基盤研究S　（2017-2021）
次世代音声翻訳の研究

研究代表者：奈良先端科学技術大学院大学
中村　哲

研究分担者：河原達也（京大），猿渡 洋，高道慎之介（東大），
森島繁生（早大），戸田智基（名大），吉野幸一郎（理化学研究所）
須藤克仁，S. サクティ，田中宏季（奈良先端大）

連携研究者：水野 的（青山学院大），A.Waibel(CMU/KIT)，G. Neubig(CMU)

自動同時通訳／同時音声翻訳

- 漸進的音声認識結果を受け取り，漸進的に翻訳／通訳する
  - 漸進性：入力の終わりを待たずに，高速に処理する
  - 頑健性：入力の曖昧性や誤りの影響を抑える

- 「次世代」音声翻訳への挑戦
  - 発話終了を待たない漸進的な音声認識，機械翻訳，音声合成　（Incremental）
  - 異なる言語構造の言語間（日英）の自動同時音声翻訳のための言語運用
    - 先を予測する，文の構造を変える
  - End-to-end のSpeech-to-speech Translation
  - パラ言語の音声翻訳
  - 独自コーパスの構築
    - 400時間を目標
  - 通訳者の訳出との比較，通訳者の脳活動の分析
  - 実用向けに有意義な評価方法
    - IWSLTにおけるShared Task

本提案の研究課題

雑音源、音響特性



課題１：自動音声同時通訳と音声翻訳の高度化

音源推定
雑音抑圧

常時・同時複数
話者音声認識

同時
言語翻訳

同時
音声合成

パラ言語
情報抽出

同時パラ
言語翻訳

同時パラ言語
音声合成

字幕生成

課題２：パラ言語音声翻訳

発話顔画像
抽出、３Dモデル

発話顔
画像変換

発話顔画像
制御、生成

課題３：ビデオコンテンツ字幕翻訳，音声画像翻訳

課題4：脳活動を含むセンシングによる実時間コミュニケーション測定

視線検出装置Tobi，モバイル心拍計，32ch脳波計を用いて同時通訳作業時，および同時通訳ユーザの聴取負荷の測定を行う．

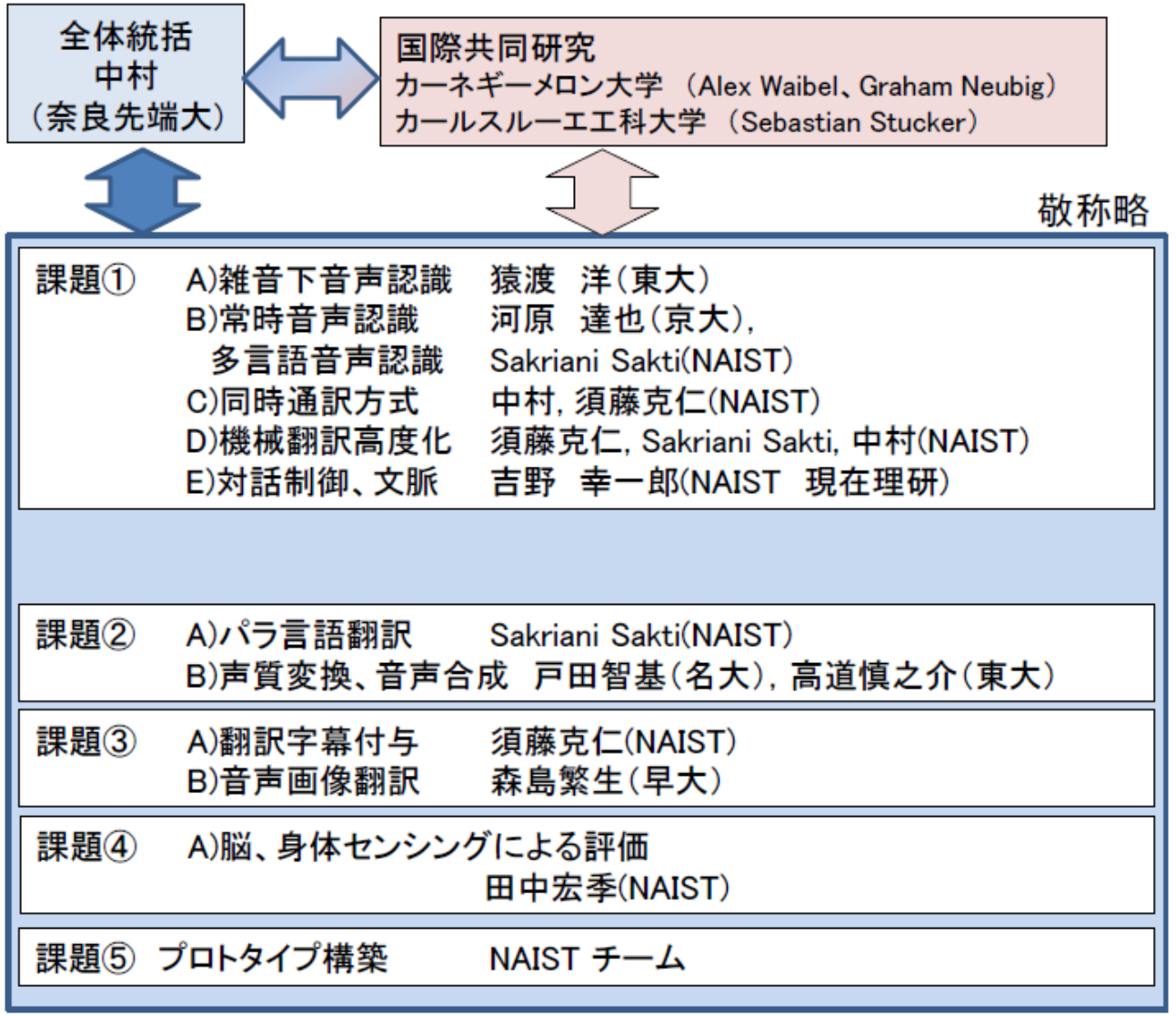
課題５：　コーパス構築とプロトタイプシステム

400時間以上の研究用の日英双方向同時通訳コーパス，およびビデオ翻訳コーパスを構築する．

次世代音声翻訳システムのプロトタイプを構築する．研究代表者，分担者の大学の講義，講演，会議の同時通訳とアーカイブ翻訳に適用し，評価，コーパスとしての蓄積，モデルの学習，改良を継続的に行う．

本提案の研究体制

本日はNAIST担当分について紹介



内容

1. 同時音声翻訳
   - 漸進的音声認識
   - 漸進的音声合成
   - 漸進的機械翻訳
   - 同時通訳，同時音声翻訳研究のためのコーパス構築
   - 同時音声翻訳プロトタイプ
2. パラ言語の取り扱い
   - パラ言語音声翻訳,音声表現とテキスト表現の等価性
3. End-to-end speech-to-speech translation
4. Machine Speech Chain
   - Multi-speaker Machine Speech Chain, Incremental Machine Speech Chain
5. まとめと今後の展開

内容

1. 同時音声翻訳
   - 漸進的音声認識
   - 漸進的音声合成
   - 漸進的機械翻訳
   - 同時通訳，同時音声翻訳研究のためのコーパス構築
   - 同時音声翻訳プロトタイプ
2. パラ言語の取り扱い
   - パラ言語音声翻訳,音声表現とテキスト表現の等価性
3. End-to-end speech-to-speech translation
4. Machine Speech Chain
   - Multi-speaker Machine Speech Chain, Incremental Machine Speech Chain
5. まとめと今後の展開

同時音声翻訳

- 同時音声翻訳
  - 発話終了を待たない漸進的な音声認識，機械翻訳，音声合成　（Incremental）
  - 異なる言語構造の言語間（日英）の自動同時音声翻訳のための言語運用
    - 順送り翻訳
    - 先を予測する，文の構造を変える

- 同時音声翻訳の構成要素
  - 漸進的音声認識(Incremental ASR)
  - 漸進的機械翻訳(Incremental MT/Simultaneous MT)
  - 漸進的音声合成(Incremental TTS)

人間の同時通訳モデル[水野2016]

英日通訳の場合の例

(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.

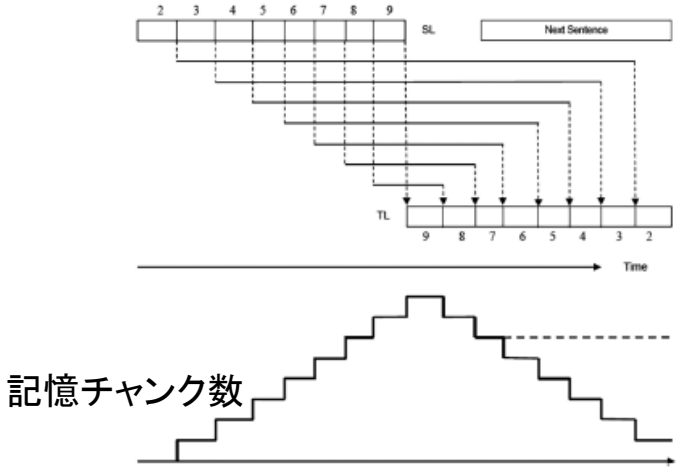
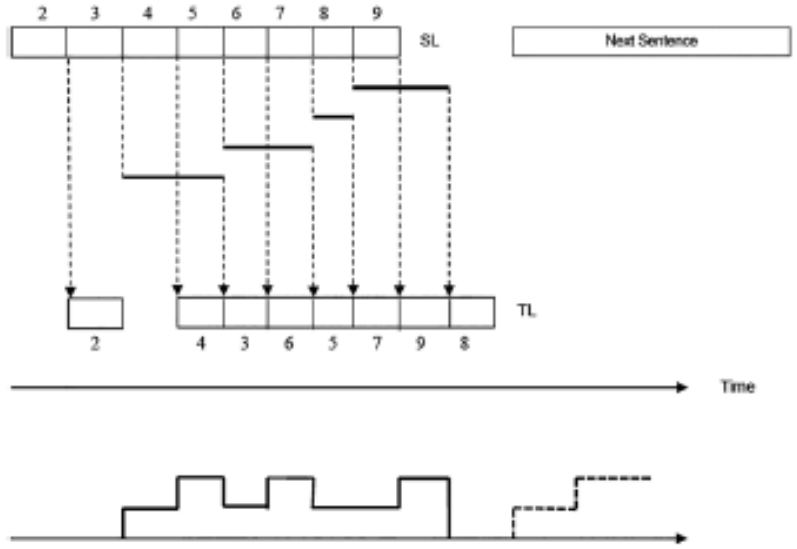


Fig.4 Translation to seek syntactic correspondence and its load
The dotted line of lower right indicates assumed load when next sentence comes in before the completion of translation of previous sentence.

必要短期記憶＞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.



必要短期記憶＜３！

逐次性向上に向けて

- 伝達遅延と精度低下のtrade-off：速く＆正確に

- 訳出開始判定
  - 語順を崩さずに訳出できるか否かを逐次予測
    - その時点までの統語要素＋直後の要素の予測結果を利用（Oda+ ACL-IJCNLP2015）
    - 音声認識と統合して実用的な効果を確認する

- 洗練された訳出方略
  - 文構造を変えて訳出開始を早める（関係詞節等）
  - 述語等を予測してさらに早く（特に日英通訳）
  - 同時通訳の理論やノウハウの活用
    （水野，同時通訳の理論―認知的制約と訳出方略）

内容

1. 同時音声翻訳
   - 漸進的音声認識
   - 漸進的音声合成
   - 漸進的機械翻訳
   - 同時通訳，同時音声翻訳研究のためのコーパス構築
   - 同時音声翻訳プロトタイプ
2. パラ言語の取り扱い
   - パラ言語音声翻訳,音声表現とテキスト表現の等価性
3. End-to-end speech-to-speech translation
4. Machine Speech Chain
   - Multi-speaker Machine Speech Chain, Incremental Machine Speech Chain
5. まとめと今後の展開

漸進的音声認識

Novitasari, Tjandra, Sakti, and Nakamura:

“Sequence-to-Sequence Learning via Attention Transfer for Incremental Speech Recognition”,

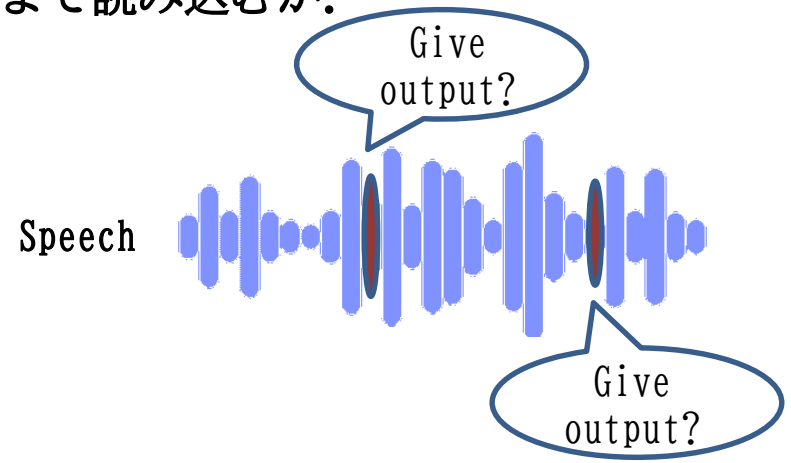
Proc. Interspeech 2019, pp. 3835-3839 (2019)

ニューラルネットによる漸進的音声認識

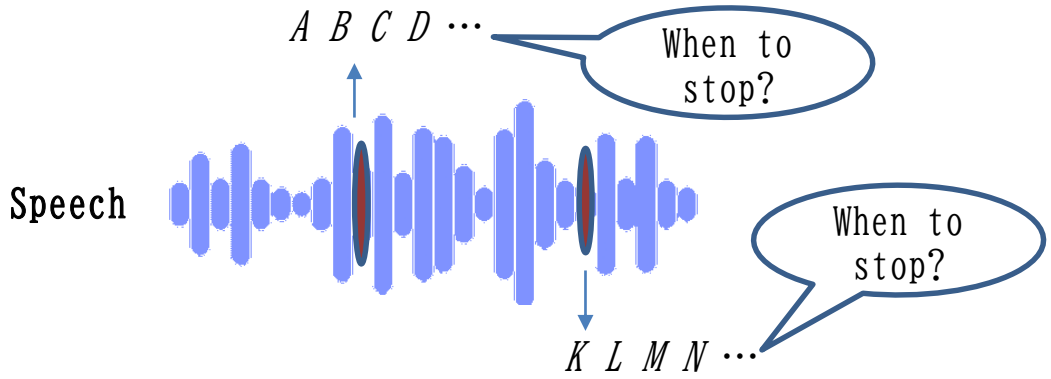
漸進的音声認識　ISR: Incremental Speech Recognition

- **ISRは発話終了を待たずに音声認識出力**（low delay）
  - 入力音声の部分系列から漸進的に出力
  - 出力も可変長：Encoder-decoderの特徴
  - HMMベースのものは存在するがEnd-to-endで実現したい.
- **課題：入力からの部分系列をどう決めるか.**

1)　**入力音声の境界の決定**
どこまで読み込むか.



2)　**出力系列の境界の決定**
どこまで出力するか



短い入力，短い出力の間のアライメントを学習する必要がある.

13

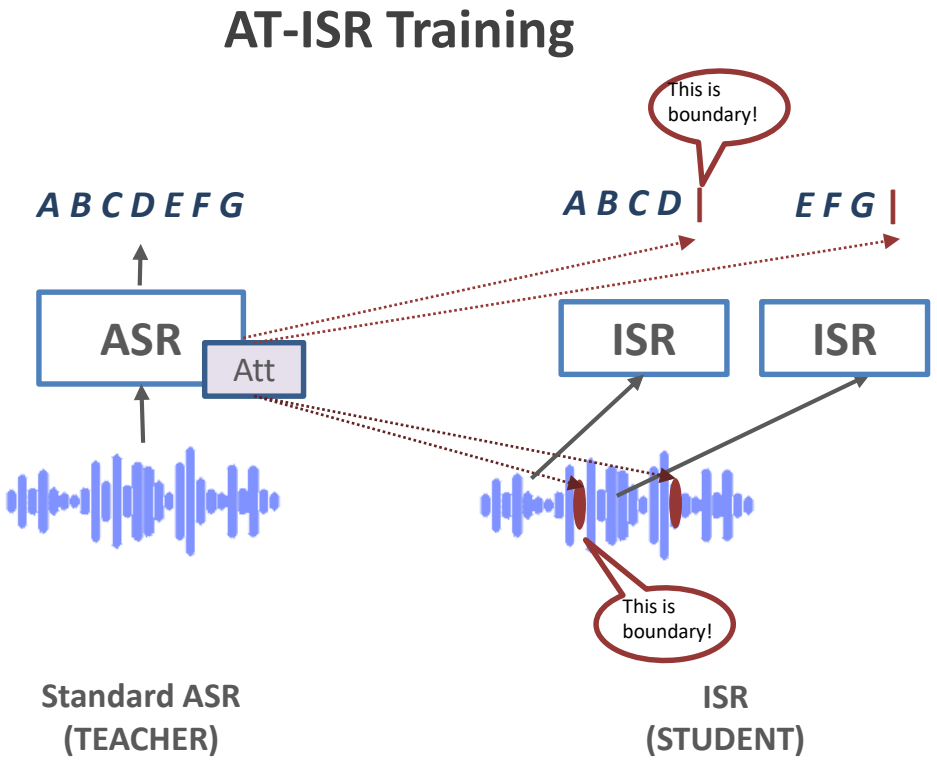
Attention-Transfer ISR

Attention-Transfer Incremental Speech Recognition (AT-ISR)

[Novitasari et al., 2019]

- **目的**
  従来のSeq2Seq ASRを教師モデル(Teacher)，短い単位で動作するISR ASRを生徒モデル(Student) とし，教師モデルのアライメントを使いながら生徒モデルのパラメータ学習を行う
  - ISR architecture　　: 教師モデルと同じ Seq2Seq
  - 漸進ステップ　　: 教師ASRモデルのアテンションから学習

- **Attention transfer** : Attention knowledge を教師モデルから生徒モデルにトランスファーする.
  - 先行研究 → 画像認識
    - Teach another model [Zaguruyko and Komodakis, 2017]
    - Domain transfer (image to video) [Li et al., 2017]
  - 漸進的音声認識への適用は初めて



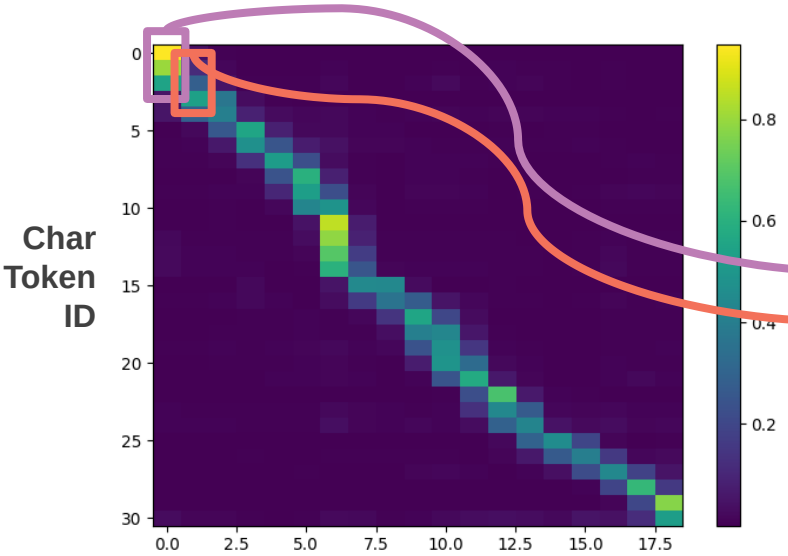
Attention-Transfer ISR

Attention Transfer

教師モデルにおけるAttentionベースのアライメントを使ってISR生徒モデルを学習する．(teacher)

1) 学習時の音声とテキスト（文字列）のアライメントをアテンションマトリクスから抽出する．
(alignment pair = high attention score):

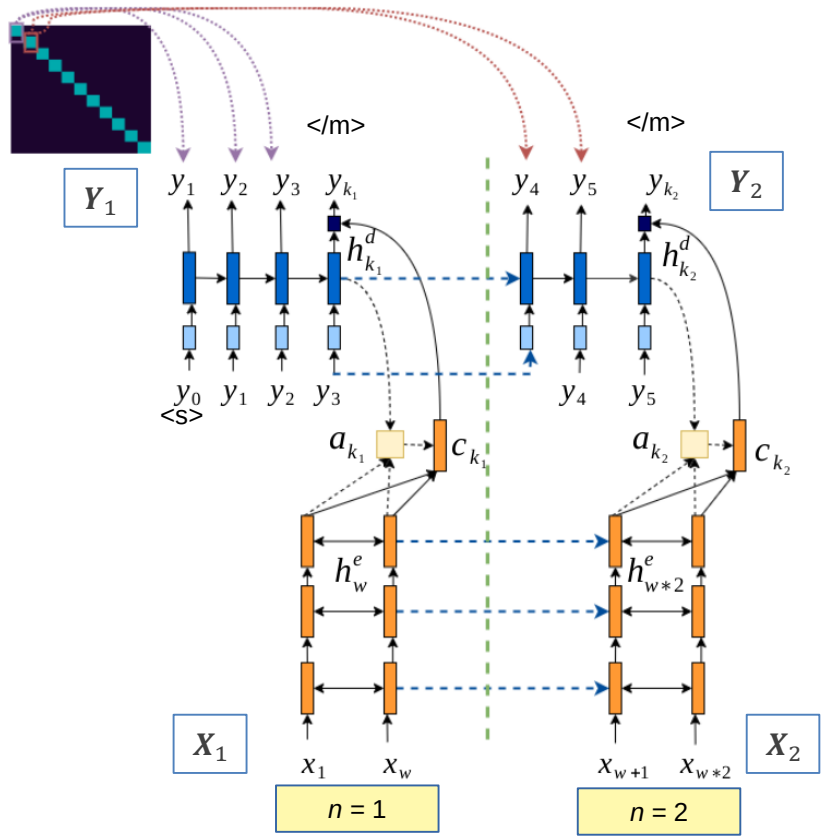
Teacher ASR attention matrix



Alignment

Seg. ID (n)	Speech seg. (X_n)	Text seg. (Y_n)
1	$x_1 - x_w$	$y_1 - y_3$
2	$x_{w+1} - x_{w+2}$	$y_4 - y_5$
(etc.)		

2) ISR生徒モデルY_n + </m> を教師モデルのX_nを用いて学習（/m は処理ブロックの文字列終了タグ）



ISR delay can be managed by changing X_n and Y_n size during training
e.g. higher delay : combine several segments into one

15

AT-ISR Performance

WSJデータによる評価．文字列の誤り率

Model	Delay (sec)		CER (%)
	Input	Computation	
Non-incremental ASR (Topline)			
Att Enc-Dec (ours)	7.88 (avg)	0.32 (avg)	6.26
BiLSTM-CTC [1]			8.97
Joint CTC+Att [1]			7.36
Baseline neural ISR			
Input/step: 1 *m* + 1 *la*	0.24	0.02	20.15
Input/step: 1 *m* + 4 *la*	0.54	0.05	11.95
Proposed AT-ISR			
Input/step: 1 *m* + 1 *la*	0.24	0.02	**18.37**
Input/step: 1 *m* + 4 *la*	0.54	0.05	**7.52**
Other existing neural ISR			
LSTM-CTC beam search [2]	-	-	10.96

CER diff.: **1.3%**

*Note
m = main input block
la = look-ahead block (contextual input)
1 block = 8 frames = 0.14 sec

Result

- 平均的な発話長: 7.88 sec
- Machine: Intel® CoreTM i7-9700K CPU @ 3.60GHz (NVIDIA GeForce RTX 2080Ti GPU)
- 文脈をみるためLook-a-head (*la*)も利用．

AT-ISR はより短い遅延で小さなCER

[1] Suyoun Kim, Takaaki Hori, and ShinjiWatanabe. Joint CTC-attention based end-to-end speech recognition using multitask learning. In Proceedings of ICASSP, pages 4835-4839, New Orleans, USA, 2017.
[2] Kyuyeon Hwang and Wonyong Sung. Character-level incremental speech recognition with recurrent neural networks. In Proceedings of ICASSP, pages 5335 - 5339, Shanghai, China, 2016.

内容

1. 同時音声翻訳
   - 漸進的音声認識
   - 漸進的音声合成
   - 漸進的機械翻訳
   - 同時通訳，同時音声翻訳研究のためのコーパス構築
   - 同時音声翻訳プロトタイプ
2. パラ言語の取り扱い
   - パラ言語音声翻訳,音声表現とテキスト表現の等価性
3. End-to-end speech-to-speech translation
4. Machine Speech Chain
   - Multi-speaker Machine Speech Chain, Incremental Machine Speech Chain
5. まとめと今後の展開

漸進的テキスト音声合成

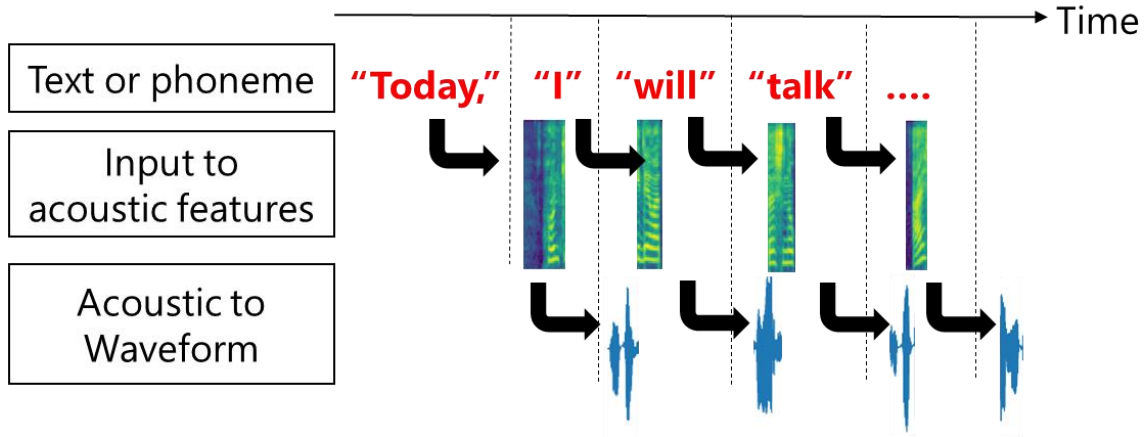
Yanagita, Sakti, and Nakamura:

Neural iTTS: Toward Synthesizing Speech in Real-time with End-to-End Neural Text-to-Speech Framework,

Proc. of the 10th ISCA Speech Synthesis Workshop, pp. 183-188 (2019)

漸進的音声合成 Text-to-speech

漸進的音声合成 iTTS：Incremental Text-To-Speech



課題

- 音声品質の確保
- 韻律を部分文からどうやって推定するか.
  - 単語やアクセントフレーズを単位として合成
  - 先読み　look-a-head を行う.

先行研究

- **HMM音声合成をベースにしたもの**

  Hidden Markov model TTS[Baumann et al., 2014.],[Pouget et al., 2015],[Yanagita, et al., 2018]

- **End-to-end 型で実現したい.**

提案手法

<s>: sentence start
</s>: sentence end
<m>: middle sentence start
</m>: middle sentence end

- 文を3つの部分に分割して学習.
  - 先頭（<s> …… </m> ）
  - 中間（<m> …… </m> ）
  - 終了（<m> …… </s> ）
- iTTSでは直前の結果を初期値として再利用
- 単位
  - 英語は単語単位
  - 日本語はアクセント句

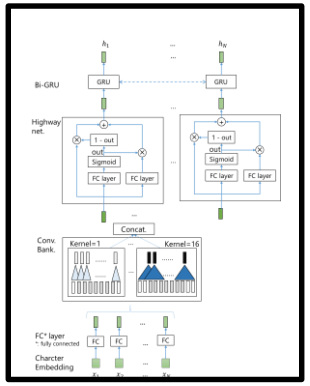
Text and acoustic features

<s>This is Tacotron.</s>
<s>This is an inital step.</s>
<s>Nice to meet you.</s>

<s>This </m>
<m>is </m>
<m>Tacotron.</s>
<s>This is </m>
<m>an </m>
<m>inital step.</s>
<s>Nice </m>
<m>to meet </m>
<m>you.</s>

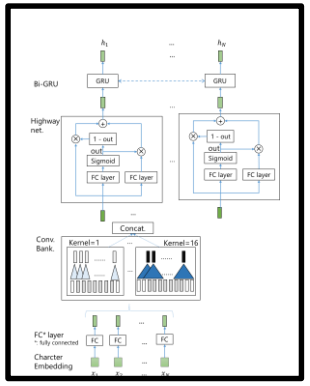
Inference: Ex. “Today we talk about TTS.”

Encoder



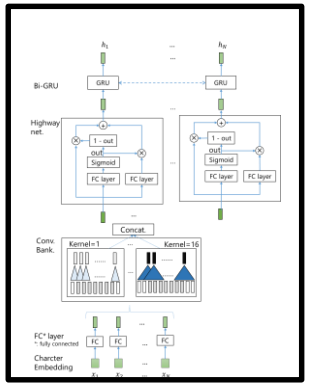
<s> Today </s>

Encoder



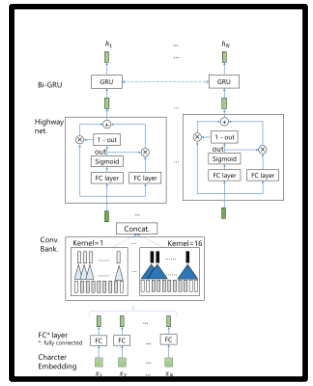
<m> we </s>

Encoder



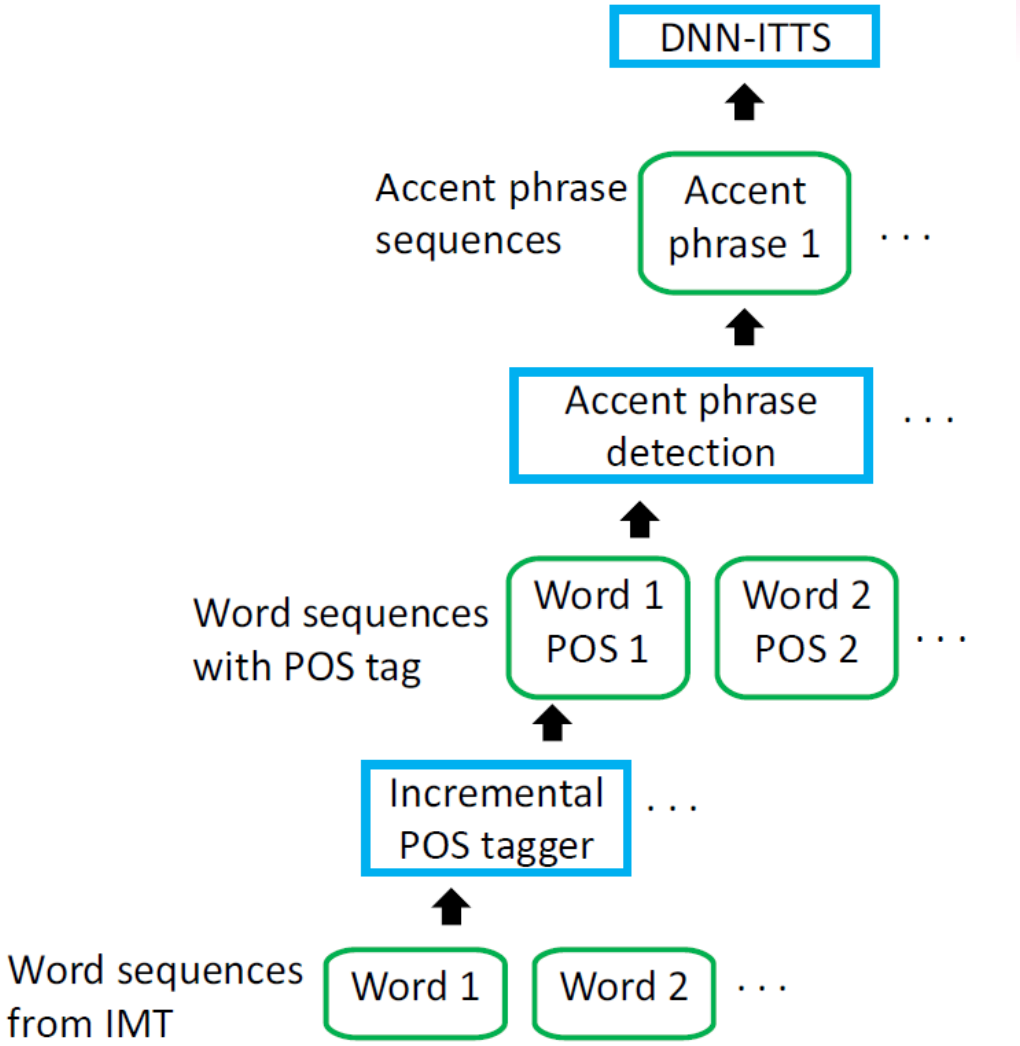
<s> Today </m>

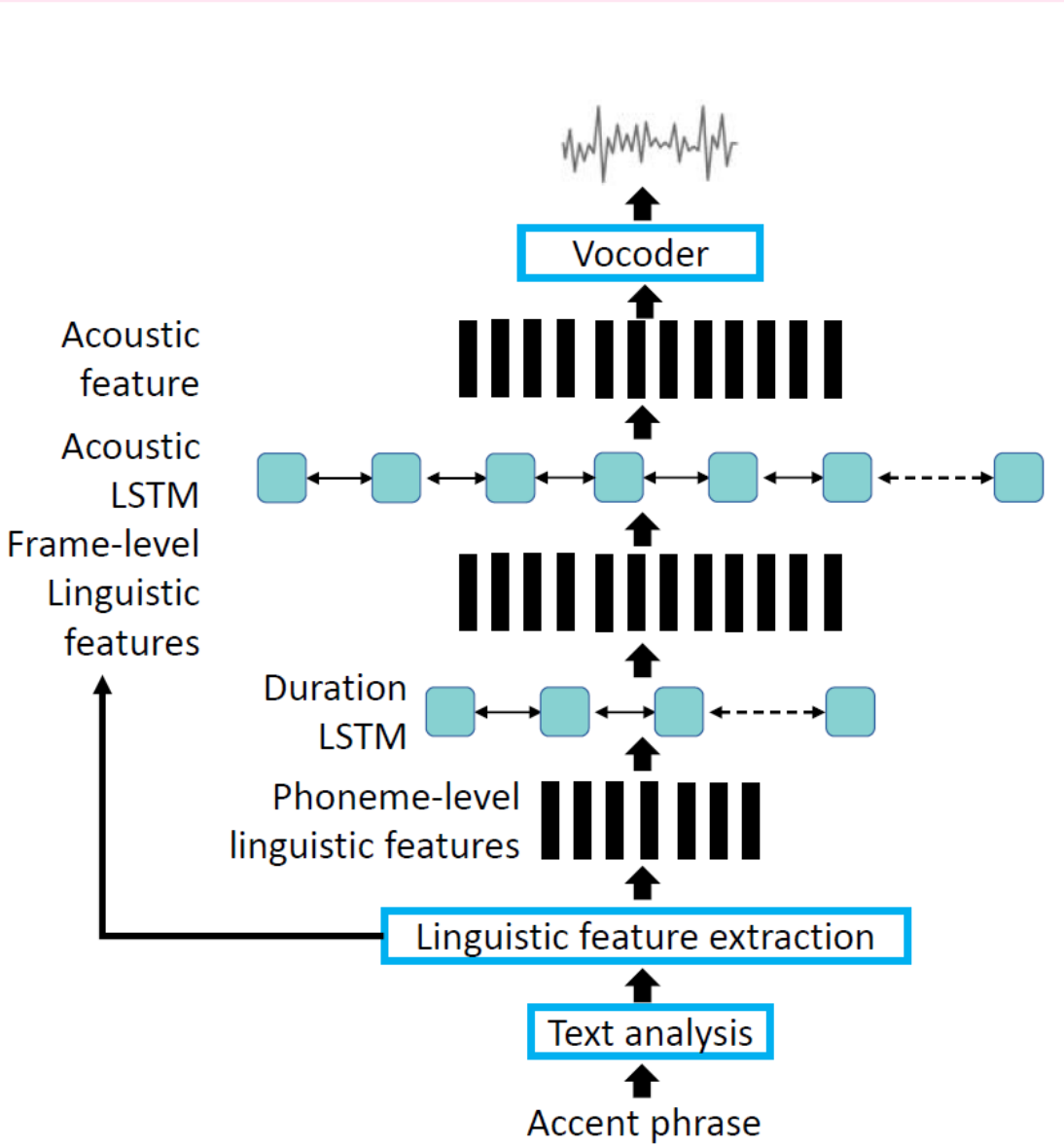
Encoder



<m> we </m>

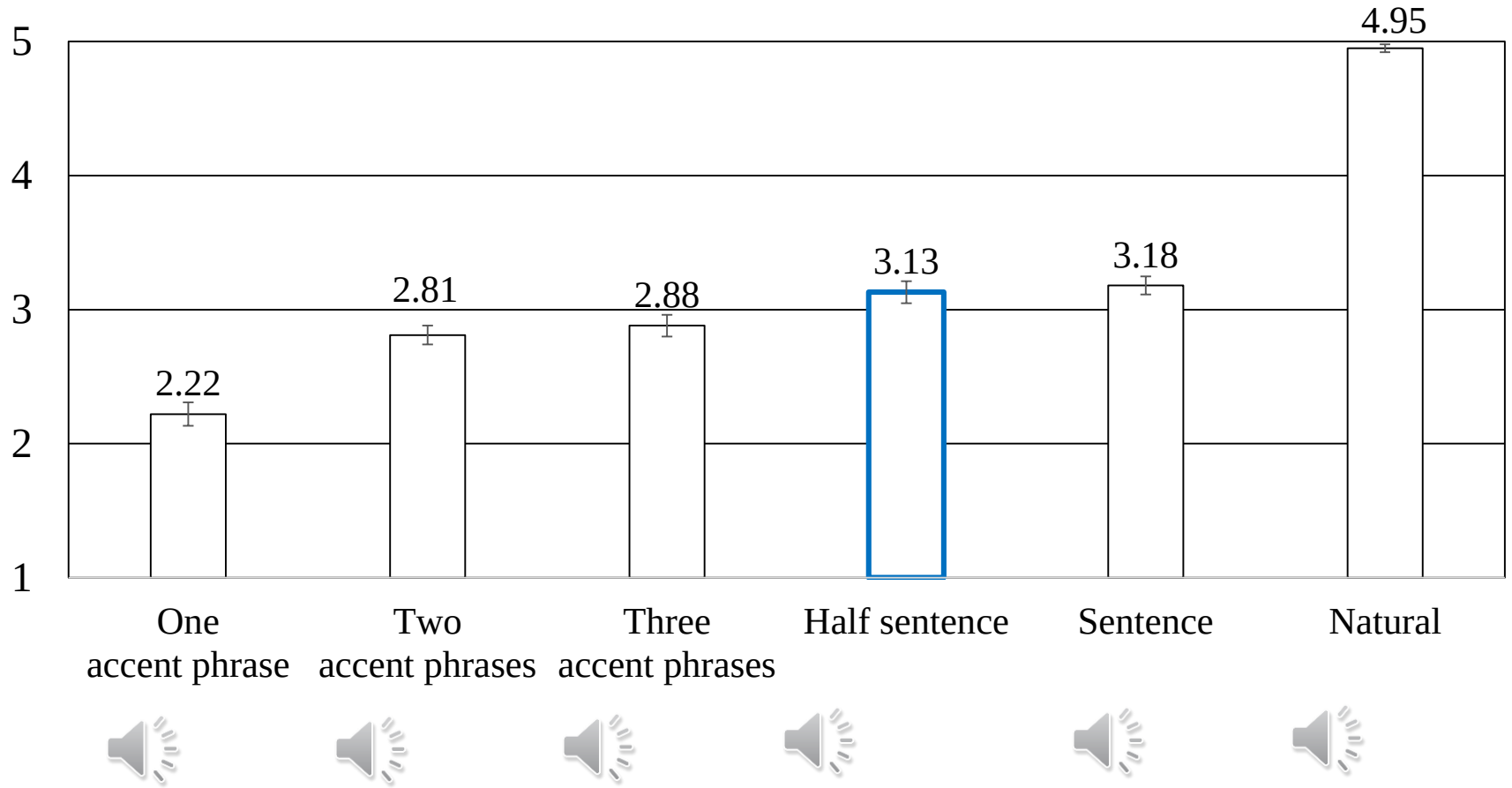
iTTS合成プロセス


DNN-ITTS
Accent phrase sequences
Accent phrase 1
Accent phrase detection
Word sequences with POS tag
Word 1 POS 1
Word 2 POS 2
Incremental POS tagger
Word sequences from IMT
Word 1
Word 2


Vocoder
Acoustic feature
Acoustic LSTM
Frame-level Linguistic features
Duration LSTM
Phoneme-level linguistic features
Linguistic feature extraction
Text analysis
Accent phrase

日本語iTTSのMOS評価

Still big gap between natural speech and synthesized speech.
Half sentence unit ≒ the full sentence units (Ja.).



内容

1. 同時音声翻訳
   - 漸進的音声認識
   - 漸進的音声合成
   - 漸進的機械翻訳
   - 同時通訳，同時音声翻訳研究のためのコーパス構築
   - 同時音声翻訳プロトタイプ
2. パラ言語の取り扱い
   - パラ言語音声翻訳,音声表現とテキスト表現の等価性
3. End-to-end speech-to-speech translation
4. Machine Speech Chain
   - Multi-speaker Machine Speech Chain, Incremental Machine Speech Chain
5. まとめと今後の展開

漸進的機械翻訳

帖佐，須藤，中村：
英日同時翻訳のためのConnectionist Temporal Classificationを用いたニューラル機械翻訳，
情報処理学会研究報告 2019-NL-241 (2019)

音声翻訳で実現するには？

４つの課題:

- 分割：いつ翻訳を開始するか?
- 予測：次の発話をどう予測するか？
- 言い換え：同時通訳用に言い換えることが出来るか?
- 評価：どの通訳結果が良いかどう評価するか?

分割してみよう (Fujita, et al., Interspeech 2013)

- 統計的機械翻訳で用いられる翻訳モデルに着目
  - ＋データから自動構築可能
  - ＋言語情報を利用
  - ＋翻訳と同じ情報を利用するため相性が良い
- 具体的には
  - 「フレーズ」と呼ばれる、翻訳に用いる単語列の区切りで翻訳開始
  - 「並べ替え確率（右確率）」で同時性と精度のバランスを調整
  - 「言語モデル適応」を行い、精度の低下を防ぐ

Tomoki Fujita,, Graham Neubig, Sakriani Sakti, Tomoki Toda, Satoshi Nakamura
"Simple, Lexicalized Choice of Translation Timing for Simultaneous Speech Translation"
Proc. of INTERSPEECH 2013, 3487-3491, 29. Aug. 2013

実験（IWSLT2013）

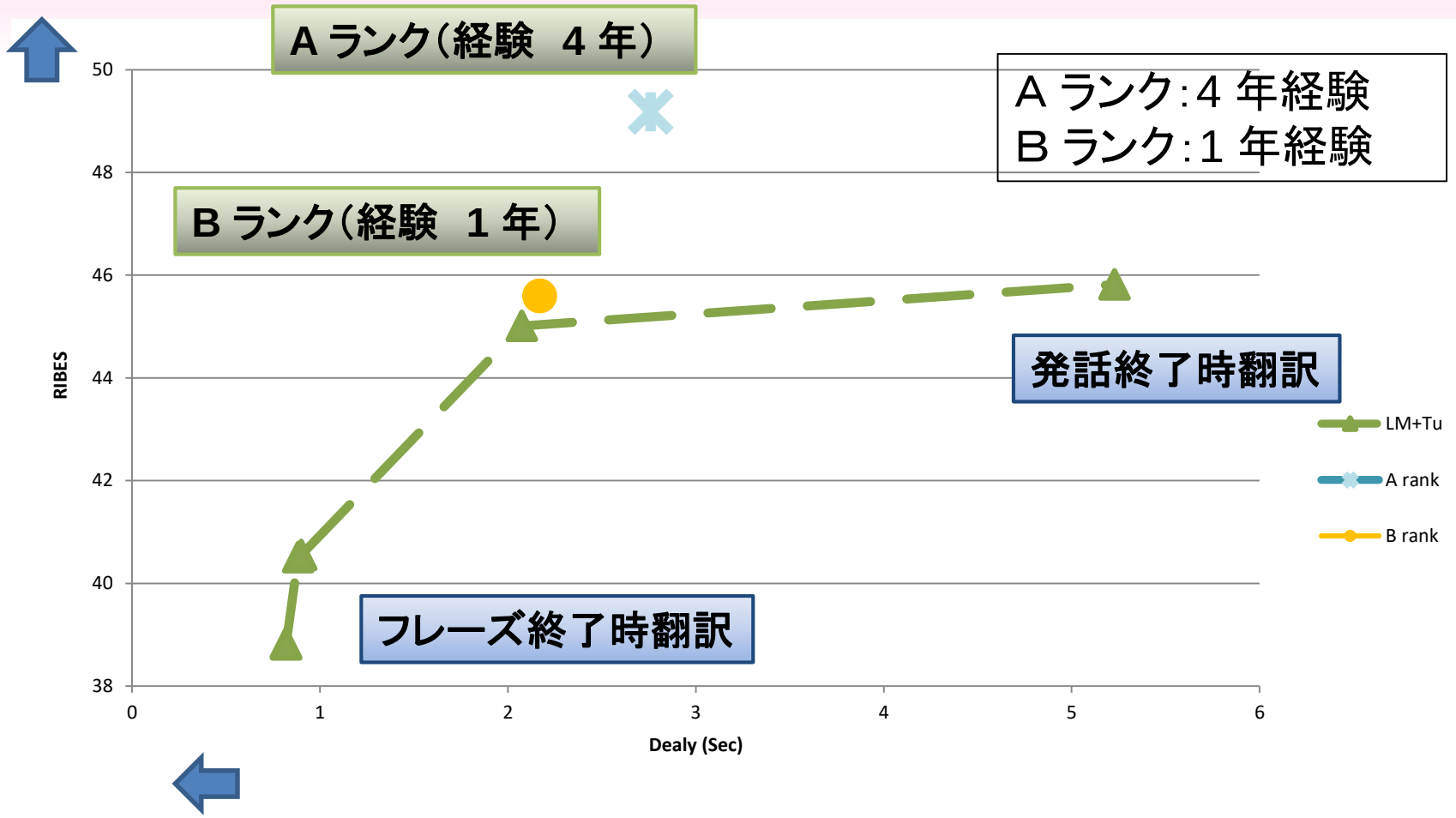
- 対象データ：TED Talk（英語⇒日本語）
  - 翻訳(キャプション)
    vs. 通訳

- 異なるスキルの通訳者

スキルのレベル	# 経験年数
S	15 年
A	4 年
B	1 年

Hiroaki Shimizu, Graham Neubig, Sakti Sakriani Watiasri, Tomoki Toda, Satoshi Nakamura
"Constructing a Speech Translation System using Simultaneous Interpretation Data"
Proc. of International Workshop on Spoken Language Translation (IWSLT), Dec. 2013

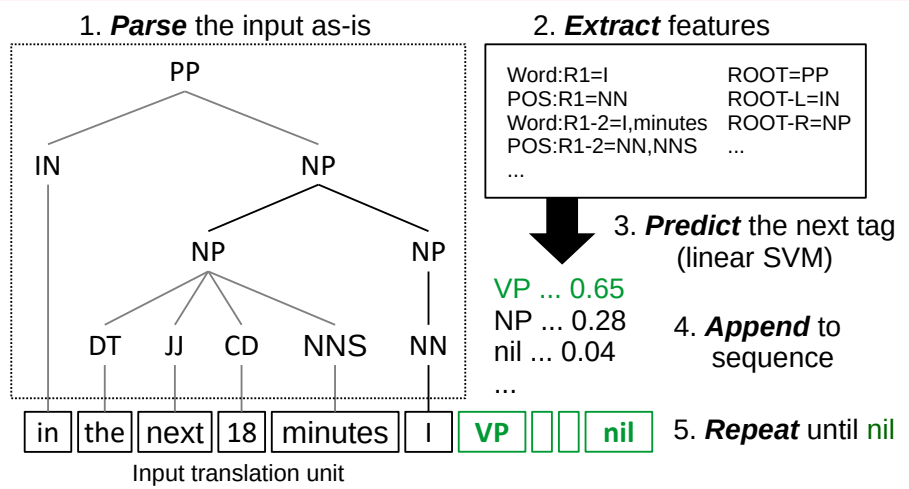
結果



≒　経験年数1年のB ランク通訳者と同等

統語要素予測に基づく訳出開始判定

- 未観測の統語要素（ラベル）を予測
  - 既観測部を構文解析
  - 素性抽出＆要素予測



- 翻訳途中に統語要素ラベルが現れたら「待機」
  - 句の並べ替えの有無によって訳出タイミングを変更する

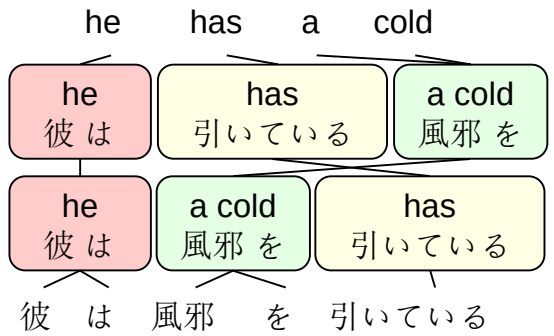
タグ推定後の入力文	in the next 18 minutes i 'm going to take **[NP]** (待機) i 'm going to take you on a journey
翻訳結果	18 分 で あ る [NP] を 行 っ て い ま す 皆さん を 旅 に お連れ し ます

Oda, Yusuke *et al.*, Syntax-based Simultaneous Translation through Prediction of Unseen Syntactic Constituents, Proc. of ACL-IJCNLP 2015.

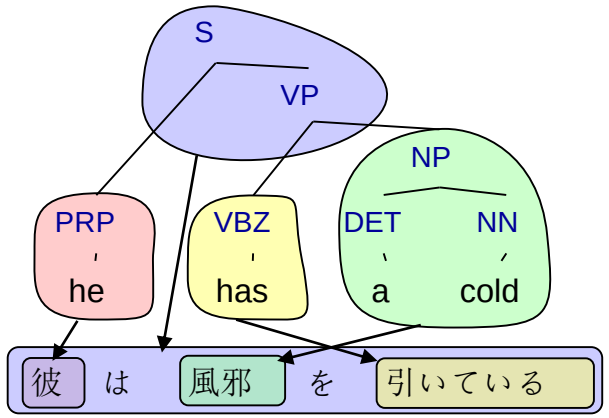
Statistical Machine Translation Frameworks

Symbolic Models

Phrase-based MT [Koehn+ 03]

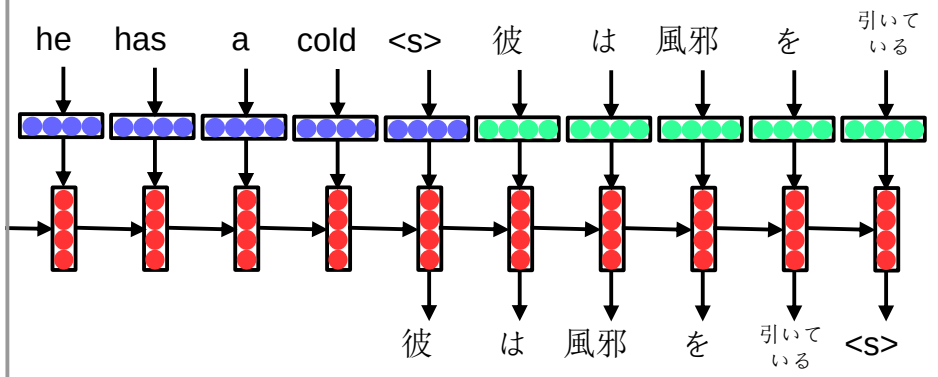


Tree-to-String MT [Liu+ 06]

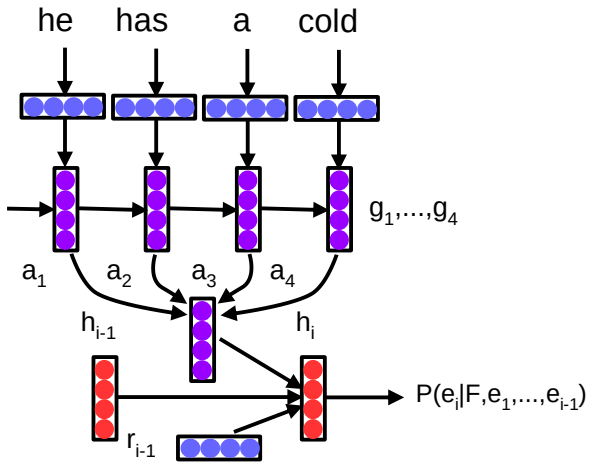


Continuous-space (Neural) Models

Encoder-Decoder [Sutskever+ 14]

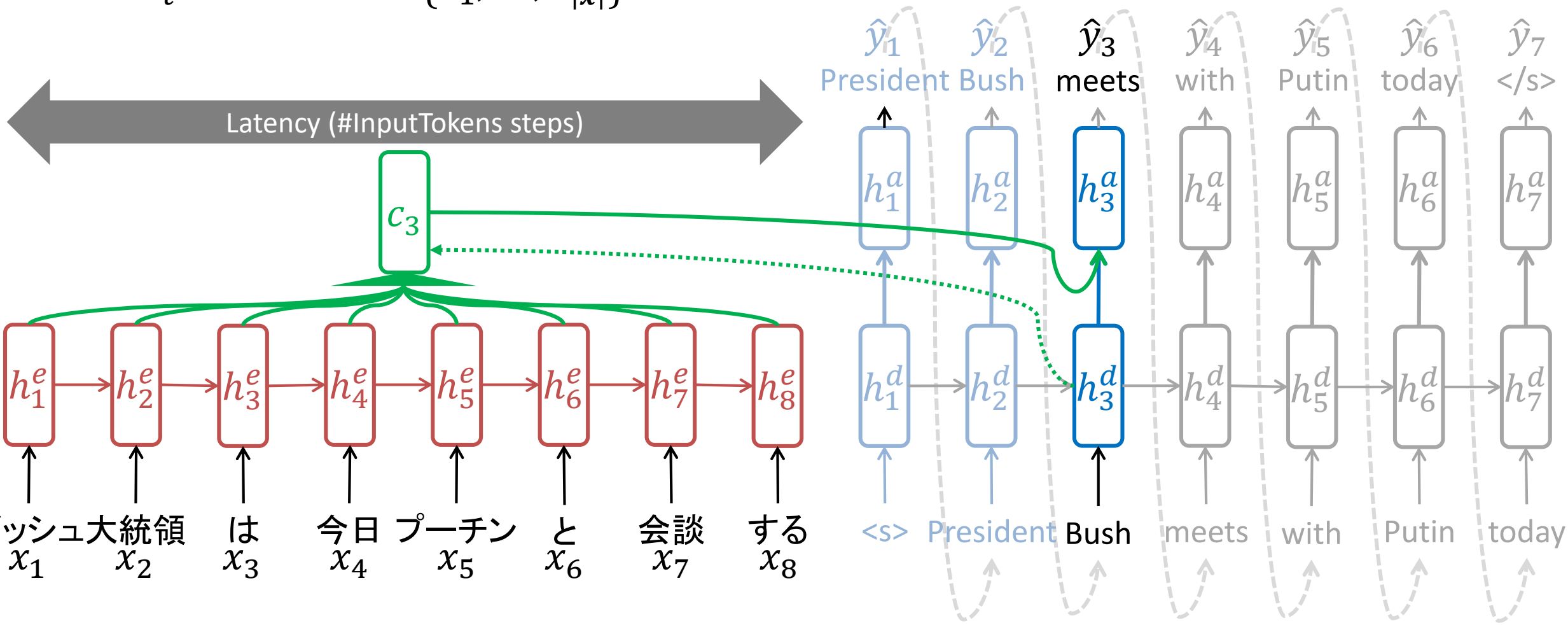


Attentional [Bahdanau+ 15]



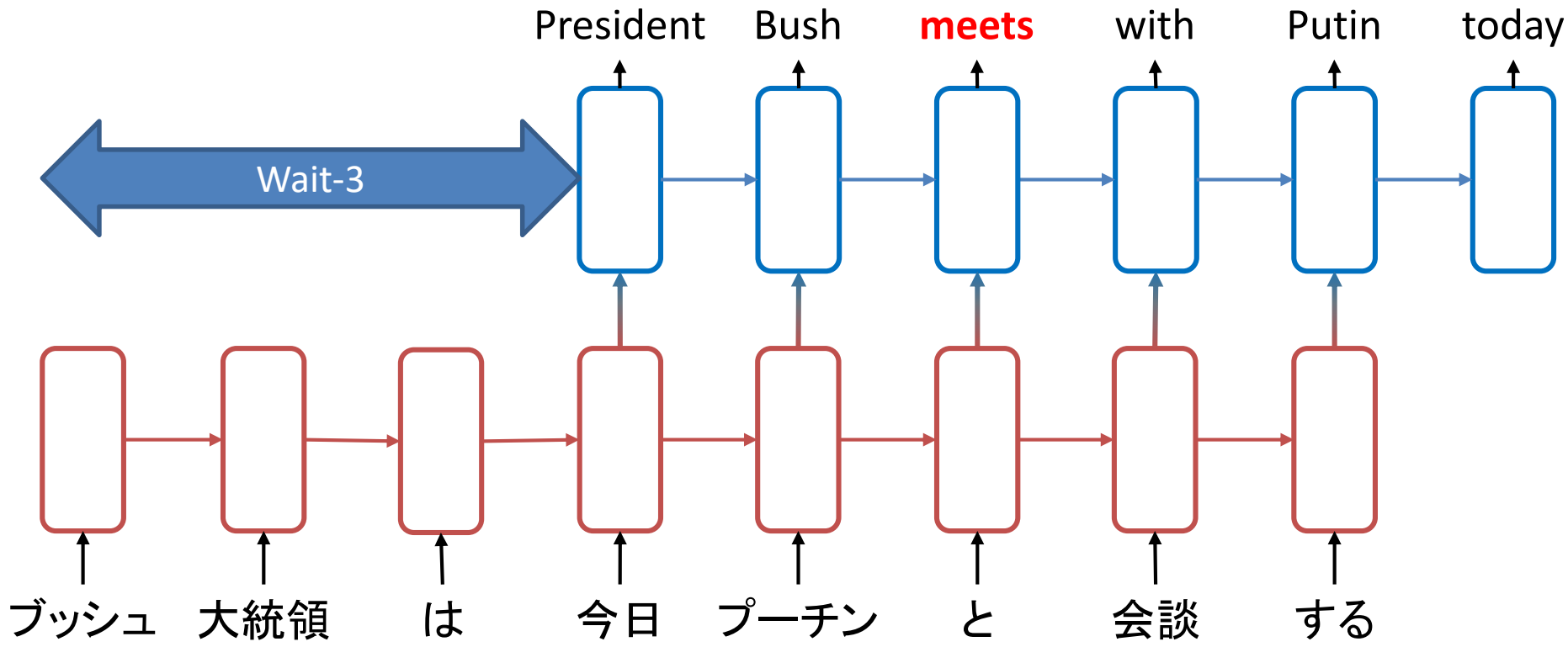
Attention-based Representation

- Decoder takes attentional context vectors
  - c_t looks over $\{x_1, \cdots, x_{|x|}\}$



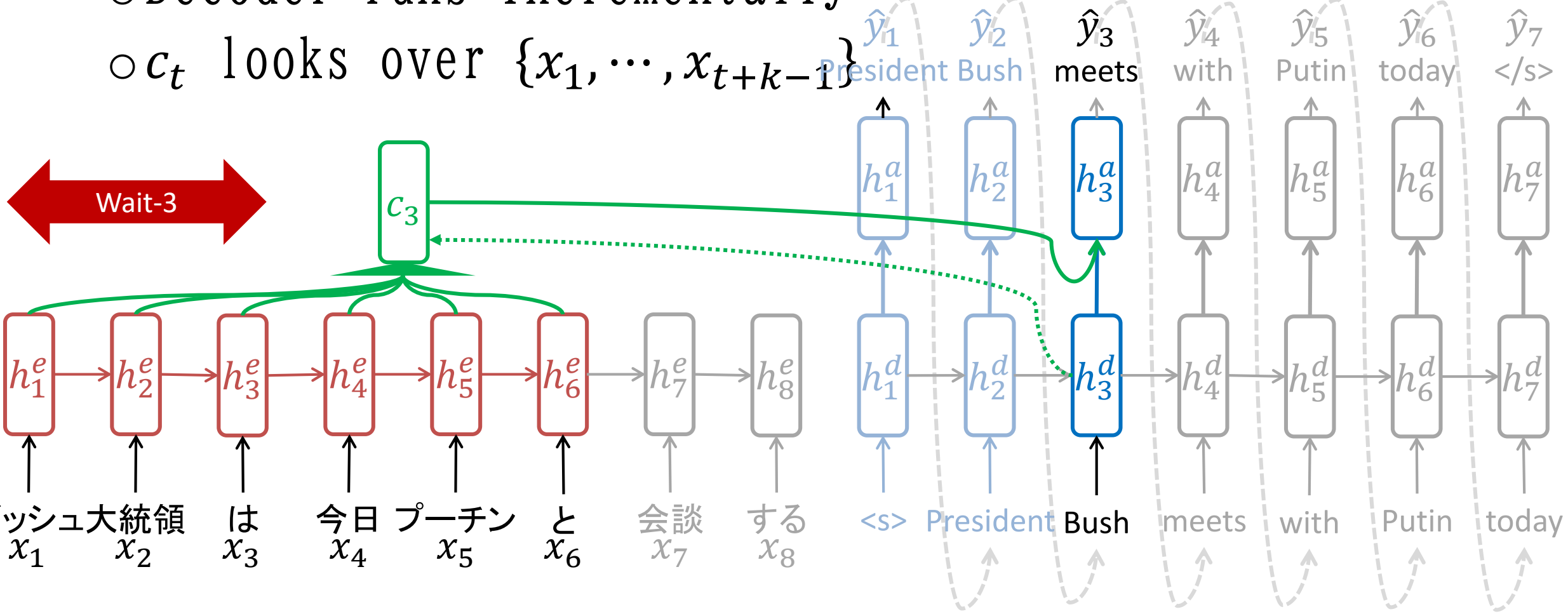
従来手法：訳出開始の固定ステップ数遅延

- Wait-k (Ma+, ACL 2019)



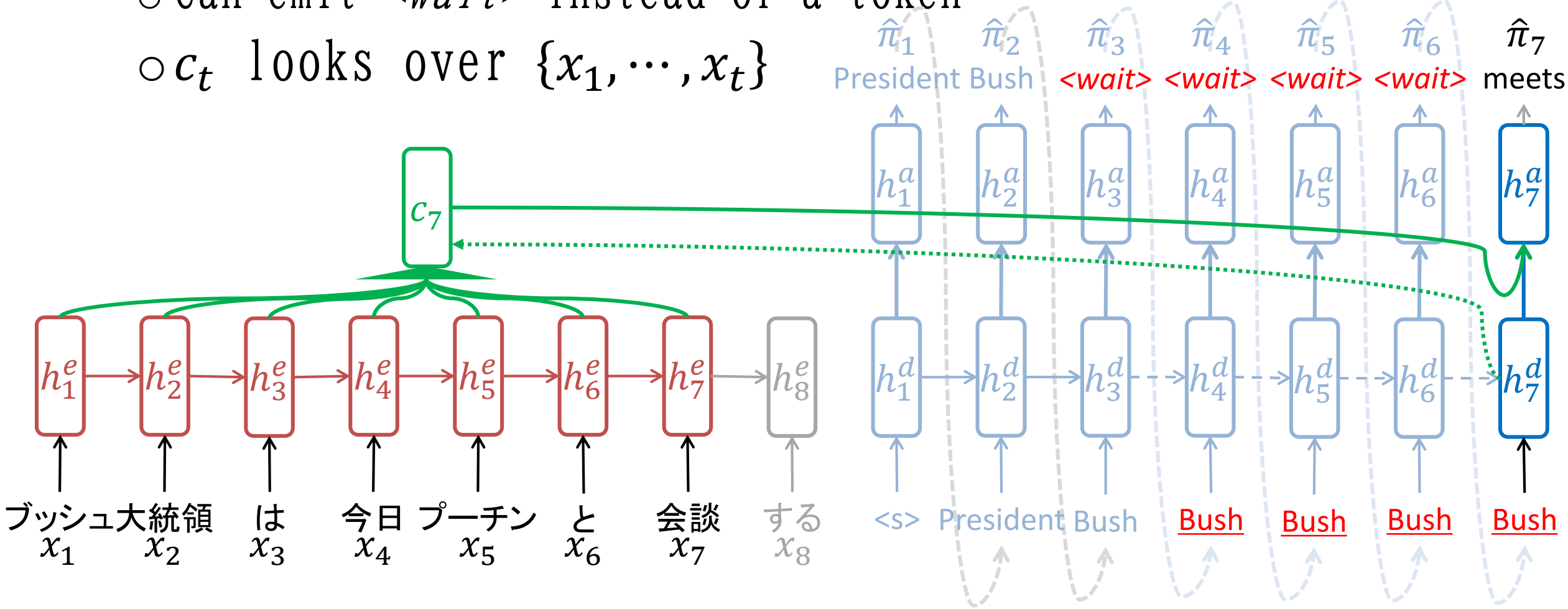
Previous Work

- Wait-k (Ma+, ACL 2019)
  - Decoder runs incrementally
  - c_t looks over $\{x_1, \cdots, x_{t+k-1}\}$



Proposed Method

- Adaptive Waits (Chousa+, IPSJ SIG-NL 241)
  - Can emit *<wait>* instead of a token
  - c_t looks over $\{x_1, \cdots, x_t\}$



Results

- Achieved adaptive waits but also observed some degradation in BLEU (En-Ja translation)

Method		Delay	BLEU
Full Sentence		9.75 (±2.69)	34.53
Wait-k	k=5	5.00 (±0.00)	33.29
	k=3	3.00 (±0.00)	31.06
Proposed	α=0.00	4.32 (±3.14)	28.01
	α=0.01	4.29 (±3.16)	30.42
	α=0.03	2.88 (±2.95)	26.47
	α=0.05	0.80 (±1.96)	22.60

Table 1: Results using very small corpus (50K sents.)

Method		Delay	BLEU
Full Sentence		29.81 (±14.30)	32.22
Wait-k	k=7	7.00 (±0.00)	23.20
	k=5	5.00 (±0.00)	21.53
Proposed	α=0.03	23.03 (±14.08)	24.86
	α=0.05	21.96 (±13.88)	22.45
	α=0.1	17.13 (±12.69)	23.66

Table 2: Results using mid-scale corpus (1M sents.)

内容

1. 同時音声翻訳
   - 漸進的音声認識
   - 漸進的音声合成
   - 漸進的機械翻訳
   - 同時通訳，同時音声翻訳研究のためのコーパス構築
   - 同時音声翻訳プロトタイプ
2. パラ言語の取り扱い
   - パラ言語音声翻訳,音声表現とテキスト表現の等価性
3. End-to-end speech-to-speech translation
4. Machine Speech Chain
   - Multi-speaker Machine Speech Chain, Incremental Machine Speech Chain
5. まとめと今後の展開

NAIST 同時通訳コーパス①

- 2012-2016年度収録
  - 元音声: MP4 (TED), MP3 (CNN), PCM
  - 通訳音声: 24bit 48kHz PCM
    - 通訳者ランク：S (10年+), A(3年+), B
    - 複数の通訳音声が収録されたものも一部あり

言語方向	ドメイン	原音声		同時通訳収録済み	
		ファイル数	時間	ファイル数	時間
英日	TED	74	15.2	58	12.3
	CNN	13	0.731	7	0.389
	合計	**87**	**15.9**	**65**	**12.7**
日英	TED	60	11.9	60	11.9
	CSJ	31	5.51	31	5.51
	NHK	10	0.304	10	0.304
	合計	**101**	**17.7**	**101**	**17.7**

NAIST 同時通訳コーパス①　(2017-2021)

- 400 hours by March 2022

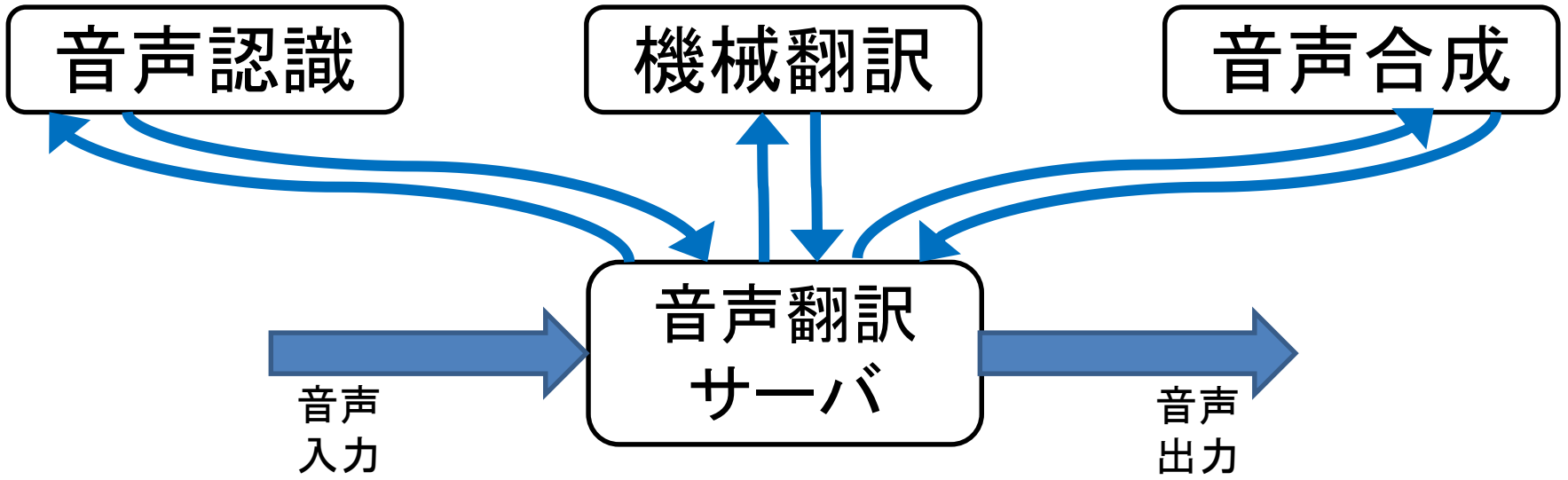
Direction	Domain	2017	2018	2019	2020	2021 (plan)
En-Ja	TED	0	**67 + 12 (4 x 3 SIs)**	50	*50*	*30*
Ja-En	CSJ	0	**33**	0	*0*	*0*
	TEDx	0	**12 (4 x 3 SIs)**	40	*0*	*30*
	Press conference	0	4	36.5	*0*	*0*
Mixed	Symposium	5.5	0	0	*0*	*0*
Total		5.5	128	126.5	*50*	*60*
Cum.		5.5	133.5	260	*310*	*400*

Table 4: Recoding hours of simultaneous interpretations (2017-)

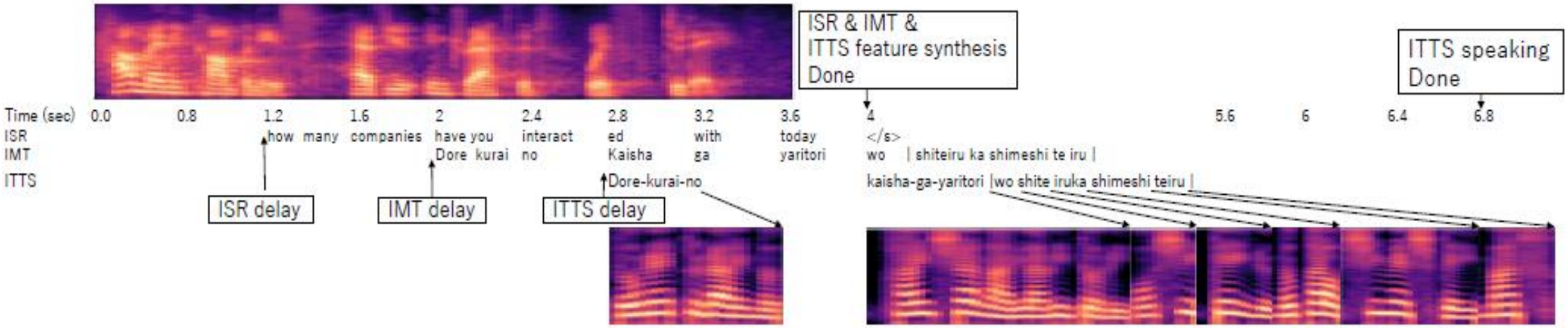
内容

1. 同時音声翻訳
   - 漸進的音声認識
   - 漸進的音声合成
   - 漸進的機械翻訳
   - 同時通訳，同時音声翻訳研究のためのコーパス構築
   - 同時音声翻訳プロトタイプ
2. パラ言語の取り扱い
   - パラ言語音声翻訳,音声表現とテキスト表現の等価性
3. End-to-end speech-to-speech translation
4. Machine Speech Chain
   - Multi-speaker Machine Speech Chain, Incremental Machine Speech Chain
5. まとめと今後の展開

同時音声翻訳システム構成



ISR, IMT, iTTS 出力タイミング



Katsuhito Sudoh, Takatomo Kano, Sashi Novitasari, Tomoya Yanagita, Sakriani Sakti, Satoshi Nakamura
“SIMULTANEOUS SPEECH-TO-SPEECH TRANSLATION SYSTEM WITH NEURAL INCREMENTAL ASR, MT, AND TTS”
arXiv:2011.04845v2 [cs.CL] 11 Nov 2020

デモ映像：同時音声翻訳

内容

1. 同時音声翻訳
   - 漸進的音声認識
   - 漸進的音声合成
   - 漸進的機械翻訳
   - 同時通訳，同時音声翻訳研究のためのコーパス構築
   - 同時音声翻訳プロトタイプ
2. パラ言語の取り扱い
   - パラ言語音声翻訳,音声表現とテキスト表現の等価性
3. End-to-end speech-to-speech translation
4. Machine Speech Chain
   - Multi-speaker Machine Speech Chain, Incremental Machine Speech Chain
5. まとめと今後の展開

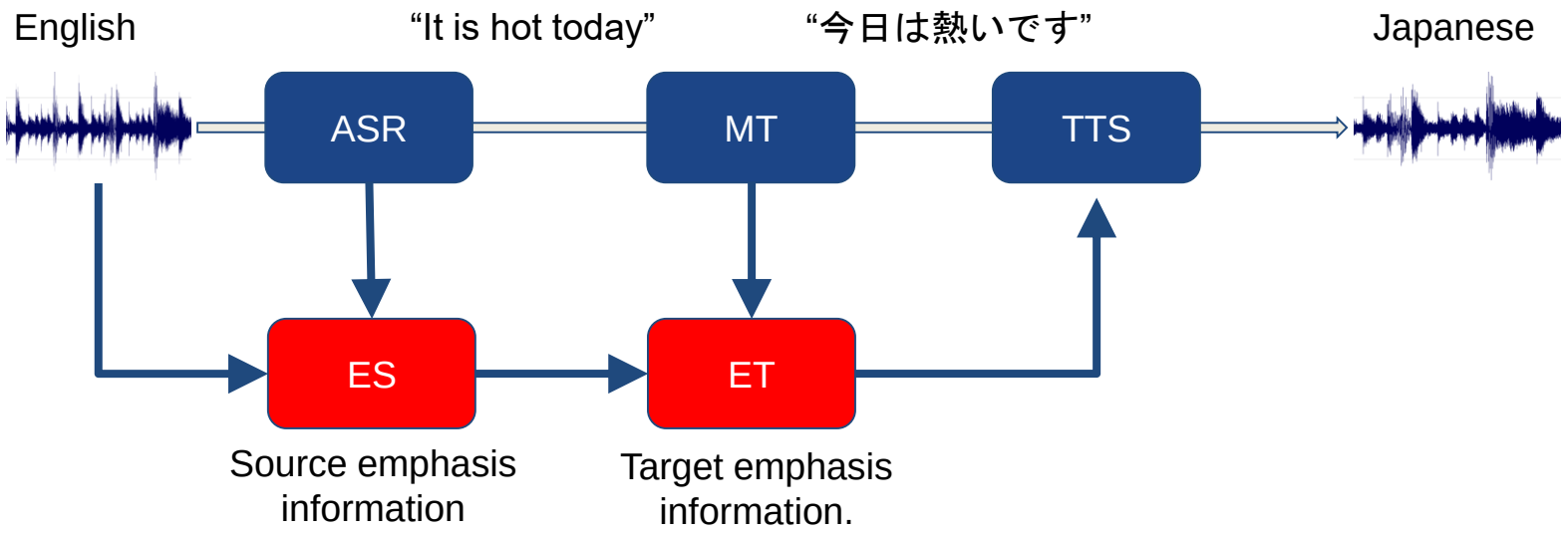
パラ言語情報

- パラ言語情報：
  - 個人性：声質に個人性が含まれる
  - 強調，感情：
    - 意図，話題の焦点が含まれる．
    - F0周波数パターンの変化や，ポーズ，強勢の違いに現れる．
    - 文構造，Phrase boundaryの影響を受ける．
    - 音響情報の言語情報の組み合わせで表現される．

    ⇒　音声処理と言語処理をさらに融合する必要．

パラ言語を考慮した音声翻訳

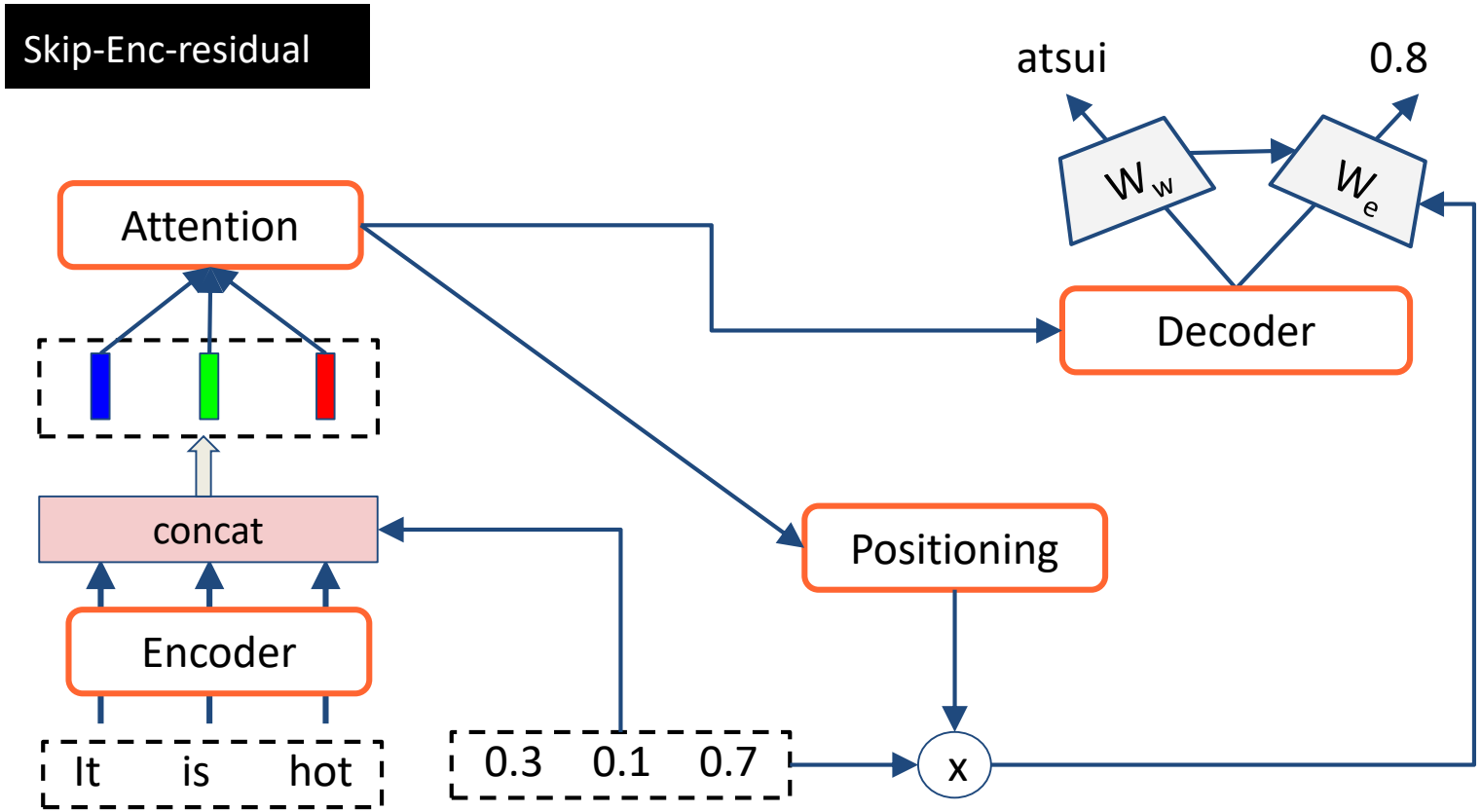
- **強調の音声翻訳**



(1) 強調の推定　Emphasis estimation (ES) systems:
音声と単語列から強調度を推定
(2) 強調の翻訳　Emphasis translation (ET) systems:
強調を目的言語に翻訳する

（2）強調情報の翻訳

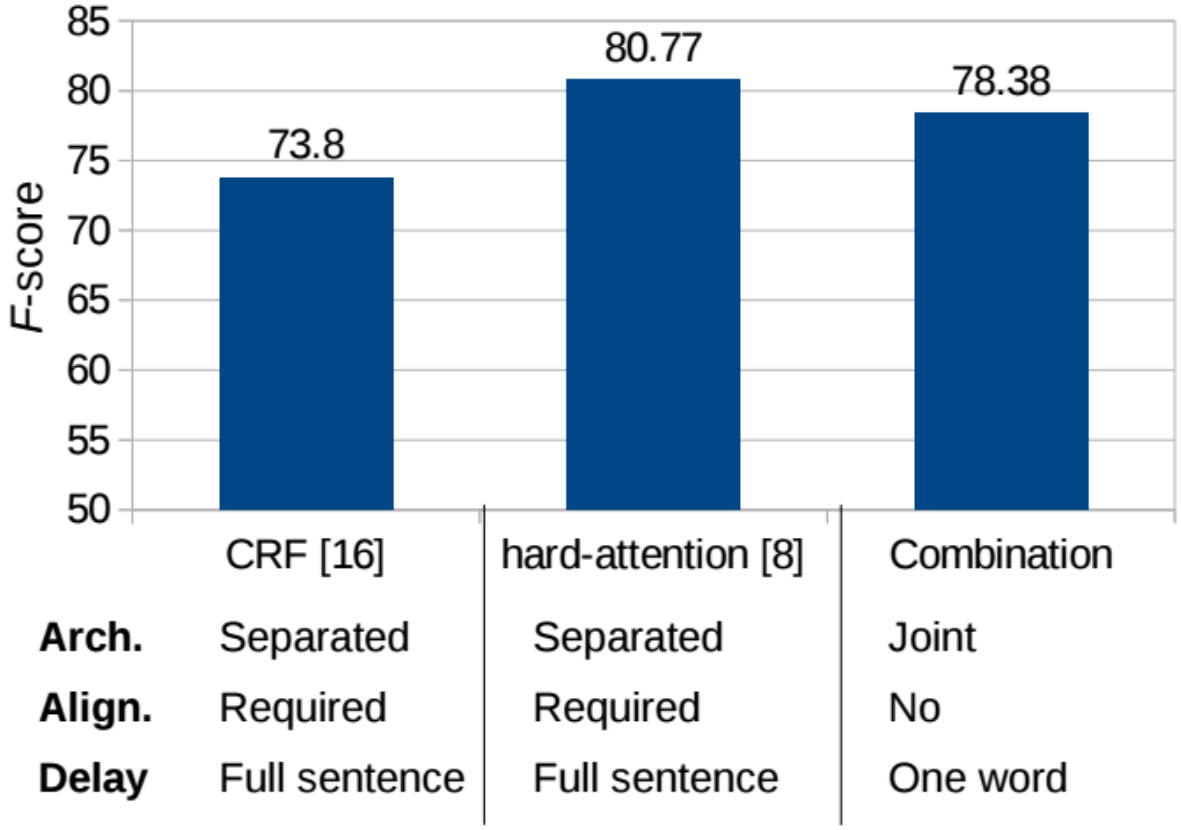
(c)言語と強調情報の統合翻訳 *[Do et al., 2017]*



The joint translation model

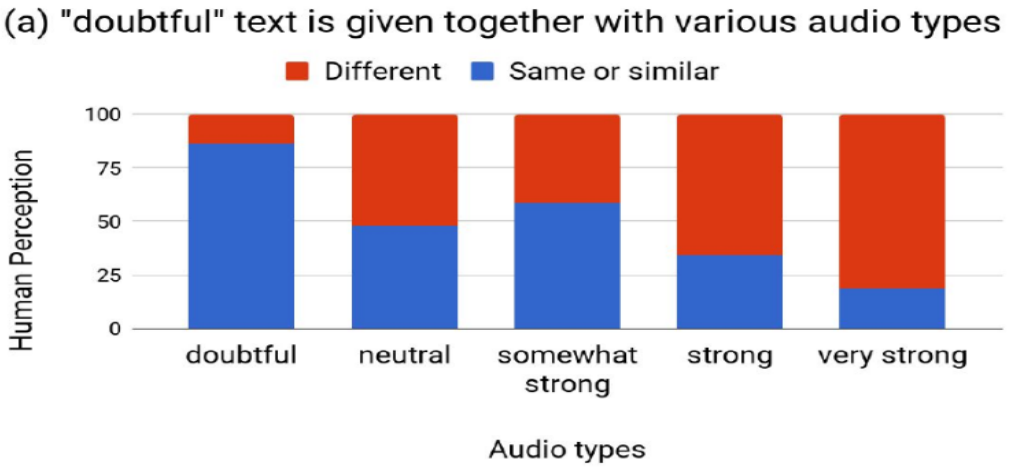
Experimental Results

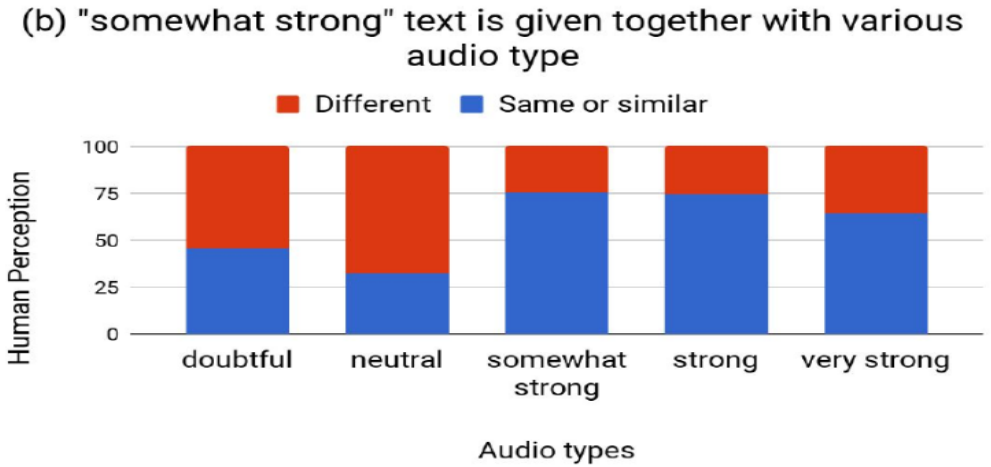
- English-Japanese Emphases Translation Performance

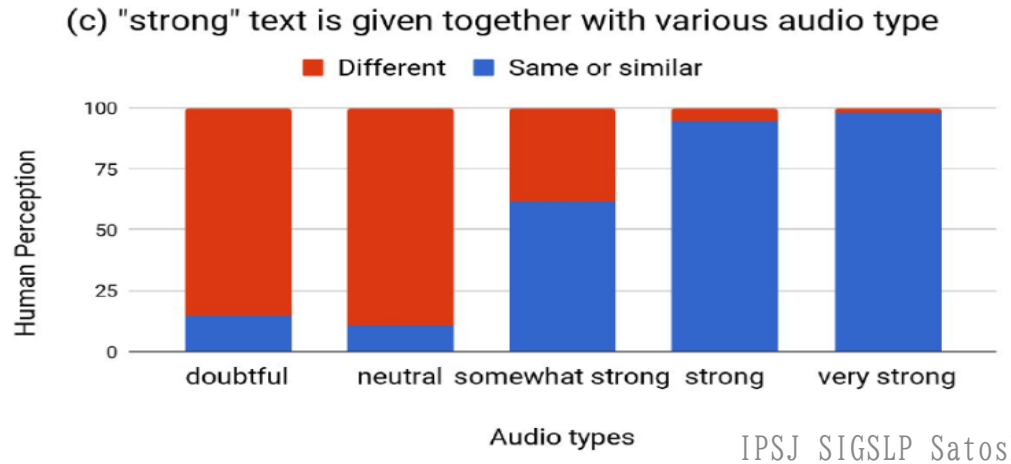


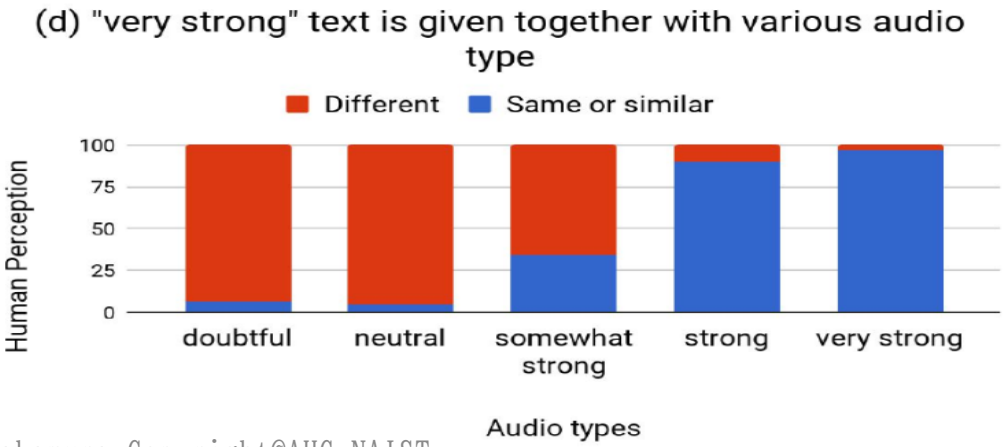
クラウドソーシングによる評価

- タスク: 音声とテキストの等価性についての評価
  - 被験者に音声とテキストを提示
  - 被験者は提示された音声とテキストの強調度が同じかどうかを回答.







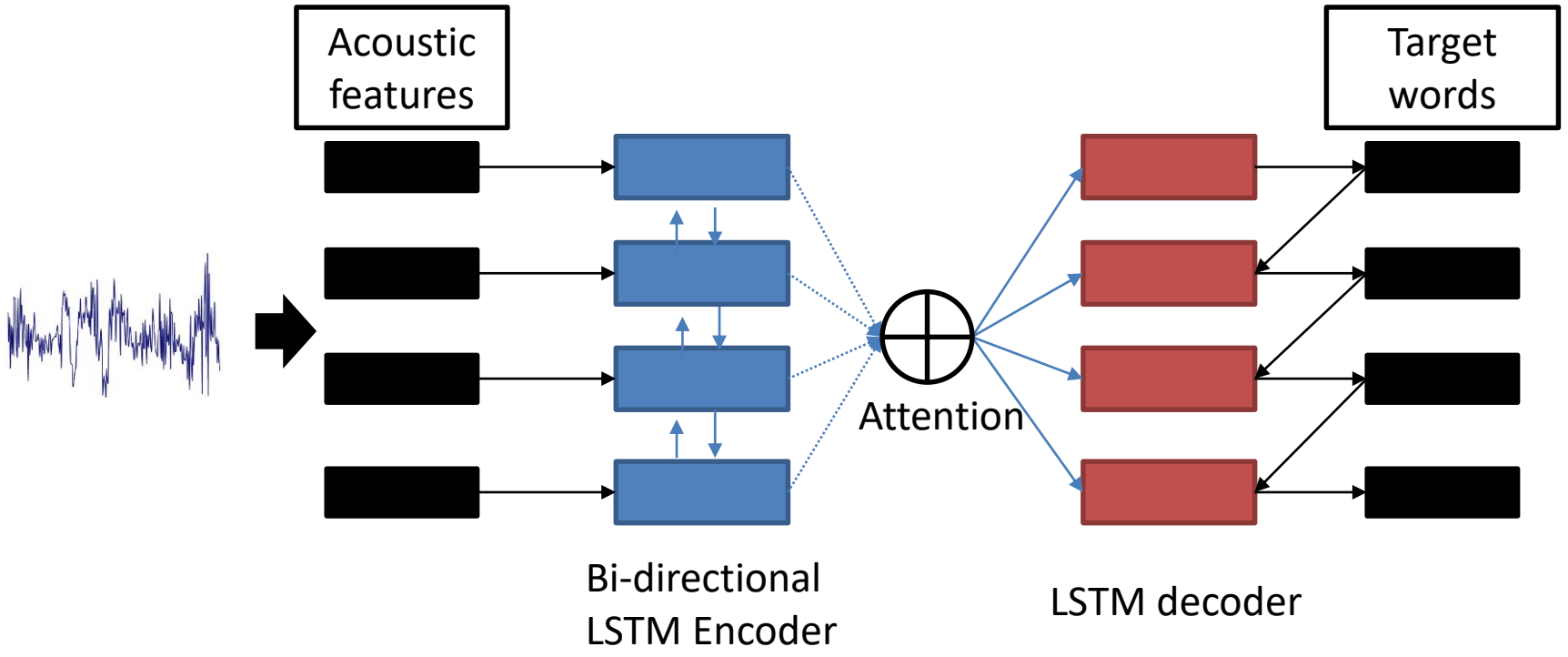


内容

1. 同時音声翻訳
   - 漸進的音声認識
   - 漸進的音声合成
   - 漸進的機械翻訳
   - 同時通訳，同時音声翻訳研究のためのコーパス構築
   - 同時音声翻訳プロトタイプ
2. パラ言語の取り扱い
   - パラ言語音声翻訳,音声表現とテキスト表現の等価性
3. End-to-end speech-to-speech translation
4. Machine Speech Chain
   - Multi-speaker Machine Speech Chain, Incremental Machine Speech Chain
5. まとめと今後の展開

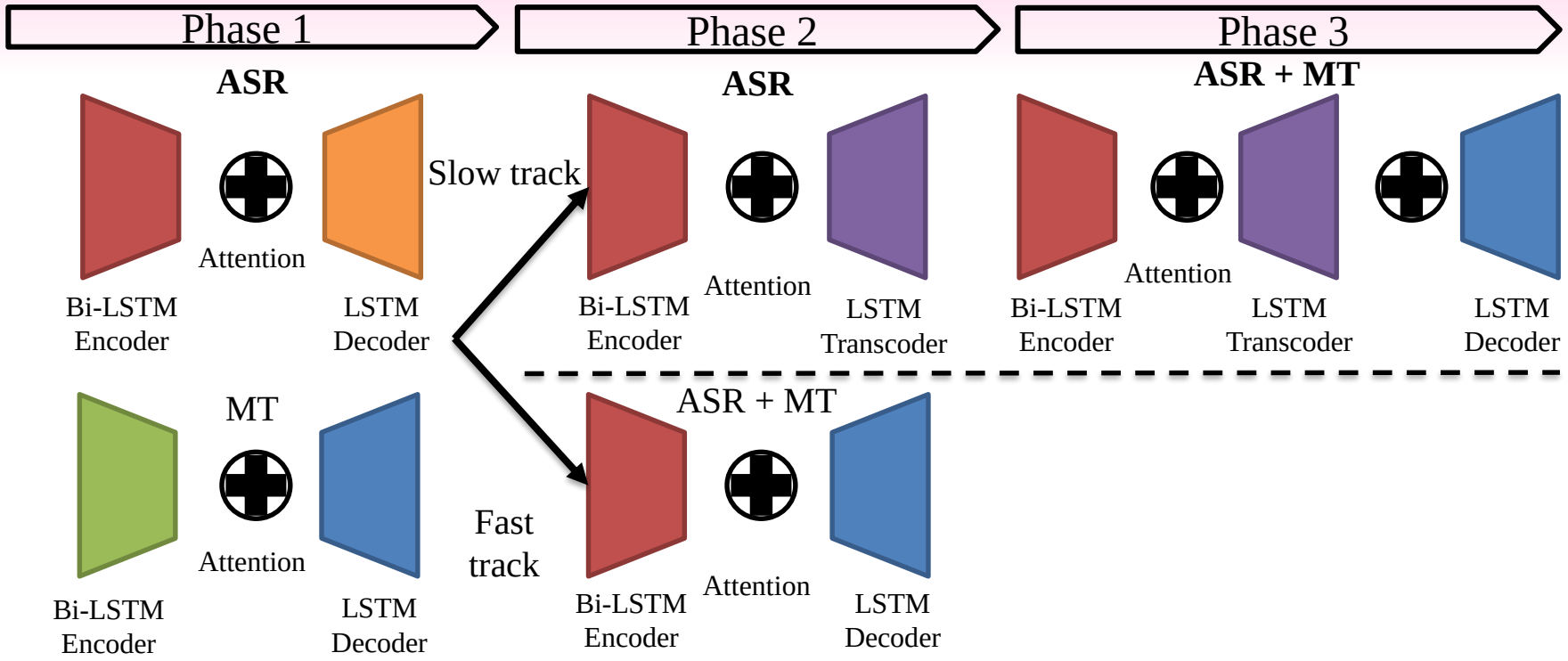
Deep Learningによる直接音声翻訳

- 音声入力、翻訳出力をEnd-to-endで学習できないか？
- Attention LSTMを用いたEnd-to-end 音声翻訳*



- L.Duong et al. NAACL 2016
- Alexandre Berard et. al “Listen and Translate: A Proof of Concept for End-to-End Speech-to-Text Translation” , NIPS workshop 2016

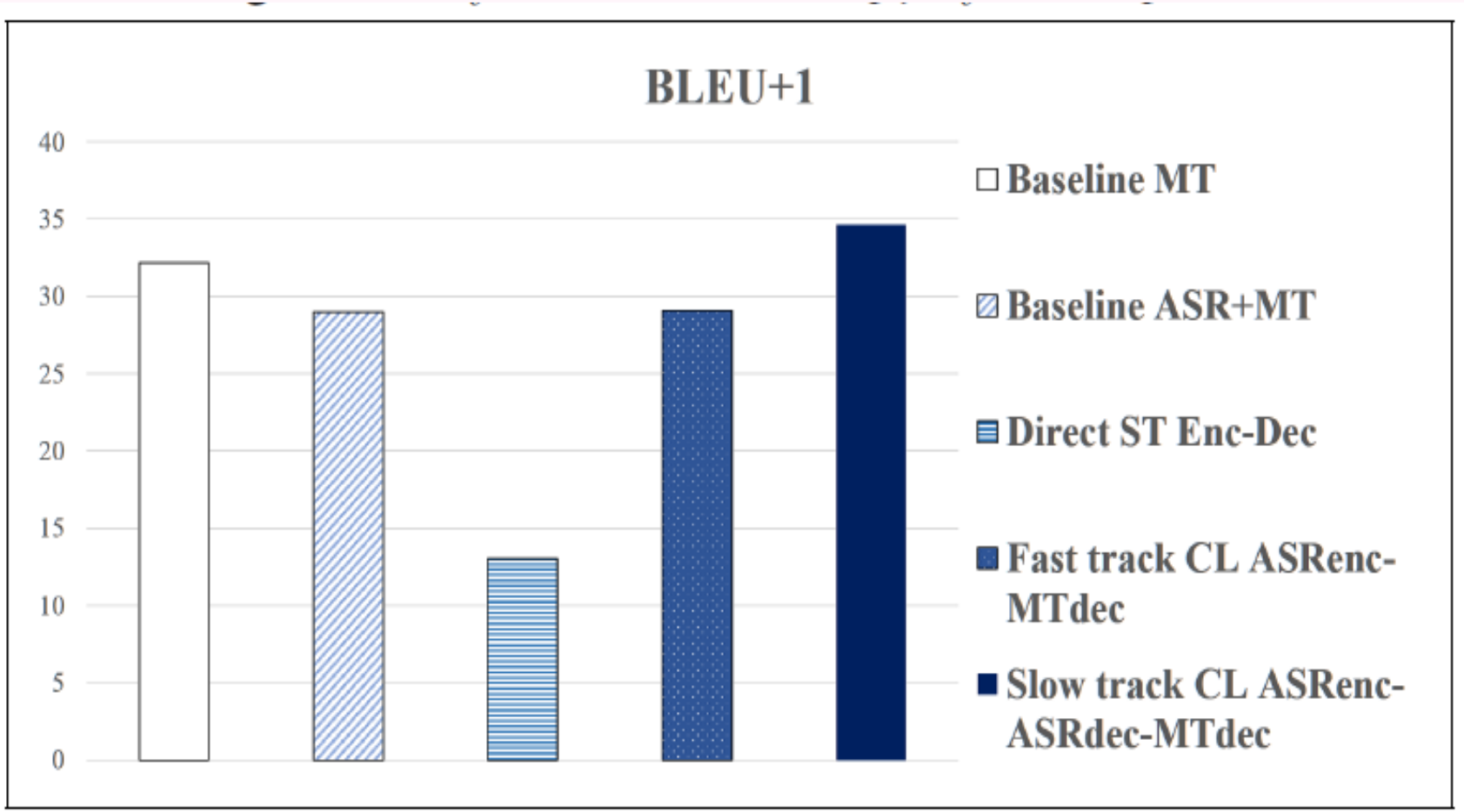
Curriculum Learningを用いたEnd-to-end 音声翻訳



Attentional-based LSTMをASR,MT用に学習しておき、逐次End-to-end音声翻訳にカリキュラム学習する

叶 高朋，サクリアニ サクティ，中村 哲，“カリキュラムラーニングを用いた日英直接翻訳システムの提案”、音響講論 2-10-5

Curriculum Learningによる音声翻訳結果

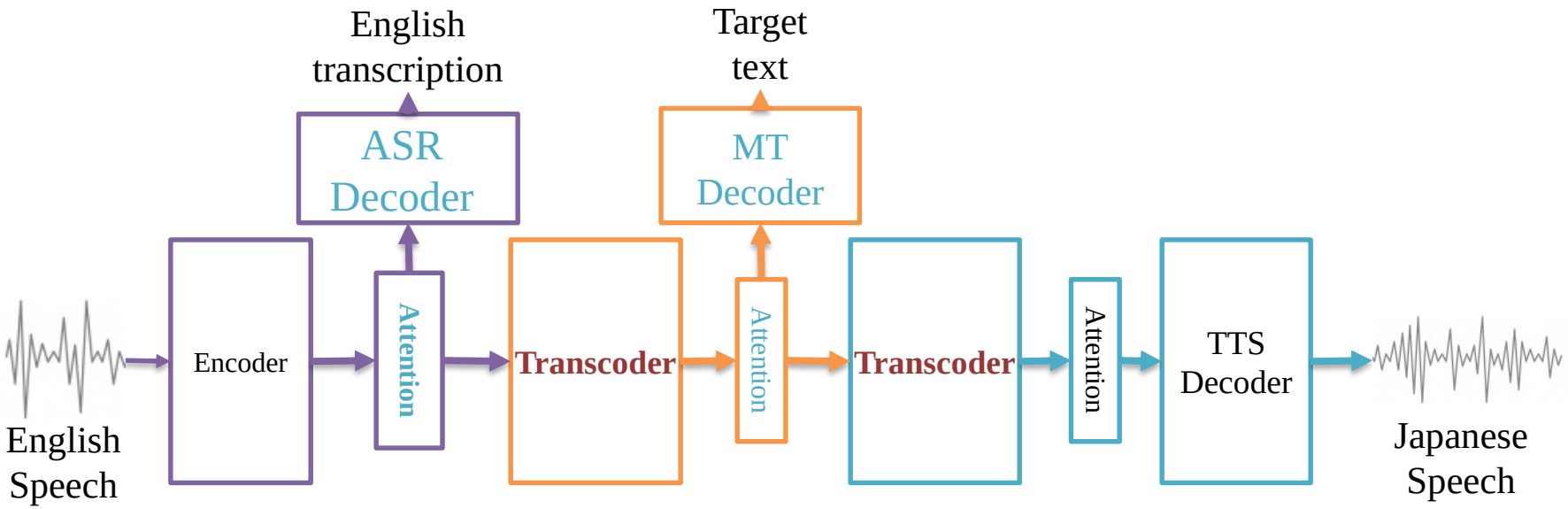


End-to-end音声翻訳の性能

Proposed: Transformer-based Direct Speech-to-speech Translation with Transcoder

Proposed

✓ **The model utilizes all trained modules to translate.**



- ✓ **ASR and MT decoders provide attention alignment necessary in the next block.**
  - ✓ The model uses a combination of 3 attentions for alignment source to target speech.
- ✓ **Transcoder converts different hidden states** in speech translation.
  - ✓ The model converts ASR and MT hidden states into MT and TTS hidden states.
- ✓ We revamp the **RNN**-based speech-to-speech translation model by the **Transformer**.

Takatomo Kano, Sakriani Sakti, Satoshi Nakamura, “TRANSFORMER-BASED DIRECT SPEECH-TO-SPEECH TRANSLATION WITH TRANSCODER”, IEEE Spoken Language Technology Workshop, Jan. 20, 2021

Experiment Results

BLEU and METEOR scores of speech-to-speech translation

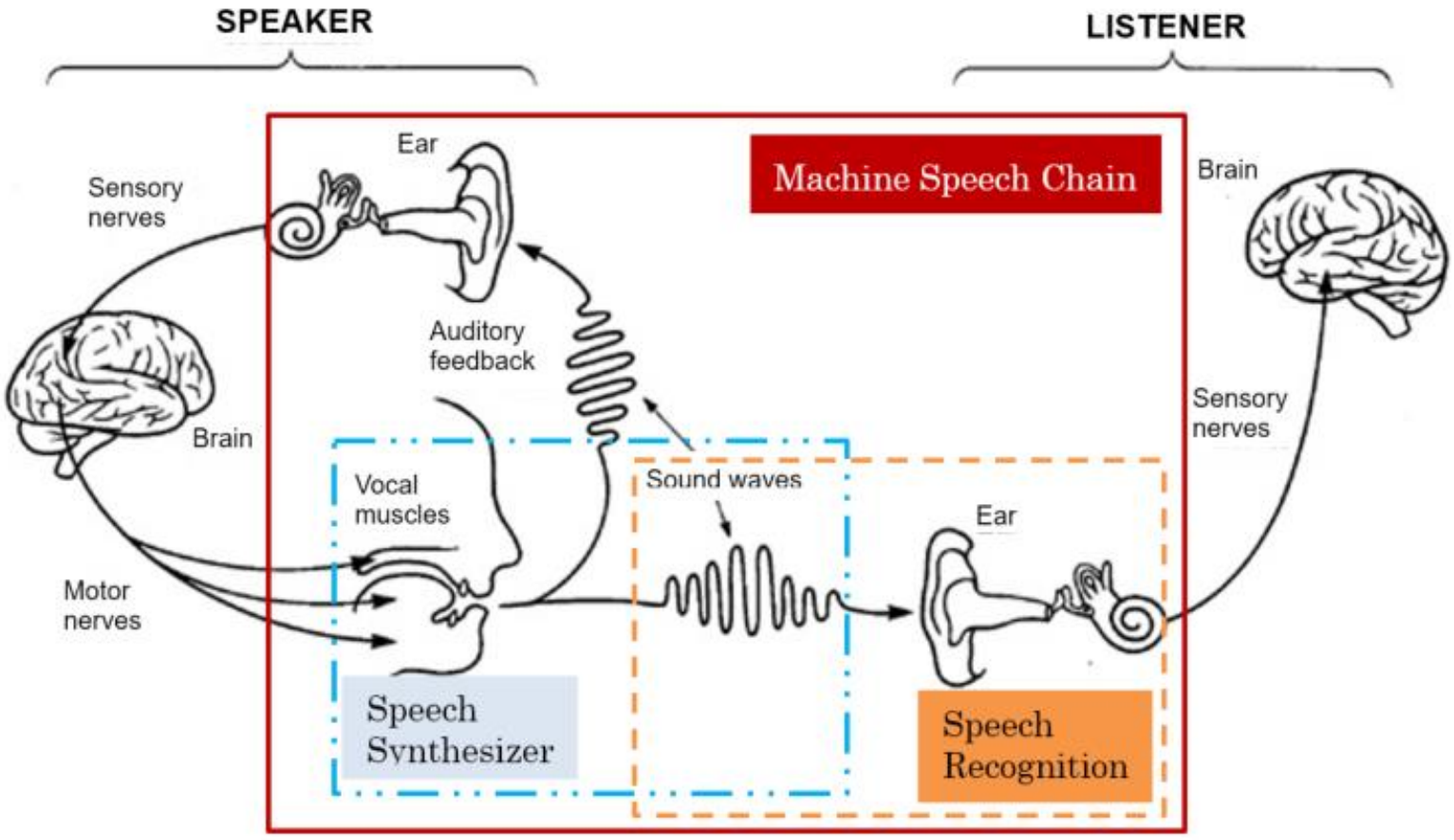
Model	Syntactic similar				Syntactic distant			
	En to ES		Ja to Ko		En to Ja		Ja to En	
	BLEU		METEOR		BLEU		METEOR	
Baseline: Cascade (RNN)	38.9	47.7	38.7	49.1	32.5	44.2	32.0	43.2
Baseline: Cascade (Transformer)	41.3	52.1	41.0	51.1	34.1	45.2	35.0	45.3
Multi-task (RNN)	38.8	48.2	39.1	49.9	33.2	45.5	34.2	45.0
Multi-task (Transformer)	43.1	58.8	42.5	58.3	36.9	52.6	38.3	48.4
Transcoder(Transformer)	44.0	59.3	42.9	58.8	40.6	56.6	41.0	55.8

- ✓Our proposed approach outperforms Multi-task speech translation.
- ✓The proposed method with Transformer further improves performances.

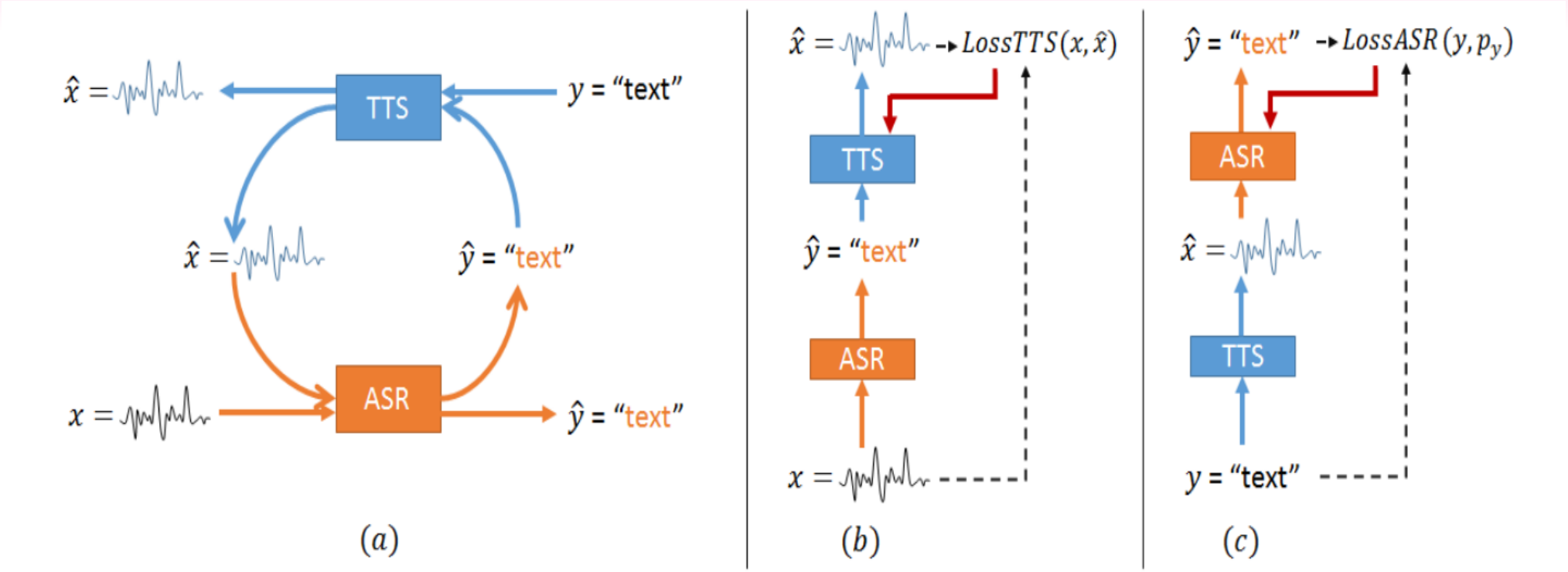
内容

1. 同時音声翻訳
   - 漸進的音声認識
   - 漸進的音声合成
   - 漸進的機械翻訳
   - 同時通訳，同時音声翻訳研究のためのコーパス構築
   - 同時音声翻訳プロトタイプ
2. パラ言語の取り扱い
   - パラ言語音声翻訳,音声表現とテキスト表現の等価性
3. End-to-end speech-to-speech translation
4. Machine Speech Chain
   - Multi-speaker Machine Speech Chain, Incremental Machine Speech Chain
5. まとめと今後の展開

①Human vs Machine Speech Chain



①Machine speech chain の構成



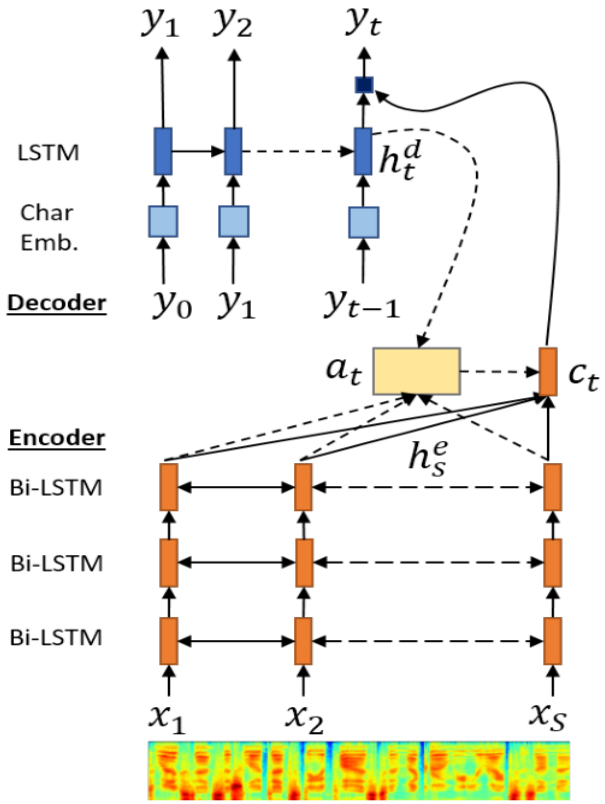
a. Machine speech chain の構成

b. ASR から TTS
c. TTSからASR

A.Tjandra,et al., "Listening while Speaking: Speech Chain by Deep Learning" , arXiv:1707.04879, 2017
Accepted for IEEE ASRU 2017

①Machine speech chain の構成

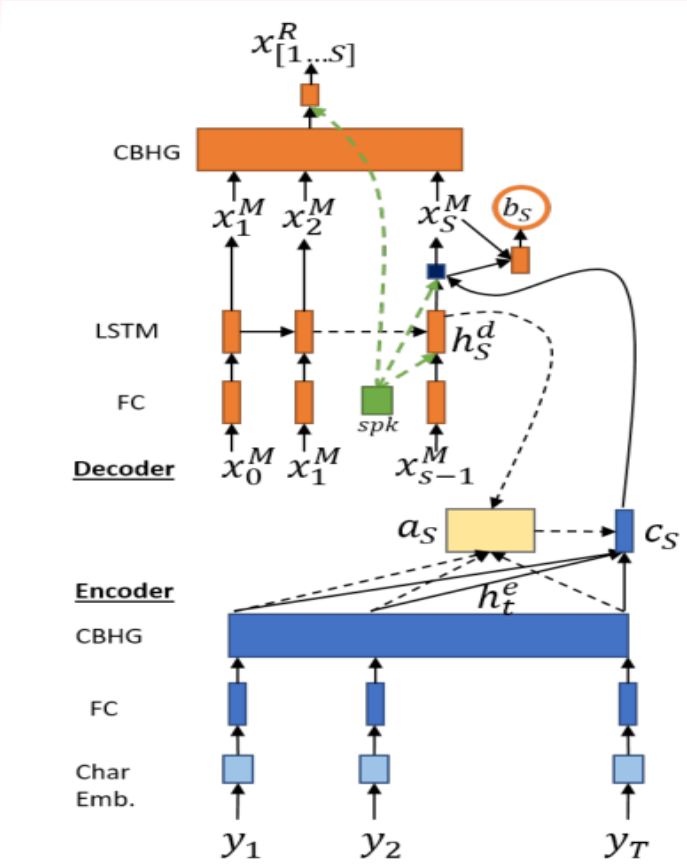
ASRの基本構造



Encoder with BiLSTM & sequence subsampling
Decoder with LSTM & attention module

Reference : Chan et al (2015), Listen Attend Spell

TTSの基本構造

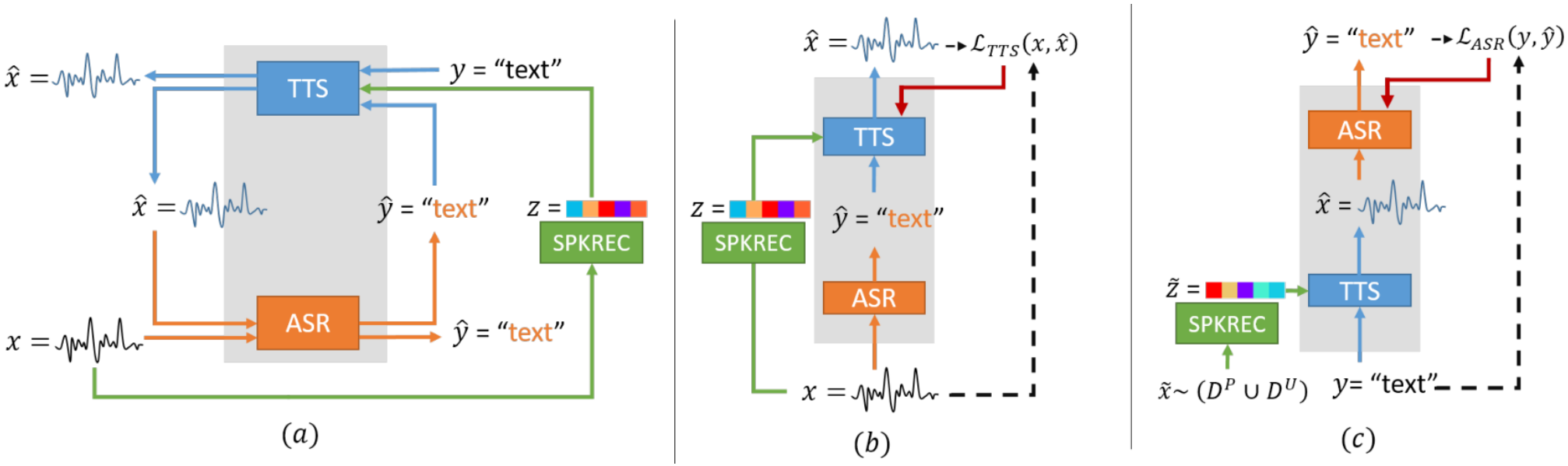


Modified Tacotron with additional speaker embedding + frame ending binary prediction

Reference : Wang et al (2017), Tacotron

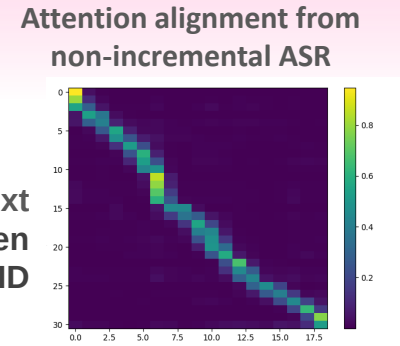
Sequel: Speech Chain with One-shot Speaker Adaptation

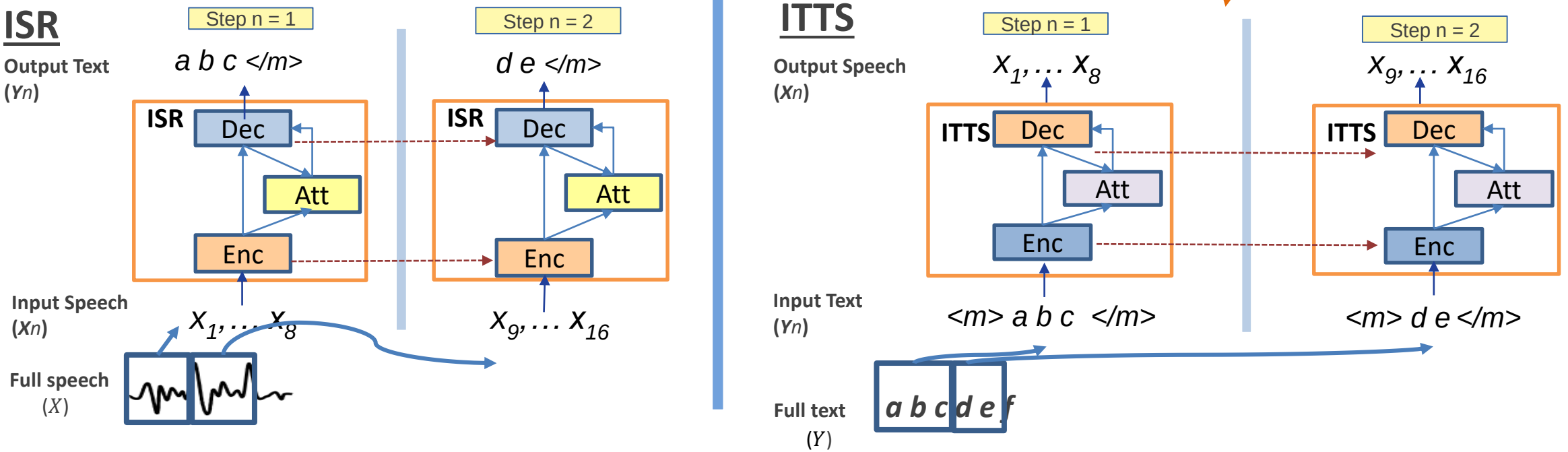
- 課題
  - 単一話者から複数話者へ
  - 未知の話者への対応（discrete speaker embedding）
- 提案モデル



ISR and ITTS Independent Training

- Incremental : Predict a complete output sequence in *N* steps, for each step *n* :
  1. Encode a segment of input from input window
  2. Decode and predict a segment of output
  3. Shift the input windows
- ISR and ITTS training by attention transfer from non-incremental ASR [Novitasari et al., 2019] → same alignment for ISR and ITTS

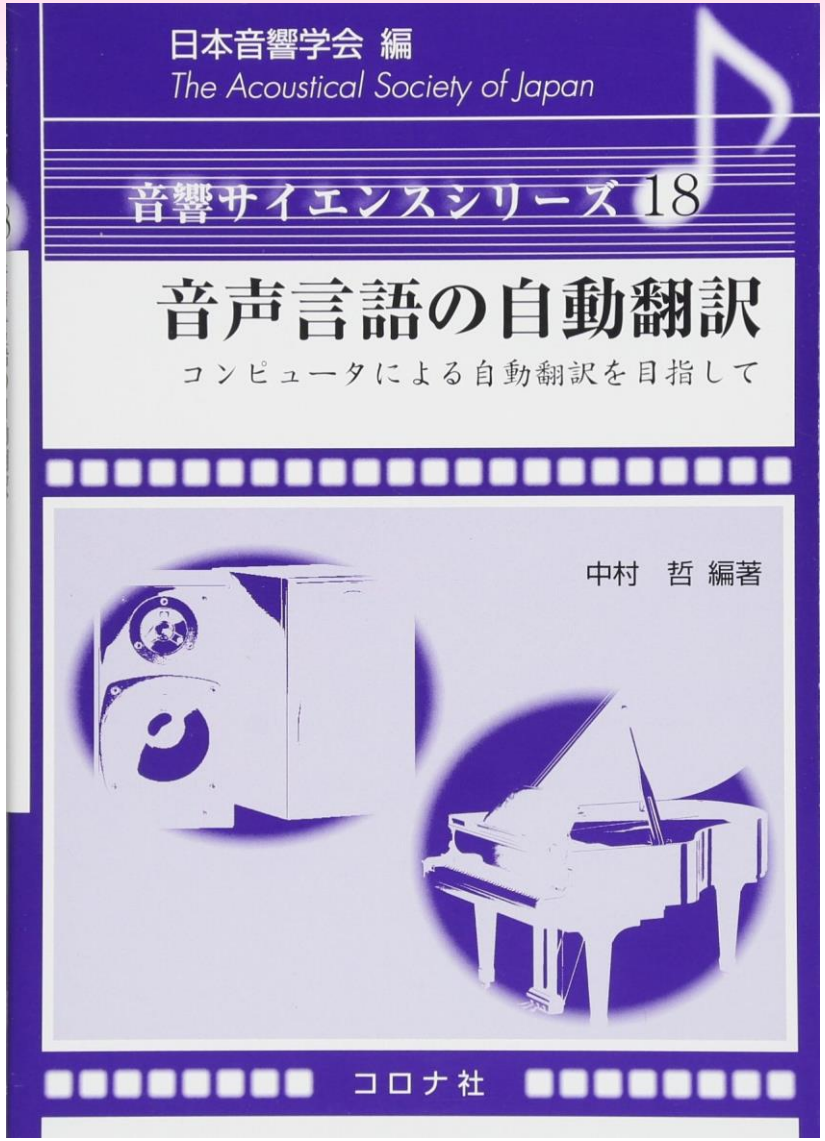




まとめ

1. 基盤Sによる次世代音声翻訳研究
    1. 同時音声翻訳
        - 漸進的音声認識
        - 漸進的音声合成
        - 漸進的機械翻訳
    - 同時通訳，同時音声翻訳研究のためのコーパス構築
    - 同時音声翻訳プロトタイプ
2. パラ言語の取り扱い
    - パラ言語音声翻訳,音声表現とテキスト表現の等価性
3. End-to-end speech-to-speech translation
4. Machine Speech Chain
    - Multi-speaker Machine Speech Chain, Incremental Machine Speech Chain

音声言語の自動翻訳　コロナ社



課題と研究課題

- 伝達遅延と精度低下のtrade-off：速く＆正確に
  - 洗練された訳出方略
    - 語順を崩さずに訳出できるか否かを逐次予測
    - 文構造を変えて訳出開始を早める（関係詞節等）
    - 予測機能により訳出開始を早める．述語等を予測してさらに早く（特に日英通訳）
    - 同時通訳の理論やノウハウの活用
  - 通訳品質，同時自動音声翻訳品質の評価法
- 音声認識，音声合成との統合
  - End-to-end 同時音声翻訳,パラ言語情報の取り扱い
  - 雑音・残響対応、遠隔認識
  - 全体最適化
- 持続的なデータ自動収集と教師無し学習
  - 方言・アクセント対応
- 対話構造・談話構造の考慮
- 知識表現、意味表現とその利用
- IWSLT SHARED TASK

とても5年では終わりそうにない．．．．