A Method for Translation of Paralinguistic Information

Takatomo Kano, Sakriani Sakti, Shinnosuke Takamichi, Graham Neubig,

Tomoki Toda, Satoshi Nakamura

Nara Institute of Science and Technology,Japan

Nakamura SRG
Augmented Human Communication (AHC) Laboratory

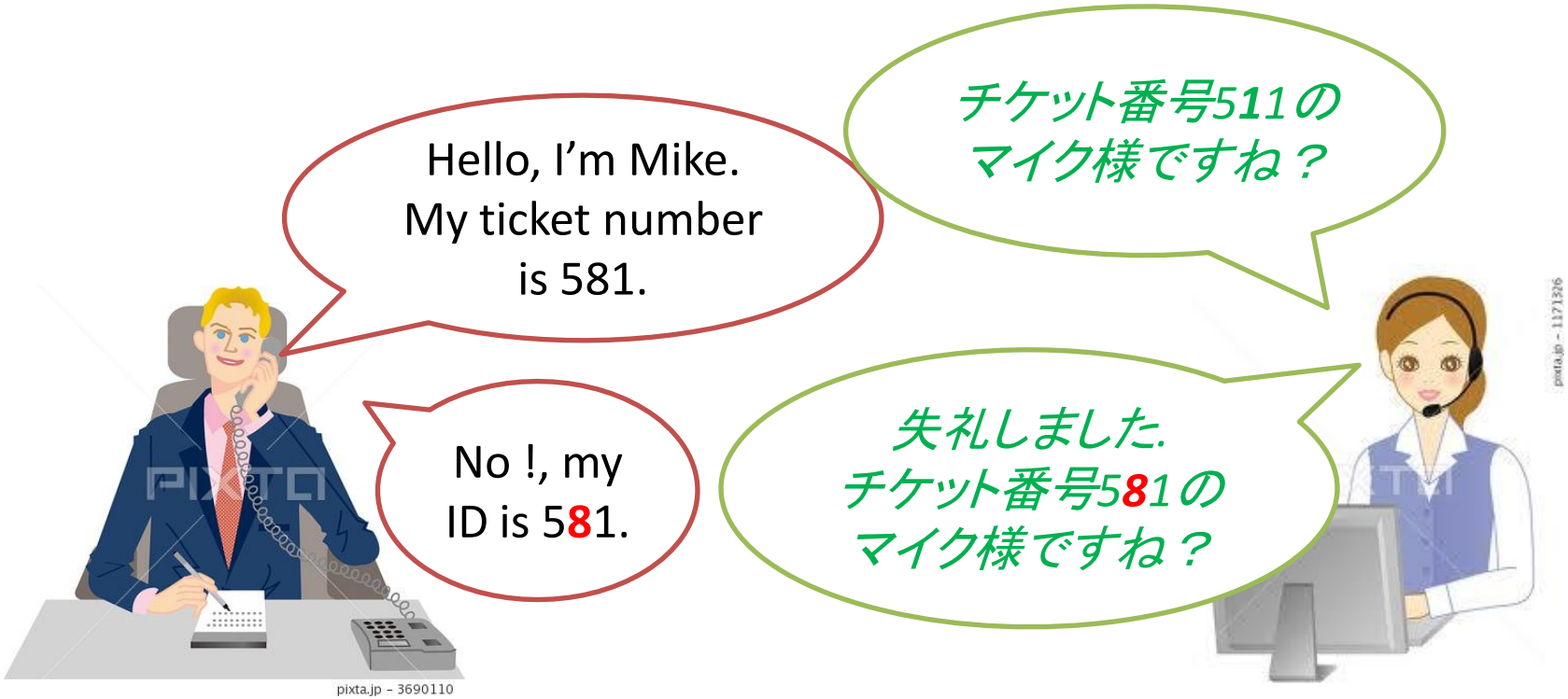
NAIST

Background

Paralinguistic information is important!!

Example : Digit translation

How paralinguistic information affects communication.

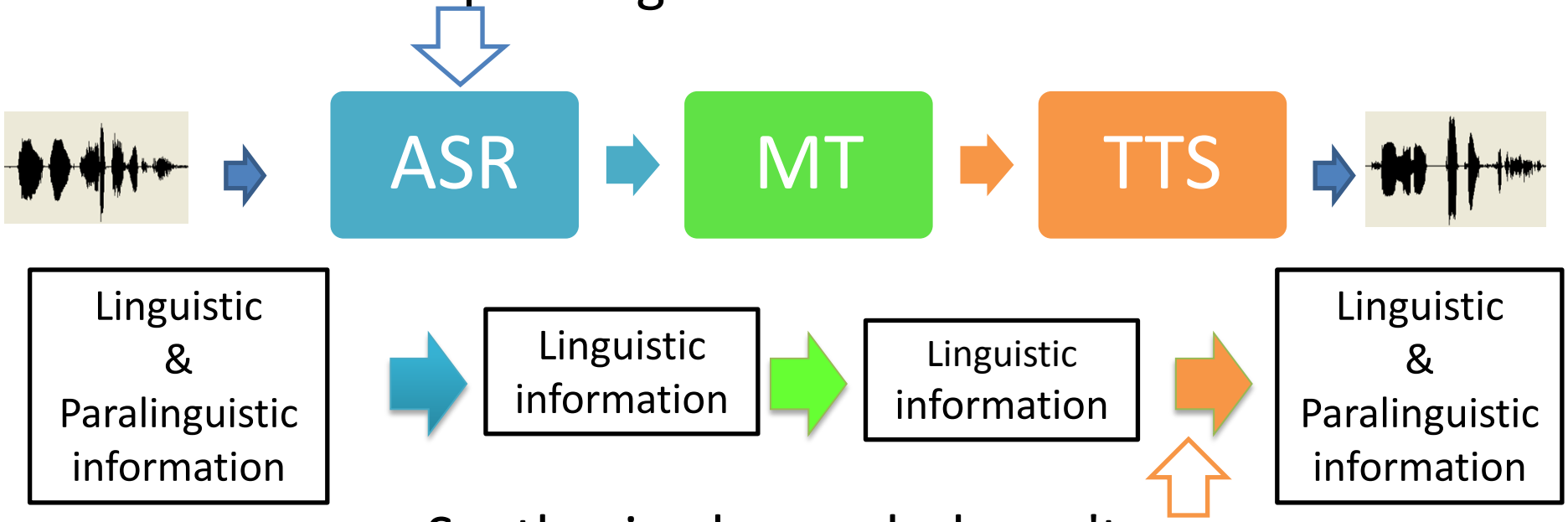


"Five **Eight** One"
(emphasis)

*"Go **Hachi** Go"*
(emphasis)

Problem in traditional method

Lose all paralinguistic information in ASR

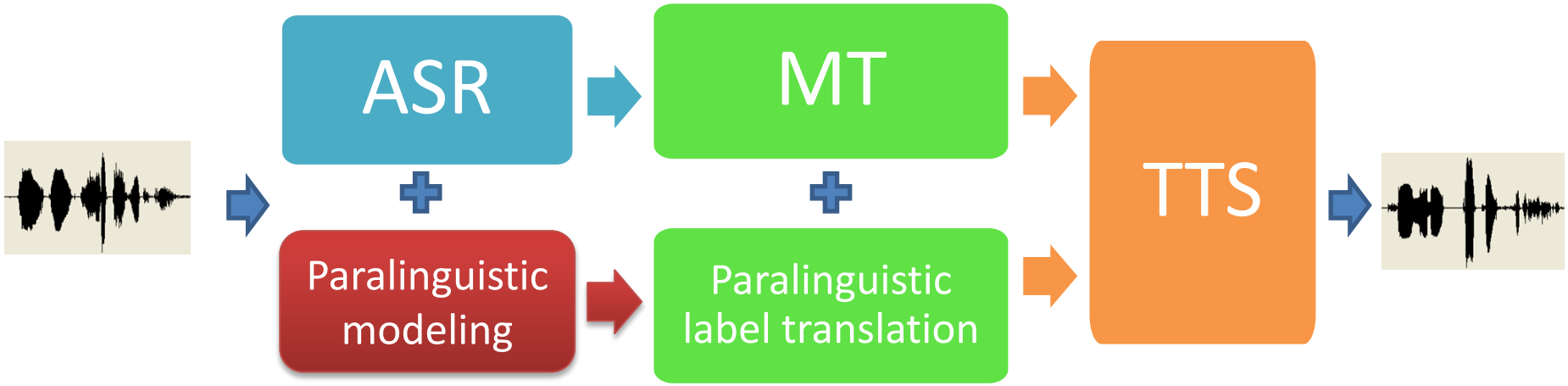


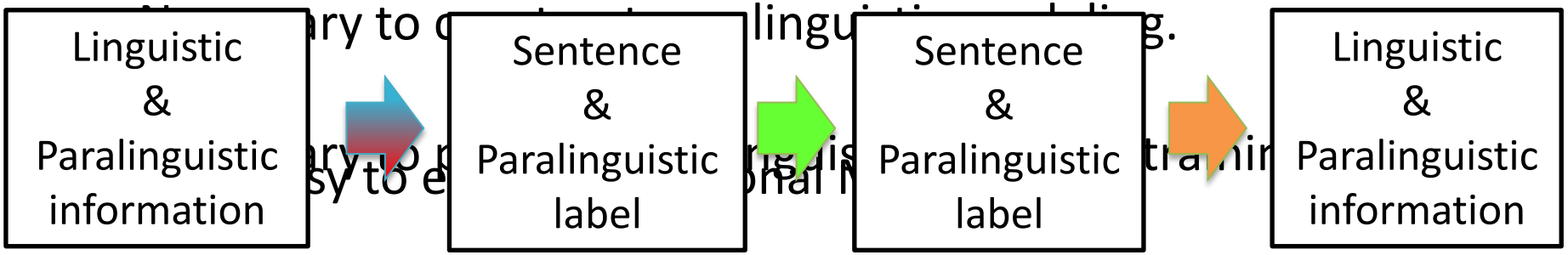
Synthesized speech doesn't reflect input paralinguistic information in TTS

Paralinguistic information is not translated!

Existing method with paralinguistic translation

(Agüero, P. D. et al,2006)
(V. Kumar, et al,2011)

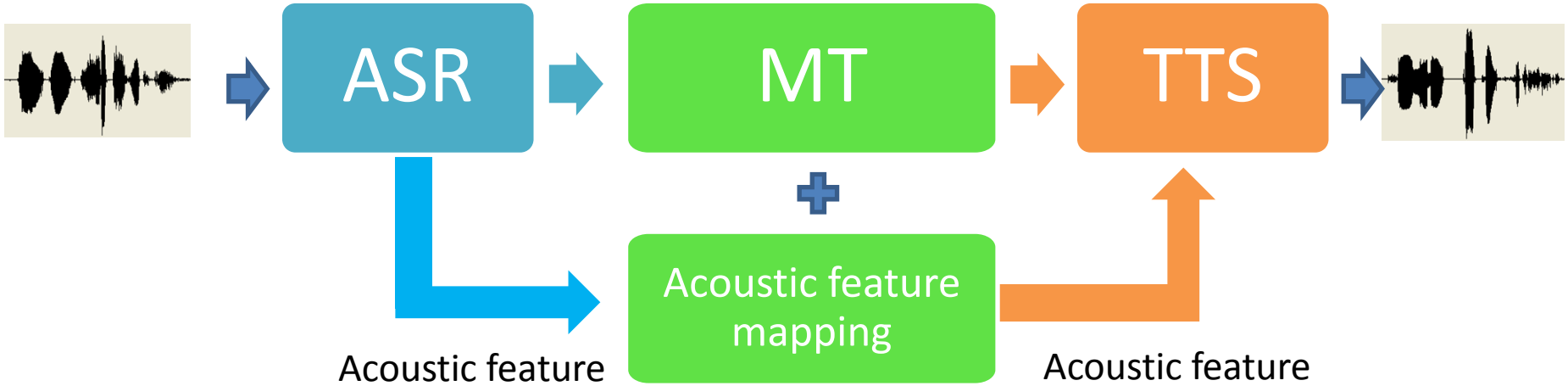




Difficult to represent sensitive differences in the same paralinguistic class or label.

Proposed method

Proposed method



Paralinguistic modeling not necessary

Linguistic & Paralinguistic information → sentence & Acoustic feature → sentence & Acoustic feature → Linguistic & Paralinguistic information

Paralinguistic labels for training data not necessary

Possible to represent sensitive differences in the similar prosody

ASR module

Traditional method

Transcribe speech to text

Text to text translation

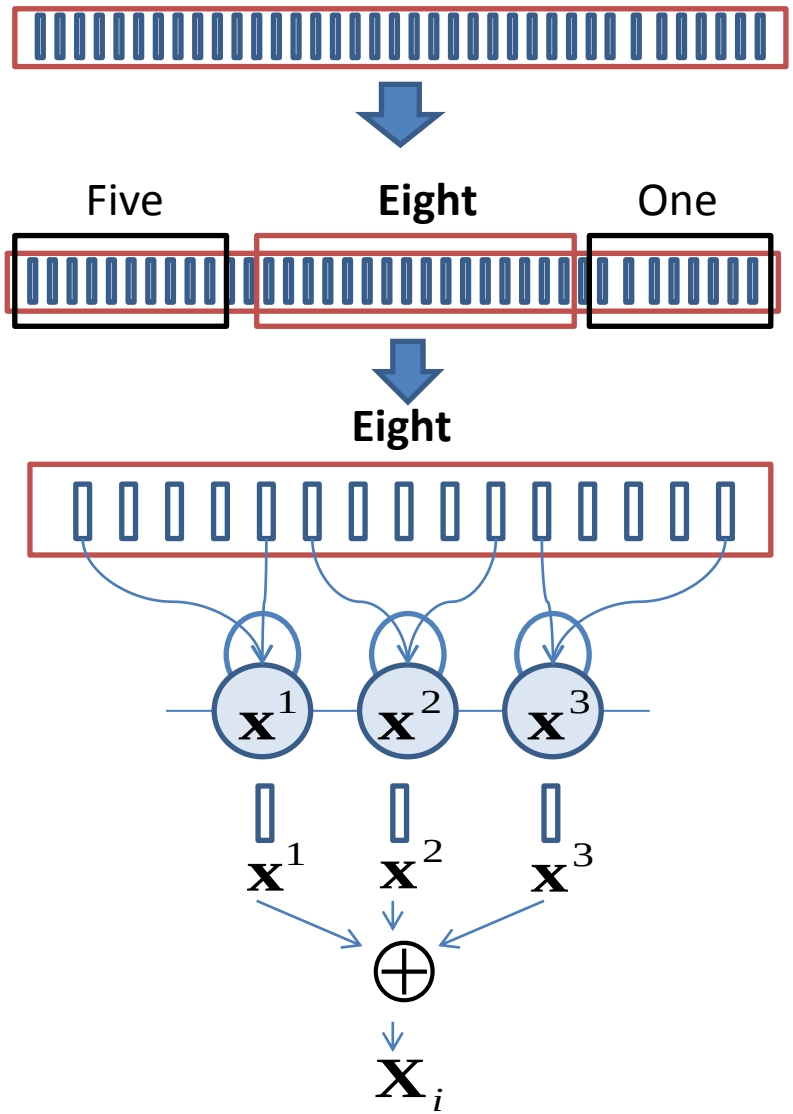
Synthesize speech from text

ASR → MT → TTS

Transcript speech to text .
Extract acoustic features in a word level

Proposed method

Acoustic features mapping



Construct word-based acoustic models for speech recognition

↓

Amount of time in each state
→ state duration

↓

Average power of all frames in each state
→ state power

↓

Combine these into a word acoustic feature super vector

MT module

Traditional method

Transcribe speech to text

Text to text translation

Synthesize speech from text

ASR → MT → TTS

Transcript speech to text .
Extract acoustic features in word level

Text to text.
Map acoustic feature in word level

Proposed method

Acoustic feature mapping

Learn the relationship between English and Japanese acoustic features

"Eight"

Super vector of "**EIGHT**" $\mathbf{x}_i$

$$P(\mathbf{y}_i|\mathbf{x}_i) = N(\mathbf{y}_i; \mathbf{W}_{\mathbf{e},\mathbf{j}}\mathbf{x}'_i, S)$$

Input speech

$$\mathbf{x}'_i = [1\mathbf{x}_i^{\mathrm{T}}]^{\mathrm{T}}$$

Super vector of "***HACHI***" $\mathbf{y}_i$

Reference acoustic features

Regression matrix **W** is optimized by

$$\hat{\mathbf{W}}_{\mathbf{e},\mathbf{j}} = \arg\min_{\mathbf{W}_{\mathbf{e}_i,\mathbf{j}_i}} \sum_{n=1}^{N} \left\| \mathbf{y}_n^* - \mathbf{y}_n \right\|^2 + \alpha \left\| \mathbf{W}_{\mathbf{e}_i,\mathbf{j}_i} \right\|^2$$

TTS module

Traditional method

Transcribe speech to text

Text to text translation

Synthesize speech from text

ASR → MT → TTS

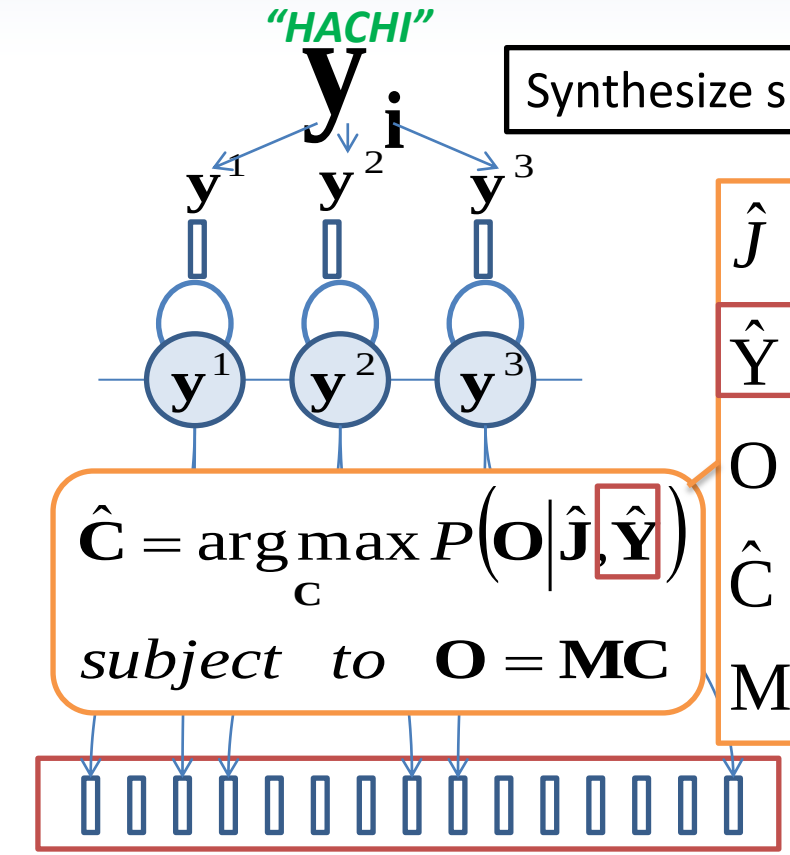
Transcript speech to text . **Extract acoustic features in word level**

Text to text. **Map acoustic features in word level**

Synthesize speech from **text and acoustic features**

Proposed method

Speech synthesis



Synthesize speech based on translated acoustic features Y

$$\hat{\mathbf{C}} = \arg\max_{\mathbf{C}} P\left(\mathbf{O} \middle| \hat{\mathbf{J}}, \boxed{\hat{\mathbf{Y}}}\right)$$

$$\textit{subject to} \quad \mathbf{O} = \mathbf{MC}$$

$\hat{J}$ is a target word label

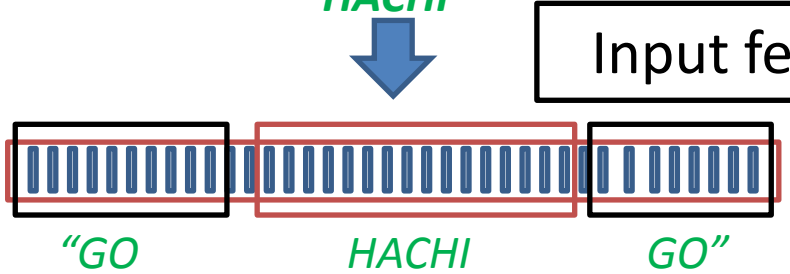
$\hat{Y}$ is a translated acoustic feature vector

O is the synthesized features

$\hat{C}$ is the optimized parameters

M is a matrix that maps C to O

"HACHI"



Input features are reflected in output speech!

Experiment

- English to Japanese digit speech translation task
- Recorded parallel corpus of emphasized speech

Corpus	
Vocabulary size	11 words
Recorded utterances	455for train set 55 for test set
Speaker	1 male
TIDigit:AURORA2	8440 (TIDigit:AURORA2)
Speaker	55 male and 55 female

ASR and TTS Settings

ASR	
Training utterances	TIDigit:AURORA2
HMM states	16

TTS	
Training utterances	Recorded utterances
HMM states	16

MT	
Training utterances	Recorded utterances
Feature	Duration. Power, ΔPower, , ΔΔPower
Regularization term	10

Experiment

- Evaluation
  - **Automatic evaluation** of paralinguistic information in Root Mean Squared Error (RMSE)
  - **Manual evaluation** of emphasis prediction rate, emphasis subjective strength evaluation by 3 subjects

- Systems
  - 【Baseline】Traditional lexical translation
  - 【+dur】Paralinguistic translation of **duration**
  - 【+dur&pow】Paralinguistic translation of **duration** and **power**

Automatic evaluation

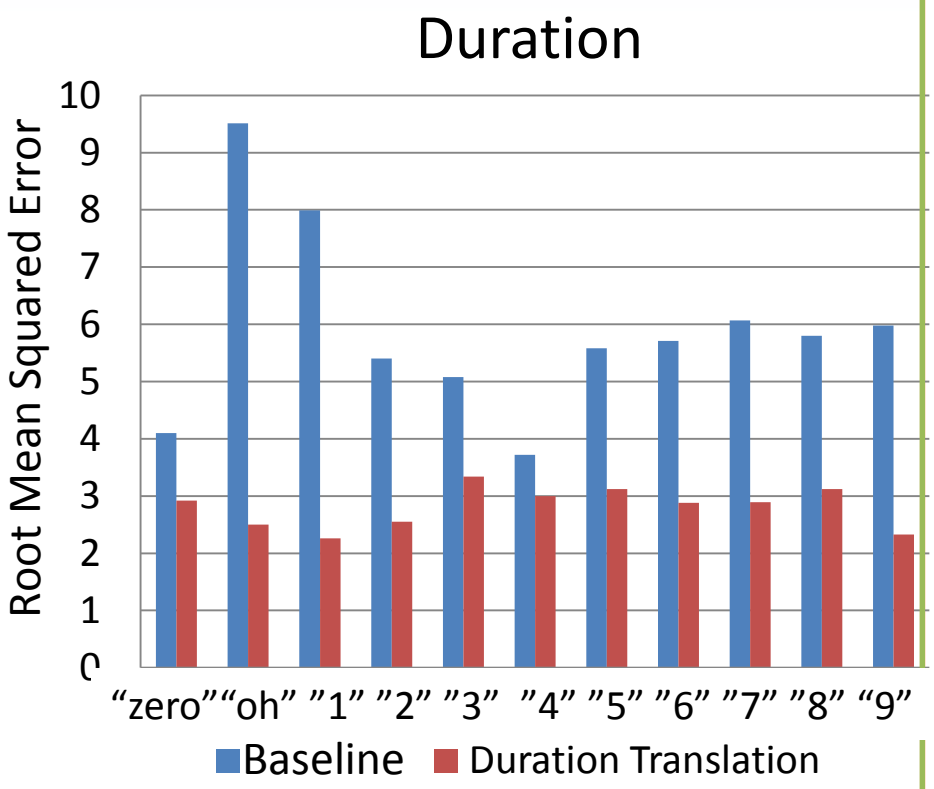


Figure 1:RMSE between the reference target duration and the system output

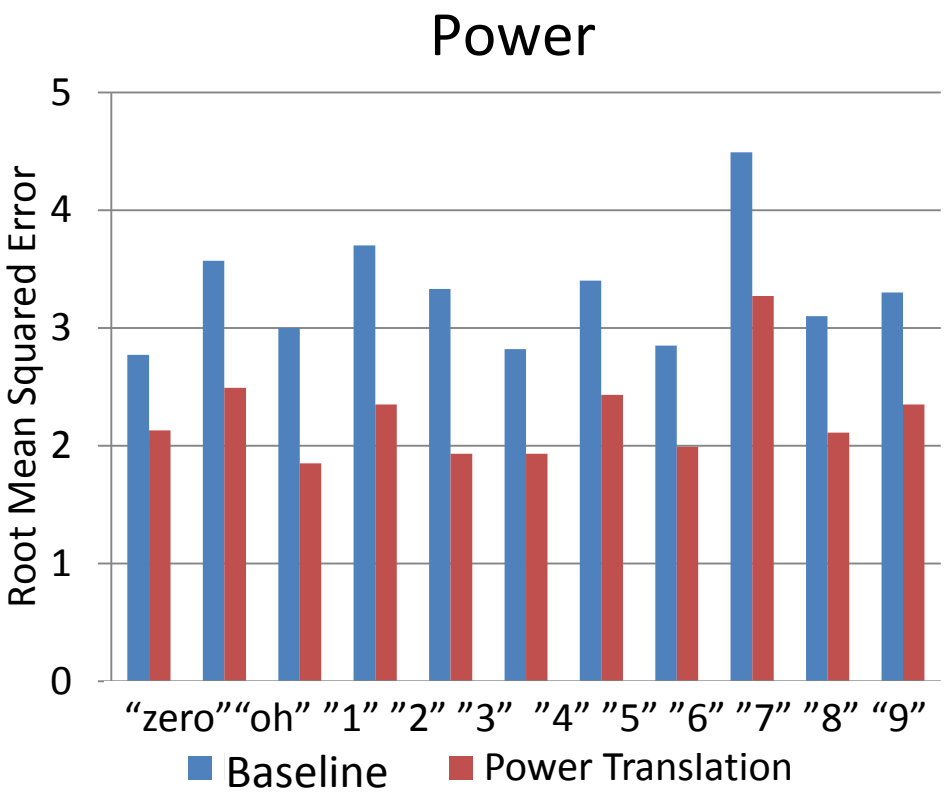
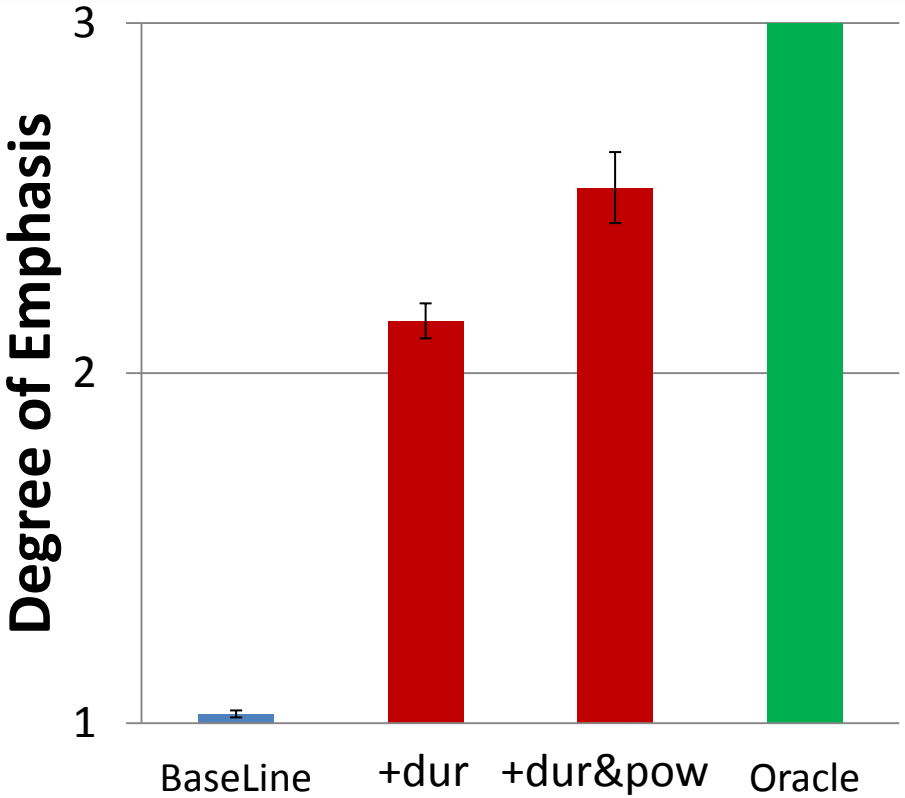
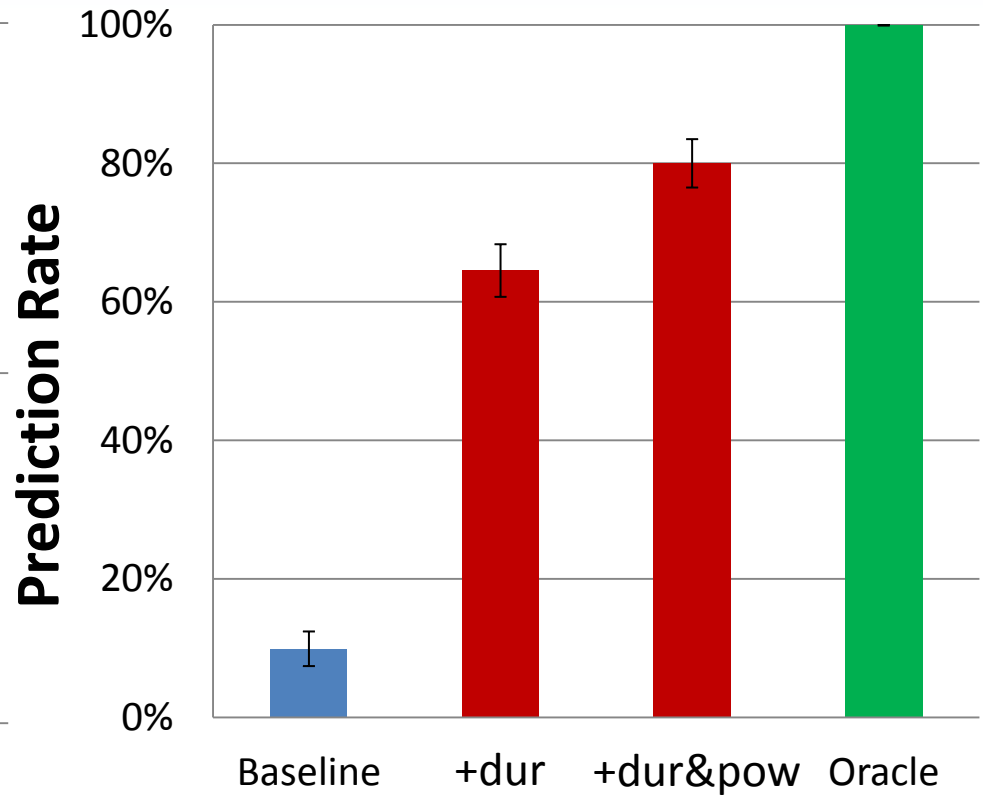


Figure 2: RMSE between the reference target power and the system output

Manual evaluation





115

320

Input | Oracle | Baseline | +dur | +dur&pow

1: Not emphasized
2: Slightly emphasized
3: Emphasized

Conclusion

- We propose a speech translation method using direct acoustic feature mapping to translate paralinguistic information

- This proposed method outperforms traditional lexical speech translation system in represent emphasis.

Future works

- Expand this method towards large vocabulary speech translation tasks

Thank you very much!
If you have some question
please ask me **slowly**.