

Wed-A-V-6-7: Transcribing Paralinguistic Acoustic Cues to Target Language Text in Transformer-based Speech-to-Text Translation

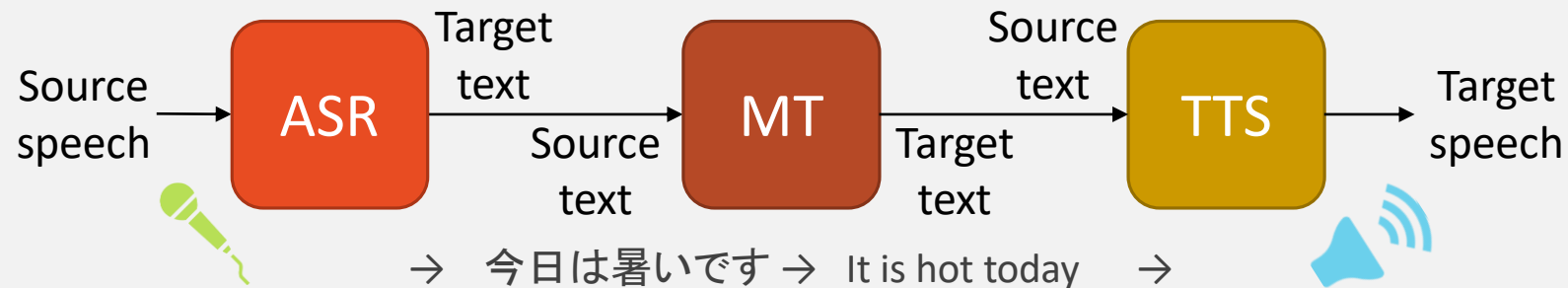
Hiroataka Tokuyama¹, Sakriani Sakti^{1,2}

Katsuhito Sudoh^{1,2} Satoshi Nakamura^{1,2}

¹Nara Institute of Science and Technology, Japan

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

Background: Speech-to-Speech Translation



- Speech-to-Speech (S2S) Translation has 3 systems

- ASR : Automatic Speech Recognition
- MT : Machine Translation
- TTS : Text-to-Speech

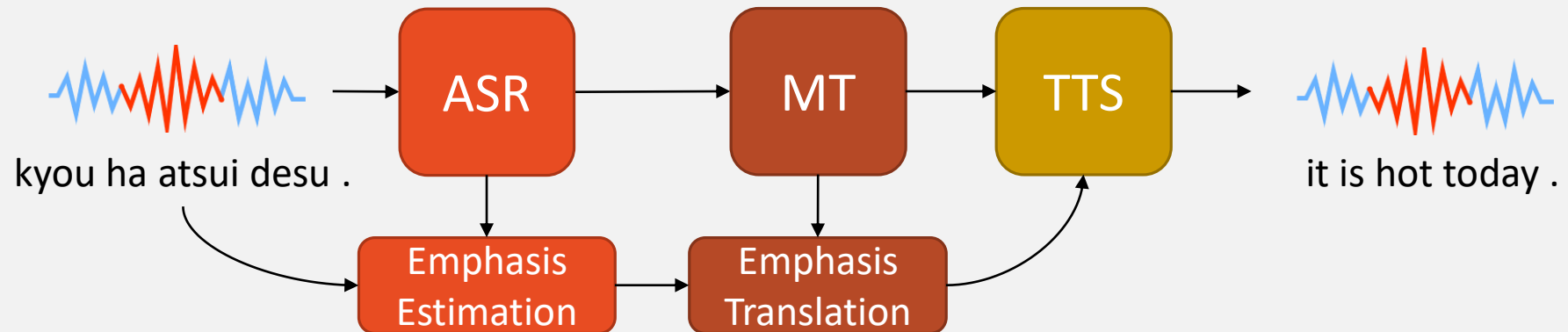
- But, conventional system has the same problems

- These systems based on linguistic information
 - What is being said and the actual words
- They cannot reflect Paralinguistic information
 - Emphasis, Emotion etc...
- Different intonation or emphasis can bring different meaning to the linguistic contents



Background: Previous Cascade S2S Structure

- Acoustic-to-Acoustic Emphasis translation [1]
 - Emphasis in input speech reflects output speech



- This framework is complicated and suboptimal
 - Estimate and translate emphasis separately
 - Neural MT with HMM-based ASR-TTS
 - Paralinguistic information was kept in acoustic form
- Is emphasis expressed in only acoustic (intensity) ?
 - Some language may be expressed in other ways[2]

[1] Sequence-to-Sequence Models for Emphasis Speech Translation [Q. T. Do 2018]

[2] Toward Multi-Features Emphasis Speech Translation: Assessment of Human Emphasis Production and Perception with Speech and Text Clues [Q. T. Do 2018]






Background: Paralinguistic Researches

- Court transcription of evidentiary audio recordings [3]
 - the contribution of prosodic and paralinguistic cues to the translation of evidentiary audio recordings was critical
 - It suggests that creating “written to be read as if spoken” text, instead of “written to be read”
 - “written to be read” text cannot reflect speakers' intentions, moods, power and attitudes
- It is important to constructing speech-to-text translation systems that convey acoustic information from speech into text-based communication
- The contributions of this topic
 - Expressing paralinguistic acoustic cues to text
 - Constructing cascade and direct neural speech-to-text translation
 - Performing Japanese-to-English linguistic and paralinguistic speech-to-text translation framework

[3] Contribution of prosodic and paralinguistic cues to the translation of evidentiary audio recordings [R. Chakhachiro 2016]

Propose: “Emphasis” of speech and text

- “Emphasis” is different between spoken and written form
 - Spoken: changed pitches, durations, and levels of power
 - Written: manifested by intensifier in text
- Ex. “It is hot today” with emphasis
 - Spoken: linguistic contexts are the same, but acoustic form is different
 - Written: adding intensifiers

Level	Emphasized Speech	Emphasized Text
0: Normal		It is hot today.
1: Light		It is a little bit hot today.
2: Medium		It is quite hot today.
3: Strong		It is very hot today.
4: Very strong		It is extremely hot today.

- Need to transcribe paralinguistic acoustic cues to text

Propose: Transcribing Paralinguistic Acoustic Cues

- Emphasis acoustic cues to embedded emphasis text

Method	Description	Ex. When emphasis level of "hot" is 3
No Emphasis	Original text	it is hot today .
Emphasis Separated	original text and emphasis symbols separately as in the previous method	it is hot today .
		0 0 3 0 0
Emphasis Tags	Add emphasis tags to each word	it ⁰ is ⁰ hot ³ today ⁰ . ⁰
Emphasis 1-Token	Add a token before emphasized word	it is <to3> hot today .
Emphasis All Token	Add a token before every word	<to0> it <to0> is <to3> hot <to0> today <to0> .

- Emphasis embedding text to natural text

it⁰ is⁰ hot³ today⁰ .⁰

it is **<to3>** hot today .

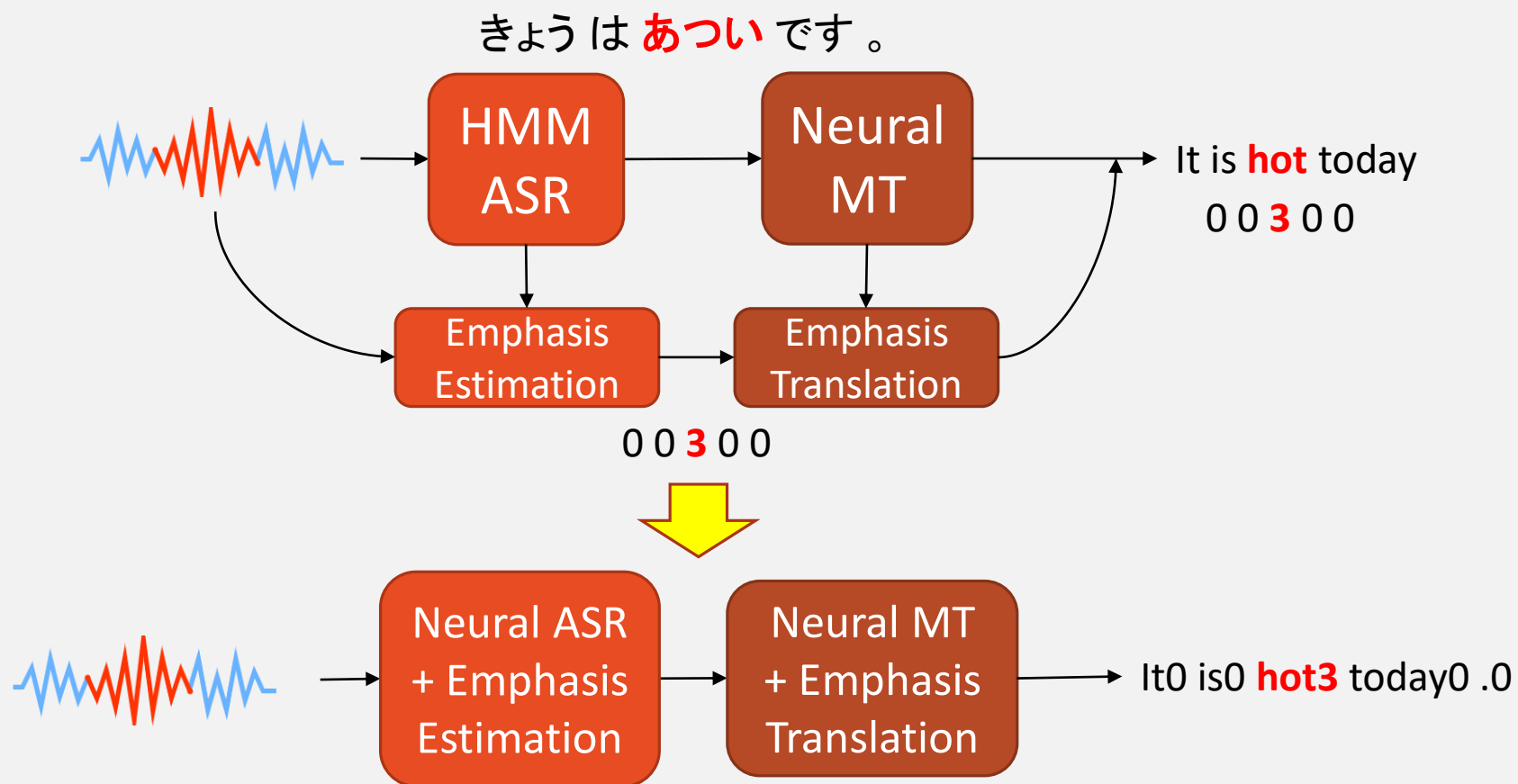
<to0> it **<to0>** is **<to3>** hot **<to0>** today **<to0>** .



It is **very hot** today.

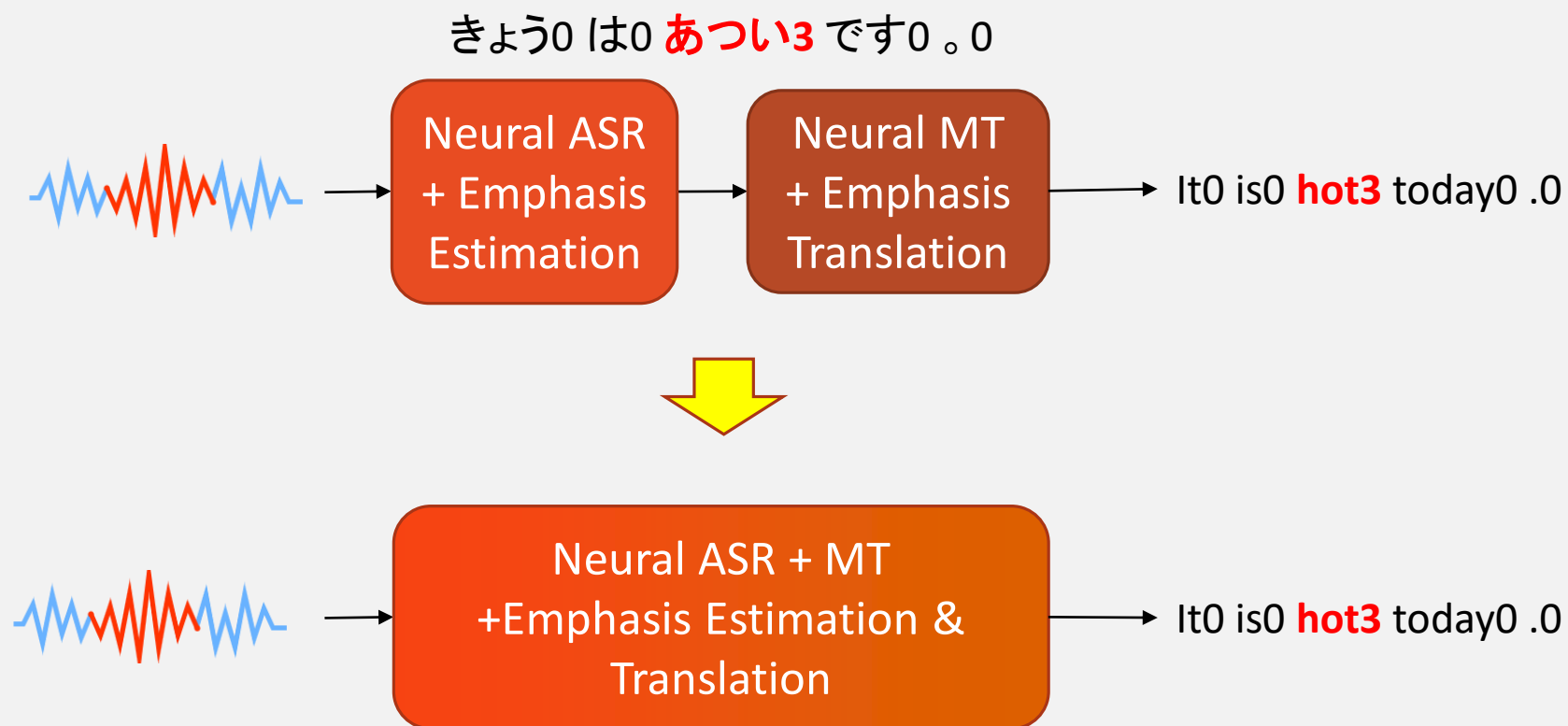
Propose Structure1 : Cascade Neural-S2T

- Simplify by using Neural Network in all systems
 - Emphasis Estimation and Translation are done in ASR and MT
 - Neural ASR and MT systems are trained separately earlier in each embedded emphasis text



Propose Structure2 : Direct Neural-S2T

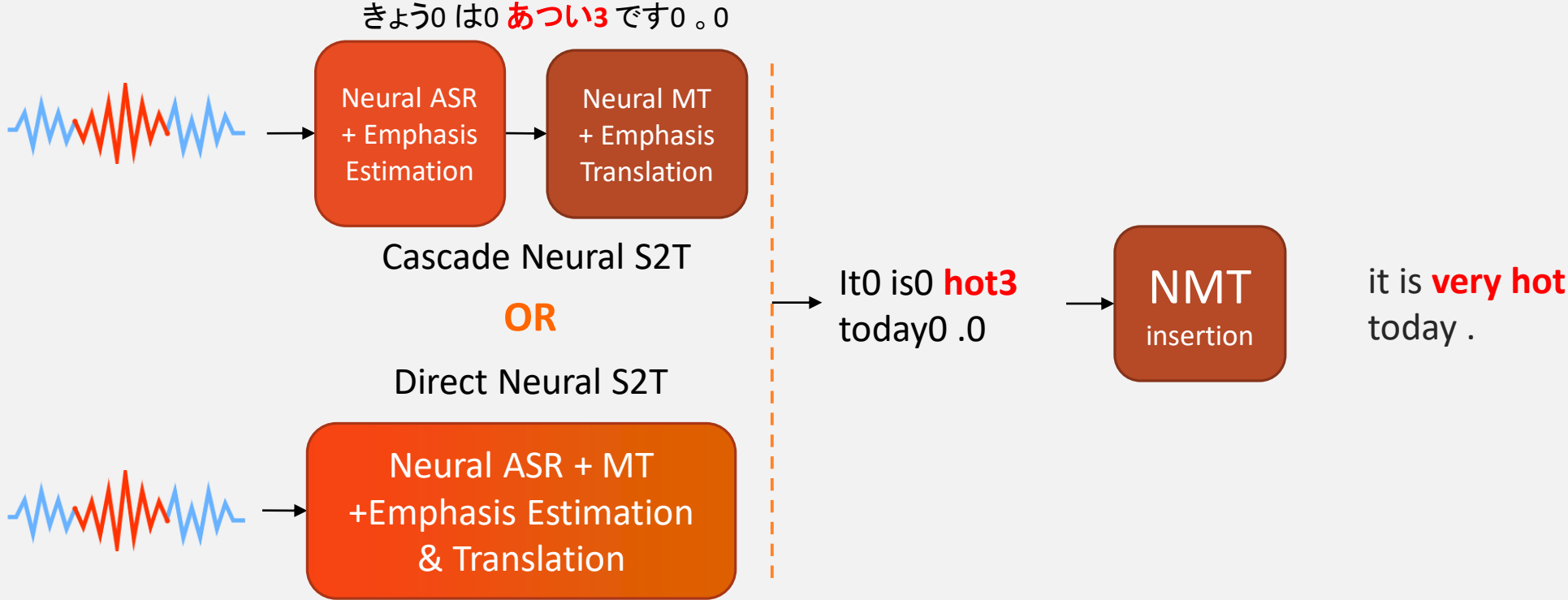
- Direct Speech-to-Text (S2T) translation using Neural Network
 - Combining ASR and MT systems including emphasis processes as single model
 - Translation that reflects the emphasis of speech
 - Using Jia et al.'s direct model as a reference [4]



[4] Direct speech-to-speech translation with a sequence-to-sequence model [Ye Jia 2019]

Propose : Acoustic-to-Linguistic Emphasis

- Acoustic Emphasis to Linguistic Emphasis (NMT-insertion)
 - Transform embedded emphasis text into natural text with intensifiers



Datasets and Models

• Dataset

- Natural Speech Dataset
 - Japanese-English natural speech and text dataset including emphasis
 - Text is prepared with emphasis part and intensifiers
 - 5,145 utterances (1,029 sentences x 5 levels)
 - 20x over sampling for training dataset
- BTEC (Basic Travel Expressions Corpus)
 - Over 400k English-Japanese Parallel Corpus
- ATR Speech Database of Many Speakers (APP-BLA)
 - 3,700 Japanese speakers reading 503 ATR's phonetically balanced sentences (without emphasis)
 - 127 speech hours in total

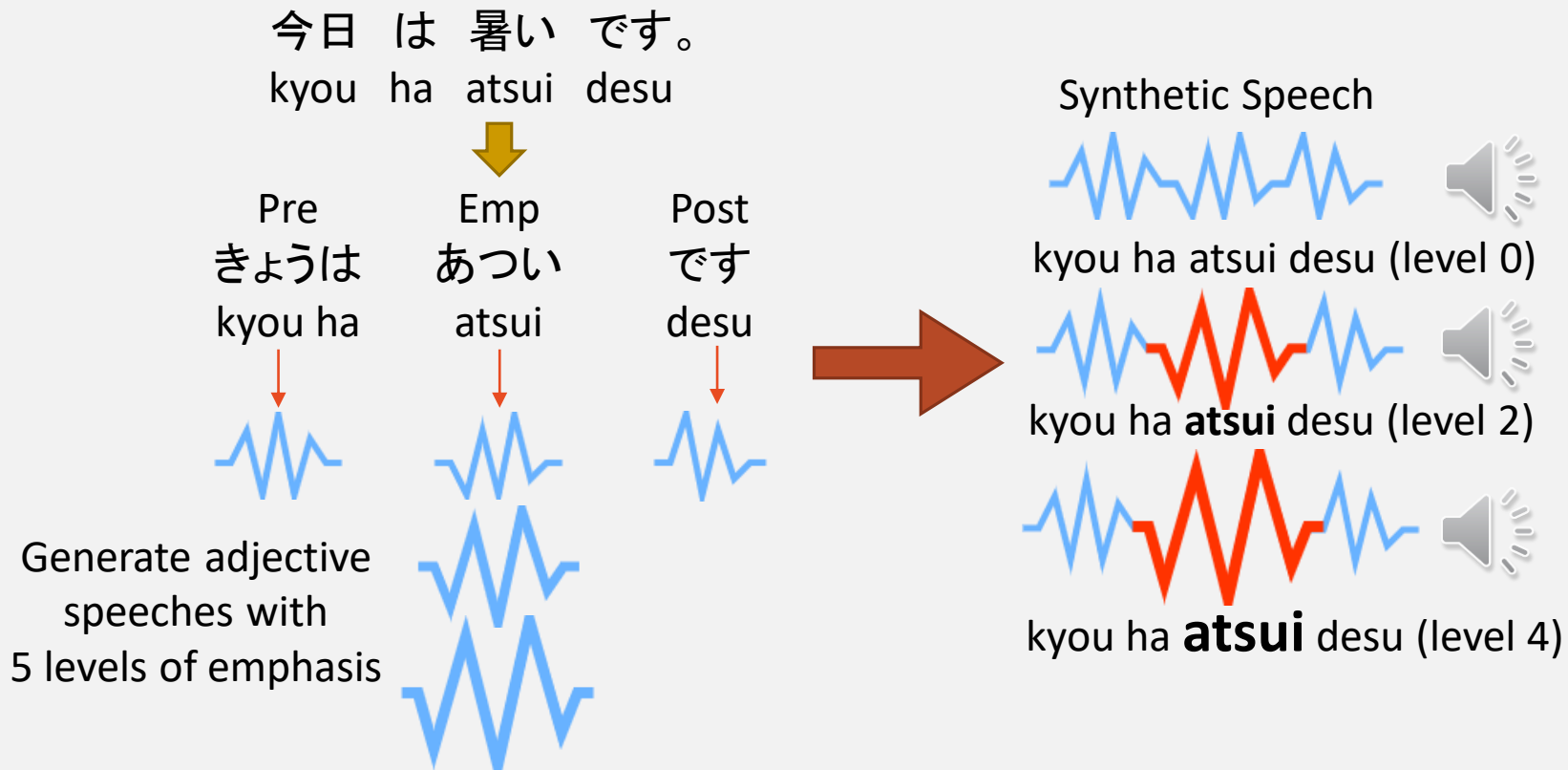
Training Dataset	Data Amount
Natural Speech	5,145
BTEC	465,466
APP-BLA	143,171

• Data Augmentation

- Natural speech with emphasis is too smaller than BTEC and APP-BLA
- Augment emphasized speech and text dataset by BTEC

Create Synthetic Emphasized Speech

1. Split before and after the first adjective
2. Generate speech each part by google TTS (gTTS)
3. Make emphasized speeches only adjective part for 5 levels
 - Normal, Light, Medium, Strong and Very strong
4. Concatenate each generated speech



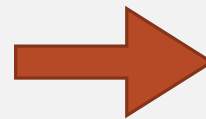
Create Synthetic Emphasized Text

1. Split before and after the first adjective
2. Add emphasis information in each word (5 levels just like speech)
3. The same process in English text
4. Align English and Japanese text
 - Some adjectives may not be used only in either Japanese or English text

今日 は 暑い です。
kyou ha atsui desu

↓

Pre	Emp	Post
きょうは	あつい	です
kyou ha	atsui	desu



Emphasis Tag

きょう⁰ は⁰ あつい³ です⁰ 。⁰
It⁰ is⁰ hot³ today⁰ .⁰

Emphasis Token

きょうは <to3> あついです。
It is <to3> hot today .

All Token

<to0> きょう <to0> は <to3> あつい
<to0> です <to0> 。
<to0> it <to0> is <to3> hot <to0> today <to0> .

Create Synthetic “Inserted” Text

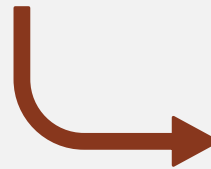
- Analyze and extract how many adverbs using in natural speech texts
 - Use top 5 of them based on respective emphases

Natural Speech Text (it is existed in advance)

Ex.1
Normal (level 0) : This is loose. (plain text)
Light (level 1) : This is **a tad** loose.
Medium (level 2) : This is **noticeably** loose.
Strong (level 3) : This is **so** loose.
Very strong (level 4) : This is **completely** loose.


Ex.2
My cholesterol is high.
My cholesterol is **a bit** high.
My cholesterol is **quite** high.
My cholesterol is **certainly** high.
My cholesterol is **extremely** high.

Analyze the inserted word



Light : **a tad, a bit ...**
Medium : **noticeably, quite ...**
Strong : **so, certainly ...**
Very strong : **completely, extremely ...**

- Insert one of them before the adjectives in target data randomly

we feel sick .  Normal : we feel sick. (plain text)
Light : we feel **a bit** sick.
Medium : we feel **quite** sick.
Strong : we feel **so** sick.
Very strong: we feel **completely** sick.

Experimental Set-up

• Dataset

◦ Train Dataset

- ASR : 718,142 speeches and texts
- MT / S2T : 574,971 text pairs
 - Natural speech dataset is 20x oversampling

◦ Test Dataset

- Natural : 500 sentences (100 sentences x 5 levels with Natural speeches and multi-speaker task)
- Synthetic : 510 sentences with 5 levels (from BTEC test set with Synthetic speeches and single-speaker task)

Training Dataset	ASR	MT	S2T	Data Amount
BTEC (emphasis)	O	O	O	184,620
BTEC (no emphasis)	O	O	O	297,451
Natural Speech	O	O	O	92,900
APP-BLA	O	X	X	143,171

• Model

- Implement OpenNMT-py using Transformer
- Parameter is the same each model

Transformer parameter	
Layer	3
RNN size	512
Word vec size	512
Transformer_ff	2048
Heads	8

• Evaluation

- ASR : word error rate (WER)
- MT, S2T: multi-bleu, sacrebleu
- Emphasis: F-score

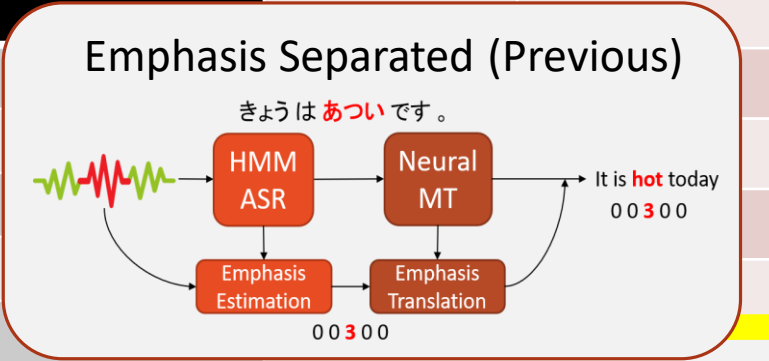
Evaluation : ASR & MT (Independently)

Emphasis Tags : it0 is0 hot3 today0 .0
 Emphasis Token : it is <to3> hot today .
 Emphasis All Token : <to0> it <to0> is <to3> hot <to0> today <to0> .

it is hot today . → Linguistic evaluation
 0 0 3 0 0 → Emphasis evaluation

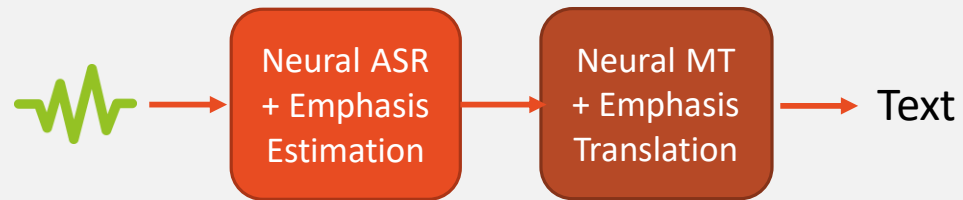
Test data	Data type		Linguistic Evaluation			Emphasis Evaluation (F-score)	
			ASR	MT		ASR	MT
			WER↓	Multi-bleu↑	Sacrebleu↑		
Synthetic (BTEC)	Previous	No Emphasis	0.91	44.84	47.80	---	---
		Emphasis Separated	0.91	43.99	46.71	90.91	23.18
	Proposed	Emphasis Tags	0.80	42.64	45.57	100.00	64.52
		Emphasis 1-Token	0.76	42.31	45.64	93.93	70.77
		Emphasis All Token	0.83	44.84	47.71	96.97	74.63

Natural	Method	Previous	
		Method	Result
Natural	Previous	No Emphasis	it is hot today .
		Emphasis Separated	it is hot today .
	Proposed	Emphasis Tags	0 0 3 0 0
		Emphasis 1-Token	it0 is0 hot3 today0 .0
		Emphasis All Token	it is <to3> hot today .
		Emphasis All Token	<to0> it <to0> is <to3> hot <to0> today <to0> .

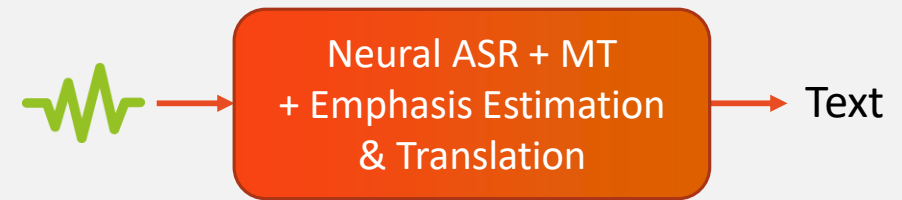


Evaluation : Speech-to-Text Translation

• Cascade Neural S2T

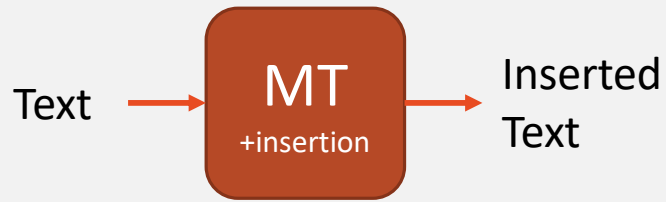


• Direct Neural S2T



Test data	Data type		Linguistic Evaluation				Emphasis Evaluation (F-score)	
			Cascade		Direct		Cascade	Direct
			Multi-bleu	Sacrebleu	Multi-bleu	Sacrebleu		
Synthetic (BTEC)	Previous	No Emphasis	42.98	45.92	44.62	47.94	---	---
		Emphasis Separated	42.46	45.40	---	---	22.22	---
	Proposed	Emphasis Tags	41.76	44.67	44.55	47.28	58.06	47.13
		Emphasis 1-Token	44.13	46.77	44.82	47.84	67.69	49.01
		Emphasis All Token	43.88	46.93	39.79	43.09	69.70	41.58
Natural	Previous	No Emphasis	19.53	21.72	7.56	9.54	---	---
		Emphasis Separated	18.41	20.55	---	---	34.81	---
	Proposed	Emphasis Tags	15.28	16.43	9.07	9.88	61.54	15.28
		Emphasis 1-Token	15.12	18.07	12.00	13.71	58.46	15.12
		Emphasis All Token	17.66	20.31	5.93	7.62	57.97	17.66

Evaluation : NMT insertion



Test data	System	Emphasis Level					Total
		0	1	2	3	4	
Synthetic (BTEC)	Emphasis Tags	98.18	87.88	93.94	94.94	90.91	96.81
	Emphasis 1-Token	99.20	90.91	100.00	100.00	100.00	98.90
	Emphasis All Token	94.33	93.94	96.97	90.90	93.94	94.25
Natural	Emphasis Tags	84.53	67.03	81.11	80.90	75.28	77.85
	Emphasis 1-Token	93.62	82.98	92.55	91.30	91.30	90.34
	Emphasis All Token	88.04	73.91	79.35	81.11	82.22	80.92

• Evaluation

- Calculate the percentage of Inserted word is matched each emphasis level without changing the original text

Evaluation Summary

- Recognition and Translation of paralinguistic information
 - is possible by using emphasized text data
 - Compared to no emphasis, the score is almost about the same or better
- Considering S2T and insertion, using tokens is better performance
 - Especially 1-token works the best in proposed method
- But, there is a gap between Natural and Synthetic data evaluation
 - The performance of ASR affects S2T
 - Synthetic evaluation is a single speaker task while natural is a multi speaker task
 - A large portion of training data is synthetic data
 - More natural data is necessary

Conclusion and Future Work

- Purpose
 - Translate with acoustic and linguistic emphasis
 - Use neural Transformer-based model in Cascade and Direct Speech-to-Text Translation
- Method :
 - Transcribe paralinguistic acoustic cues to text
 - Generate synthetic speech dataset with emphases by google TTS
 - Perform neural Transformer-based paralinguistic speech-to-text translation
- Results
 - Our propose performed better than the baseline in terms of linguistic and paralinguistic evaluation
- Future Work :
 - Analyze how humans change acoustic cues to emphasize words
 - Evaluate results and speech data subjectively
 - Implement TTS for paralinguistic speech-to-speech translation

Appendix

Vocabulary Size

- Difference of vocabulary in each method

Method	Japanese	English
No emphasis	25471	29091
Emphasis Tag	27767	37503
Emphasis 1-Token	25475	29095
Emphasis All-Token	25476	29096

- In Emphasis Tag, emphasized word are different from original word
 - Ex. “hot0” and “hot1” are recognized as different word
 - Lighter structure is indispensable because of larger model

Adverb List from Natural Speech Text

- Light (Level 1)

- 'a little', 'slightly', 'a tad', 'a bit', 'kind of', 'sort of', 'vaguely', 'almost', 'marginally', 'mildly', 'passably', 'faintly', 'a tiny', 'a touch', 'something of', 'just a', 'bit of', 'a slightly', 'more or', 'perceptibly', 'hardly', 'even a', 'imperceptibly', 'a mildly' ...

- Medium (Level 2)

- 'pretty', 'fairly', 'rather', 'quite', 'somewhat', 'relatively', 'moderately', 'reasonably', 'significantly', 'tolerably', 'noticeably', 'considerably', 'uncommonly', 'more than', 'unusually', 'admittedly', 'comparatively', 'visibly', 'partly', 'mostly', 'probably', ...

- Strong (Level 3)

- 'so', 'very', 'really', 'truly', 'super', 'largely', 'definitely', 'clearly', 'much', 'considerably', 'decidedly', 'certainly', 'undeniably', 'positively', 'deeply', 'greatly', 'undoubtedly', 'unquestionably', 'unmistakably', 'so very', 'such', 'vastly', 'prohibitively', 'assuredly', 'mostly'.....

- Very strong (Level 4)

- 'seriously', 'terribly', 'extremely', 'completely', 'impossibly', 'perfectly', 'supremely', 'obviously', 'shockingly', 'totally', 'awfully', 'exceptionally', 'horribly', 'dreadfully', 'unbearably', 'wildly', 'powerfully', 'entirely', 'amazingly', 'wonderfully', 'hideously' ...

Analysis : ASR Results

- Input speech (Natural Speech)

- このきかくはせいさんのめんですずい。(この企画は生産の面ですずい。)
- “まずい” has emphasis level 3

Data type		Output
Previous	No Emphasis	このきたくはせいさんのえんですずい。
	Emphasis Separated	このきかくはせいさんよんねんですずい。 (Emphasis Estimation) 000000030
Proposed	Emphasis Tags	この0きたく0は0せいふ0の0ねんれい0まずい3。0
	Emphasis 1-Token	このきかくはそれですんねんまえで<to3>まずい。
	Emphasis All Token	この<to0>ちかく<to0>は<to0>せーるすまん<to0>の<to0>ねんれい<to0>が<to3>まずい<to0>。

- Emphasis recognition is mostly correct but language recognition is often different
 - More natural speech data is needed because the percentage of natural speech is small
- When testing natural speech data, sometimes recognize different emphasis level

Analysis : MT Results

- Input text -> Correct output
 - このきかくはせいさんのめんでまずい。 -> This scheme is clumsy production wise .
 - Red word is emphasis part (level 3)

Data type		Output
Previous	No Emphasis	This project lacks control .
	Emphasis Separated	This project is going to be tasteless (Emphasis Translation) 0 0 3 3 0 0 0
Proposed	Emphasis Tags	Something0 is0 wrong3 with0 this0 plan0 .0
	Emphasis 1-Token	This project is <to3> tasteless .
	Emphasis All Token	<to0> This <to0> project <to0> is <to0> completely <to3> wrong <to0> .

- Translate depends on train dataset, but emphasis part is translated even if used different word
- Separated depends on Attention and sometimes translate emphasis to wrong emphasis part or other token (ex. <unk>)
- Emphasized “clumsy” word is not in train data
- Including Emphasis translation, proposed is better

Analysis : S2T Task with Synthetic Speech

- Input speech (Synthetic Speech) -> Correct output text
 - やすいれすとらんをしょうかいしていただけますか。
-> could you recommend an inexpensive restaurant ?
 - Red word is emphasis part (level 3)

	Data type		Output
Cascade	Previous	No Emphasis	could you recommend a cheap restaurant ?
		Emphasis Separated	could you recommend a cheap restaurant ? (Emphasis Translation) <s> <s> <s> 3 3 0 0 0
	Proposed	Emphasis Tags	could0 you0 recommend0 a0 inexpensive3 restaurant0 ?0
		Emphasis 1-Token	could you recommend a <to3> cheap restaurant ?
		Emphasis All Token	<to0> could <to0> you <to0> recommend <to0> an <to3> inexpensive <to0> restaurant <to0> ?
	Direct	Previous	No Emphasis
Proposed		Emphasis Tags	could0 you0 recommend0 a0 cheap3 restaurant0 ?0
		Emphasis 1-Token	could you recommend a <to3> cheap restaurant ?
		Emphasis All Token	<to0> could <to0> you <to0> recommend <to0> an <to3> inexpensive <to0> restaurant <to0> ?

- When ASR has good results, S2T Task works well

Analysis : S2T Task with Natural Speech

- Input speech (Natural Speech) -> Correct output text
 - このきかくはせいさんのめんでまずい。-> This scheme is clumsy production wise .
 - Red word is emphasis part (level 3)

	Data type		Output
Cascade	Previous	No Emphasis	i have to pay my bill now .
		Emphasis Separated	i have to pay my bill three times before i go home . (emphasis translation) 0 0 0 0 0 0 0 0 0 0 0
	Proposed	Emphasis Tags	the0 neighbors0 are0 tasteless3 .0
		Emphasis 1-Token	these are the <to3> tastes bad around here .
		Emphasis All Token	<to0> salesman <to0> 's <to3> events <to0> in <to0> this <to0> neighborhood <to0> .
	Direct	Previous	No Emphasis
Proposed		Emphasis Tags	this0 question0 is0 difficult3 to0 answer0 .0
		Emphasis 1-Token	this tea is <to3> unusual .
		Emphasis All Token	<to0> this <to0> question <to0> is <to3> difficult <to0> to <to> answer <to0> .

- S2T task depends on ASR results
 - When ASR outputs wrong text, MT outputs worse texts ...
- We need more natural speech dataset for ASR

Analysis - NMT insertion

- Input text (English text) -> Correct output text (English inserted text)
 - this scheme is clumsy production wise . -> this scheme is so clumsy production wise .
 - Inserted words are 'so', 'very', 'really', 'undeniably', 'positively', 'unquestionably' and etc...

Data type	Output
Emphasis Tags	this scheme is detail production wise.
Emphasis 1-Token	this scheme is so clumsy production wise .
Emphasis All Token	this scheme is so clumsy production wise .

- Almost good results when using token
 - Guessing there are correspondence between the token and the word to be inserted
- Sometimes other words occurs when using tags
 - Emphasized “clumsy” is not in train data, so it may make to replace other word
 - Ex. “clumsy0” is in train data, but “clumsy1” is not in train data
 - But, some emphasized word works well
 - Ex. this0 is0 loose4 .0 -> this is stupendously loose. (呆気なく,凄まじく)