

Incremental TTS for Japanese Language

TOMOYA YANAGITA¹, SAKRIANI SAKTI^{1,2}, SATOSHI NAKAMURA^{1,2}

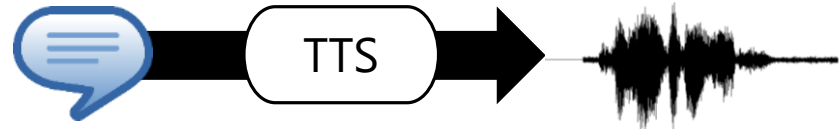
¹NARA INSTITUTE OF SCIENCE AND TECHNOLOGY, JAPAN

²RIKEN, CENTER FOR ADVANCED INTELLIGENCE PROJECT AIP, JAPAN

Background

Text-To-Speech (TTS) needs a full sentence as input

“私の名前は柳田です。”
(I am Yanagita.)



If text length is long, TTS has to wait till the end and cause a delay

“今日私はインクリメンタル音声合成
について話しますが、もちろん皆さん
ご存知の方がおいと”
(Today, I will talk about incremental
speech synthesis, but I think that there
are many people who know))



No response yet
(waiting the end
of sentence)

Require to synthesize speech in smaller chunks
→ Incremental TTS (ITTS)

Related Works

Limited existing works on ITTS (Mostly based on HMM)

Problems:

- Some contextual linguistic features become unknown
- Speech quality may deteriorate compared to standard HMM-TTS

Related Works

Limited existing works on ITTS (Mostly based on HMM)

Problems:

- Some contextual linguistic features become unknown
- Speech quality may deteriorate compared to standard HMM-TTS

[Baumann et al., 2014]

Analysis the impact of potentially missing features on the quality of the estimated prosody.

- Investigation only base on word-by-word synthesis extension.
-> German, English

Related Works

Limited existing works on ITTS (Mostly based on HMM)

Problems:

- Some contextual linguistic features become unknown
- Speech quality may deteriorate compared to standard HMM-TTS

[Baumann et al., 2014]

Analysis the impact of potentially missing features on the quality of the estimated prosody.

- Investigation only base on word-by-word synthesis extension.
-> German, English

[Pouget et al., 2015]

HMM training strategy for ITTS

- French
-> Word-by-words synthesis (potentially with a delay of one word)

Research Objectives

Most investigations focus only on German, English, French

- Focus on Word-by-word synthesis and its improvement.
- ITTS for tonal language has not been studied yet.

Example of tonal language → Japanese

- Major linguistic features: phoneme, accent phrase, and breath group (several accent phrases)
- Only part-of-speech (POS) tag is used as word-level information
- Tonal feature is important
- It may use longer unit than word as linguistic features and synthesis unit.

Necessary to investigate the effect on speech quality in linguistic and temporal locality choices in tonal language

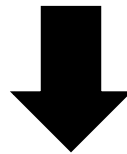
Proposed Approach

Investigates the effect on speech quality in linguistic and temporal locality choices for a Japanese ITTS system

1st

Investigation the quality of synthesized speech on various linguistic and temporal locality

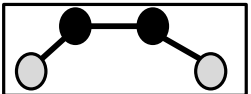
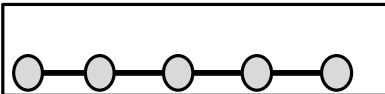
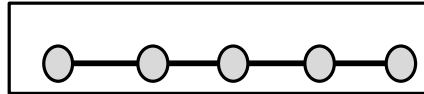
- To estimate the optimum synthesis unit



2nd

Investigation of chunk connection

- To maintain the smoothness between chunks
- To produce more natural speech

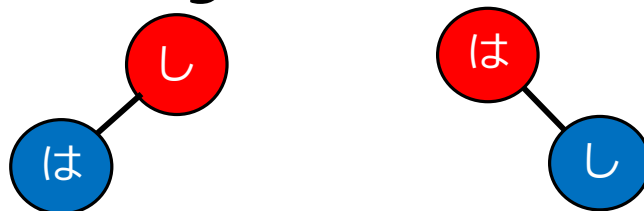
Sentence	あらゆるげんじつを, ... 曲げたのだ				
Breath group	Breath group 1				
Accent phrase			...		High pitch Low pitch
Word POS*	POS1	POS2	POS3	...	POS10 POS11 POS12 POS13
Phoneme	a ra yu ru	ge N ji ts u	wo	...	ma ge ta no da

*Part-Of-Speech tag

Example of different accent type

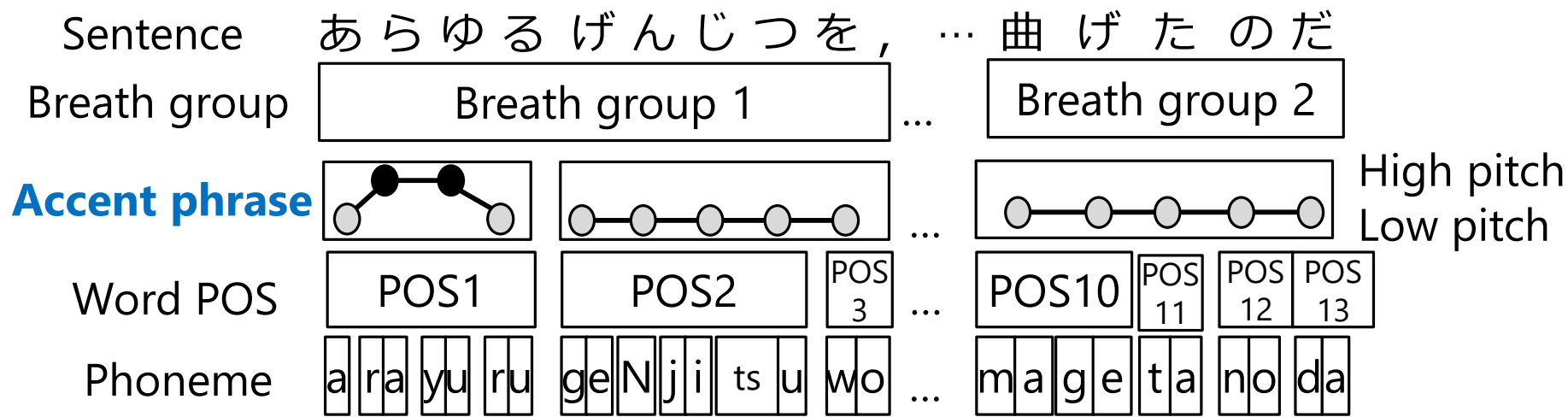
はし (ha-shi)
(Bridge) (chopsticks)

High pitch
Low pitch

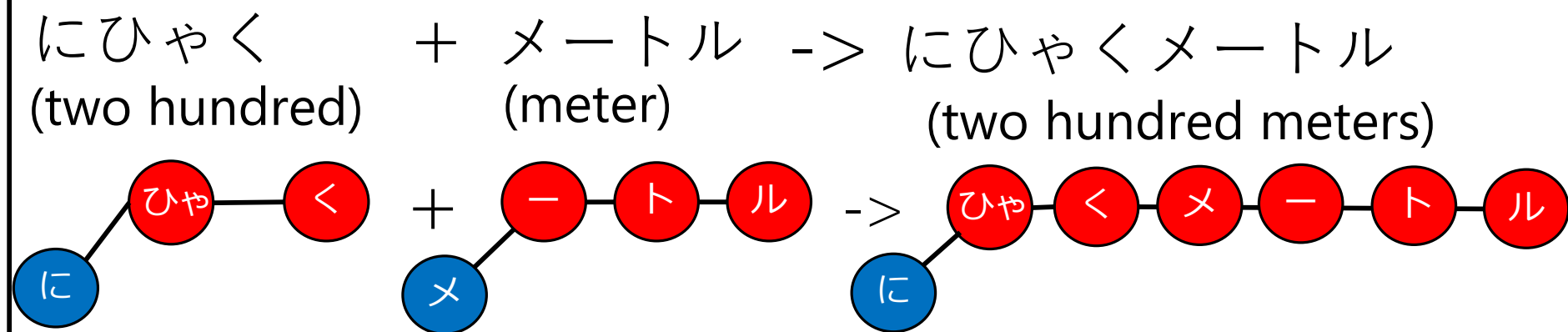


- Word has each accent type.
- Different accent type, Different meanings.

Japanese Linguistic Information

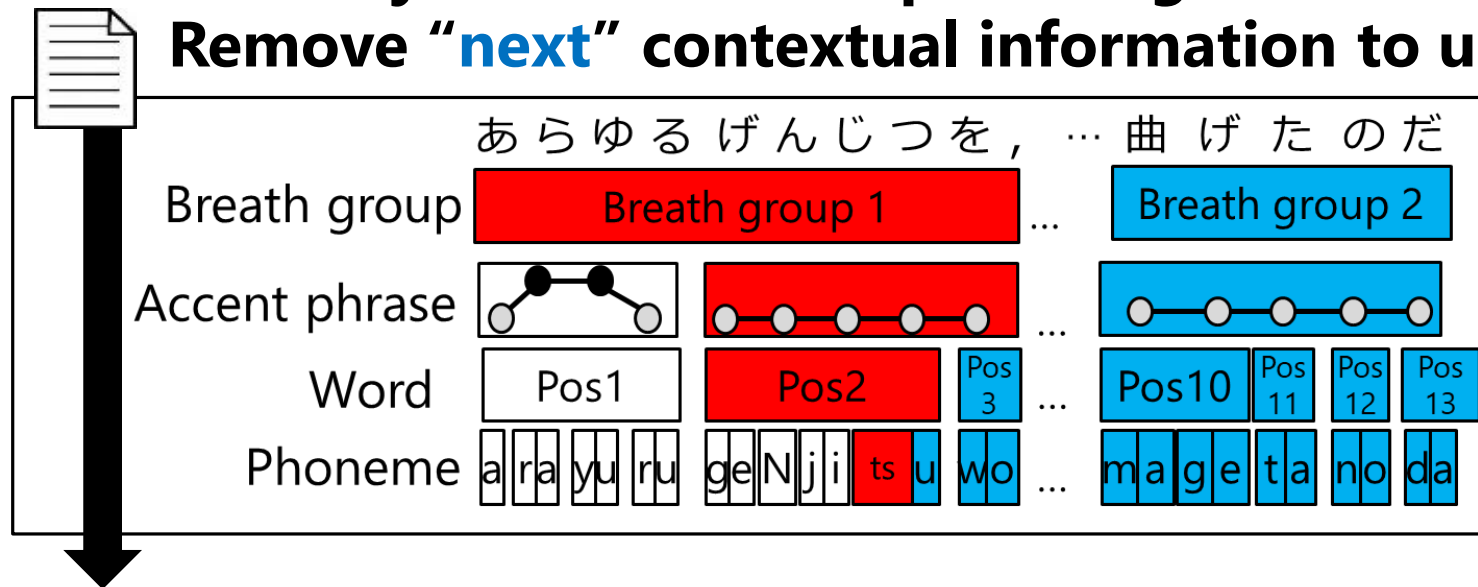


Example of connecting accent type (accent phrase)

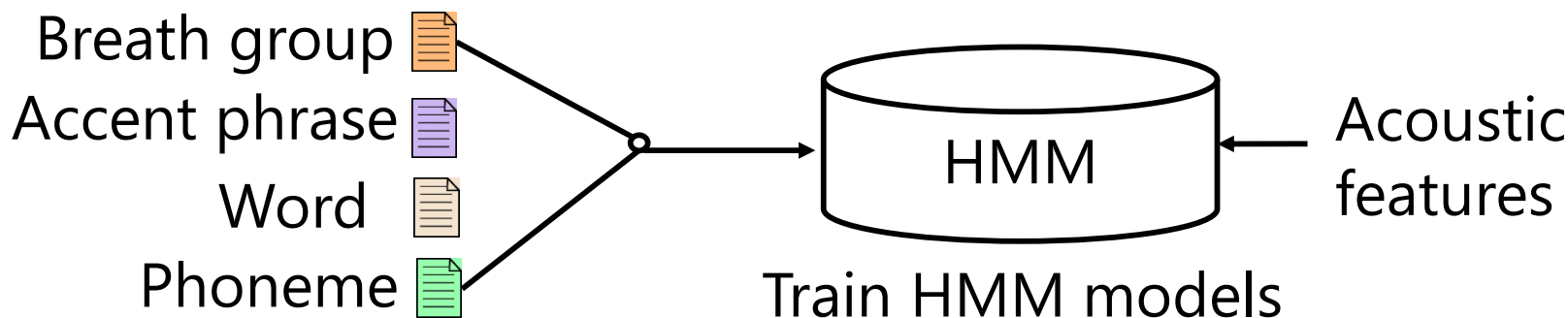


Linguistic Features Locality

Use only **“current”** and **“past”** linguistic information
Remove **“next”** contextual information to unknown



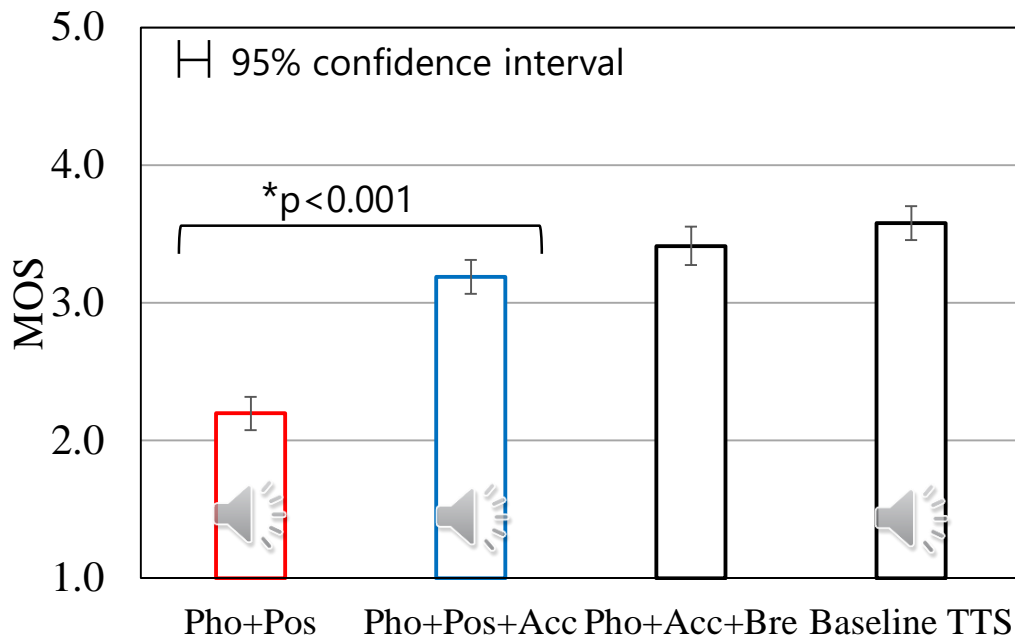
Investigate several possible linguistic locality choices



Experimental Set-up

Setting	Details
Dataset	ATR 503 phonetically balanced sentences (HTS) (Training 450 sentences • Test 53 sentences)
Acoustic feat.	F0 (1dim), Mel cepstrum coefficient (39 dim), Aperiodic component (5 dim), and dynamic feature
Analysis method	STRAIGHT [Kawahara et al. 1999]
Linguistic Information	Phoneme identity, Part-of-speech tag, relative pitch position, # of accent phrases and moras, # of breath group, position of breath group, etc.
Objective evaluation	Mean opinion score of naturalness (1: very bad, 2:bad, 3:normal, 4:good, 5:Very good) 16 Japanese native speakers, each 15 samples

Subjective Evaluation



Pho: phoneme
Pos: word POS tag
Acc: accent phrase
Bre: breath group

Linguistic Phoneme+POS

The naturalness close to bad (MOS=2)

Linguistic Phoneme+POS+accent phrase

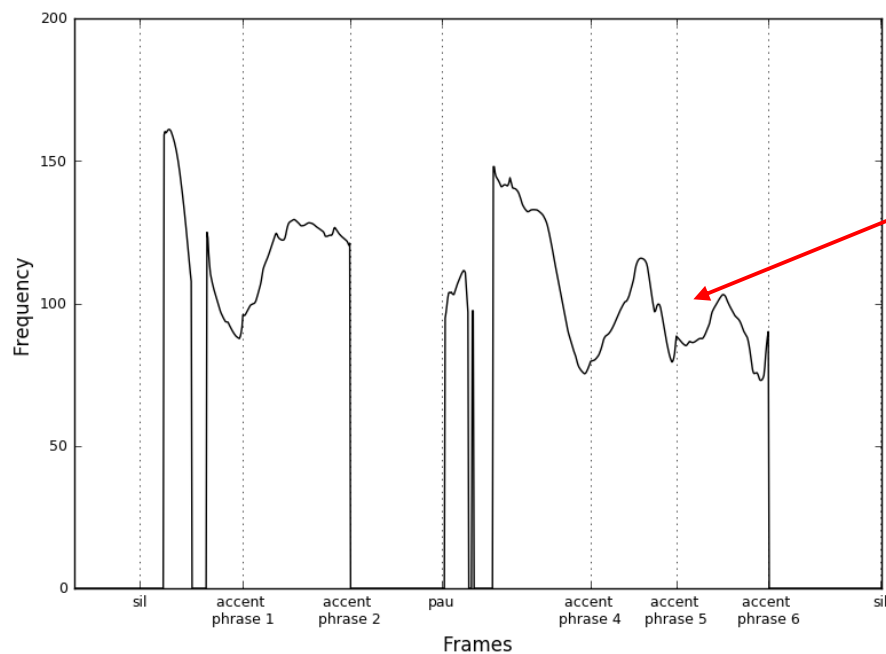
The naturalness close to normal (MOS=3)

**Based on the results,
we decided to accent phrase as a synthesis unit**

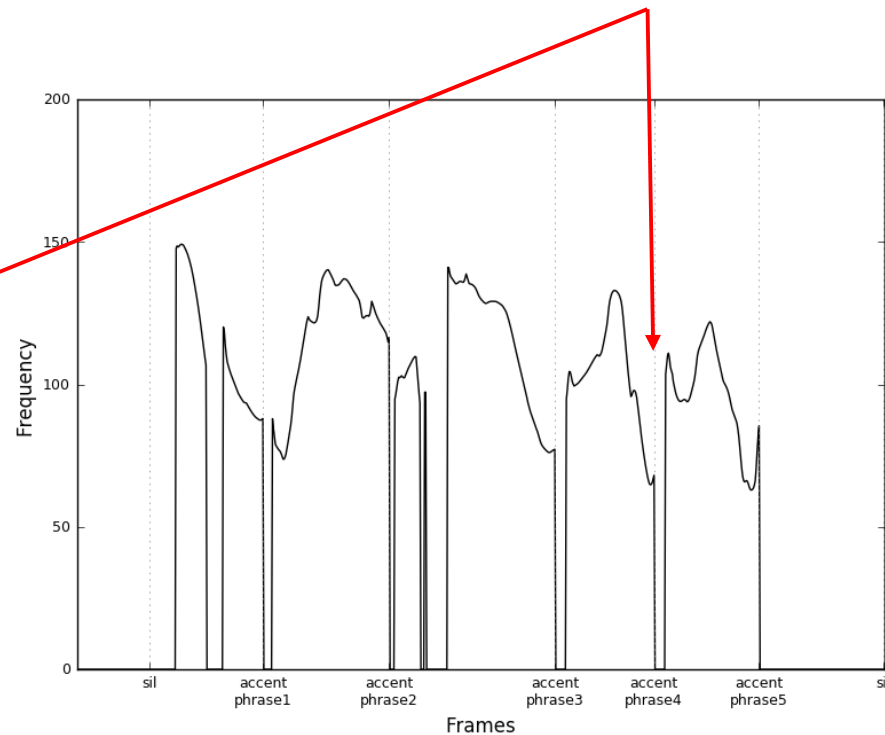
F0 Sequence Between Accent Phrases

Problem

- Prosody breaks occurs when using only current accent phrase units



Sentence unit synthesis



Accent phrase unit synthesis

Chunk connection approach for smoothing [Timo et al., 2012]

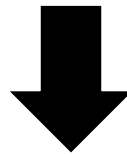
Proposed Approach

Investigates the effect on speech quality in linguistic and temporal locality choices for a Japanese ITTS system

1st

Investigation the quality of synthesized speech on various linguistic and temporal locality

- To estimate the optimum synthesis unit



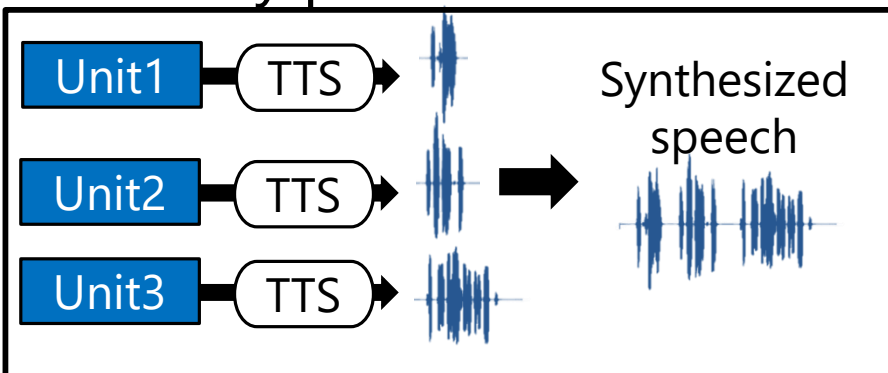
2nd

Investigation of chunk connection

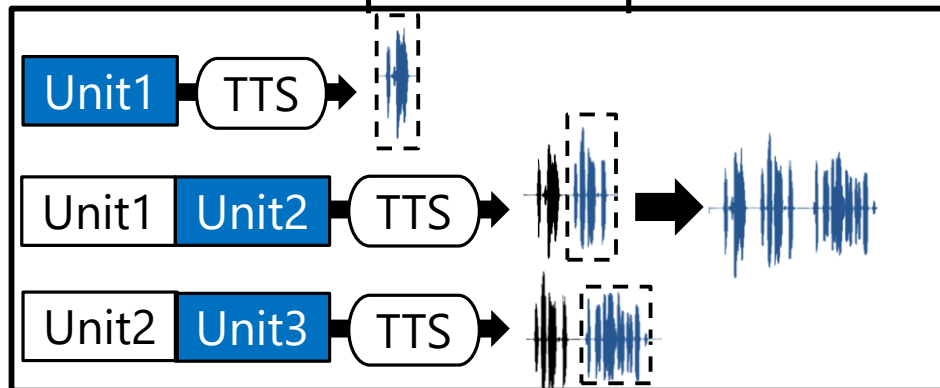
- To maintain the smoothness between chunks
- To produce more natural speech

Method of Chunk Connection

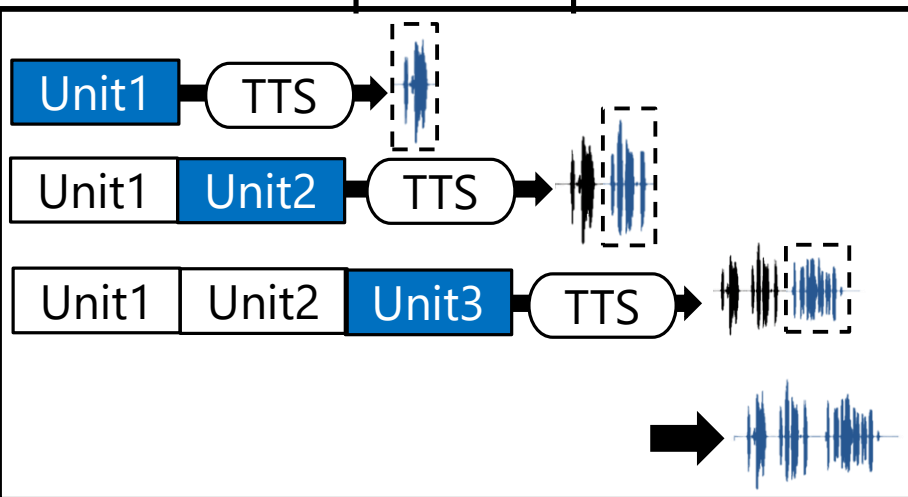
Phrase by phrase



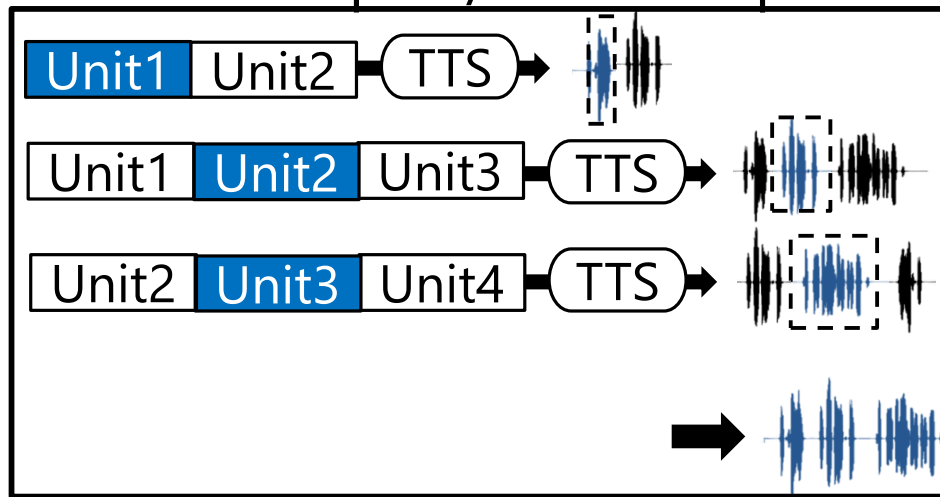
Phrase with past one phrase



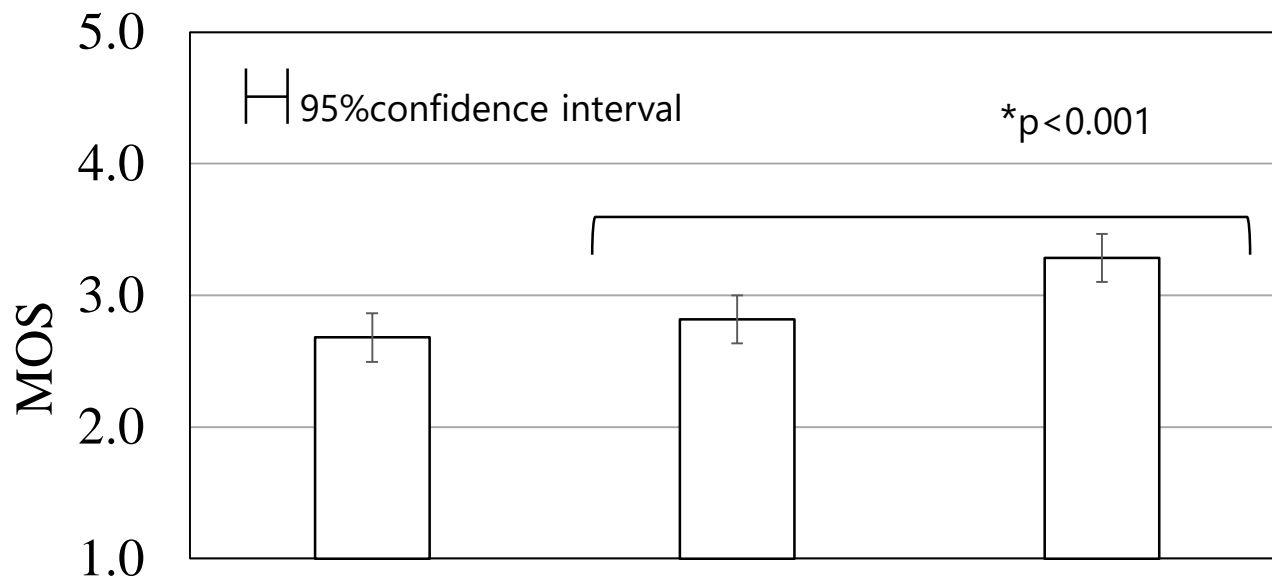
Phrase with past all phrases



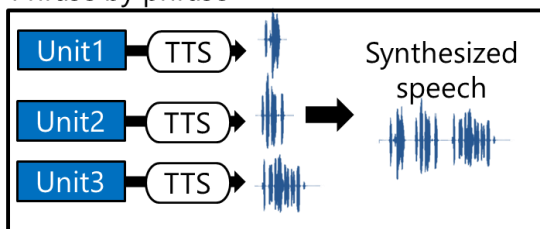
Phrase with past/next one phrase



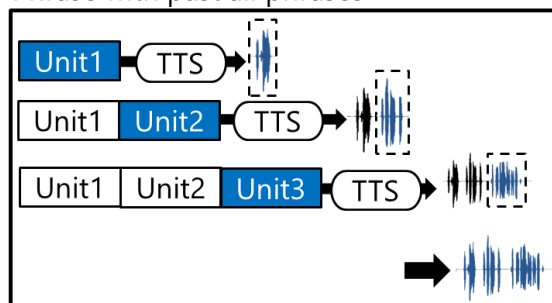
Subjective Evaluation



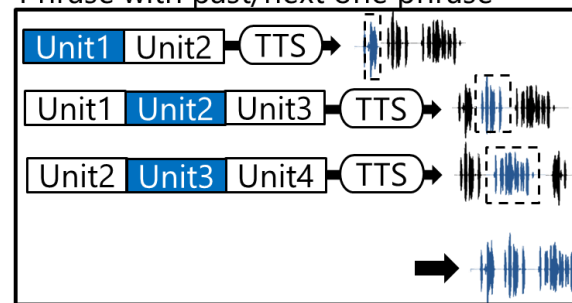
Phrase by phrase



Phrase with past all phrases

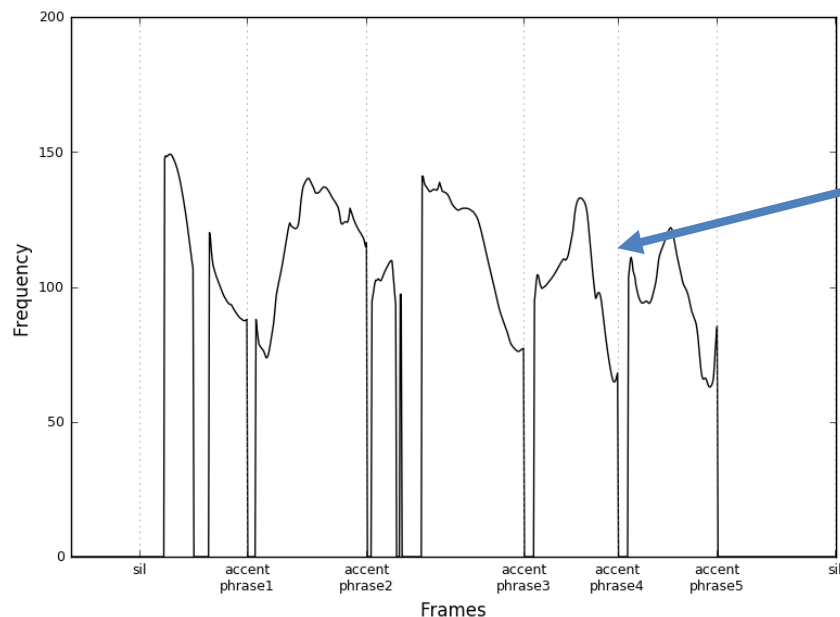


Phrase with past/next one phrase

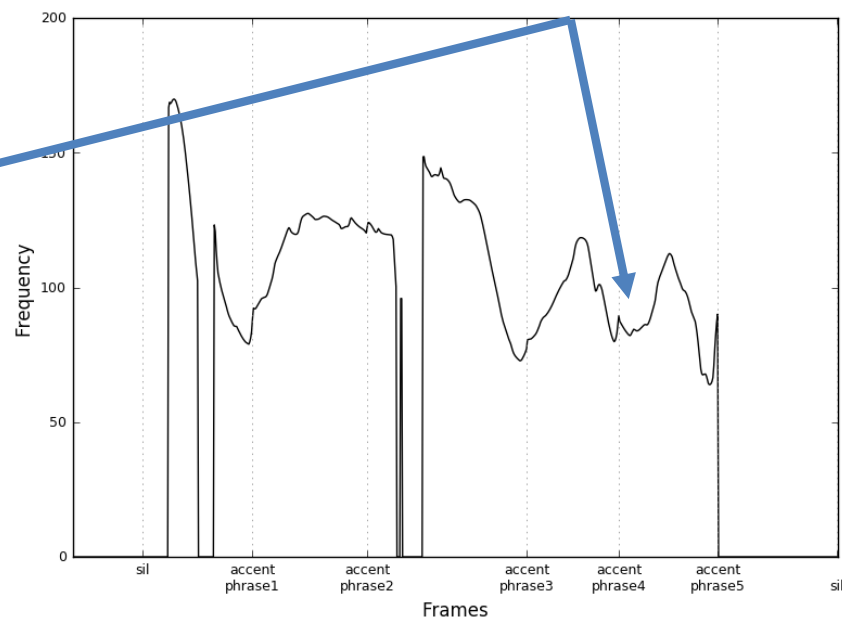
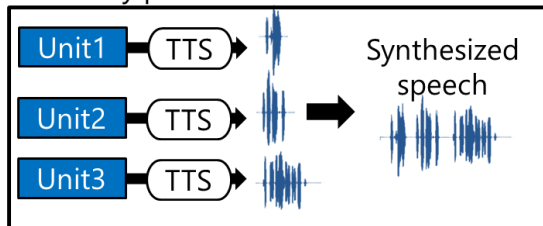


Connecting one past and next accent phrase chunks
could improve naturalness

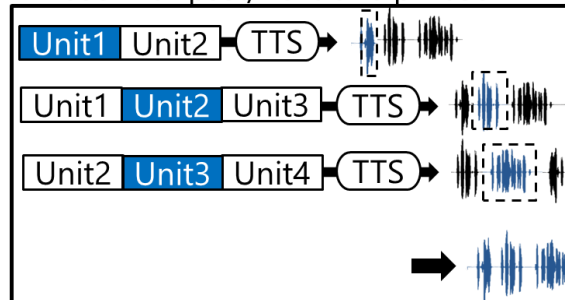
F0 Sequence Between Accent Phrases



Phrase by phrase



Phrase with past/next one phrase



F0 sequences can be smoother by waiting for one more chunk before starting the synthesis process

Conclusion

Conclusion

Japanese language as tonal language

- Accent phrase (tonal information) unit required as synthesis unit
- It's effective to wait for one accent phrase for improving quality

Future work

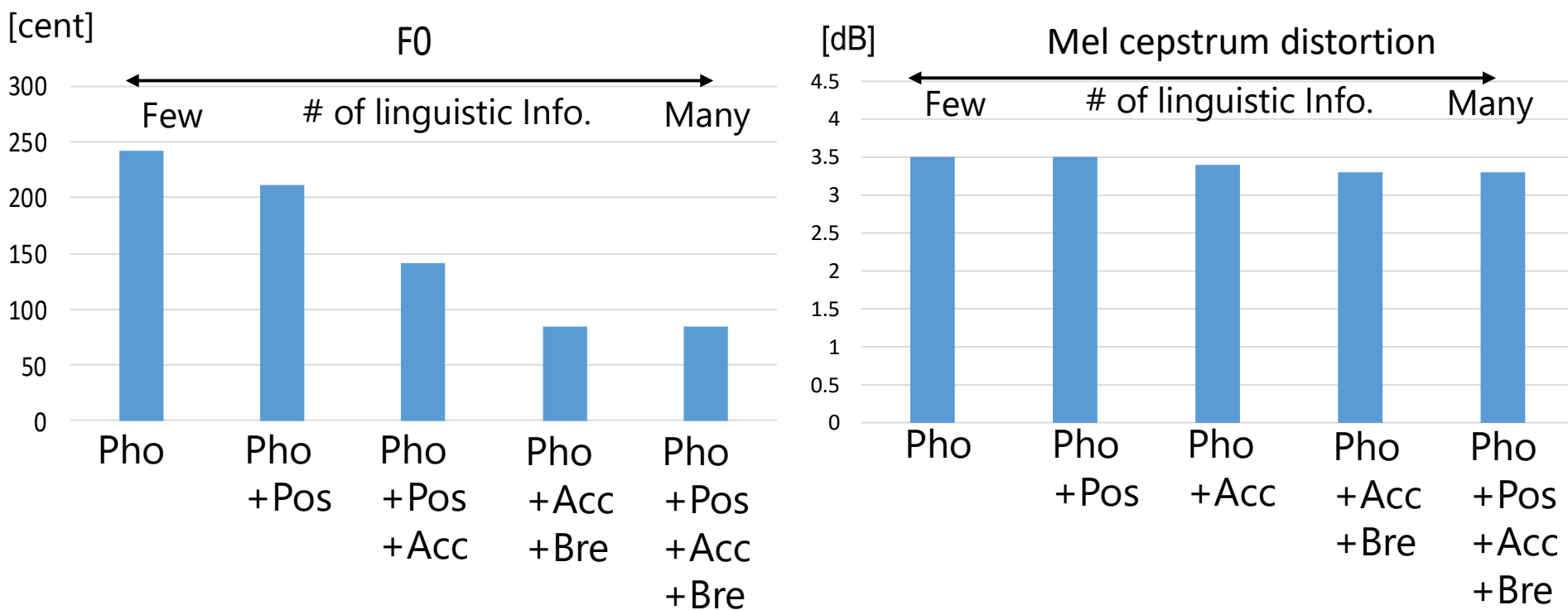
- Implementation of full-pledge Japanese ITTS System
- Experiment of other tonal language (i.e., Chinese..)
- DNN based incremental TTS

End of slide

Thank you
Q & A

Subjective result of Investigation of linguistic Information Locality

The result tends to improve when linguistic information is added.



Pho : Linguistic set of phoneme
Pos : Linguistic set of word

Acc : Linguistic set of accent phrase
Bre : Linguistic set of breath group