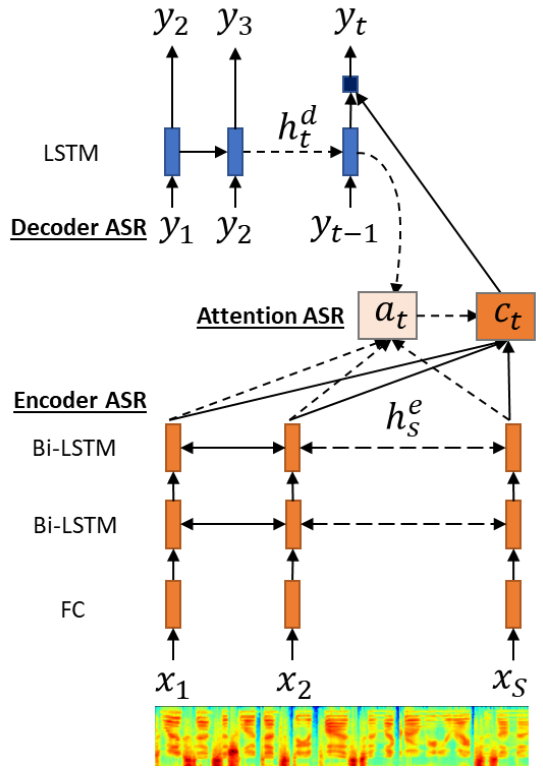
Template matching dynamic time warping



Statistical-based HMM



End-to-end ASR



Speech Recognition (2)

- Speech recognition model (Y= text; X= speech feature)

$$\arg\max p(Y|X) = \arg\max p(X|Y)\, p(Y)$$

- Conventional speech recognition (statistical-based HMM)
  - Acoustic model p(X |L)
  - Lexicon p(L |Y)
  - Language model p(Y)

- End-to-end ASR
  - Directly model p(Y|X) with a single network
  - Integrate acoustic p(X|L), lexicon p(L|Y), and language model p(Y) models

Connectionist temporal classification (CTC)

[Graves et al. 2014, Miao et al., 2015]

- Speech (X) and text (Y) can vary in length
- Use bidirectional RNNs to predict frame-based labels (including blanks)
- Find speech-text alignments using dynamic programming
- CTC objective

$$p(Y|X) = \sum_{A \in A_{X,Y}} \prod_{t=1}^{T} p_t(a_t|X)$$

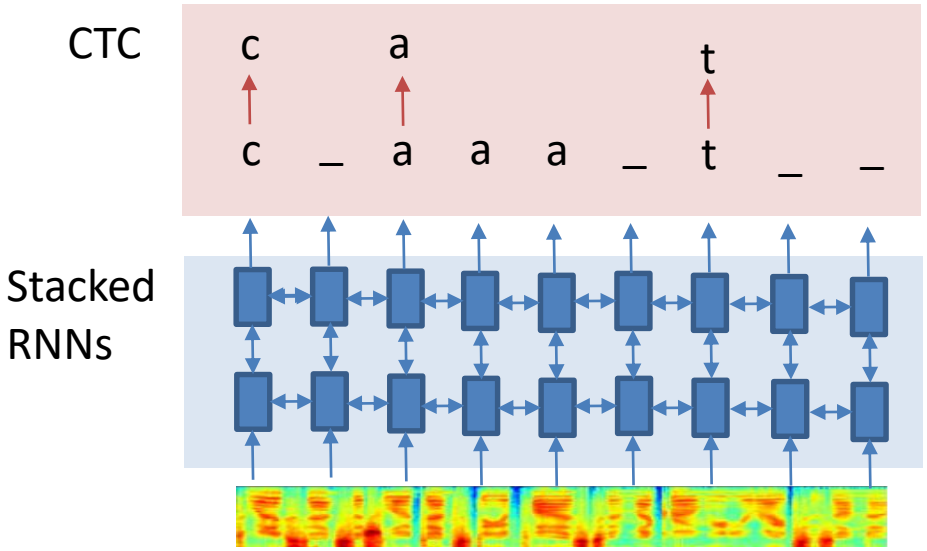
CTC conditional probability

Marginalizes over the set of output candidates

Computing the probability for a single alignment step-by-step

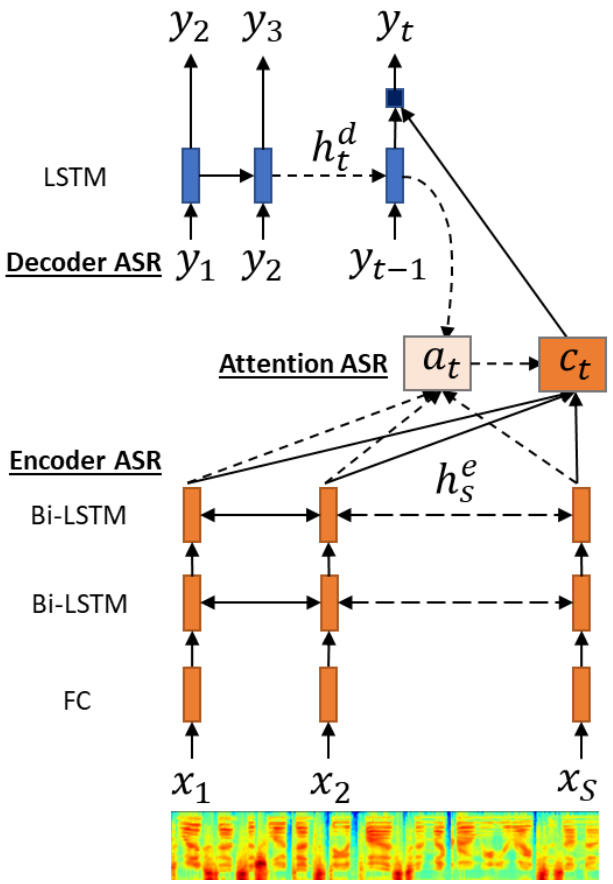
- ASR loss: negative log-likelihood

$$\sum_{(X,Y)} -\log p(Y|X)$$



Attention-based encoder-decoder

[Chorowski et al., 2014; Chan et al., 2015]



Input: Mel-spectrogram (continuous)
Output: Character / sub-word (discrete)

- Encoder: stacked bidirectional RNNs
  - Encode speech features into hidden representation
- Decoder: stacked RNNs
  - Decode the encoded representation using the previous output and attention information
- Attention:
  - Alignments between decoder state - encoder state
  - Decoder finds only relevant information from encoder states.
- ASR loss: negative log-likelihood

$$\ell_{ASR} = L_{ASR}(\mathbf{y}, \mathbf{p}_y) = -\frac{1}{T}\sum_{t=1}^{T}\sum_{c=1}^{C} \mathbb{1}(y_t = c) * \log p_{y_t}[c]$$

Recent Performance

- Saon, et al. “English Conversational Telephone Speech Recognition by Humans and Machines”, INTERSPEECH 2017

Table 1: *Word error rates on SWB and CH for human transcribers before and after quality checking contrasted with the human WER reported in [1].*

	WER SWB	WER CH
Transcriber 1 raw	6.1	8.7
Transcriber 1 QC	5.6	7.8
Transcriber 2 raw	5.3	6.9
Transcriber 2 QC	**5.1**	**6.8**
Transcriber 3 raw	5.7	8.0
Transcriber 3 QC	5.2	7.6
Human WER from [1]	5.9	11.3

[1] R. P. Lippmann, “Speech recognition by machines and humans,” Speech communication, vol. 22, no. 1, pp. 1–15, 1997.

Table 3: *Word error rates for LSTMs, ResNet and frame-level score fusion results across all testsets (36M n-gram LM).*

Model	SWB	CH	RT’02	RT’03	RT’04
LSTM (baseline)	7.7	14.0	11.8	11.4	10.8
LSTM1 (SA-MTL)	7.6	13.6	11.5	11.0	10.7
LSTM2 (Feat. fusion)	7.2	12.7	10.7	10.2	10.1
ResNet	7.6	14.5	12.2	12.2	11.5
ResNet+LSTM2	6.8	12.2	10.2	10.0	9.7
ResNet+LSTM1+LSTM2	6.7	12.1	10.1	10.0	9.7

Table 4: *WER on SWB and CH with various LM configurations.*

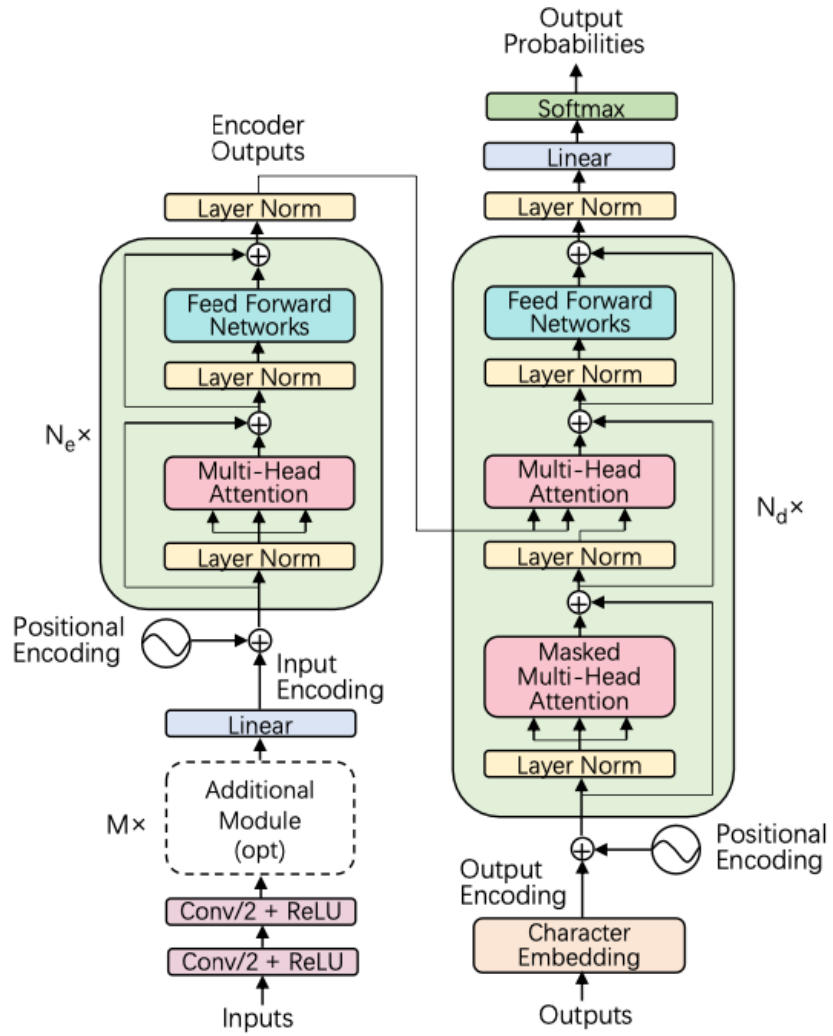
	WER [%]	
	SWB	CH
n-gram	6.7	12.1
n-gram + model-M	6.1	11.2
n-gram + model-M + Word-LSTM	5.6	10.4
n-gram + model-M + Char-LSTM	5.7	10.6
n-gram + model-M + Word-LSTM-MTL	5.6	10.3
n-gram + model-M + Char-LSTM-MTL	5.6	10.4
n-gram + model-M + Word-DCC	5.8	10.8
n-gram + model-M + 4 LSTMs + DCC	**5.5**	**10.3**

Attention-based encoder-decoder with Transformer

[Dong et al., 2020]

- Replace RNN with Transformer
- Encoder and decoder with
  - **Self-attention** to generate inner sequence alignment
  - **Multi-head attention** to leverage different attending representation jointly
  - **Positional encoding** to represent positinal information (Transformer ≠ RNN)
- ASR loss: negative log-likelihood

$$\ell_{ASR} = L_{ASR}(\mathbf{y}, \mathbf{p}_y) = -\frac{1}{T}\sum_{t=1}^{T}\sum_{c=1}^{C} \mathbb{1}(y_t = c) * \log p_{y_t}[c]$$



[Dong et al., 2020]

Talk Outline

- Recent advances
  - Machine translation (text-to-text)
  - Speech recognition (Speech-to-text)
  - Speech synthesis (Text-to-speech)

- Speech translation
  - Speech translation research history
  - End-to-end speech translation
  - Simultaneous speech translation

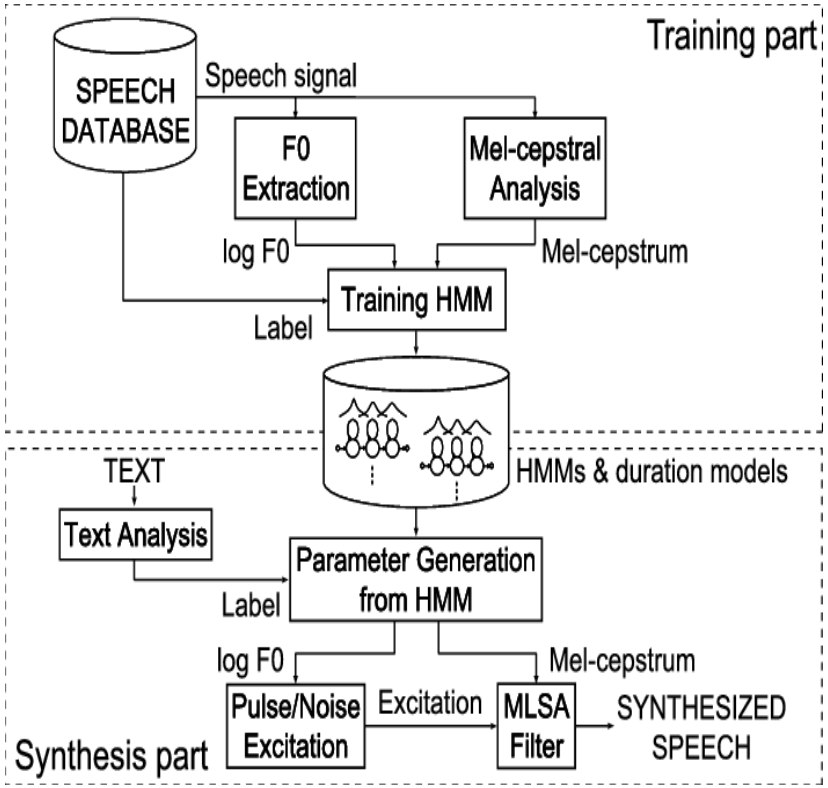
- Summary and future directions

Recent Speech Synthesis

- Formant-based Synthesis, Waveform Concatenation
- Statistical Speech Synthesis: HTS
  - Speech Synthesis by HMM
    - Tokuda, et al., “Speech parameter generation algorithms for HMM-based speech synthesis”, ICASSP 2000
- WaveNet
  - Waveform Convolution
    - van den Oord et al., “WAVENET: A GENERATIVE MODEL FOR RAW AUDIO”, arXiv:1609.03499v2 [cs.SD] 19 Sep 2016
- Tacotron (Encoder-decoder DNN TTS+ Griffin Lim)
  - End-to-end speech synthesis with character input. Waveform generation by Griffin-Lim
    - Wang, et al., “TACOTRON: TOWARDS END-TO-END SPEECH SYNTHESIS”, arXiv:1703.10135v2 [cs.CL] 6 Apr 2017
- Tacotron2 (Encoder-decoder DNN TTS+ Wavenet)
    - J. Sheng, et al, “Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions”,
- Transformer TTS:
  - Multi-head attention
    - N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, “Neural speech synthesis with Transformer network,” in Proc. AAAI Conference on Artificial Intelligence, vol. 33, no. 01, 2019, pp. 6706–6713
    - M. Chen, X. Tan, Y. Ren, J. Xu, H. Sun, S. Zhao, and T. Qin,“MultiSpeech: Multi-speaker text to speech with Transformer,” in Proc. INTERSPEECH, 2020, pp. 4024–4028.

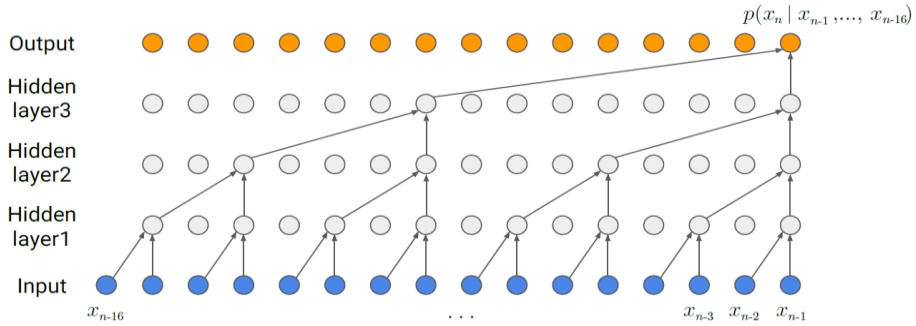
Speech Synthesis

HMM



[Zen et al. 2009]

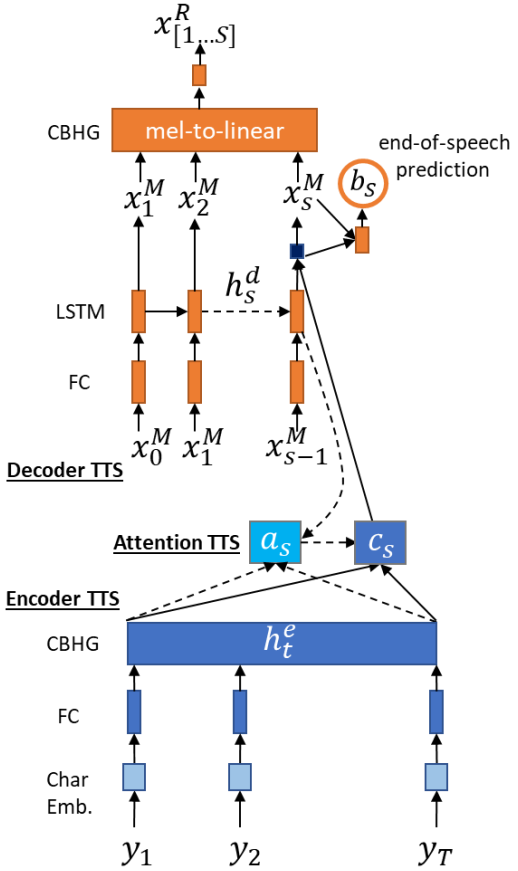
Conditional WaveNet – TTS



$$p(\mathbf{x} \mid \mathbf{h}) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1}, \mathbf{h})$$

[Oord et al., 2016]

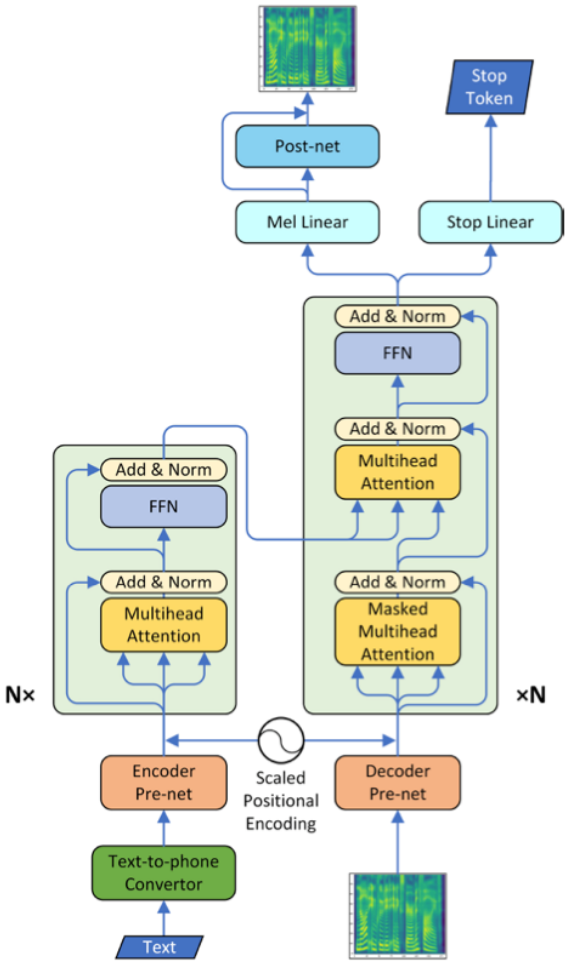
Tacotron



[Wang et al.; 2017]

Sequence-to-sequence TTS with Transformer

[Li et al., 2018]



[Li et al., 2018]

Input: Character / sub-word (discrete)

Output:

- Mel-spectrogram (continuous)
- End-of-speech flag (binary)

- Replace RNNs in Tacotron with Transformer
- Autoregressive decoding:
  - Use the previous decoder timestep output to predict the current timestep output (same as Tacotron)
- Loss TTS: L1/L2 norm + cross-entropy for end-of-speech

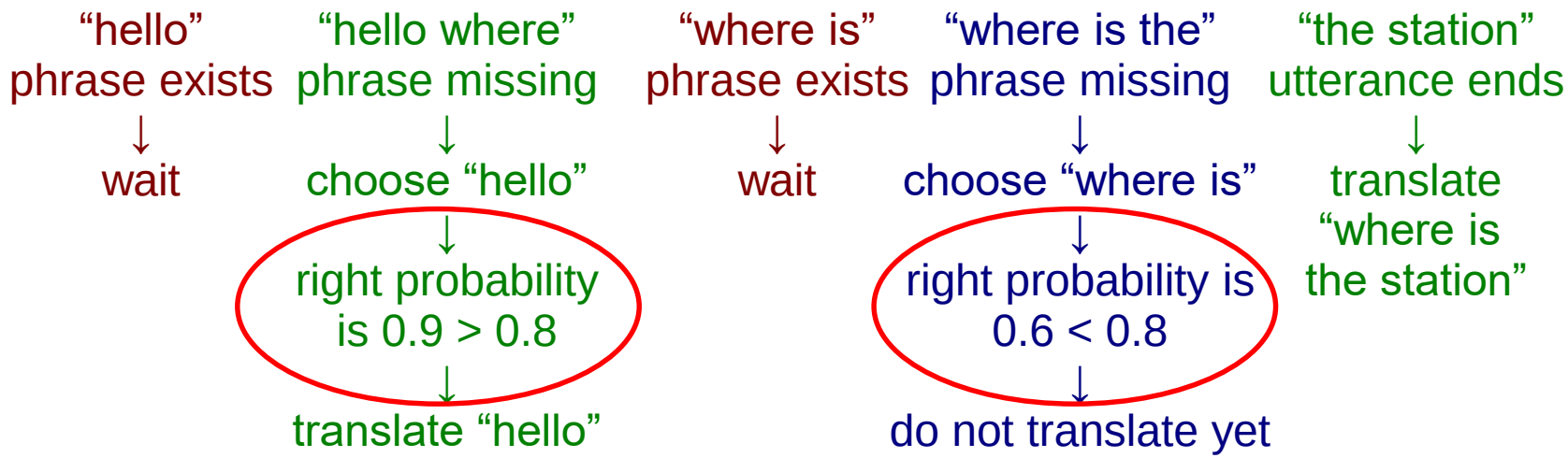
$$\ell_{TTS} = L_{TTS}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{S}\sum_{s=1}^{S} \|x_s^M - \hat{x}_s^M\|_2^2 + \|x_s^R - \hat{x}_s^R\|_2^2 - (b_s \log(\hat{b}_s) + (1-b_s)\log(1-\hat{b}_s))$$

Adjusting Timing with Reordering Probabilities in SMT[Fujita,2013]

- First, temporarily choose strings according to method one
- Next, if that phrase's right probability exceeds a threshold, actually translate the words in the cache

Example (threshold = 0.8):

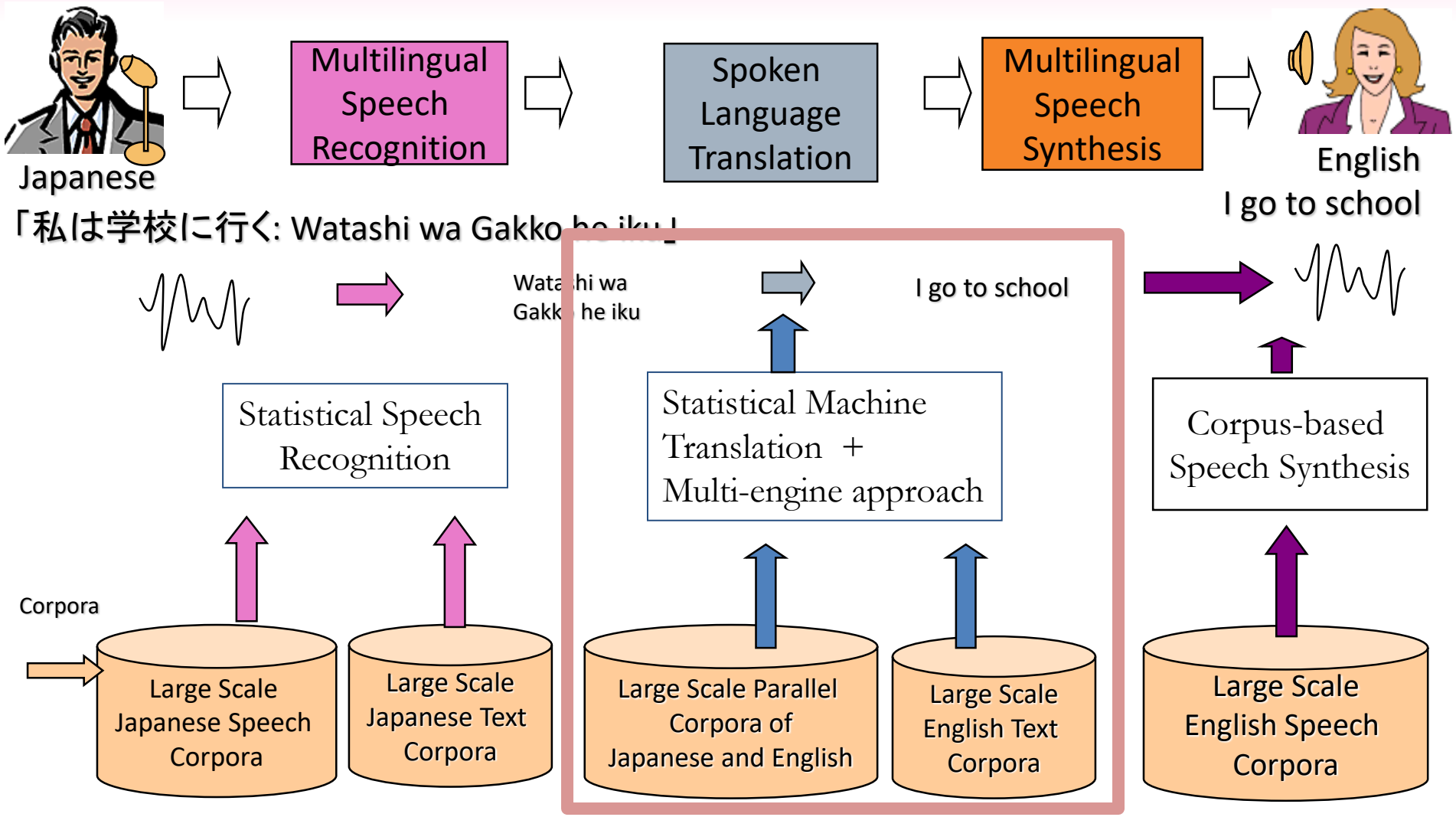
hello where is the station



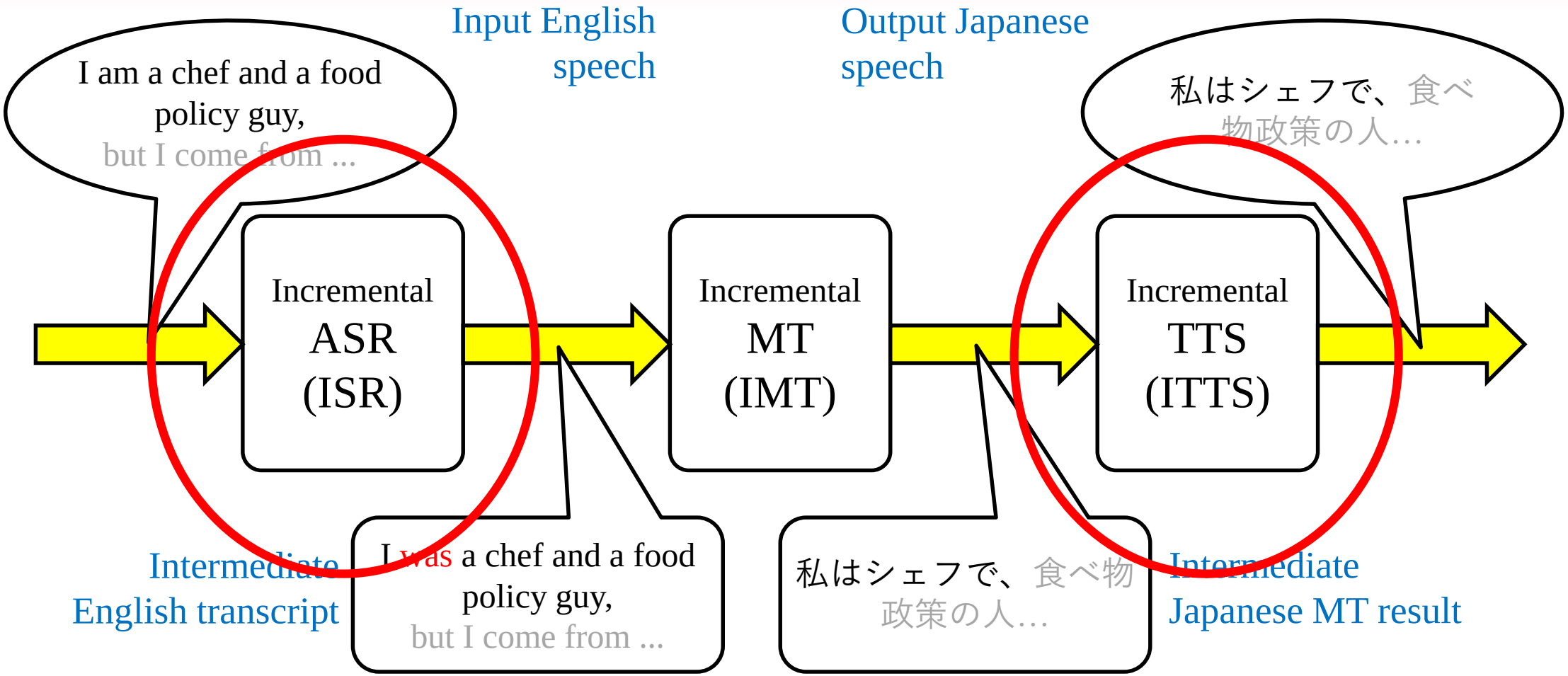
Fujita, et. al., 2013

- Threshold 1.0 = traditional, 0.0 = method one

Mechanism of speech-to-speech translation system



Cascade Simultaneous S2S Translation System



Incremental Speech Recognition (ISR) [Novitasari, 2020]

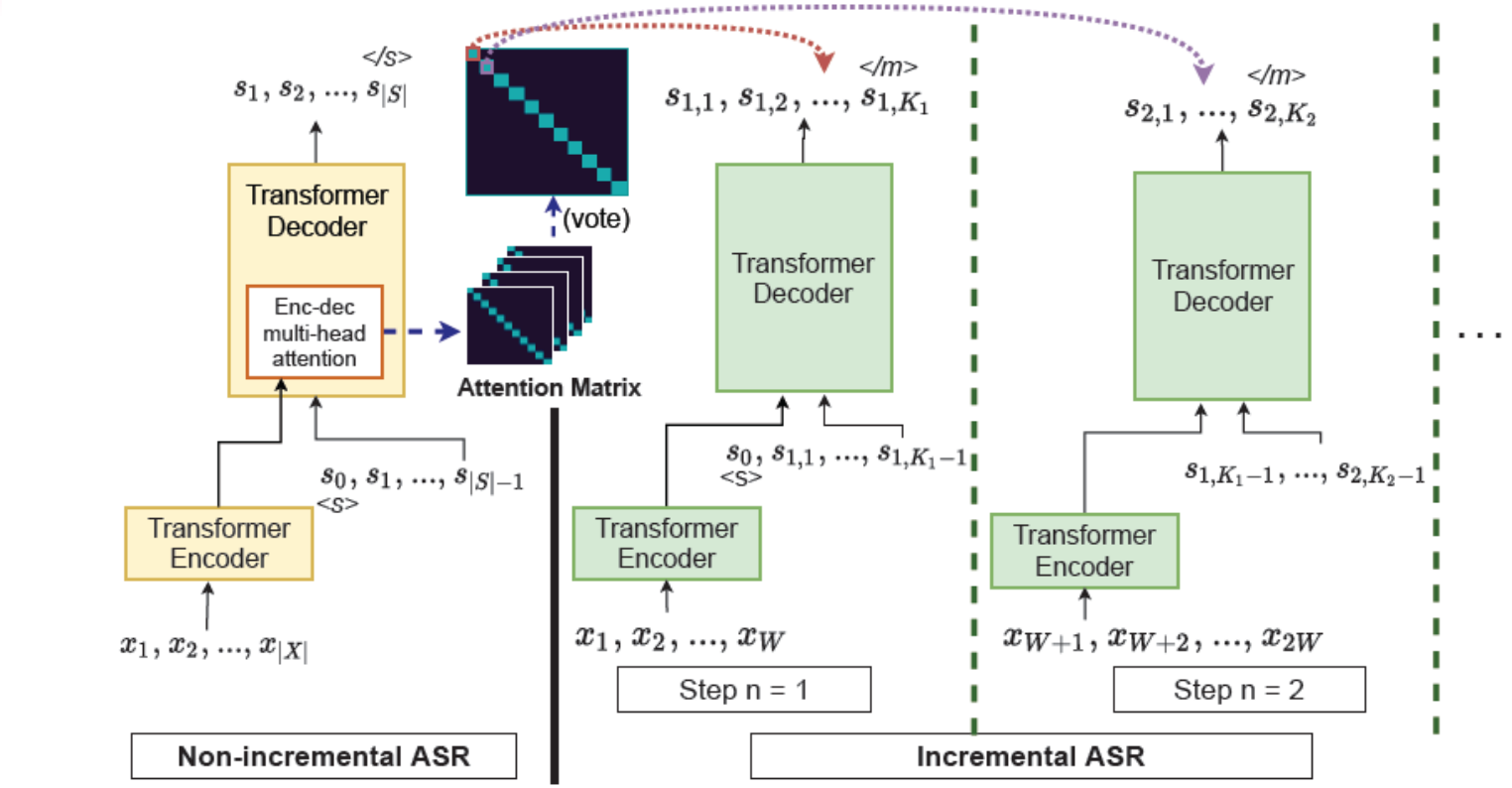


Fig. 1. Transformer-based ISR construction with attention transfer [9] from a standard Transformer-based ASR

S.Novitasari, A.Tjandra, T.Yanagita, S.Sakti, S.Nakamura, "Incremental Machine Speech Chain Towards Enabling Listening while Speaking in Real-time", Proceedings of INTERSPEECH 2020, Oct. 2020

Incremental Speech Synthesis (iTTS)

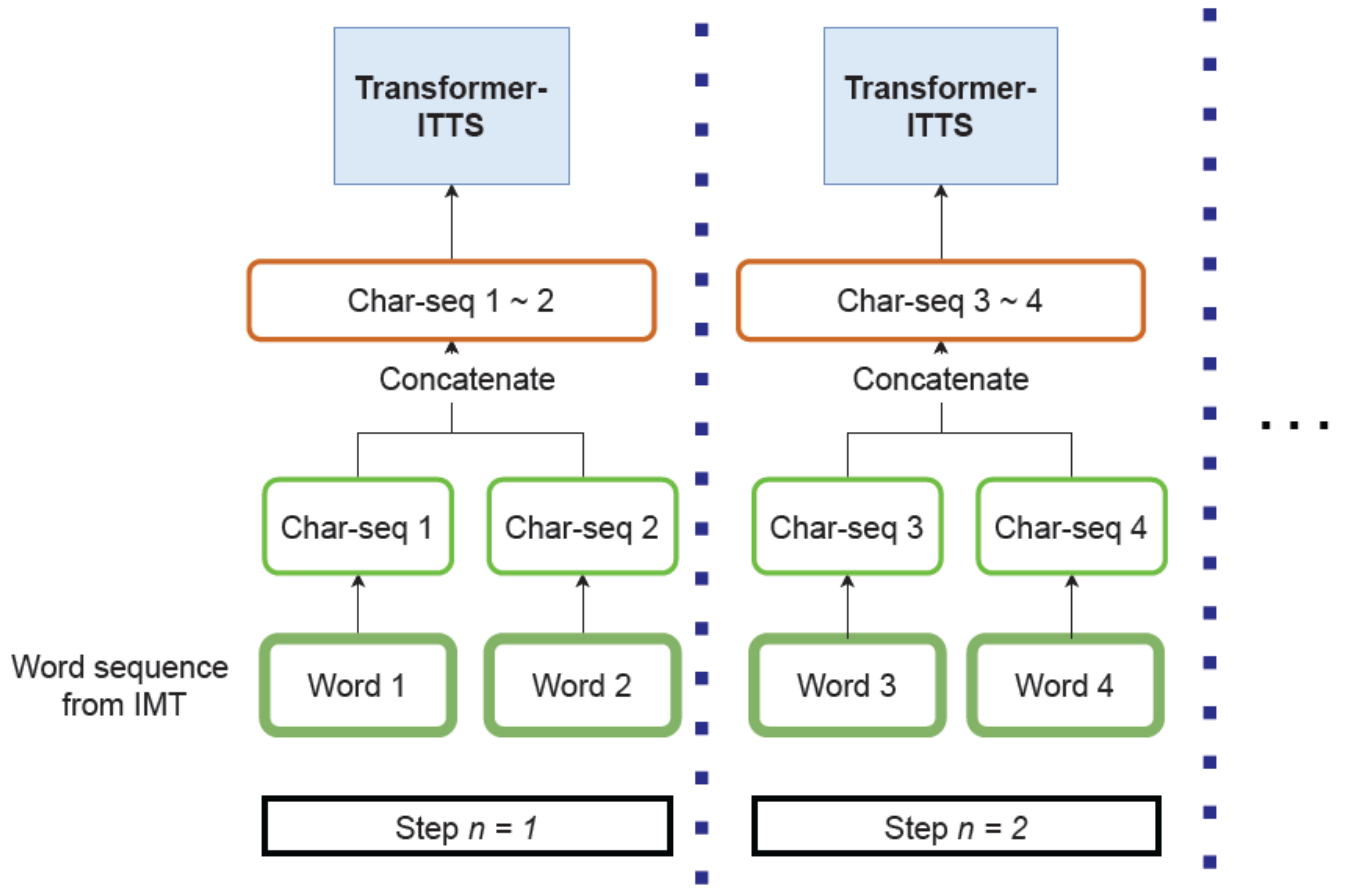


Fig. 2. Incremental text-to-speech synthesis system

S.Novitasari, S.Sakti, S.Nakamura, "Dynamically Adaptive Machine Speech Chain Inference for TTS in Noisy Environment: Listen and Speak Louder", Proc. Interspeech 2021, 4124-4128, Aug. 30, 2021

Overview of our corpus

- Over 300 hours
- * = interpreted by interpreters from all 3 ranks (4h x 3 interpreters)
- Others = interpreted by either an S- or A-rank interpreter
- About half of the SIs have been transcribed

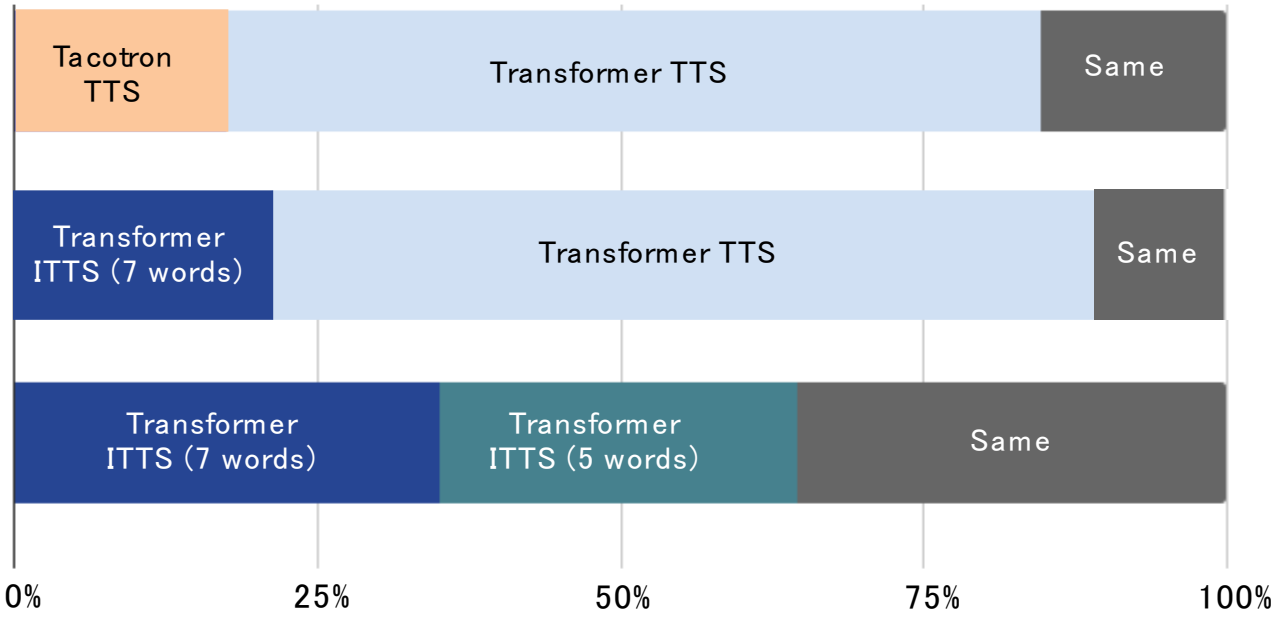
Direction	Source	2018	2019	2020
En → Ja	TED	67+12*	50	50
Jp → En	TEDx	12*	40	0
	CSJ	33	0	0
	JNPC	4	36.5	0
Total		128	126.5	50
Cum.		128	254.5	304.5

Experience	Rank
15 years	S-rank
4 years	A-rank
1 years	B-rank

Detailed analysis: K.Doi, K.Sudoh, S.Nakamura, "Large-scale English-Japanese Simultanous Interpretation Corpus: Construction and Analysis with sentence-aligned data, Proc. IWSLT 2021

Evaluation: Quality (TTS)

	Model	L2-norm loss
Non-incremental	Tacotron	0.57
	Transformer	0.51
Incremental	Tacotron (low latency)	0.77
	Tacotron (high latency)	0.58
	Transformer (low latency)	0.65
	Transformer (high latency)	0.57



R.Fukuda, et al, "SIMULTANEOUS SPEECH-TO-SPEECH TRANSLATION SYSTEM WITH TRANSFORMER-BASED INCREMENTAL ASR, MT, AND TTS", Proc. Oriental COCOSDA 2021

Evaluation Overview

- The cascade system was evaluated using TED Talks data in:
  - System-level latency
    - Ear-Voice Span (EVS)
      - Span between the start of the input & output
  - Module-level quality
    - ISR: Word Error Rate (WER)
    - ISR+IMT: BLEU
    - ITTS: L2-norm loss and subjective evaluation (AB preference test)
- Three latency regimes: low, medium, high (following IWSLT)

R.Fukuda, et al, “SIMULTANEOUS SPEECH-TO-SPEECH TRANSLATION SYSTEM WITH TRANSFORMER-BASED INCREMENTAL ASR, MT, AND TTS”, Proc. Oriental COCOSDA 2021

Problem: Delay (Ear-Voice Span)

Delay

ASR

こんにちは、駅はどこですか？

konnichiwa eki wa dokodesuka

MT

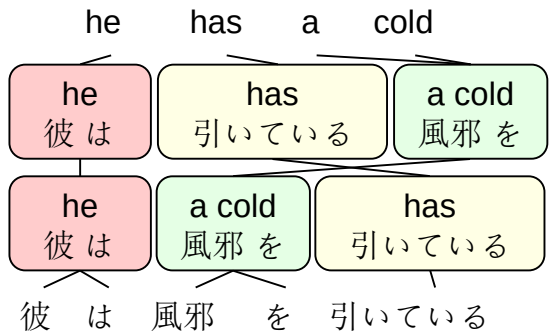
Hello, where is the station?

TTS

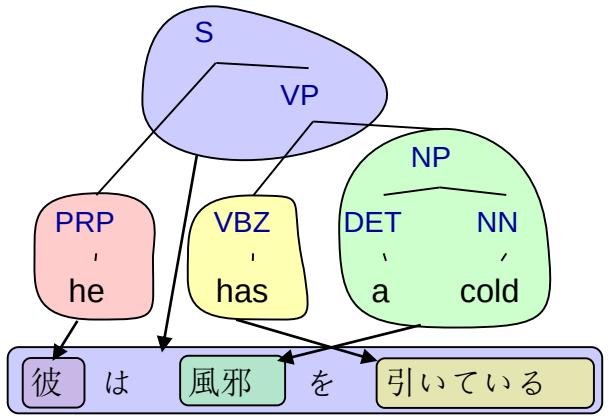
Statistical Translation Frameworks

Symbolic Models

Phrase-based MT [Koehn+ 03]

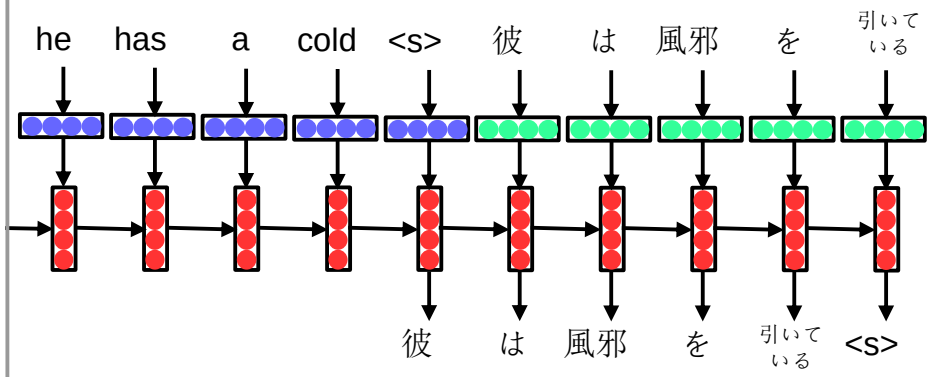


Tree-to-String MT [Liu+ 06]

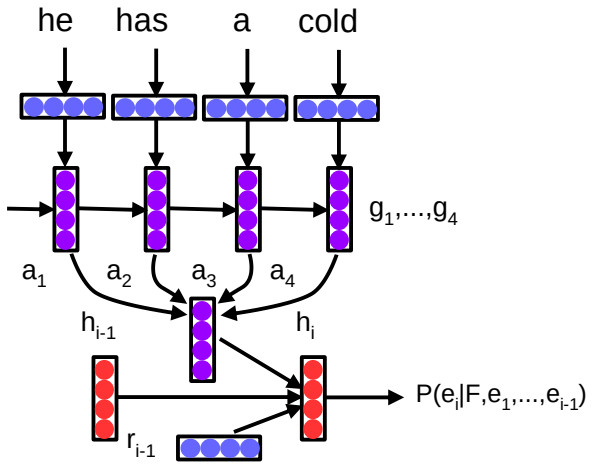


Continuous-space (Neural) Models

Encoder-Decoder [Sutskever+ 14]



Attentional [Bahdanau+ 14]



Evaluation: Latency (TED Talks)

ISR: Incremental Speech Recognition, IMT: Incremental Machine Translation, ITTS: Incremental Text-to-Speech Synthesis

Latency regime	ISR delay (sec.)	ISR+IMT delay (sec.)	Total delay (=EVS) (sec.)	Hyper-parameters
Low	0.93	4.69	8.81	ISR: 64 frames IMT: k=10 ITTS: 5 words
Medium	0.93	8.43	11.87	ISR: 64 frames IMT: k=20 ITTS: 5 words
High	1.30	11.47	16.91	ISR: 64 frames IMT: k=30 ITTS: 7 words

R.Fukuda, et al, “SIMULTANEOUS SPEECH-TO-SPEECH TRANSLATION SYSTEM WITH TRANSFORMER-BASED INCREMENTAL ASR, MT, AND TTS”, Proc. Oriental COCOSDA 2021

Simultaneous Interpretation

European Commission

Simultaneous interpreting is a mode of interpreting in which the speaker makes a speech and the interpreter reformulates the speech into a language his audience understands *at the same time (or simultaneously).*

The three main actions are also essentially the same as consecutive interpreting.

1) listen actively (understand)
2) analyse (structure the message)
3) reproduce (communicate)

The difference with consecutive interpreting is that in simultaneous interpreting all of these things need to happen *at the same time (or simultaneously).*

In addition to a special way of listening, prioritising information and distinguishing between primary and secondary information, activating short-term memory, communicating, etc. ,
a good simultaneous interpreter also has to be able to *anticipate* what the speaker might say.

Translation Timing Control by Syntactic Prediction in NMT[Kano, et al, 2021]

ICLP: Incremental Constituent Label Prediction (LSTM, BERT)

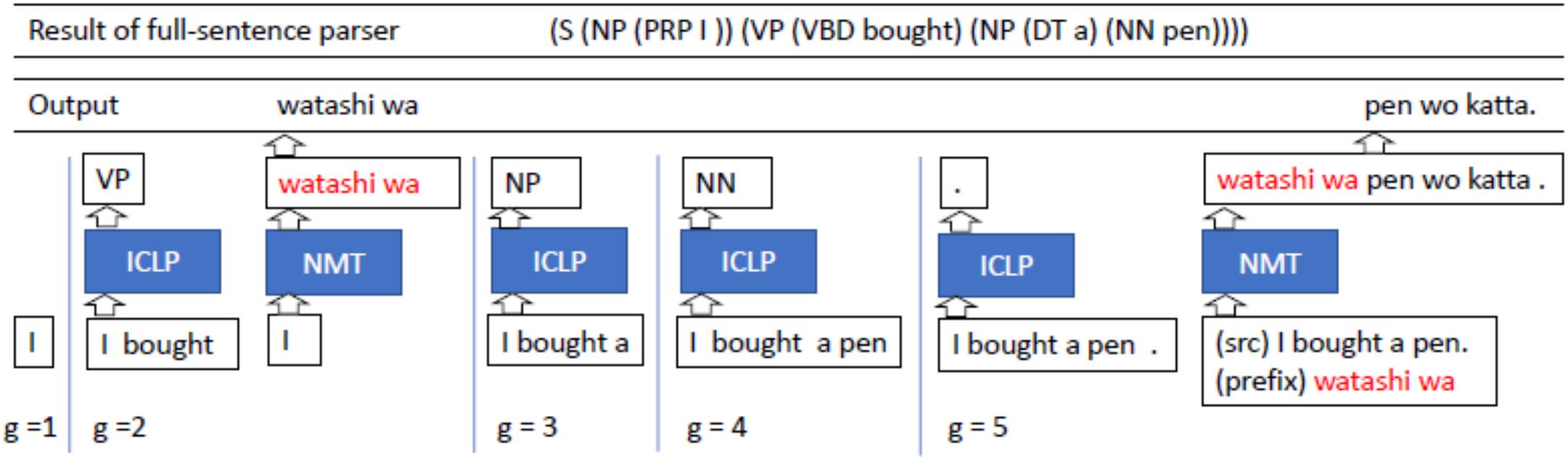


Figure 1: One look-ahead ICLP gives constituent labels. When a boundary is detected based on the label and rules, NMT starts to translate the source subsequence. The previous translation, which is red in the figure, is used as prefix words for the next translation. EOS is omitted for simplicity in the figure.

Y.Kano, K.Sudo, S.Nakamura, “Simultaneous Neural Machine Translation with Constituent Label Prediction”, Proc. of WMT 2021.