Neural Text-to-Speech @Microsoft - large scale production and ongoing exploration

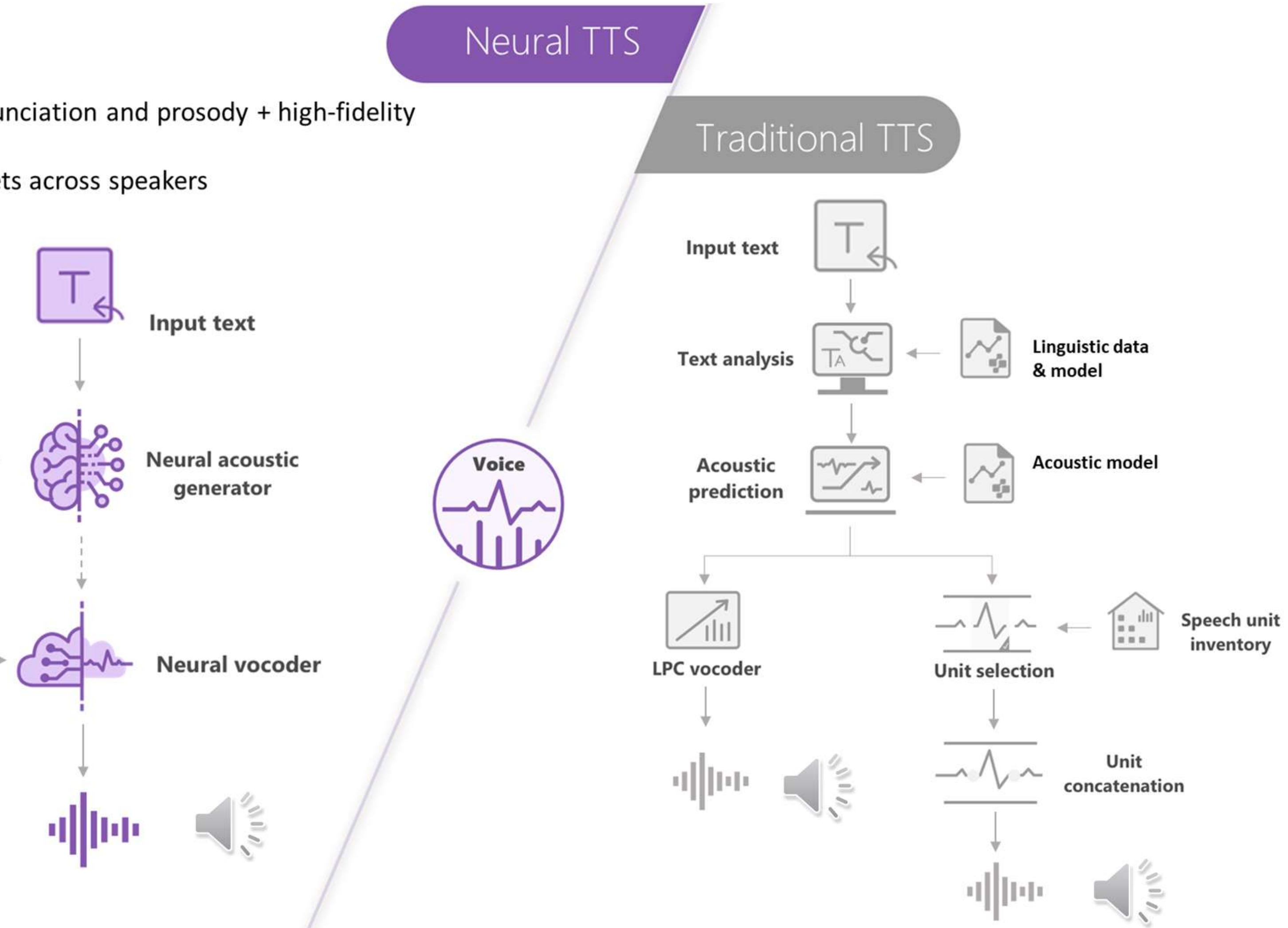
Lei He
Microsoft Azure Speech
July 24. 2021

Outline

- Neural TTS dominates TTS production
  - A recap
  - Recent progresses
    - Robustness & Cost challenges
    - Transfer learning across speakers, styles/emotions and languages
  - A brief of MS' TTS service
- What is next/ongoing – selected topics
  - Human parity TTS beyond sentence
    - Contextual aware neural TTS & Intelligent text analysis
  - Higher expressiveness
    - Cross speaker style transferring
  - Synergy of speech recognition and synthesis
    - TTS data augmentation for SR & Dual learning

A recap

- Joint optimization of pronunciation and prosody + high-fidelity audio generation
- Learning from large datasets across speakers



Human recording

Recent progress

- High fidelity, high natural neural TTS dominates the TTS products.
  - Across global tech-giants, top speech vendors, and startup companies.
- Two key contributors of Neural TTS production
  - **Cost** – rapid evolution in neural vocoder, which keeps good fidelity (even not perfect) with much less computation load and efficient inference. [1]
  - **Robustness** – complicate attention mechanisms is removed in many TTS applications, which is proved to be efficient in majority of TTS scenarios [2/3]; combine robust attention and large scale pretraining, the attention-based method is used in "rich scenarios" [4/5].

1. IS2020: An Efficient Subband Linear Prediction for LPCNet-Based Neural Synthesis
2. NIPS 2019: FastSpeech: Fast, Robust and Controllable Text to Speech
3. ICLR 2021: FastSpeech 2: Fast and High-Quality End-to-End Text to Speech
4. IS2019: Robust Sequence-to-Sequence Acoustic Modeling with Stepwise Monotonic Attention for Neural TTS
5. IS2019: Exploiting Syntactic Features in a Parsed Tree to Improve End-to-End TTS

Recent progress (cont.)

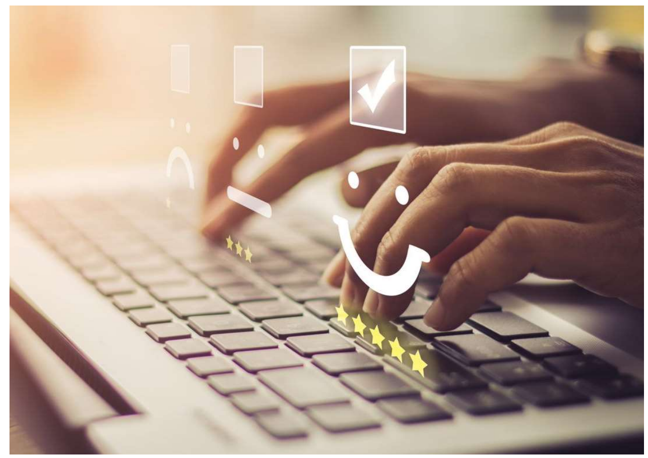
- Learning from large diverse speech data, across speakers, styles and languages, became a foundation of production at scale.
  - Across speakers training in both acoustic model and neural vocoder is widely adopted, enable high quality voice customization (and personalization). [1/2]
  - Style model is studied in different granularity, from sentence level style embedding to fine-grained prosody control. [3/4]
  - Multi-lingual neural TTS enable the cross-lingual TTS (polyglot) and scale to new language with much less data. [5/6]

1. ArXiv:1812.05253, Modeling Multi-speaker Latent Space to Improve Neural TTS: Quick Enrolling New Speaker and Enhancing Premium Voice
2. ICLR 2021: ADASPEECH: ADAPTIVE TEXT TO SPEECH FOR CUSTOM VOICE
3. ICASSP2019: Learning Latent Representations for Style Control and Transfer in End-to-End Speech Synthesis
4. Neural Networks 2021: Cycle consistent network for end-to-end style transfer TTS training
5. IS2020: Towards Universal Text-to-Speech
6. Arxiv:2103.03541v1, Multilingual Byte2Speech Text-To-Speech Models Are Few-shot Spoken Language Learners

Azure Text-to-Speech

Neural Voice

170 voices, 70+ languages (growing)
REST APIs
SDKs
Web portal
Cloud, on-prem



Custom Neural Voice

16 languages through self-service (growing)
REST APIs
SDKs
Web Portal
Cloud, on-prem

Neural TTS

Voice samples

LANGUAGE	VOICE	SAMPLE
English (UK)	Ryan	
	Mia	
English (US)	Jenny – general	
	Jenny – chat	
	Jenny – customer service	
	Jenny – Chinese *	

Style/emotion degree tuning

sad=0.1 sad=0.5 sad=1.0 sad=1.5 sad=2.0

https://azure.microsoft.com/en-us/services/cognitive-services/text-to-speech/

* Cross language transfer in preview

Custom Neural Voice

Human-like voices
custom made for your use case

300-2000 utterances (30 mins to 2 hours of speech data) to create a highly natural voice

Gating for responsible AI*

Human	Custom Neural Voice

https://speech.microsoft.com

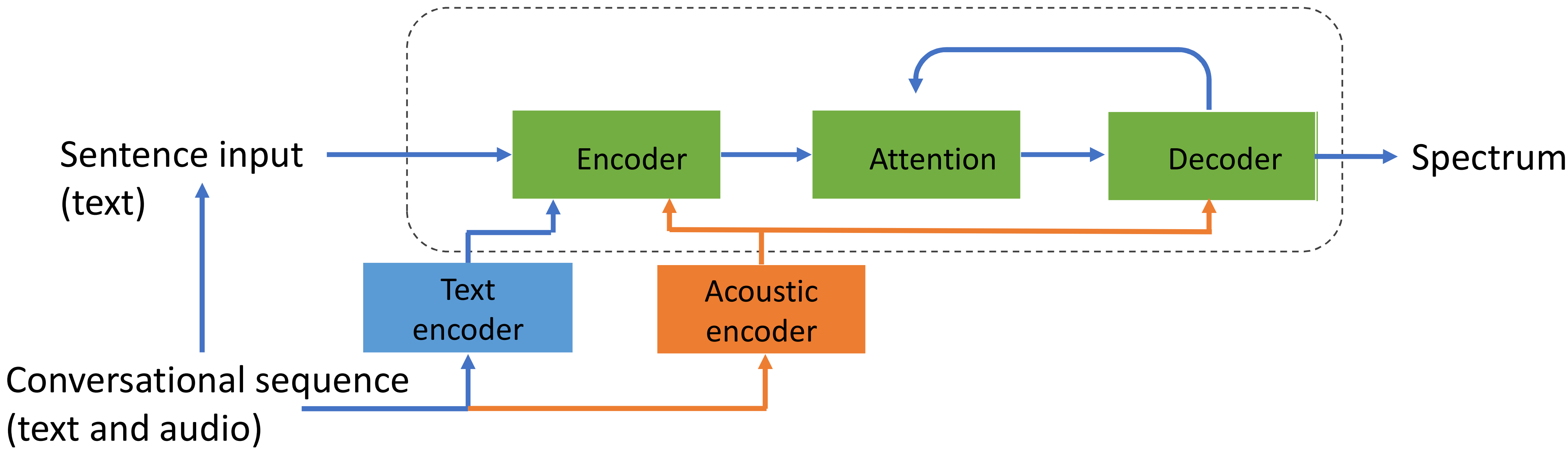
What is next (selected topics)

- Human parity TTS beyond sentence level
- Controllable higher expressiveness
- Synergy of speech recognition and synthesis
- …

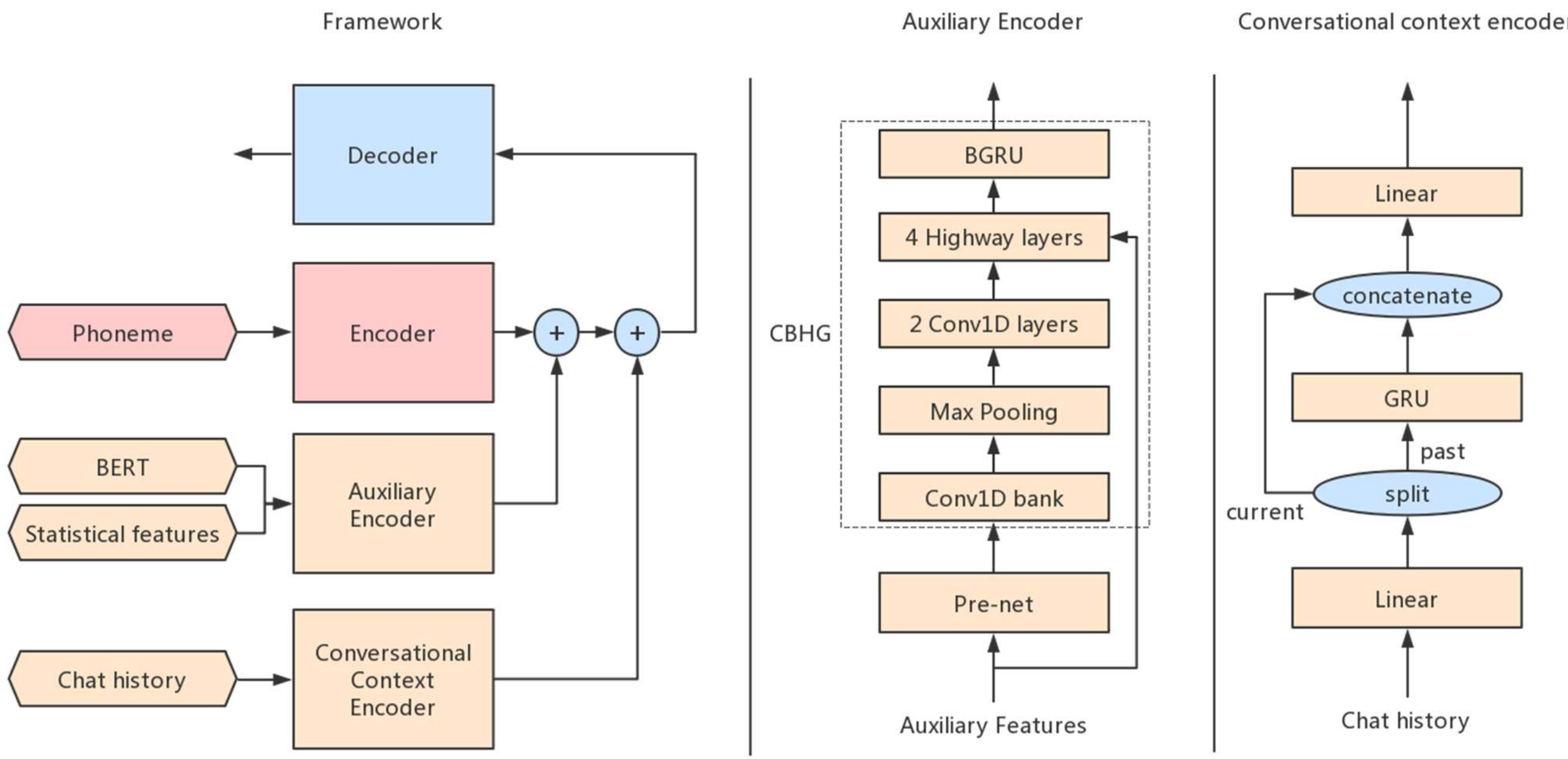
Contextual aware neural TTS - Motivation

- Neural TTS shows close human parity quality
  - In sentence level synthesis and in target domains ☺
  - With clear prosody gap in paragraph reading and dialogue conversation ☹
- Next step
  - Beyond sentence prosody – model larger context
- Context defined here
  - **Paragraph** – the context between continuous utterances in one paragraph.
  - **Dialogue conversation** – the context between continuous turns in one conversation.

High level design - conversation



Conversational TTS



Experimental results

- Models:
  - M1: Modified Tacotron2
  - M2: M1 + Auxiliary Encoder
  - M3: M2 + Conversational Context Encoder
- Training strategy:
  - Pre-train a standard model with reading style TTS data set
  - Train M1, M2 and M3 on top of the pre-trained model respectively

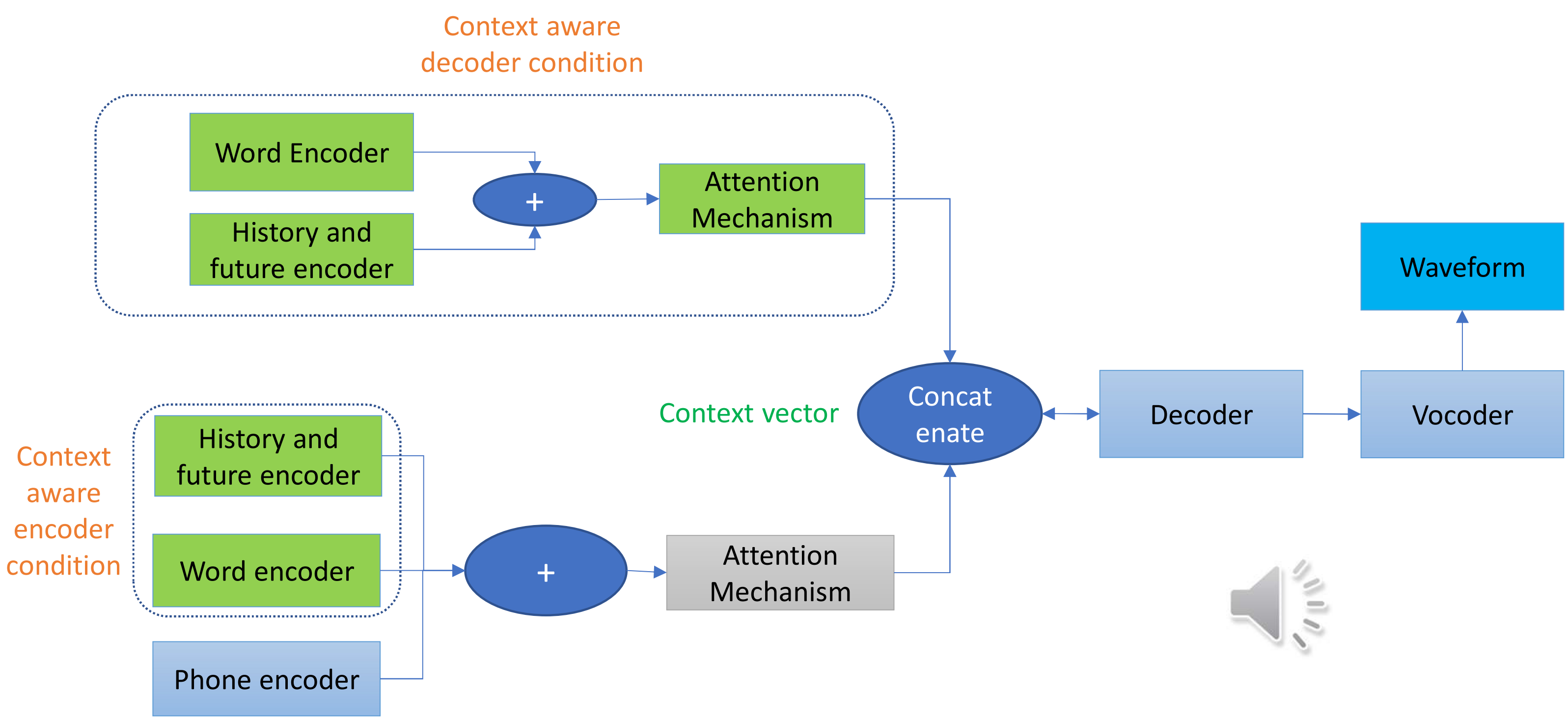
Table 4. The results of CMOS tests. "U-level" and "C-level" donate the utterance level and conversation level respectively. Preference is calculated according to CMOS scores: the score greater than 0 means bias towards B, equal to 0 means neutral and less than 0 means bias towards A.

	CMOS	Preference (%)			
		M_1	Neutral	M_2	p-value
U-level	0.22	24.4	32.7	42.9	0.0001
C-level	0.62	21.0	20.0	59.0	0.0001

	CMOS	Preference (%)			
		M_2	Neutral	M_3	p-value
U-level	0.18	28.1	29.8	42.1	0.0001
C-level	0.39	28.0	15.0	57.0	0.001

Paragraph synthesis

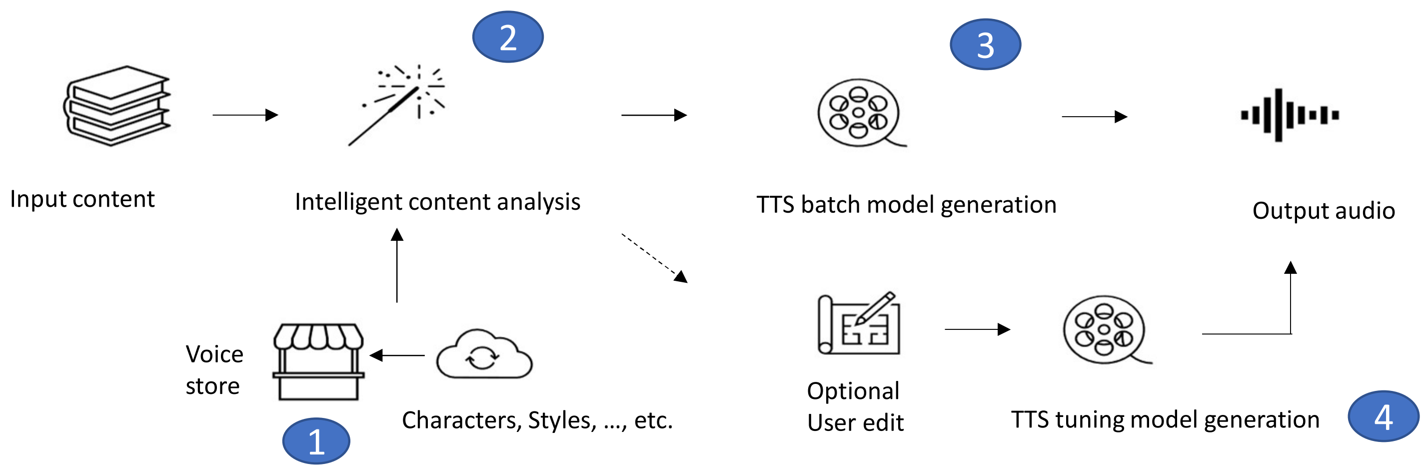
- Introduce paragraph analyzed context encoder and decoder.



https://speech.microsoft.com/audiocontentcreation

Rich audio content creation

- Long form, high expressiveness and "rich" content read out
  - Audio book, enhanced news/media reading, etc.
- Tech
  - Intelligent text analysis
  - Multi style, role and characters voice model
  - Contextual aware neural TTS



Style Transfer with Prosody Bottleneck

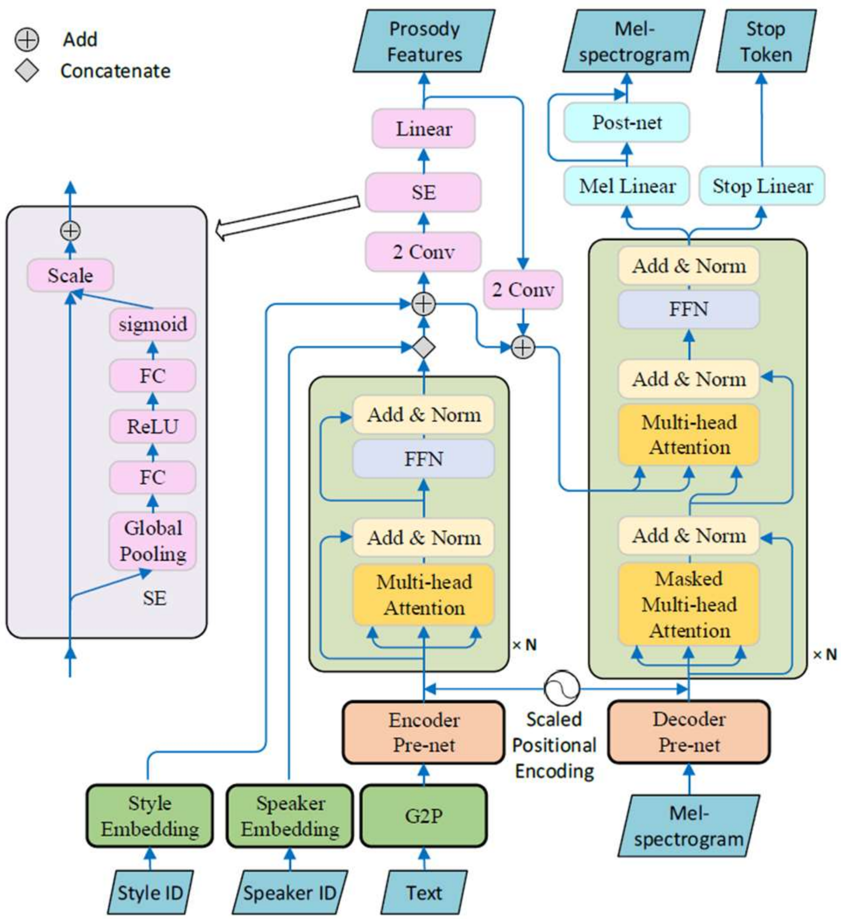
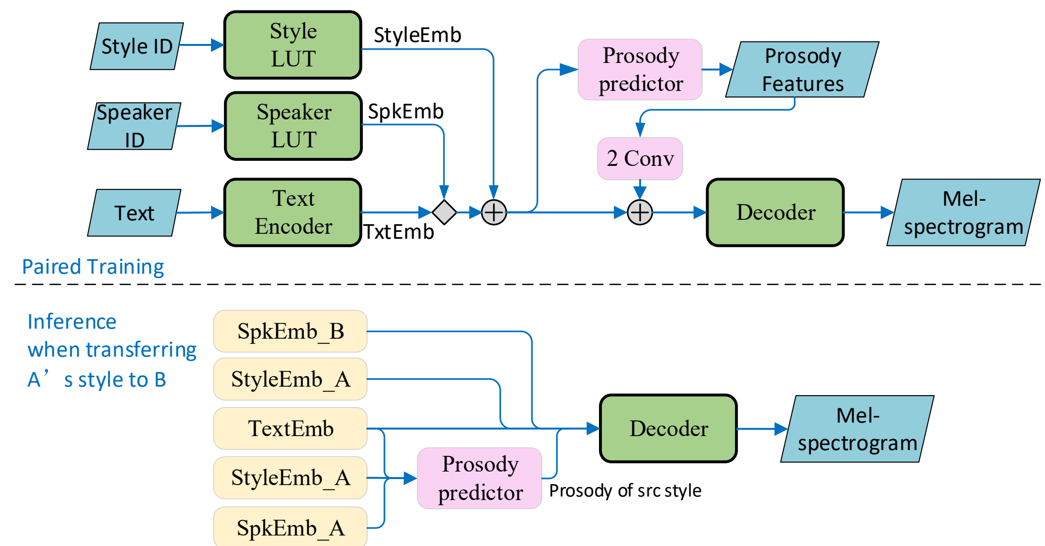


Figure 1: Overall structure of the proposed model



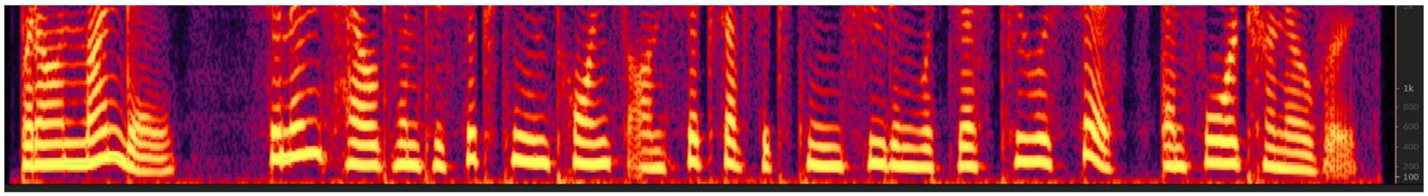
IS2021: Cross-speaker Style Transfer with Prosody Bottleneck in Neural Speech Synthesis

The Transfer Works?

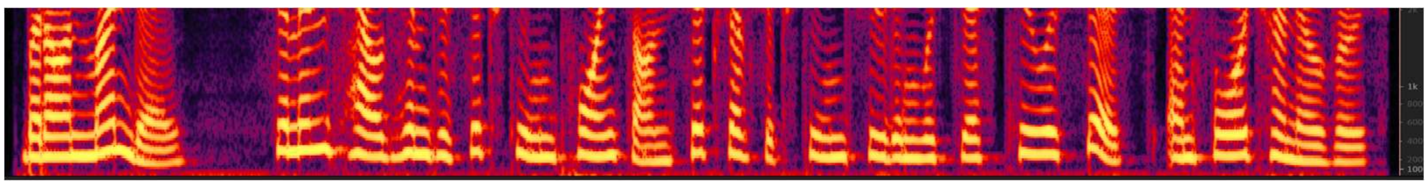
Take the transferring Spk-A to Spk-B for example. Model refined using Spk-A and Spk-B’s data.

- **Spk-A _self**: only use Spk-A for speaker id input during synthesis
- **Spk-B _trans**: use Spk-A id for prosody generation and Spk-B id during synthesis

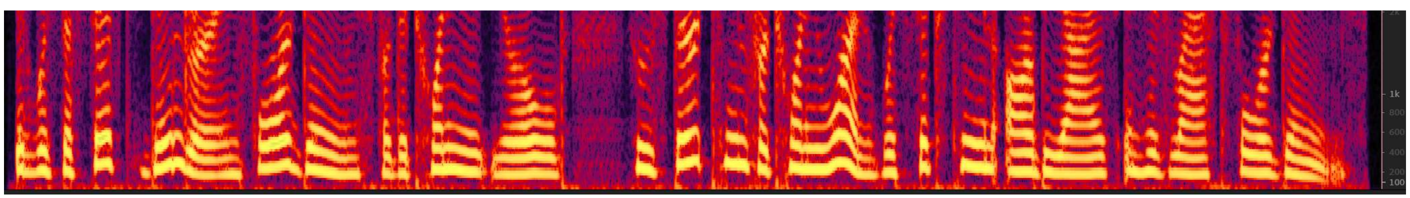
Spk-A_self



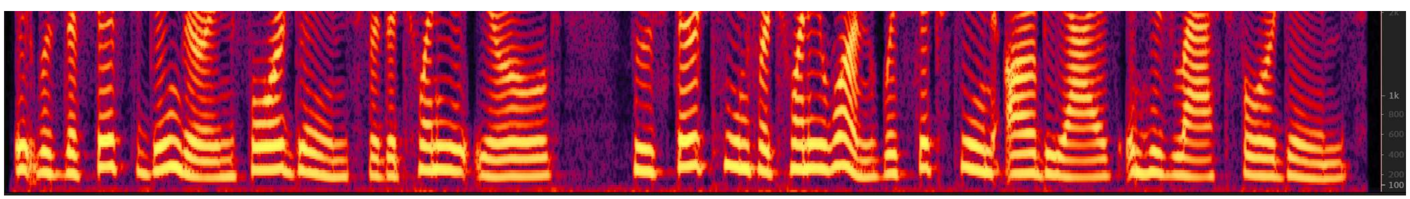
Spk-B _trans



Spk-A _self



Spk-B _trans



The transferred voice can **maintain** the source prosody very well.

Samples: https://peterpanseu.github.io/index.html

Experiment

Exp-1: training **from scratch** with sufficient data of both Source and Target speakers
Exp-2: **onboarding new source and target speakers** with less data, given a pretrained model (150h, 80 speakers, 10 styles, A/B excluded)

Experiment	Speaker	Neutral	Happy	Sad	Angry
1	A	10k	4k	4k	4k
1	B	10k	-	-	-
2	A	500	-	-	-
2	B	500	-	-	-

Voices:

- **Spk-A_Rec:** Speaker A's recording, held out for test.
- **Spk-A_SD:** Speaker A's SD model, viewed as the upper boundary for style evaluation.
- **Spk-B_SD:** Speaker B's SD model, viewed as the lower boundary for style evaluation.
- **Spk-B_Trans_CC:** A Transformer TTS version of the cycle consistency loss enhanced method in [11].
- **Spk-B_Trans_GMVAE:** GMVAE-based style transfer model
- **Spk-B_Trans_Pros:** The proposed model.

Table 4: *Prosody Measurement of Exp-1.*

Model	Lf0_ Corr	Dur_ Corr	Energy Corr	Lf0_ RMSE
Spk-A_SD	0.425	**0.848**	**0.915**	0.255
Spk-B_SD	0.226	0.749	0.659	0.302
Spk-B_Trans_CC	0.230	0.794	0.884	0.284
Spk-B_Trans_GMVAE	0.382	0.837	0.907	0.262
Spk-B_Trans_Pros	**0.439**	0.844	0.893	**0.237**

Table 5: *Prosody Measurement of Exp-2.*

Model	Lf0_ Corr	Dur_ Corr	Energy Corr	Lf0_ RMSE
Spk-A_SD	0.638	**0.843**	**0.931**	0.175
Spk-B_SD	0.402	0.786	0.892	0.222
Spk-B_Trans_GMVAE	0.503	0.821	0.91	0.198
Spk-B_Trans_Pros	**0.66**	0.842	0.917	**0.156**

Table 3: *MOS.*

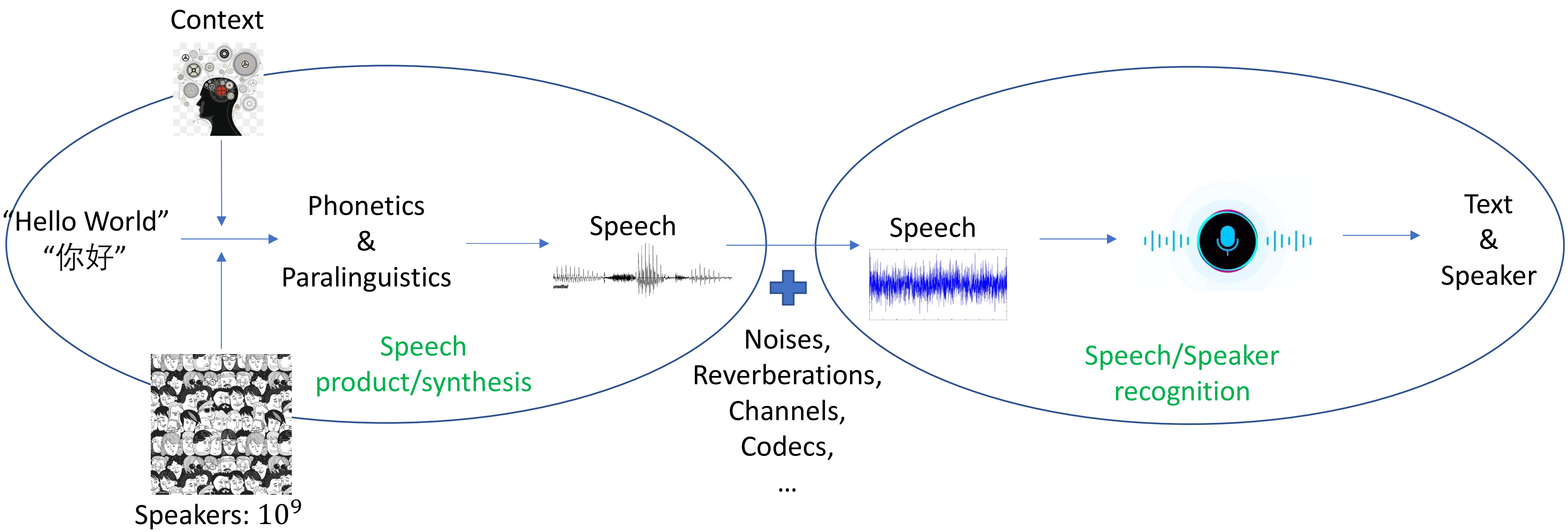
Model	Exp-1	Exp-2
Recording_Spk-B	4.29 ± 0.05	4.29 ± 0.07
Spk-B_SD	4.01 ± 0.05	4.06 ± 0.07
Spk-B_Trans_CC	$\mathbf{4.13} \pm 0.04$	-
Spk-B_Trans_GMVAE	4.08 ± 0.04	4.08 ± 0.06
Spk-B_Trans_Pros	4.07 ± 0.05	$\mathbf{4.11} \pm 0.07$

Table 7: *ABX (%).*

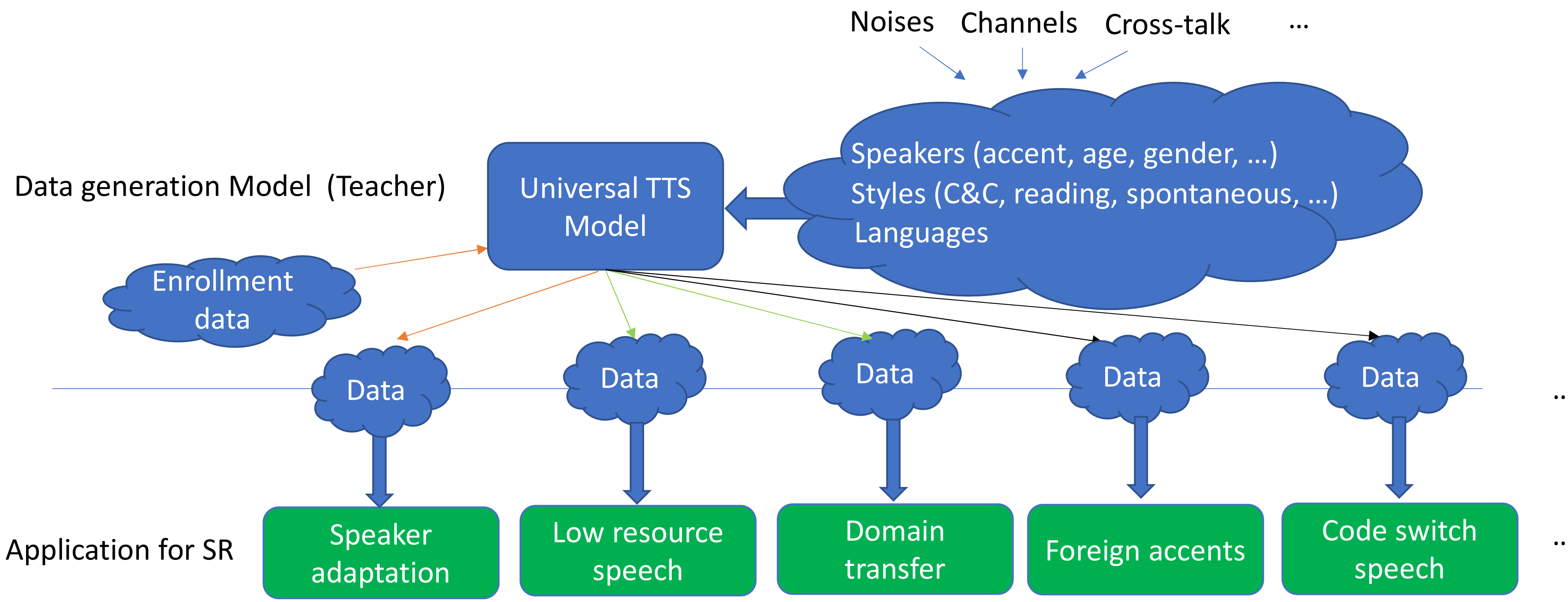
Exp	Spk-B_SD	Spk-B_Trans_CC	Spk-B_Trans_GMVAE	Spk-B_Trans_Pros	Equal
1	18.4			**77.7**	4.0
1		27.3		**62.2**	10.6
1			28.9	**63.2**	7.9
2	29.3			**66.3**	4.3
2			30.3	**63.9**	5.8

11. M. Whitehill, S. Ma, D. McDuff, Y. Song, "Multi-Reference Neural TTS Stylization with Adversarial Cycle Consistency," *INTERSPEECH 2020*

Speech production, transmission and recognition



SR data augmentation via Universal TTS model



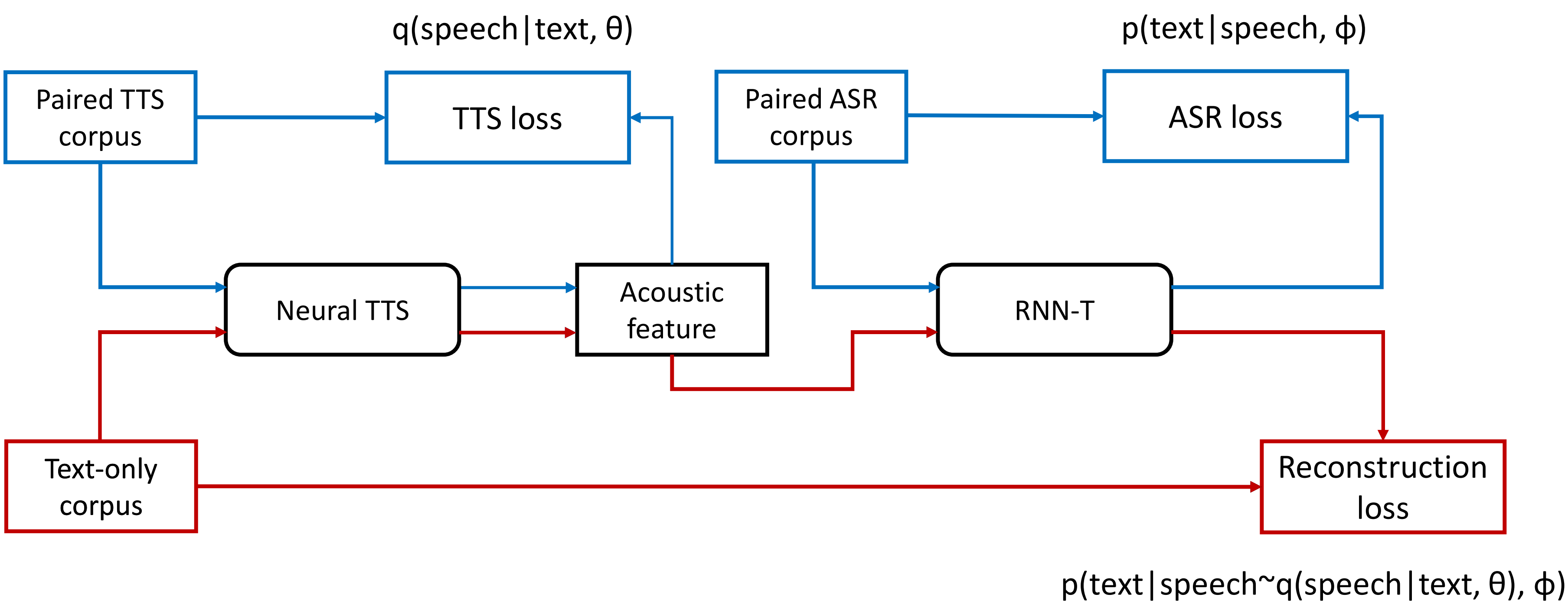
ICASSP2019: Using Personalized Speech Synthesis and Neural Language Generator for Rapid Speaker Adaptation
IS2020: Rapid RNN-T Adaptation Using Personalized Speech Synthesis and Neural Language Generator
ICASSP2020: Adaptation of RNN Transducer with text-to-speech technology for keyword spotting
IS2020: Developing RNN-T Models Surpassing High-Performance Hybrid Models with Customization Capability

E2E SR domain scaling by semi-supervised training with neural TTS

- Motivation
  - It’s challenging for an E2E SR model (e.g. RNN-T) in new domains with context or words different from training data.
  - Joint training with neural TTS is helpful for E2E ASR models in low-resource scenarios, how about the effectiveness in scaling a well-trained ASR model to new domains with text-only corpus?
  - Compared with other customization methods, are they complementary to each other or not?

Semi-supervised training with neural TTS

- Training framework



Results of RNN-T in domain scaling

- OOD task (with out-of-domain context):
  - Text-only corpus: ~75k sentences from a new domain, generated by randomly parsing the grammar and crowd sourcing.
  - Test set: 800 utterances from the same new domain.
- OOV task (with out-of-vocabulary words):
  - Text-only corpus: ~1.27M sentences generated using pre-designed OOV word list and pattern list.
  - Test set: 11k utterances from conversational data containing OOV words.

Table 1: *Performance of semi-supervised training in new domains. WERR is the relative WER reduction. OOD Task has out-of-domain context; OOV Task has out-of-vocabulary words.*

Method	WER (%)	WERR (%)
OOD Task		
Baseline	16.52	
+ Semi-supervised Training	6.37	61.4
OOV Task		
Baseline	27.50	
+ Semi-supervised Training	12.70	53.8

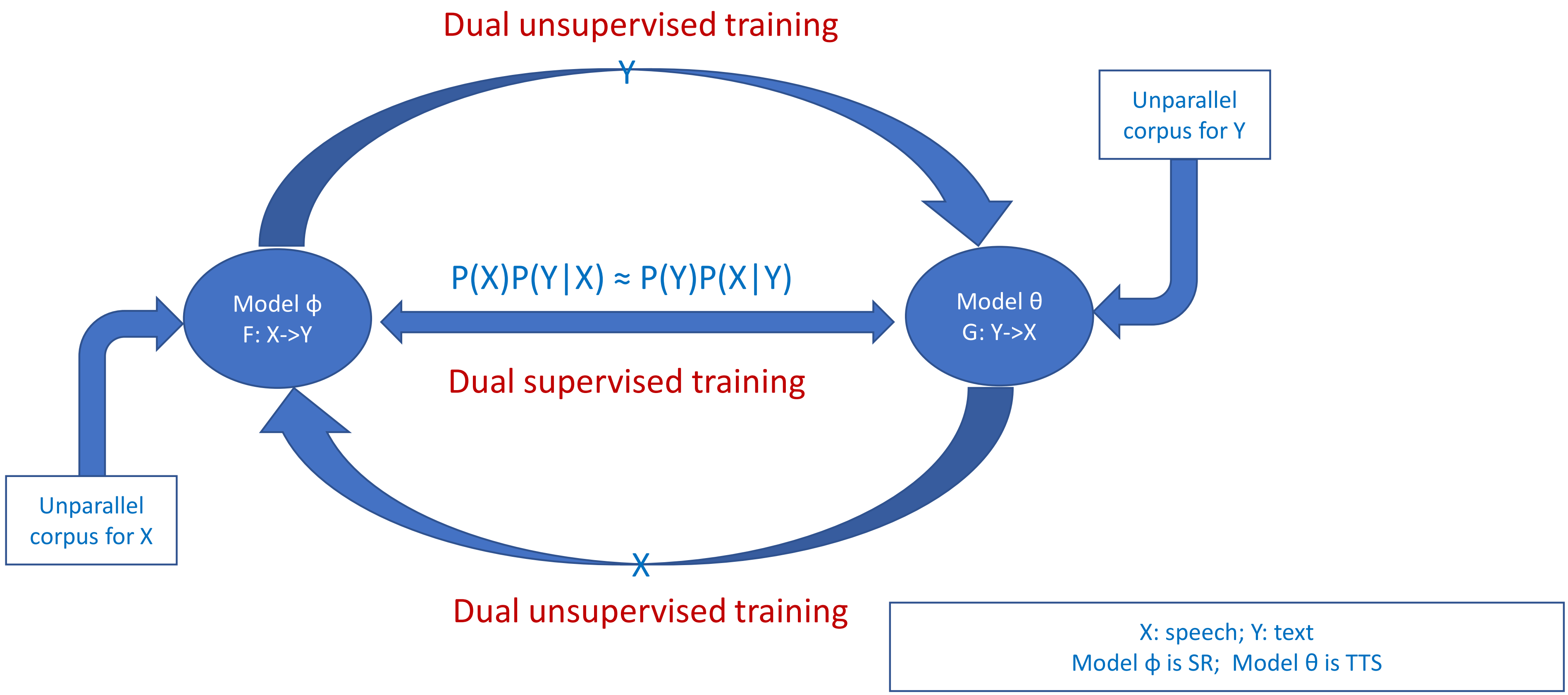
Table 5: *Comparison & combination of different customization methods. WERR is the relative WER reduction. OOD Task has out-of-domain context; OOV Task has out-of-vocabulary words.*

Method	WER (%)	WERR (%)
OOD Task		
Baseline	16.52	
+ Semi-supervised Training	6.37	61.4
+ ILME	5.71	65.4
+ Splicing Data	5.02	69.2
+ Semi-supervised Training	4.33	73.8
+ ILME	3.74	77.4
OOV Task		
Baseline	27.50	
+ Semi-supervised Training	12.70	53.8
+ Biasing	11.30	58.9

IS2021: Improving RNN_T for Domain Scaling Using Semi Supervised Training with Neural TTS

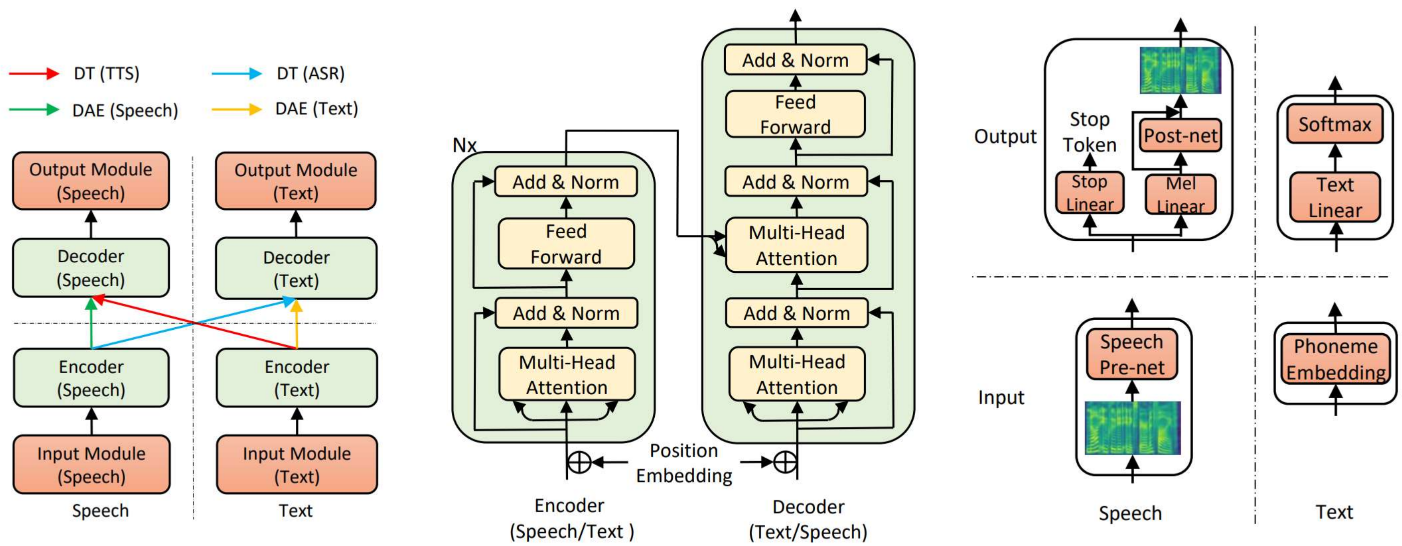
Dual learning – SR/TTS joint model

- Overall framework



Dual learning for low resource language development

ICML 2019: **Almost Unsupervised Text to Speech and Automatic Speech Recognition**



(a) Unified training flow (b) Encoder/Decoder for speech/text (c) Input/Output module for speech/text

Figure 1. The overall model structure for TTS and ASR. Figure (a): The unified training flow of our method, which consists of a denoising auto-encoder (DAE) of speech and text, and dual transformation (DT) of TTS and ASR, both with bidirectional sequence modeling. Figure (b): The speech and text encoder and decoder based on Transformer. Figure (c): The input and output module for speech and text.

A following research: KDD2020, LRSpeech: Extremely Low-Resource Speech Synthesis and Recognition

Thanks！

Q&A

helei@microsoft.com