Modeling of Paralinguistic Speech Attributes for Intelligent Speech Interaction: Emotion and Emphasis as Example

Zhiyong Wu (吴志勇)

清华大学人机语音交互实验室

清华大学-香港中文大学媒体科学、技术与系统联合研究中心

Acknowledgements

Tsinghua University

- Yishuang NING (宁义双)
- Xixin WU (吴锡欣)
- Runnan LI (李润楠)
- Mu WANG (王木)
- Liangqi LIU (刘良琪)
- Jiankun HU (户建坤)
- Dongyang DAI (代东洋)
- Huirong HUANG (黄晖榕)
- Xiong CAI (蔡雄)
- Xiang LI (李翔)
- Changhe SONG (宋长河)
- …

- Prof. Lianhong CAI (蔡莲红教授)
- Prof. Helen MENG (蒙美玲教授)
- Prof. Jia JIA (贾珈教授)
- Prof. Mingxing XU (徐明星教授)

- The Chinese University of Hong Kong
- Tencent AI Lab
- Microsoft Research Asia
- Beijing Sogou Technology Development Co., Ltd.

- National Natural Science Foundation of China (NSFC): 62076144, 61375027, 61433018
- NSFC-RGC (Research Grant Council of Hong Kong): 61531166002, N_CUHK404/15
- National Social Science Foundation of China (NSSF): 13&ZD189

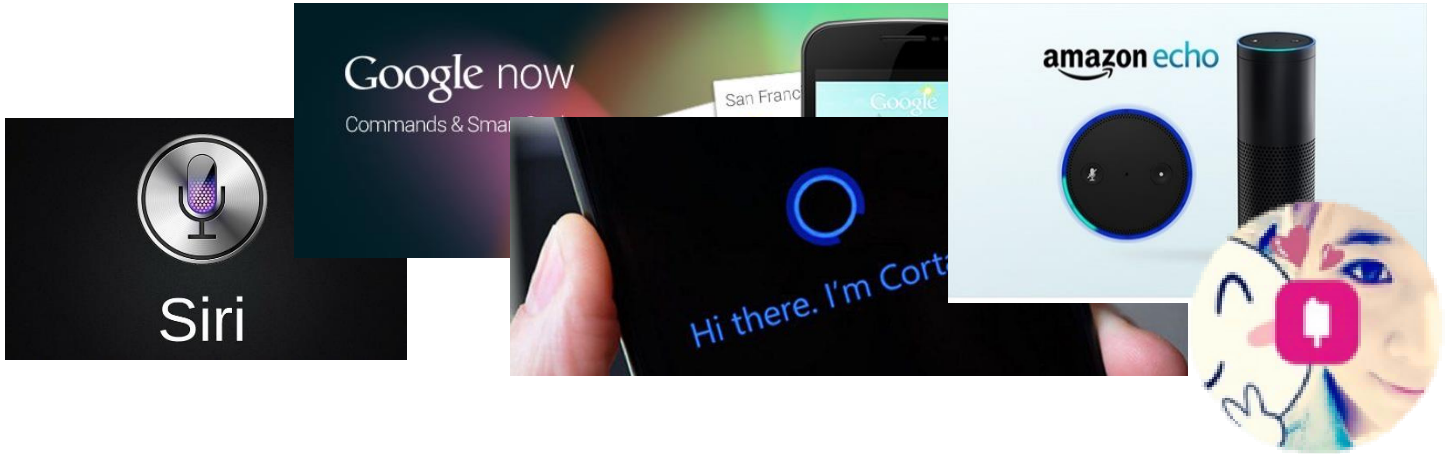
Outline

- Background & Motivation
    - Intelligent speech interaction
    - Paralinguistic attributes for intelligent speech interaction
- Modeling of Paralinguistic Speech Attributes
    - Speech Emotion
    - Speech Emphasis
- Conclusions & Further Work

Background

Intelligent Speech Interaction

- Speech interaction systems have been widely used
  - Apple Siri
  - Google Assistant
  - Microsoft Cortana
  - Xiaoice
  - Amazon Alexa . . .



- A well-experienced speech interaction system must be able to:
  - Understand user's implicit intention accurately
  - Provide speech responses with accurate semantic, high naturalness and human-like expressiveness

Motivation

Intelligent Speech Interaction

- Conventional speech interaction systems

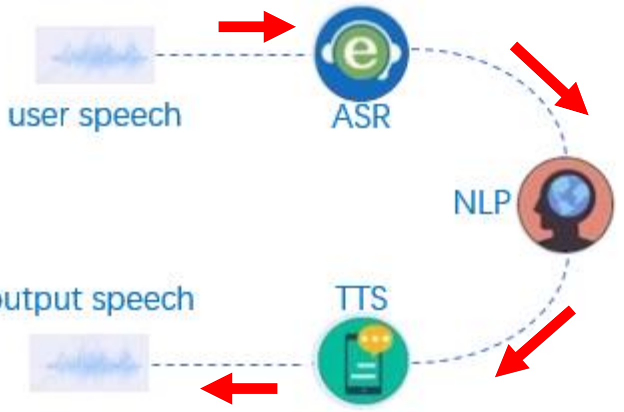


Fig1. Response flow: user speech -> ASR -> NLP -> TTS -> output speech

Motivation

清華大學 Tsinghua University

Intelligent Speech Interaction

- Conventional speech interaction systems

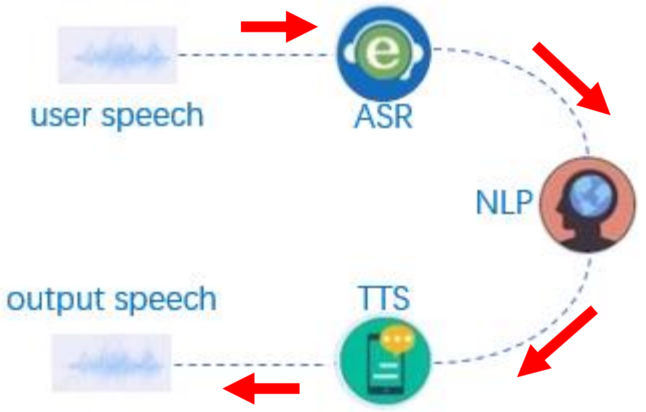


Fig1. Response flow: user speech -> ASR -> NLP -> TTS -> output speech

- Problems
  - Understanding of user intention is not accurate enough
    - At the stage of ASR, the additional intention conveyed by paralinguistic attributes in user speech is completely ignored
  - The output speech is neutral and lacks appropriate expressiveness
    - At the stage of TTS, only the naturalness and quality of speech is considered well, while the flexible control of paralinguistic attributes for speech expressiveness is usually not considered

Motivation

Tsinghua University

Intelligent Speech Interaction

- Conventional speech interaction systems

Fig1. Response flow: user speech -> ASR -> NLP -> TTS -> output speech

- Towards more **intelligent speech interaction**:
    - For more accurate understanding of user intention, the **analysis for the paralinguistic attributes** of user speech is indeed required
    - For more human-like expressiveness, the **flexibly controllable generation of paralinguistic attributes** in TTS systems need to be well considered

Motivation

Paralinguistic Attributes for Intelligent Speech Interaction

- Two main paralinguistic attributes for conveying speech intention
  - **The emotion of the speech**

neutral　　sad　　happy

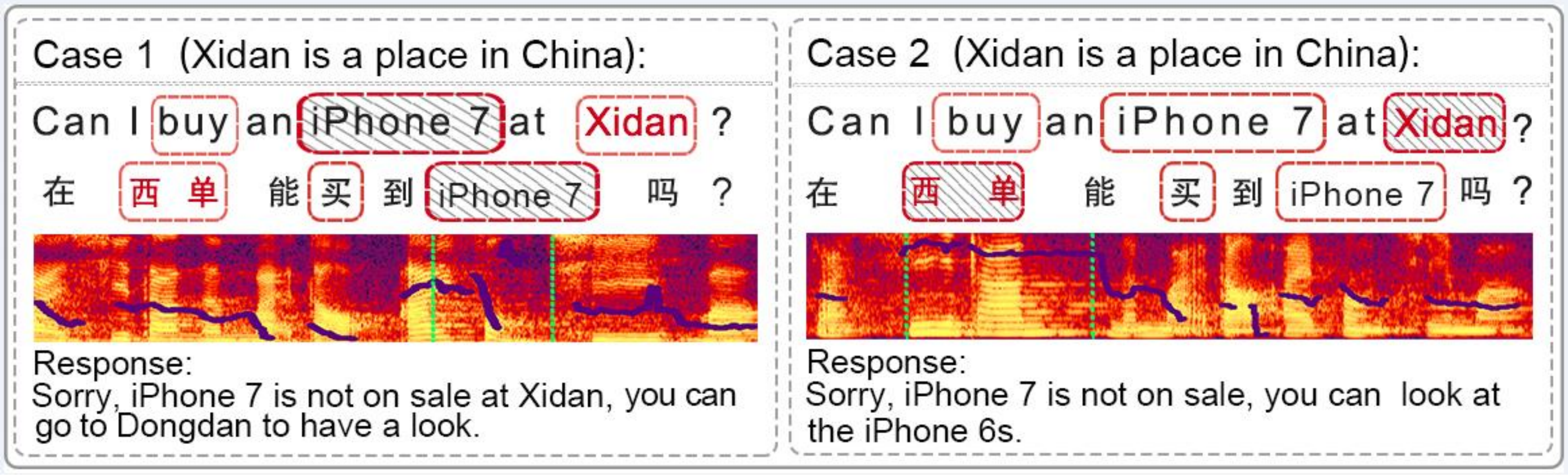
Text: 火车站就在剧院区附近
Translated text: The railway station is near the theatre

Motivation

Tsinghua University

Paralinguistic Attributes for Intelligent Speech Interaction

- Two main paralinguistic attributes for conveying speech intention
  - **The emotion of the speech**
  - **The emphasis of spoken words**



*: the words in shadowed boxes are the emphasized words

Motivation

Tsinghua University

Paralinguistic Attributes for Intelligent Speech Interaction

- Two main paralinguistic attributes for conveying speech intention
  - **The emotion of the speech**
  - **The emphasis of spoken words**
- Related speech representation learnings for intelligent speech interaction
  - For the analysis of user intention in speech
    - Speech emotion recognition
    - Speech emphasis detection
  - For the controllable generation of paralinguistic attributes of speech
    - Emotional speech synthesis
    - Emphatic speech synthesis

Framework

User Speech

Speech Emphasis Detection

Speech Emotion Recognition

User's Emotive State Changes Prediction

User Intention Understanding

Emphatic Emphasis Generation

Emotional Speech Synthesis

Intelligent Speech Interaction

Speech Emotion Recognition (SER)

Tsinghua University

Towards Discriminative Representation Learning for Speech Emotion Recognition

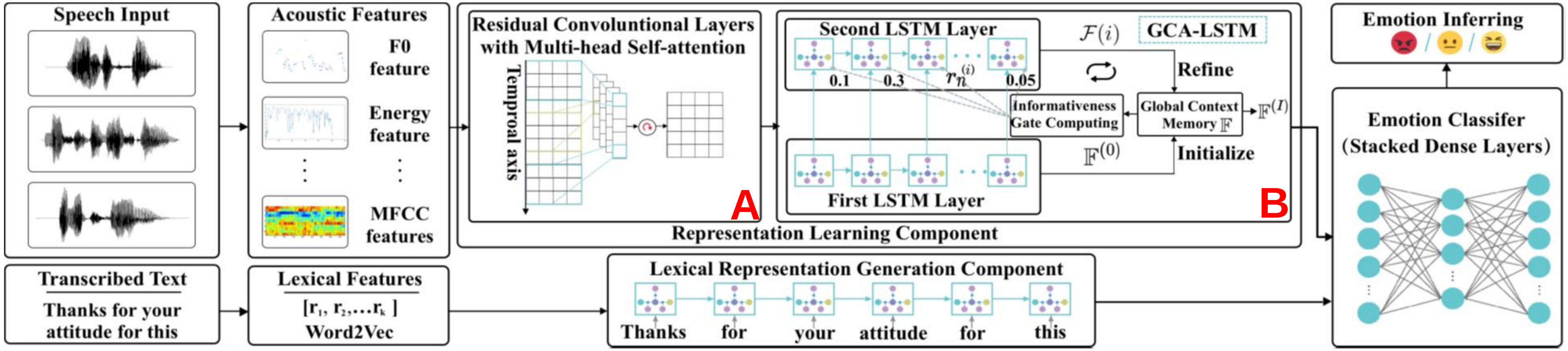
- Problem nature of SER
  - Sequence modeling problem, i.e. **temporal order information** is important
  - Context of **whole utterance** greatly contributes to the perception of emotion

- Challenge of SER
  - Deriving discriminative utterance-level representation

Li, Runnan, et al. "Towards Discriminative Representation Learning for Speech Emotion Recognition." *IJCAI*. 2019.

Speech Emotion Recognition (SER)

- Model Architecture



- Model Highlights
  - Residual CNNs extracts local patterns and preserves temporal order information
  - Multi-head self-attention builds utterance-level context dependencies
  - Global context-aware attention LSTM (GCA-LSTM) stores the global emotion-salient patterns

The capture of local and global context
+
The store of emotion-salient information

=> emotion discriminative representation

Speech Emotion Recognition (SER)

Tsinghua University

- Experiment results
  - Datasets: IEMOCAP && RID

	Method	IEMOCAP				IEMOCAP				RID		
		Input	Reported UA*	UA	F1	Input	Reported UA*	UA	F1	Input	UA	F1
[Xia and Liu, 2017]	DNN	A	60.1%	60.4%	0.597	A+L	-	72.8%	0.731	A	68.7%	0.691
[Poria *et al.*, 2016]	CNN	A	61.3%	60.7%	0.608	A+L	65.1%	69.7%	0.702	A	71.2%	0.719
[Poria *et al.*, 2017]	LSTM	A	57.1%	55.8%	0.563	A+L	74.5%	73.9%	0.740	A	62.1%	0.620
[Mirsamadi *et al.*, 2017]	RNN & Attention	A	58.8%	59.6%	0.594	A+L	-	74.3%	0.745	A	69.1%	0.687
Our approach	The proposed RLC	A	-	**69.4%**	**0.693**	A+L	-	**79.2%**	**0.791**	A	**90.2%**	**0.901**

Table 1: The performances of state-of-the-art approaches and the proposed framework on IEMOCAP and RID. Unweighted Accuracy (UA) and F1-measure score (F1) are the higher the better. A: acoustic features, L: lexical features. (*: the original performance reported in paper.)

  - Comparing with baselines
    - Our approach greatly improves the performance in the two datasets

Speech Emotion Recognition (SER)

Tsinghua University

- Experiment results
  - Datasets: IEMOCAP && RID

	Parameters	Residual CNN	Multi-head Self-attention	RNN Cell	IEMOCAP						RID		
					Input	UA (%)	F1	Input	UA (%)	F1	Input	UA (%)	F1
Baseline	9.27M	NO	NO	LSTM	A	56.4%	0.565	A+L	72.9%	0.731	A	62.1%	0.620
S1	9.15M	YES	NO	LSTM	A	61.8%	0.621	A+L	74.3%	0.745	A	74.6%	0.744
S2	9.07M	YES	YES	LSTM	A	66.1%	0.667	A+L	77.1%	0.770	A	85.3%	0.849
S3	9.11M	YES	YES	GCA-LSTM	A	**69.4%**	**0.693**	A+L	**79.2%**	**0.791**	A	**90.2%**	**0.901**

Table 2: Experimental results for component contribution evaluation. A: acoustic features. L: lexical features. Unweighted Accuracy (UA) and F1-measure score (F1) are the higher the better. The units employed in comparison systems are balanced to ensure parameters consistency.

  - Results of component contribution research
    - Indicate the rational usage of each small sub-module appeared in our model

Speech Emphasis Detection

Tsinghua University

Learning Contextual Representation with Convolution Bank and Multi-head Self-attention for Speech Emphasis Detection

- Challenge in Speech Emphasis Detection
  - Various vocal characteristics and expressions of spoken language
  - Long-range temporal dependencies in the speech utterance
  - Local context dependencies at different scope

Liu, Liangqi, et al. "Learning Contextual Representation with Convolution Bank and Multi-head Self-attention for Speech Emphasis Detection." 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC).

Speech Emphasis Detection

- Model Architecture

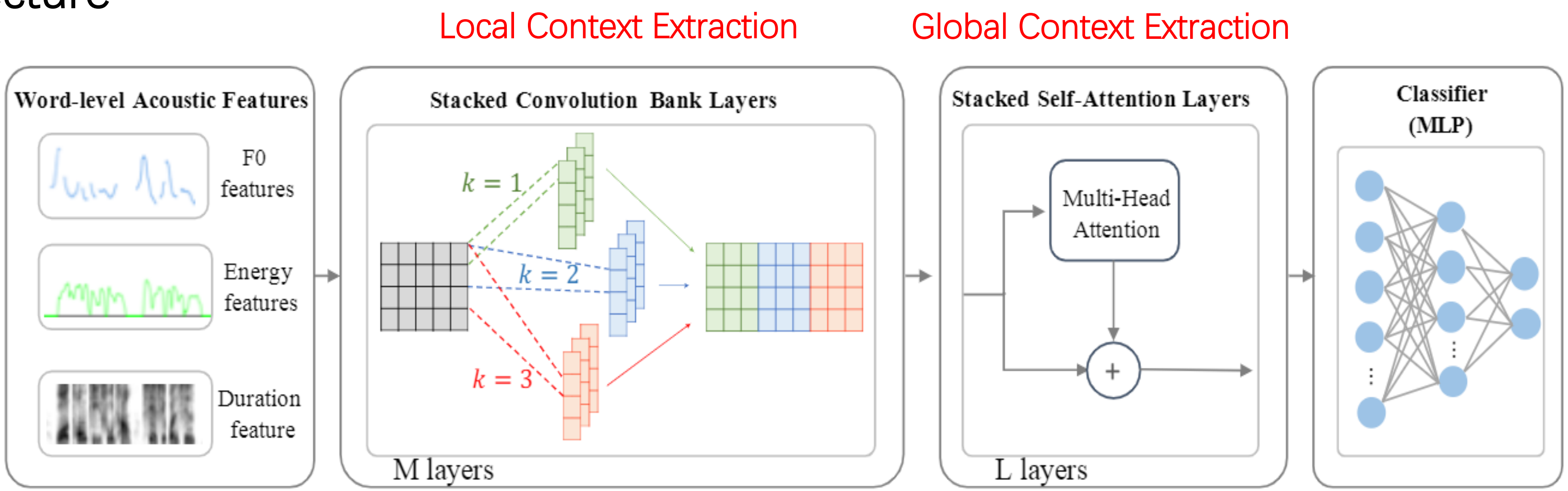


Fig. 1. The architecture of the proposed model (CB-SA) for speech emphasis detection: a stack of M convolution bank layers + a stack of L self-attention layers. k is the kernel size of convolutional filters.

- Model Highlights
  - CNN bank layers worked as k-gram to extract the local informative patterns and learn various expressions of emphasis
  - Residual multi-head self-attention layers to model the utterance-level dependencies

Speech Emphasis Detection

Tsinghua University

- Experiment results
  - Datasets: Samsung emphasis dataset

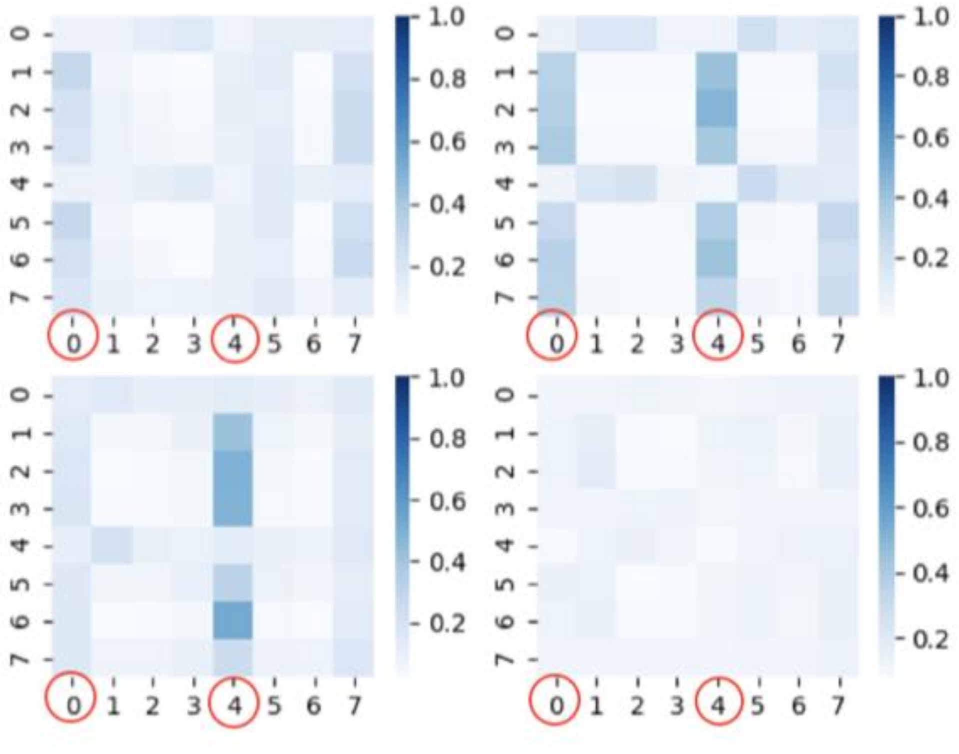
TABLE I
RESULTS OF USING DIFFERENT COMPARISON METHODS ON EMPHASIS TEST SET.

Method	Accuracy	Precision	Recall	F1
SVM	90.08%	81.41%	77.88%	0.7961
LSTM	90.48%	81.12%	80.37%	0.8074
BLSTM	91.57%	84.78%	80.46%	0.8256
CB-BLSTM	92.45%	85.04%	**84.44%**	0.8473
CB-SA	92.91%	**86.97%**	84.05%	**0.8549**

  - Comparing with baselines
    - Our approach greatly improves the performance in F1-measure metric

Speech Emphasis Detection

- Experiment results
  - Analysis of Multi-head Self-attention mechanism



Chinese Text:	只要	你	努力	工作	自然	能	赚到	钱
English Transcription:	(As long as)	(you are)	(hard)	(working)	(finally)	(you can)	(earn)	(the money)
Emphasis Label:	1	0	0	0	1	0	0	0

X-axis: the time step of keys

Y-axis: the time step of queries

User Intention Understanding

Tsinghua University

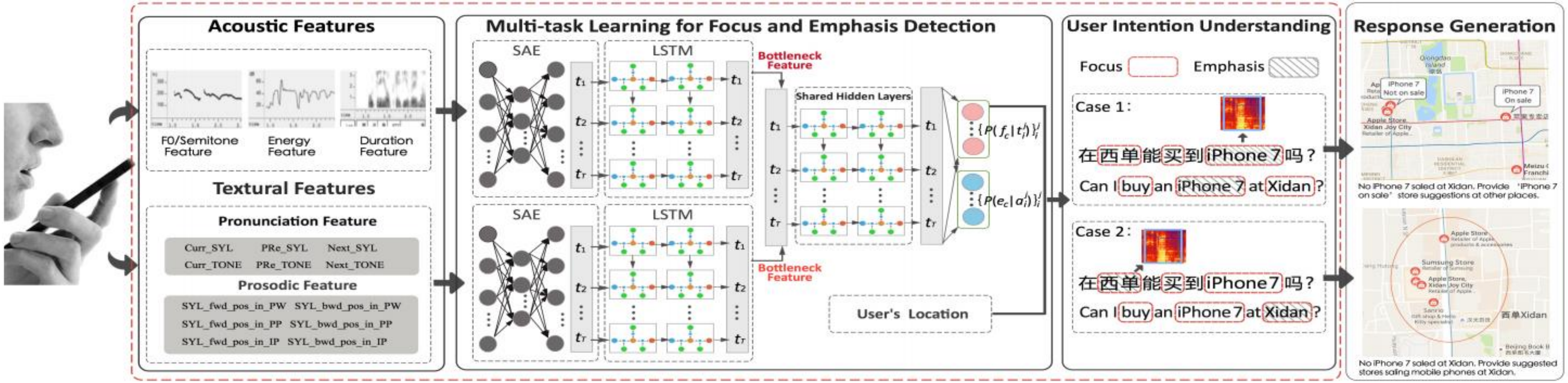
Multi-task Deep Learning for User Intention Understanding in Speech Interaction Systems

- Motivation
  - User intention is usually affected by both the focus text and emphasized spoken words
  - Focus text and emphasis spoken words are often consistent in terms of user intention
- Core Contributions
  - Address the problem of user intention understanding in real-world human-mobile interaction scenarios
  - Formulate the problem of using focus by text, emphasis by speech and location to predict Intention Prominence (IP)
  - Propose a multi-task deep learning model to integrate the three modalities

Ning, Yishuang, et al. "Multi-task deep learning for user intention understanding in speech interaction systems." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 31. No. 1. 2017.

User Intention Understanding

- Model Architecture



- Model Highlights
  - Sparse auto-encoder (SAE) is pre-trained in large unlabeled data to learn compact features
  - Multi-task learning build the correlations of prediction of *text focus* and *speech emphasis*
  - Bayesian network is used to model feature dependencies in the three modalities (*user's location* included)

User Intention Understanding

清華大學
Tsinghua University

- Experiment results
    - **Datasets**: 135,566 utterances from Sogou Voice Assistant with 2,000 labeled by 3 raters

Table 1: Comparison of results using different models.

Models	Focus			Emphasis			Intention Prominence		
	Precision	Recall	F1-measure	Precision	Recall	F1-measure	Precision	Recall	F1-measure
SVM	0.390	0.608	0.475	0.308	0.009	0.017	0.627	0.618	0.621
BN	0.704	0.760	0.731	0.462	0.272	0.343	0.797	0.789	0.791
CRF	0.724	0.755	0.739	0.457	0.036	0.066	0.769	0.754	0.761
LSTM	**0.763**	0.755	**0.759**	0.605	**0.568**	**0.575**	0.792	0.803	0.797
LSTM+BN	-	-	-	-	-	-	**0.868**	**0.865**	**0.866**

- Comparing with baselines
    - Our LSTM + BN model greatly improves the performance in all tasks

User Intention Understanding

清華大學 Tsinghua University

- Experiment results
  - Datasets: 135,566 utterances from Sogou Voice Assistant with 2,000 labeled by 3 raters

Table 3: Experimental results of the top-10 coverage ratio of the original utterances and intention prominence.

	Coverage Ratio	CI
Original Utterances	65.25%	[0.607,0.697]
Intention Prominence	72.25%	[0.687,0.758]

  - Practicality tests
    - The proposed metric Intention Prominence (IP) can greatly improve the performance of user intention understanding in real-world speech interaction scenarios

Emotional Speech Synthesis

Emotion Controllable Speech Synthesis Using Emotion-unlabeled Dataset with the Assistance of Cross-Domain Speech Emotion Recognition

- Motivation
  - Emotion-labeled TTS dataset is usually difficult to obtain, but unlabeled dataset is easily available
  - In the field of SER， there are many emotion-labeled dataset and methods for SER task

Cai, Xiong, et al. "Emotion controllable speech synthesis using emotion-unlabeled dataset with the assistance of cross-domain speech emotion recognition." *ICASSP, 2021.*

Emotional Speech Synthesis

- Model Architecture

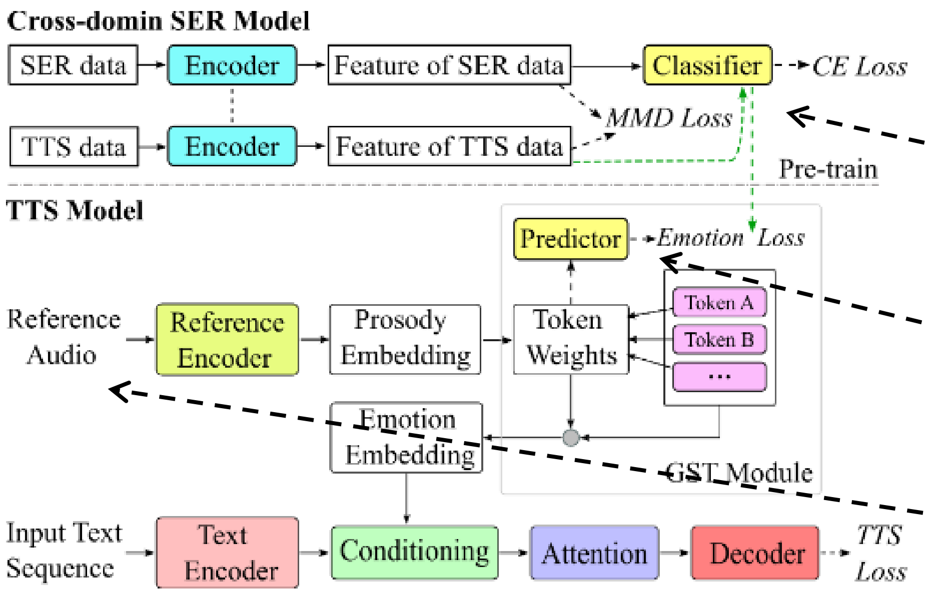


Fig. 1. The overall structure of the cross-domain SER and GST-based TTS model.

- Highlights

  Emotion controllable TTS on corpus without manually annotated emotion labels

- Model Highlights
  - Pre-train a cross-domain SER model and then label the emotion-unlabeled TTS dataset by model predictions
  - Add an auxiliary emotion prediction task based on the GST weights
  - Propose a top-K scheme to choose a high-confident reference audio set

Emotional Speech Synthesis

- Experiment results
  - Datasets: IEMOCAP && Blizzard Challenging 2013 TTS dataset

Table 2. MOS of **base-4cls** and **our-4cls** for 4 emotion categories.

model	neu	ang	hap	sad	average	p-value
base-4cls	3.90	3.84	3.45	3.74	3.73	–
our-4cls	4.12	3.80	3.11	3.61	3.66	**0.20**

Table 3. MOS of **our-2d** for arousal and valence dimensions.

model	low	high	neg	pos	average	p-value
our-2d	3.99	3.33	3.91	3.41	3.66	**0.18**

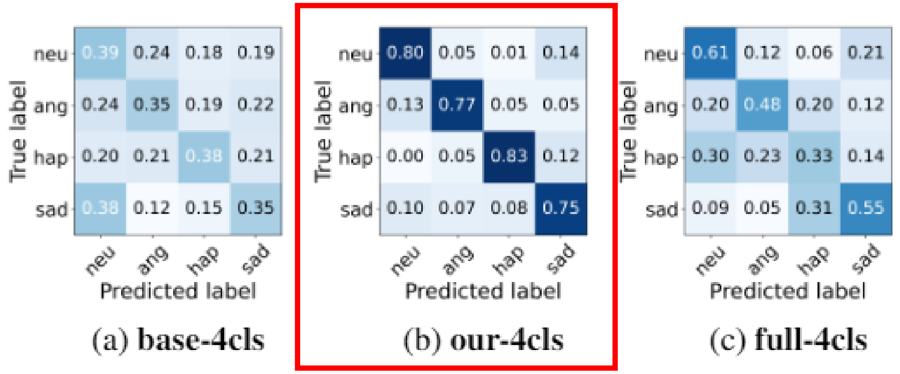


Fig. 2. Confusion matrices of 4 emotion categories for the three methods: **base-4cls**, **our-4cls** and **full-4cls**.

  - MOS scores for speech overall quality and subjective emotion prediction evaluation
    - Compared with the baseline, our model has near MOS scores but much higher subjective emotion classification accuracy

Emotional Speech Synthesis

- Experiment results
  - Datasets: IEMOCAP && Blizzard Challenging 2013 TTS dataset

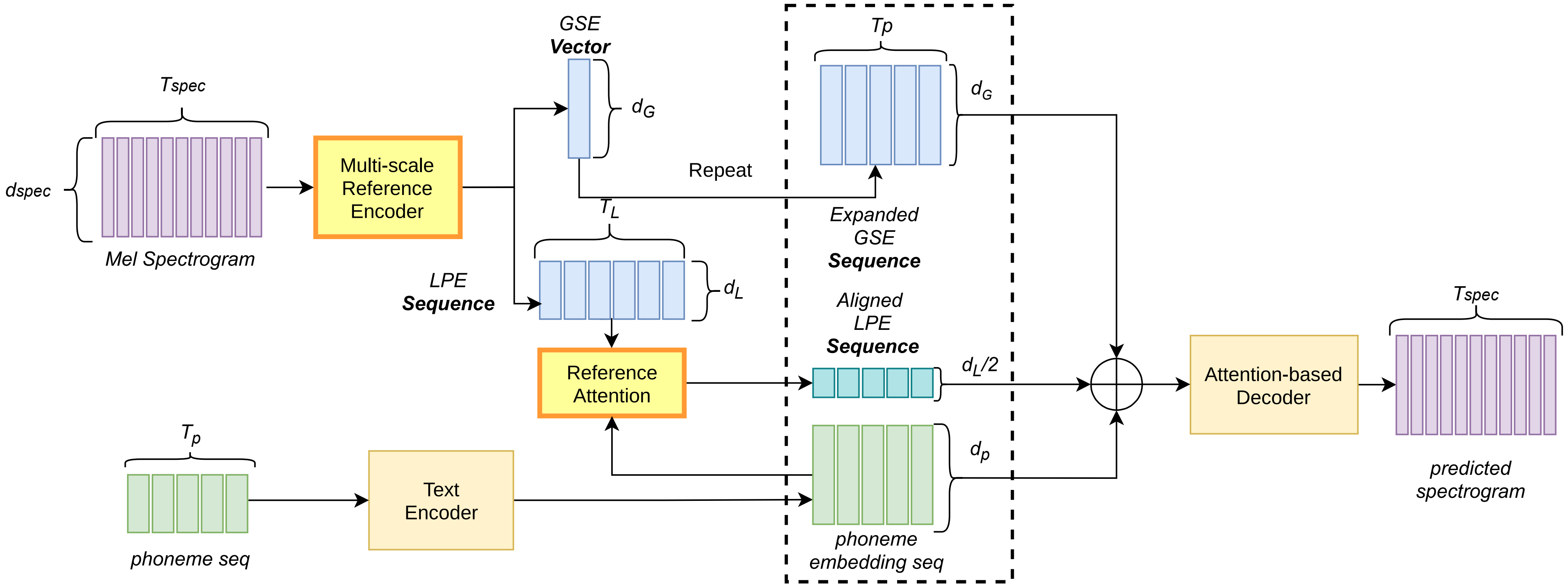
Neutral	Angry	Sad	Happy

Text: "I read a few lines, but I did not understand a word."

Controllable Emotional Speech Synthesis

Tsinghua University

Towards Multi-Scale Style Control for Expressive Speech Synthesis

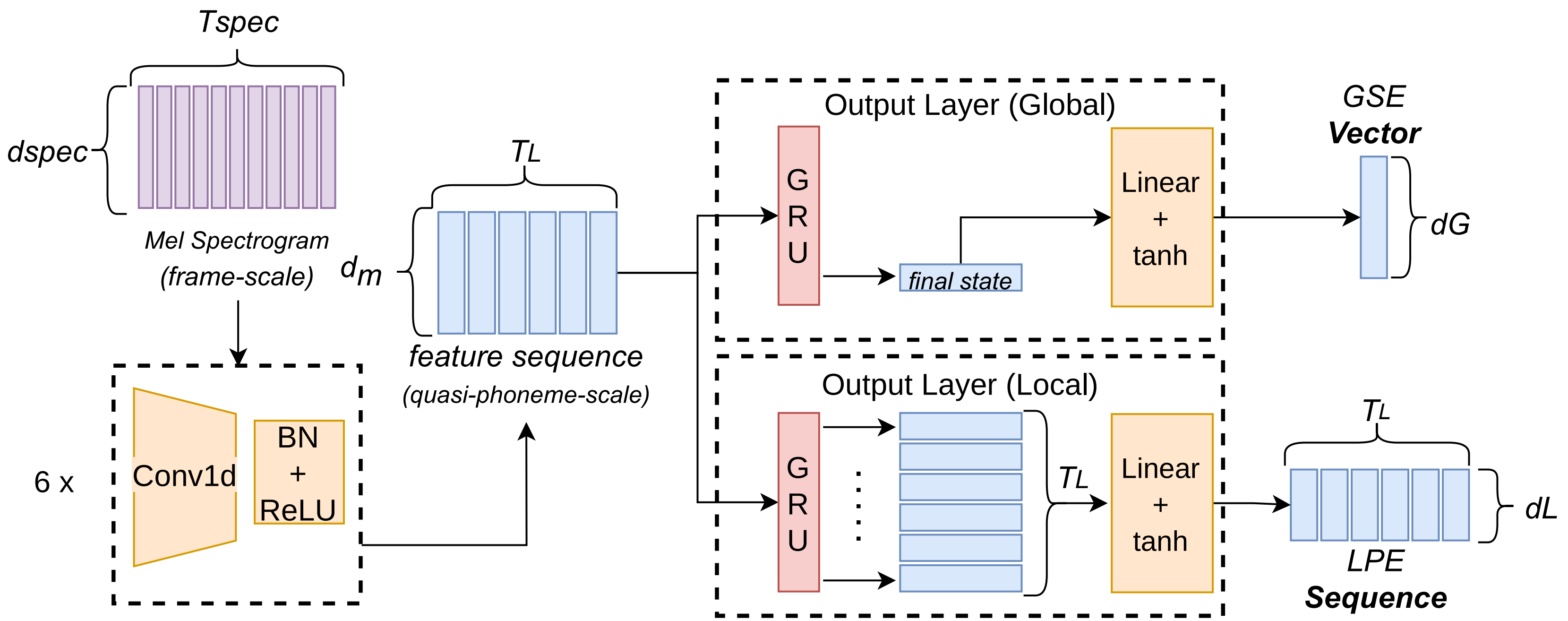


Li, Xiang, et al. "Towards Multi-Scale Style Control for Expressive Speech Synthesis." Submitted to *INTERSPEECH, 2021.*

Controllable Emotional Speech Synthesis

清華大學 Tsinghua University

Towards Multi-Scale Style Control for Expressive Speech Synthesis



Controllable Emotional Speech Synthesis

Tsinghua University

- Multi-scale Style Control
  - ✓ global-scale style embedding
  - ✓ local-scale prosody embedding

Ref (Global)						
Ref (Local)						
Result						

Emphatic Speech Synthesis

清華大學 Tsinghua University

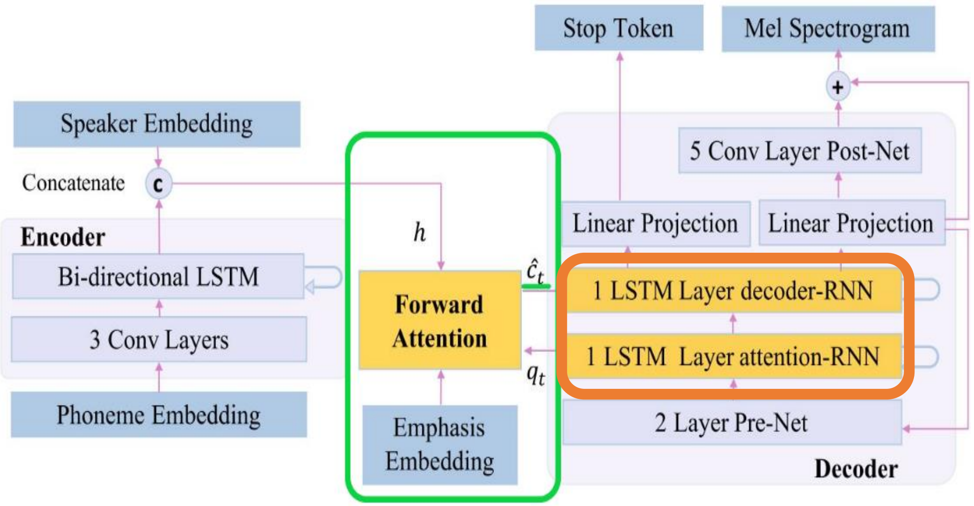
Controllable Emphatic Speech Synthesis based on Forward Attention for Expressive Speech Synthesis

- Challenge in Emphatic Speech Synthesis
  - Lack of **interpretability** for how emphatic codes affect the model
  - Lock of **controllability**: no **separate control** of emphasis on **duration and on intonation & energy**

Liu, Liangqi, et al. "Controllable Emphatic Speech Synthesis based on Forward Attention for Expressive Speech Synthesis." *SLT, 2021* (**Best Paper Finalist**).

Emphatic Speech Synthesis

- Model Architecture



- Highlights

  Interpretable and separately controlled emphatic TTS

- Model Highlights
  - Improved forward attention to explicitly control the duration of words
  - Improved Tacotron Decoder to model intonation and energy of words

Emphatic Speech Synthesis

- Experiment results
  - Datasets: Samsung emphasis dataset && DataBaker 10,000 neutral corpus

Table 1: *Emphasis identification test*

Method	Precision	Recall
Base Model	80.8%	52.5%
Proposed Model	94.2%	80.8%

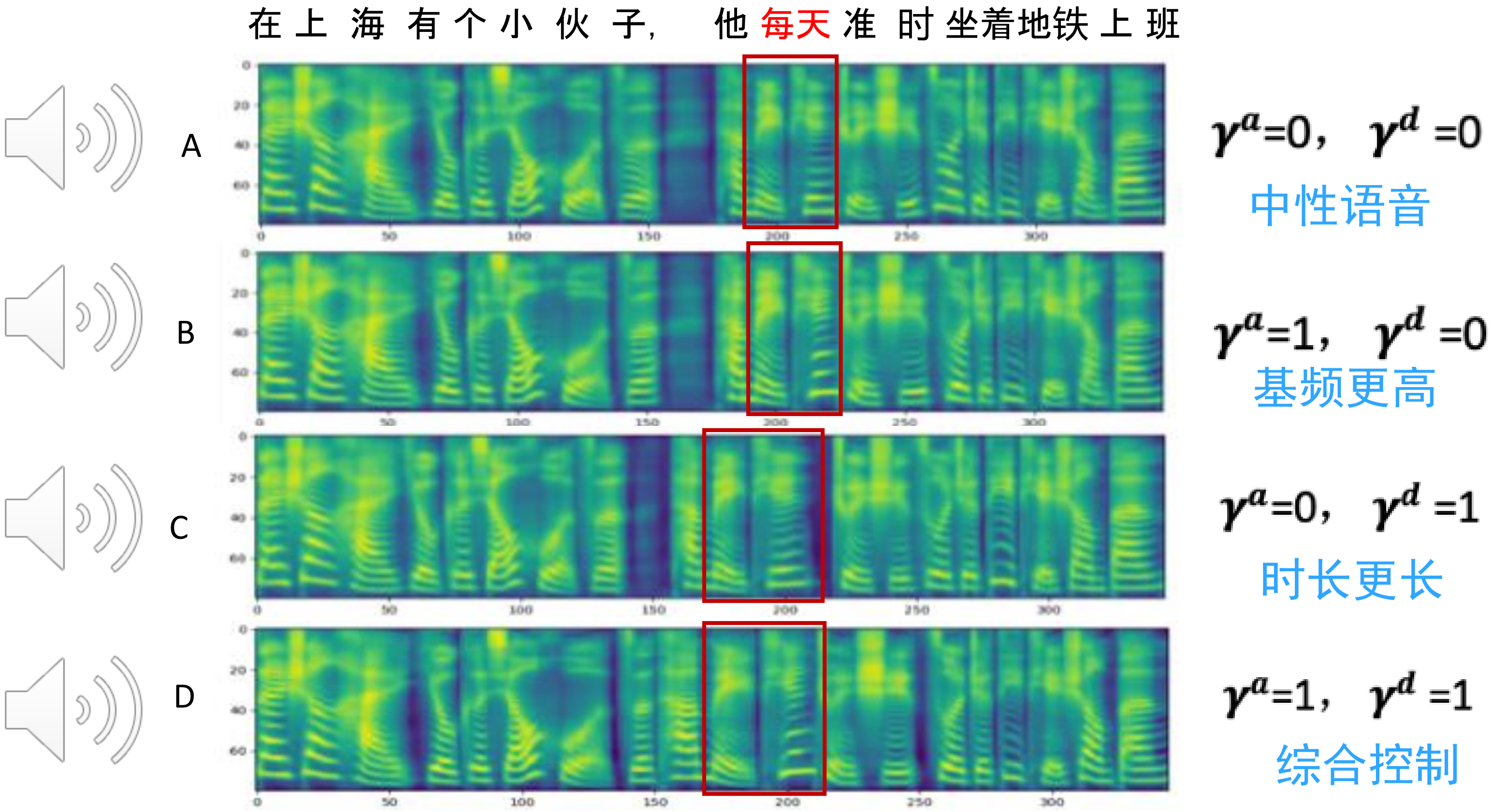
Table 2: *Naturalness test*

Method	MOS
Base Model	3.63(0.72)
Proposed Model	3.95(0.57)

  - MOS scores for speech overall quality and subjective emphasis identification test
    - Compared with the baseline, our model higher MOS scores and identification performance

Emphatic Speech Synthesis

- Experiment results
  - Datasets: Samsung emphasis dataset && DataBaker 10,000 neutral corpus



Emphatic Speech Synthesis

- Experiment results
  - Datasets: Samsung emphasis dataset && DataBaker 10,000 neutral corpus

	0	0.5	1
0			
0.5			
1			

Chinese Text:

"我想去西单买手机。"

English Translation:

"I want to go to Xidan to buy mobile phone."

Conclusions & Future Work

- Representation learning of paralinguistic speech attributes is important for intelligent speech interaction
  - **User intention understanding**
  - **Expressive speech generation**
    - Both need to model the paralinguistic attributes of speech
    - Representation learning of emotion and emphasis will help boost the performance of nowadays speech interaction systems
- Possible research focuses for improving the speech interaction
  - **Cross-domain problem** in the analysis of paralinguistic attributes of speech
    - The recording devices, speakers . . . are usually different from the training datasets
  - **Context understanding** in dialog systems
  - Generation of expressive speech with **more styles** (e.g. chat) is desired
  - Analysis and generation of **spontaneous speech** leads to more natural speech interactions

One More Thing!

Expressive Speech Driven Talking Head

Neutral

Angry

Happy

Sad

Q & A

Thanks!