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

Hirotaka Tokuyama[1], Sakriani Sakti[1,2]

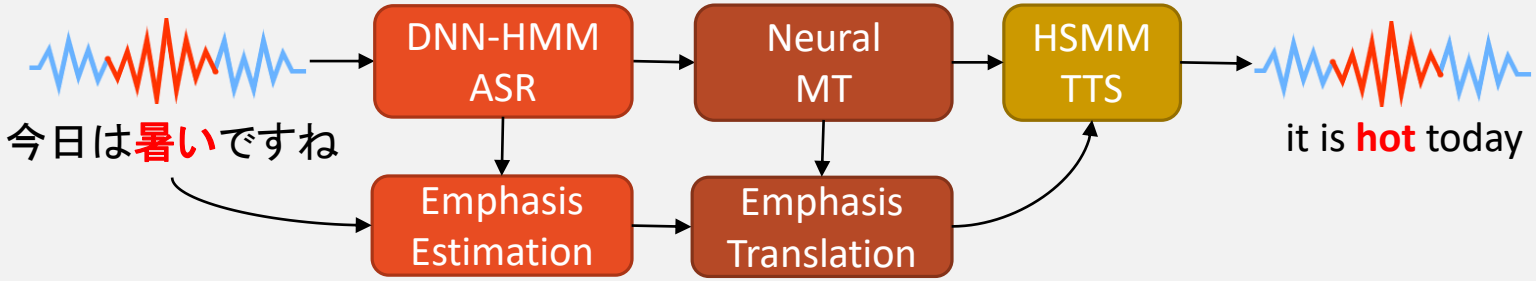
Katsuhito Sudoh[1,2] Satoshi Nakamura[1,2]

[1]Nara Institute of Science and Technology, Japan

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

Overview

- Conventional Speech-to-Text Translation system ignoring **paralinguistic information**
  - Ex. Emphases, emotion etc…
- Do et al. proposed Speech-to-Speech Translation framework focused on emphasis [1]



- BUT, this framework is complicated and suboptimal
  - Recognize and translate emphasis separately
  - Neural MT with HMM-based ASR-TTS
  - Paralinguistic information was kept in acoustic form
- Many applications require text documentation that is **"written to be read as if spoken"**
- This paper constructs a novel Transformer-based speech-to-text translation system

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

Proposed

1. Transcribing Paralinguistic Acoustic Cues
   - Embed emphasis information to emphasized word

Normal Speech

it is hot today .

Embed emphasis information

Tags :it0 is0 hot0 today0 .0
1-Token :it is hot today.
All Token :<to0> it <to0> is <to0> hot <to0> today <to0> .

Emphasized Speech (Level 3)

it is **hot** today .

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

   - Transform to natural text with intensifiers

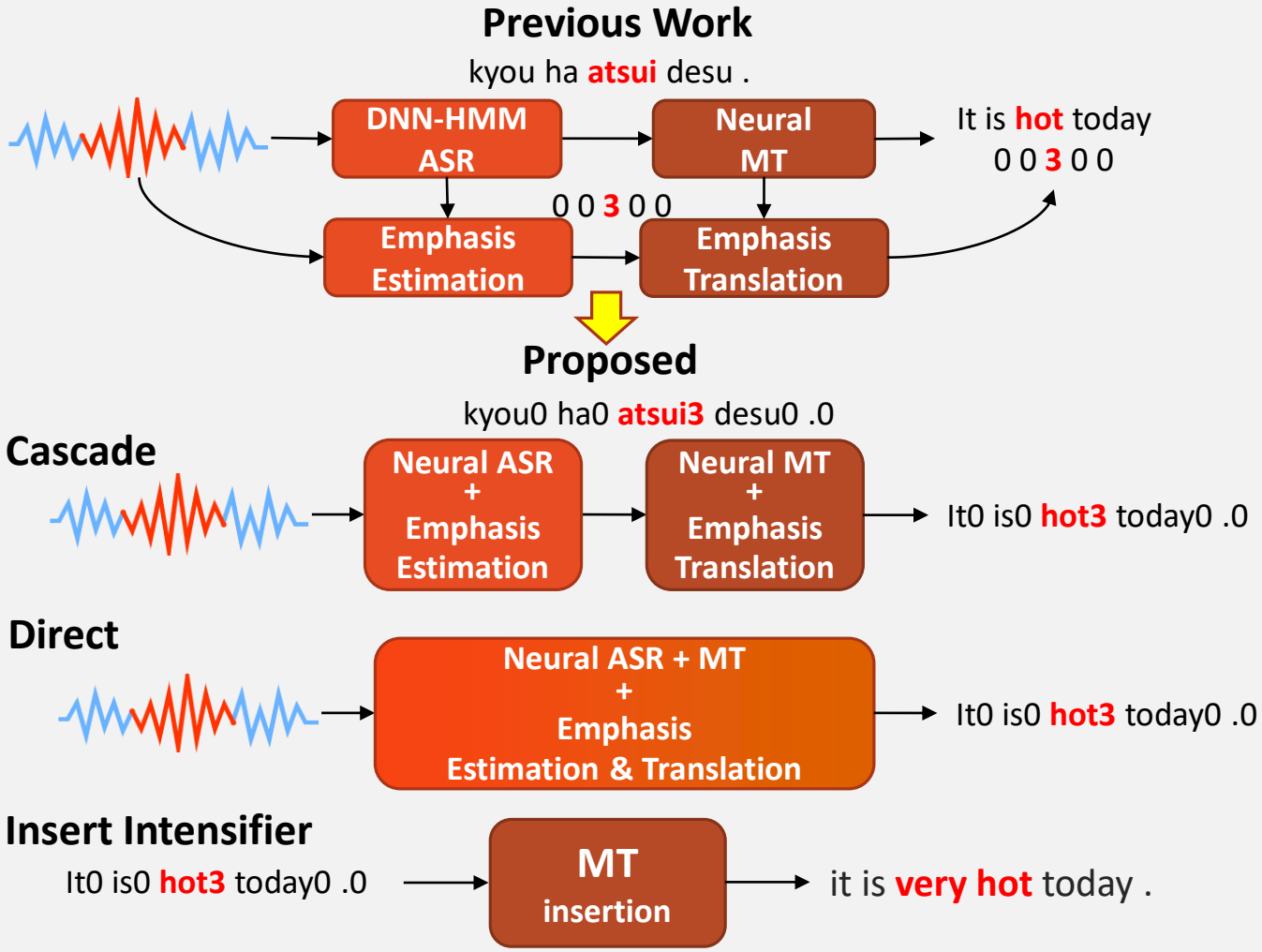
Tags :it0 is0 hot3 today0 .0
1-Token :it is <to3> hot today.
All Token :<to0> it <to0> is <to3> hot <to0> today <to0> .

→ it is **very hot** today .

Proposed

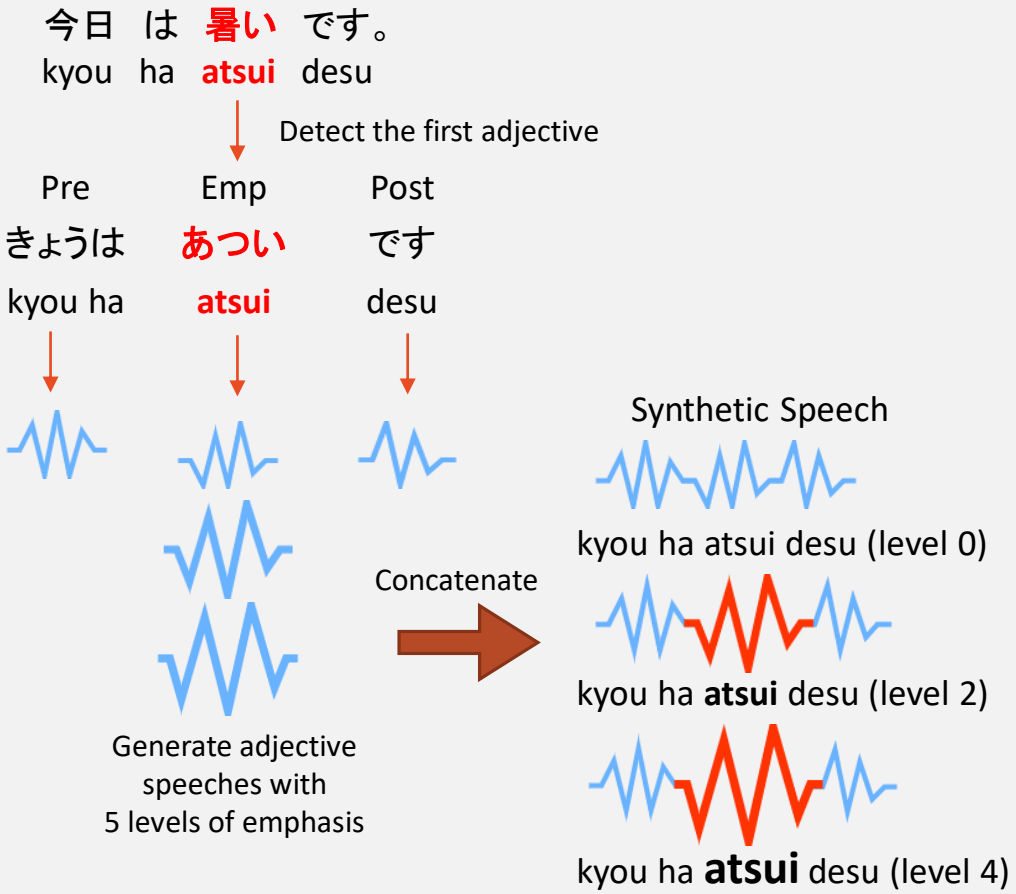
2. Speech-to-Text Translation Architecture

- Simplify framework and Insert intensifiers



3. Data Augmentation

- Make emphasized speeches from corpus data



Result

System		Linguistic Evaluation (Sacrebleu)		Emphasis Evaluation (F-score)	
		Cascade	Direct	Cascade	Direct
Previous	No Emphasis	45.92	47.94	---	---
	Emphasis Separated	45.40	---	22.22	---
Proposed	Emphasis Tags	44.67	47.28	58.06	47.13
	Emphasis 1-Token	46.77	**47.84**	67.69	49.01
	Emphasis All Token	**46.93**	43.09	69.70	41.58

System	Example	
No Emphasis	it is hot today .	
Emphasis Separated	it is **hot** today . 0 0 **3** 0 0	* Same as previous work
Emphasis Tags	it0 is0 **hot3** today0 .0	
Emphasis 1-Token	it is **<to3>** **hot** today .	
Emphasis All Token	<to0> it <to0> is **<to3>** **hot** <to0> today <to0> .	

adding token only before the emphasis word

System	Emphasis Level					Total
	0	1	2	3	4	
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

- It is possible to recognize and translate paralinguistic information by 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