NEURAL INFORMATION PROCESSING SYSTEMS

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

Microsoft

CoVoMix: Advancing Zero-Shot Speech Generation for Human-like Multi-talker Conversations

Presenter: Leying Zhang (Shanghai Jiao Tong University)

Advisor: Yanmin Qian, Yao Qian

NeurIPS 2024, Vancouver, Canada

Outline

1. Background and Motivation
2. Design of CoVoMix
3. Data Processing
4. Evaluation Results
5. Conclusion

Outline

Background and Motivation

Conversation is the most frequent form of human communication

Daily Talk

Meeting

Podcast

Human-Machine Interaction

Challenges

1) Rendering the **spontaneous-style speech**

filled pauses, interjections, repair, repetition, laughing, etc

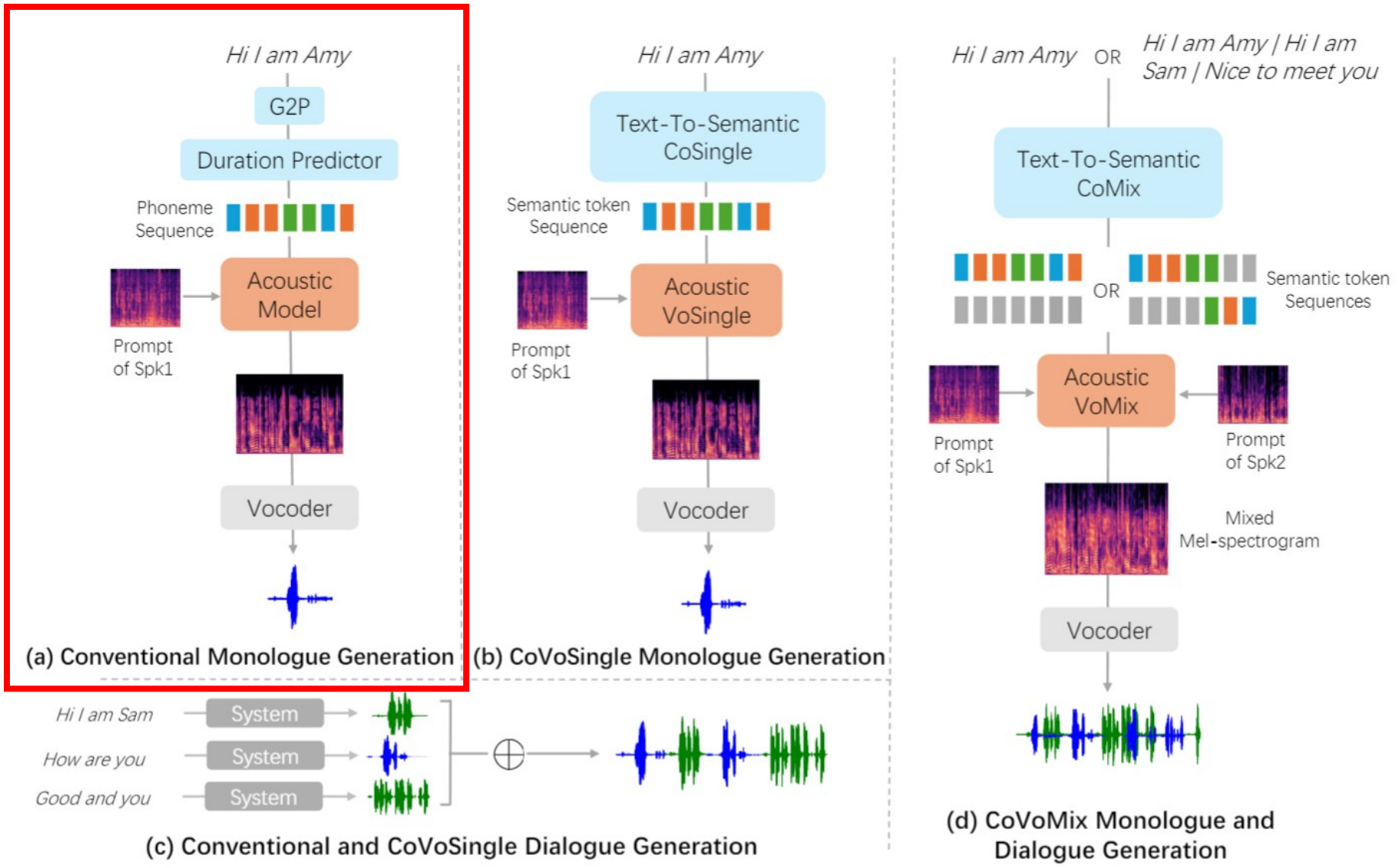
2) Managing **natural speech transition of multiple speakers**

the timing of their speech, determining when one speaks and when the other will follow

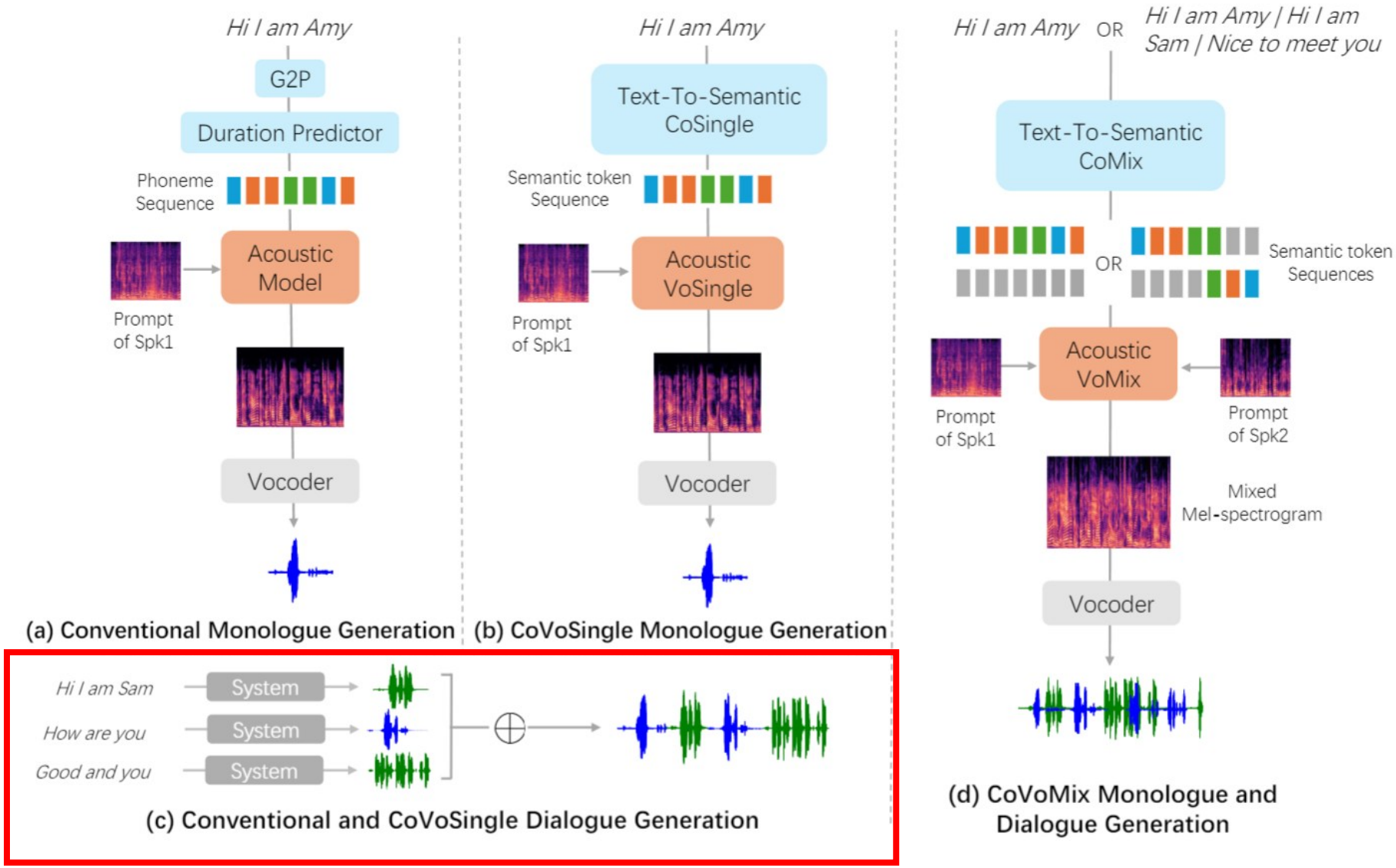
Outline

1. Background and Motivation
2. **Design of CoVoMix**
3. Data Processing
4. Evaluation Results
5. Conclusion

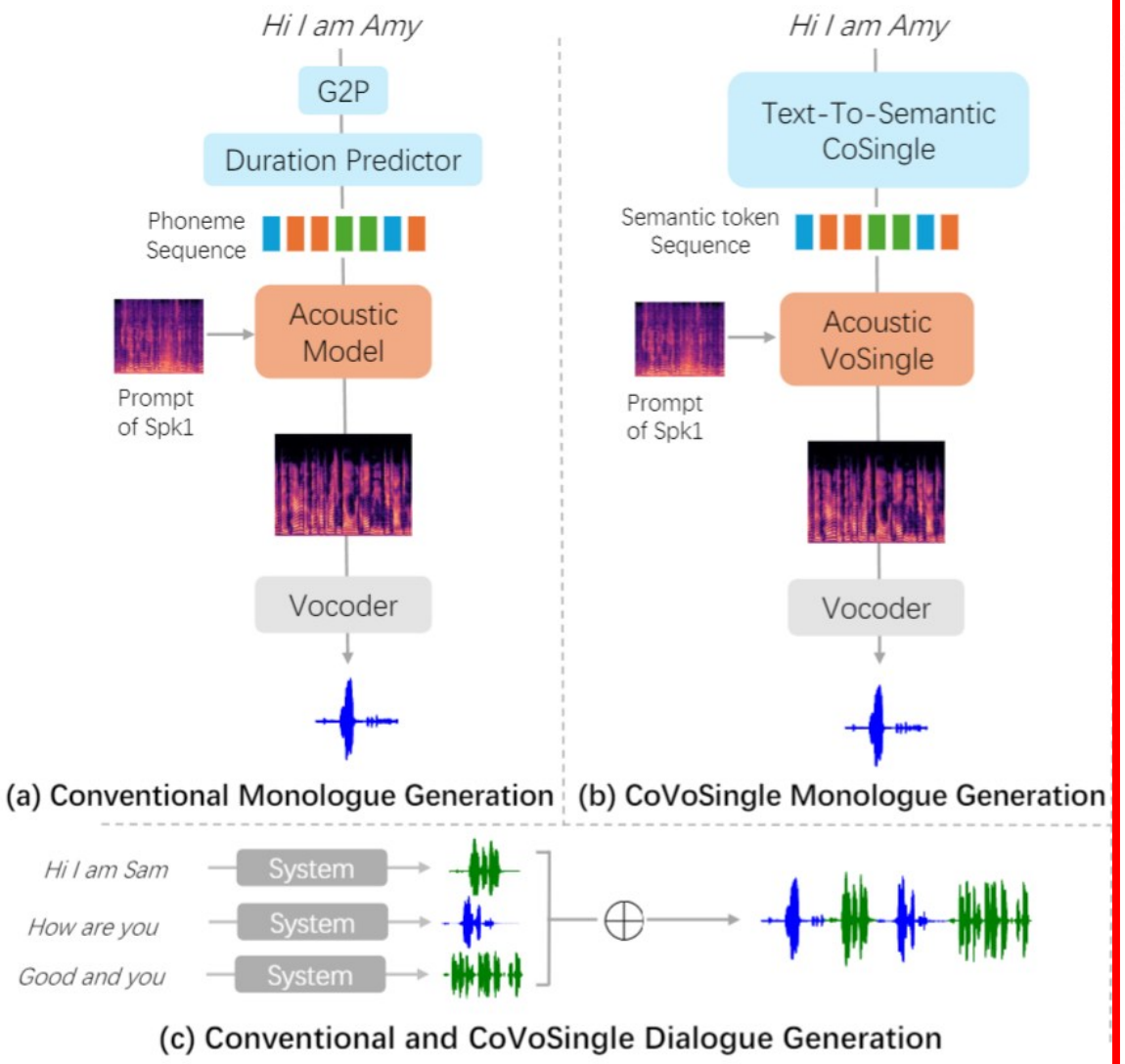
Pipeline Comparison

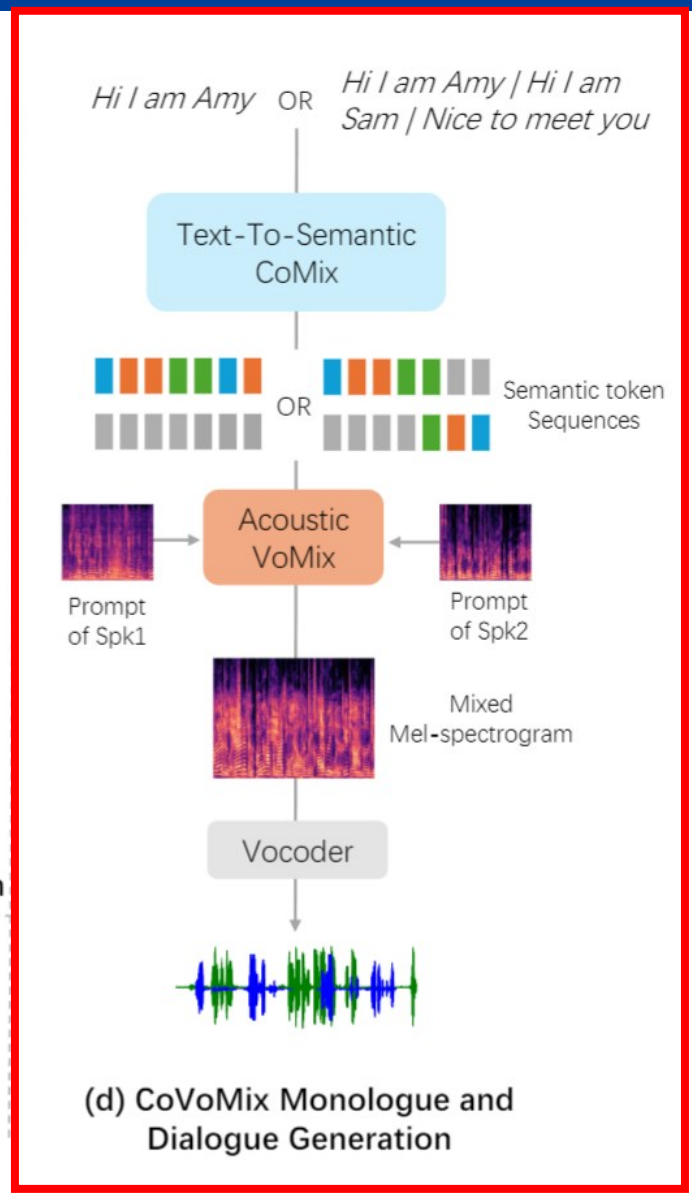


Pipeline Comparison

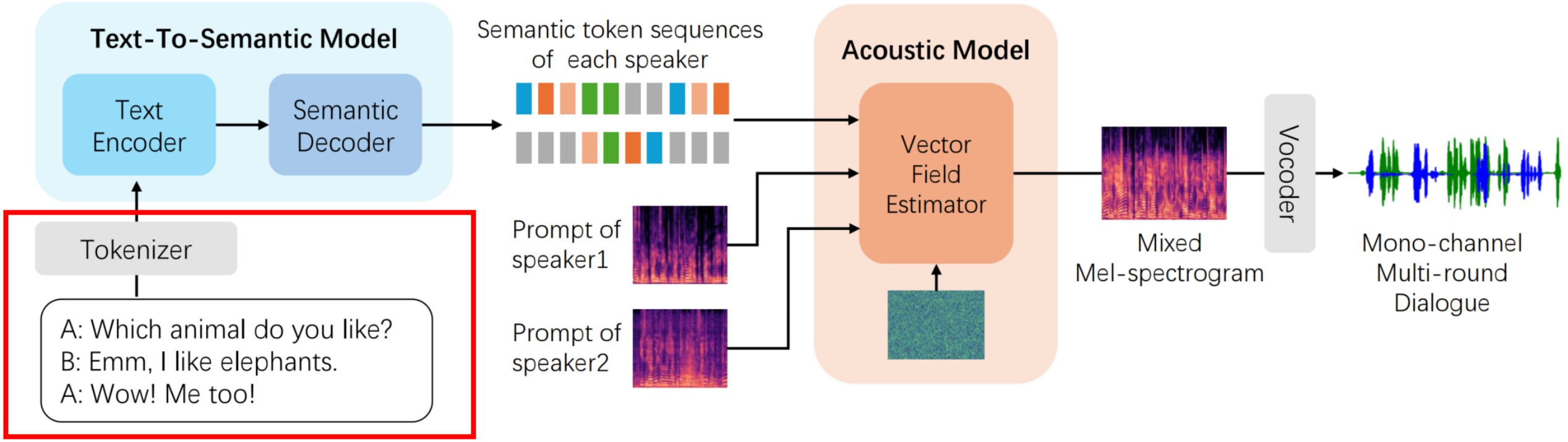


Pipeline Comparison

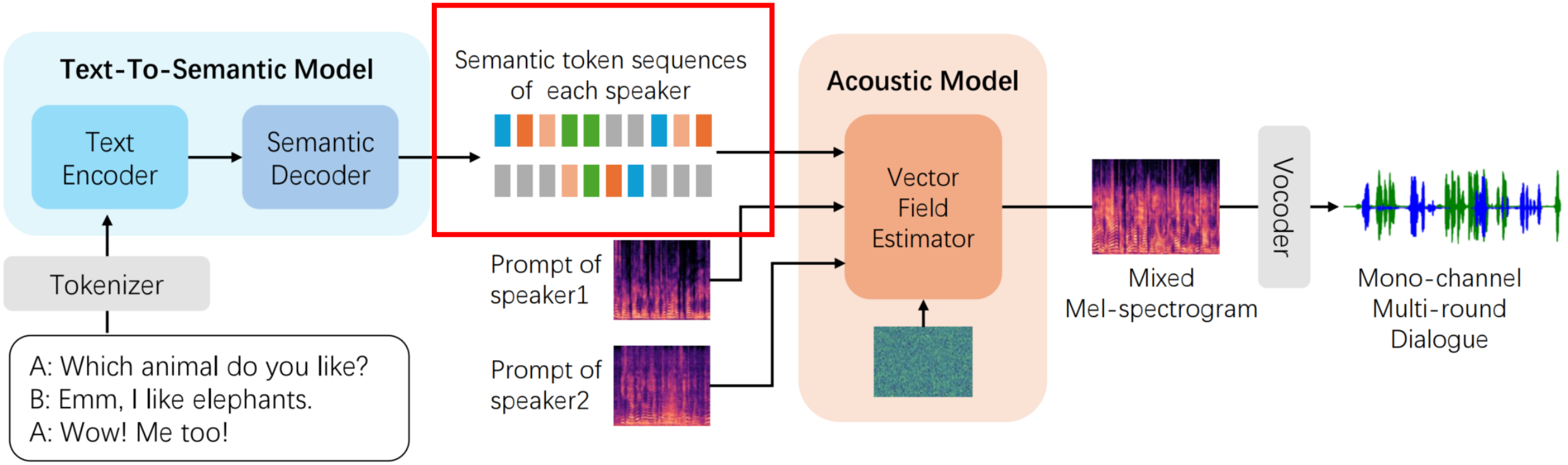




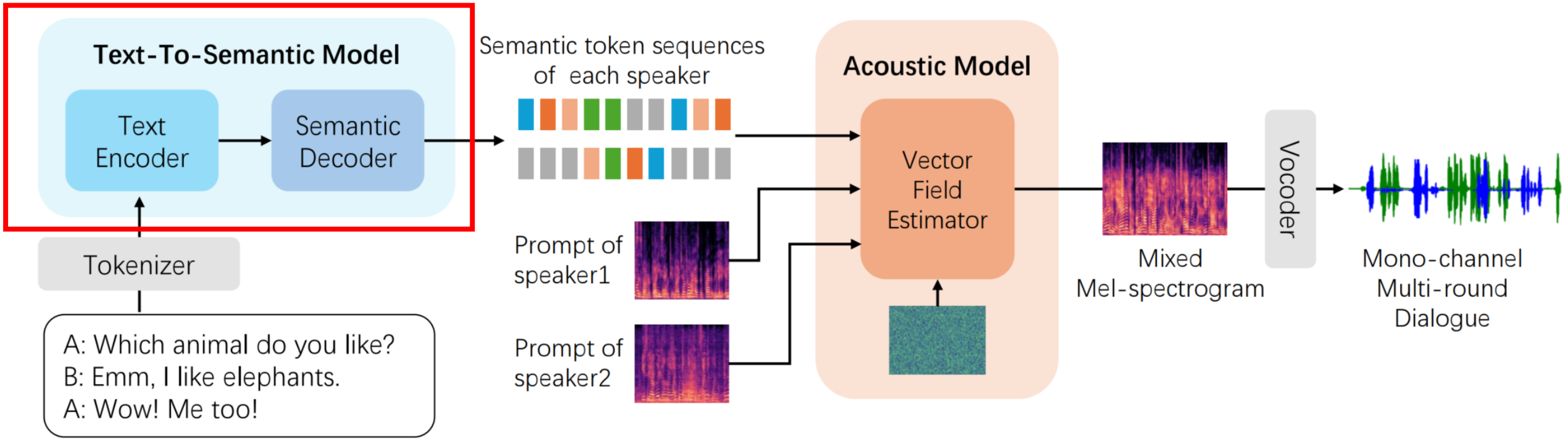
CoVoMix


Text-To-Semantic Model
Text Encoder
Semantic Decoder
Tokenizer
A: Which animal do you like?
B: Emm, I like elephants.
A: Wow! Me too!
Semantic token sequences of each speaker
Prompt of speaker1
Prompt of speaker2
Acoustic Model
Vector Field Estimator
Mixed Mel-spectrogram
Vocoder
Mono-channel Multi-round Dialogue

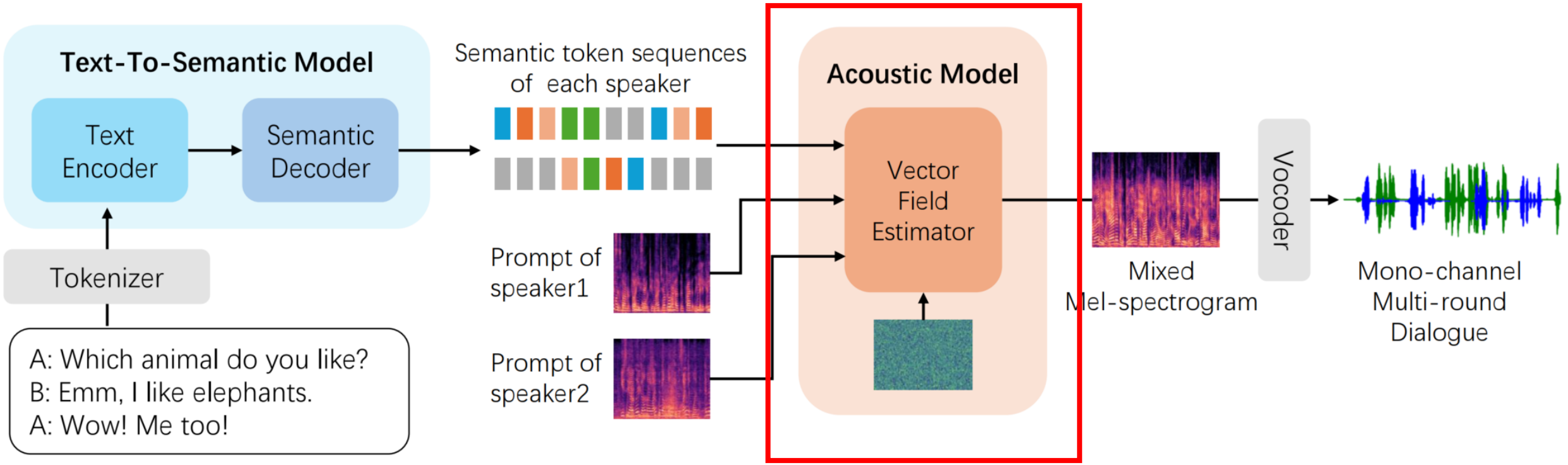
CoVoMix


Text-To-Semantic Model
Text Encoder
Semantic Decoder
Tokenizer
A: Which animal do you like?
B: Emm, I like elephants.
A: Wow! Me too!
Semantic token sequences of each speaker
Prompt of speaker1
Prompt of speaker2
Acoustic Model
Vector Field Estimator
Mixed Mel-spectrogram
Vocoder
Mono-channel Multi-round Dialogue

CoVoMix


Text-To-Semantic Model
Text Encoder
Semantic Decoder
Tokenizer
A: Which animal do you like?
B: Emm, I like elephants.
A: Wow! Me too!
Semantic token sequences of each speaker
Prompt of speaker1
Prompt of speaker2
Acoustic Model
Vector Field Estimator
Mixed Mel-spectrogram
Vocoder
Mono-channel Multi-round Dialogue

CoVoMix


Text-To-Semantic Model
Text Encoder
Semantic Decoder
Tokenizer
A: Which animal do you like?
B: Emm, I like elephants.
A: Wow! Me too!
Semantic token sequences of each speaker
Prompt of speaker1
Prompt of speaker2
Acoustic Model
Vector Field Estimator
Mixed Mel-spectrogram
Vocoder
Mono-channel Multi-round Dialogue

Outline

1. Background and Motivation
2. Design of CoVoMix
3. **Data Processing**
4. Evaluation Results
5. Conclusion

Data Processing

good morning | good morning | it's been a long time since i saw you | yeah 😀 i'll be in touch

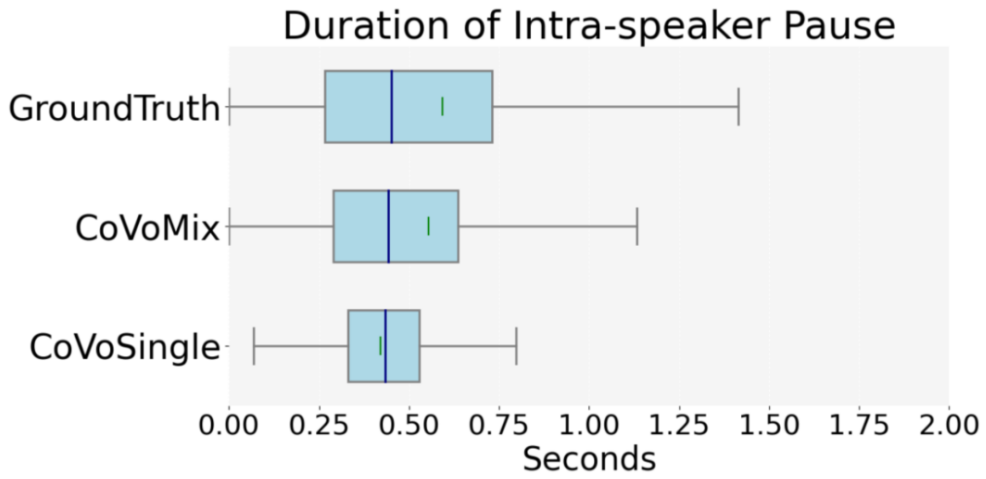
Outline

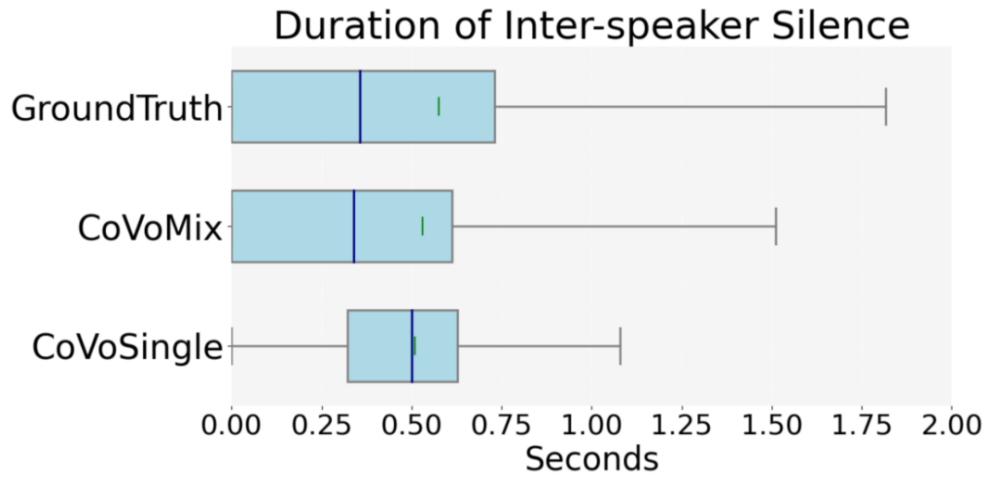
1. Background and Motivation
2. Design of CoVoMix
3. Data Processing
4. **Evaluation Results**
5. Conclusion

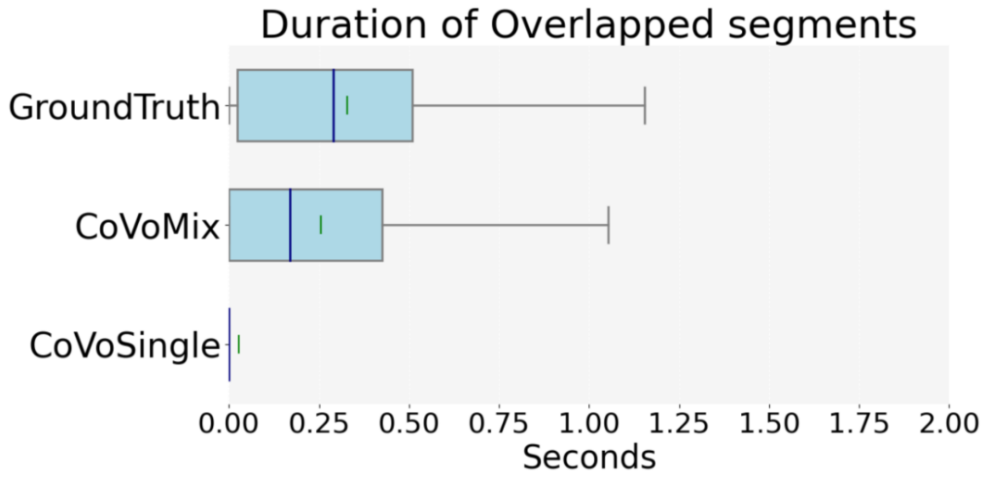
Results for Monologue and Dialogue Generation

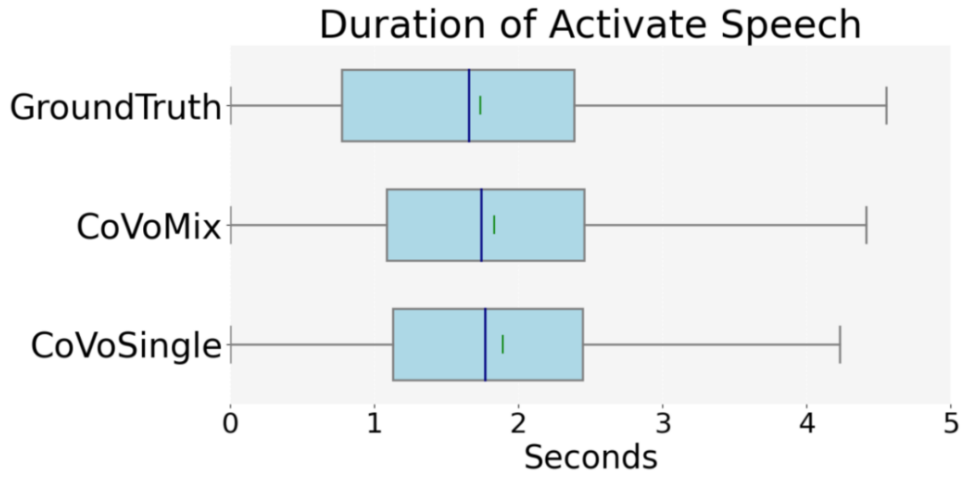
Eval Set	System	SIM ↑	WER ↓	MCD ↓	NISQA ↑	CMOS ↑	SMOS ↑
Monologue	GroundTruth	0.59	6.10	/	3.03	/	/
	Baseline	0.42	15.85	9.45	2.93	-1.60†	-1.18†
	CoVoSingle	0.49	9.99	6.15	3.04	0.00	0.00
	CoVoMix	0.49	8.95	6.04	3.01	0.83†	0.11
Dialogue	GroundTruth	/	14.91	/	2.73	/	/
	CoVoSingle	/	11.77	6.91	2.90	0.00	0.00
	CoVoMix	/	19.84	6.82	2.87	0.81†	0.60†

Turn-taking Statistics for Dialogue Generation

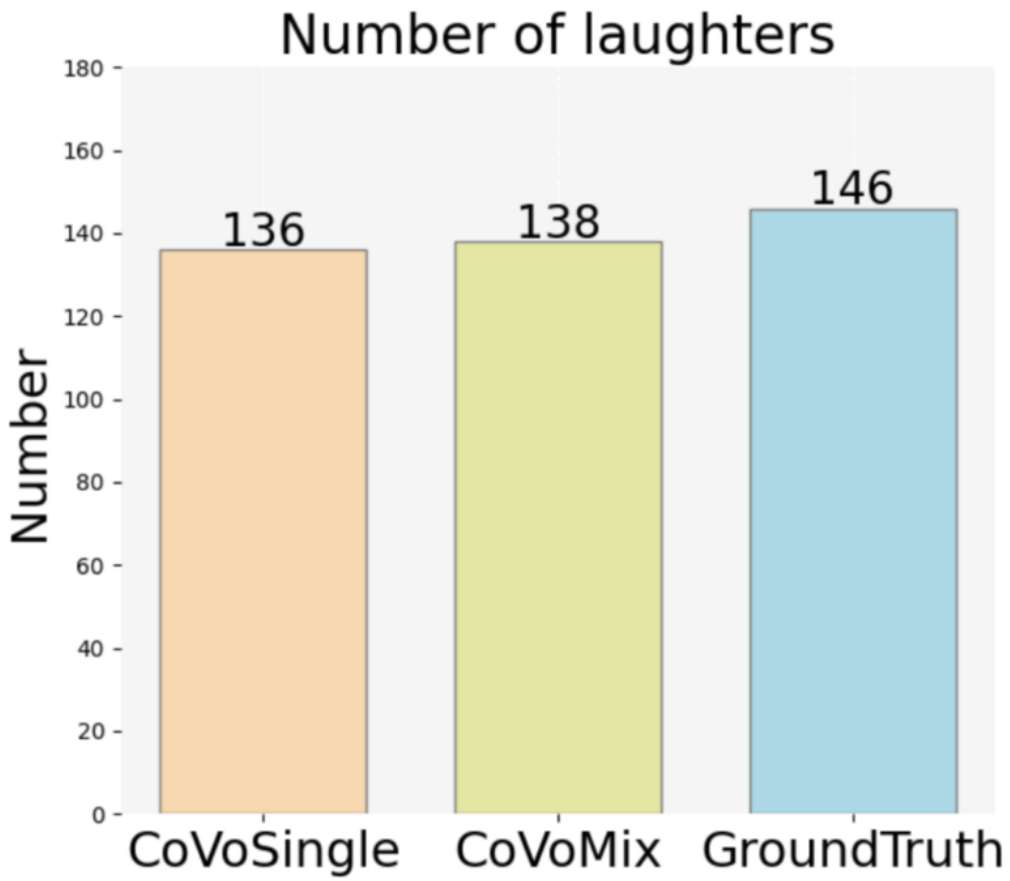


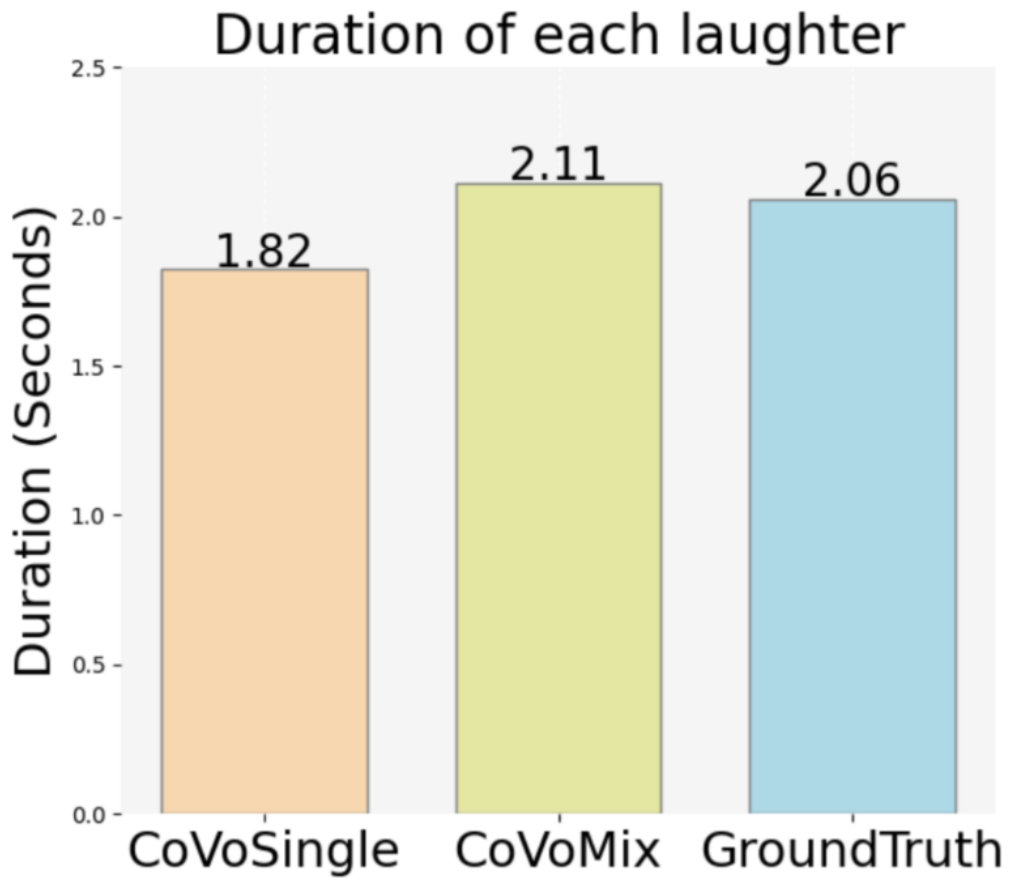




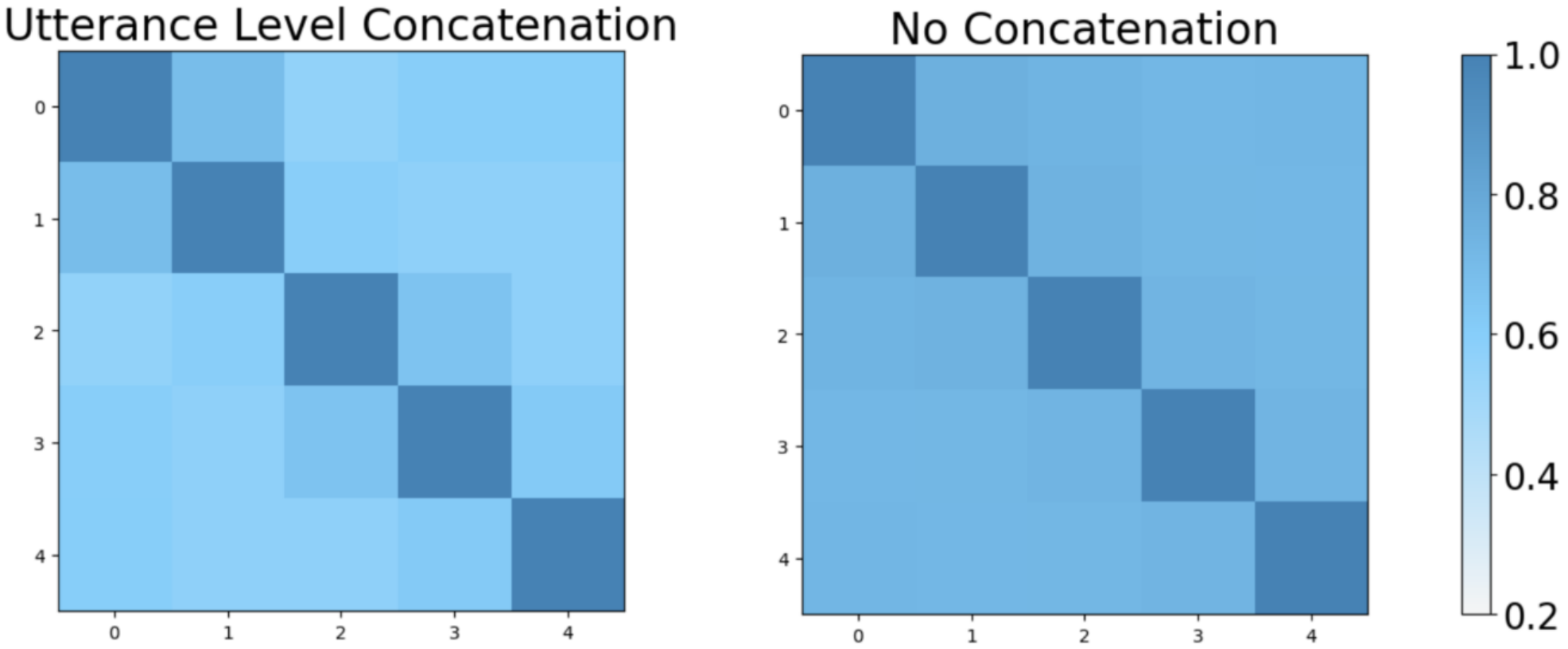


Laughter Statistics for Dialogue Generation





Speech Consistency for Multi-round Dialogue



Outline

1. Background and Motivation
2. Design of CoVoMix
3. Data Processing
4. Evaluation Results
5. Conclusion

Conclusion

- CoVoMix is composed of an auto-regressive text-to-semantic model and a flow-matching based acoustic model, with semantic token sequence as an intermediate representation
- CoVoMix achieves high naturalness and zero-shot speaker similarity in both monologue and dialogue generations
- CoVoMix demonstrates its proficiency in the fluency of dialogue turn-taking and spontaneous behavior generation

Thank You!

CoVoMix: Advancing Zero-Shot Speech Generation for Human-like Multi-talker Conversations

Discussion: Leying Zhang (Presenter), zhangleying@sjtu.edu.cn

上海交通大學

SHANGHAI JIAO TONG UNIVERSITY