A Survey on Neural Text to Speech Synthesis

https://arxiv.org/pdf/2106.15561.pdf

Xu Tan
Microsoft Research Asia

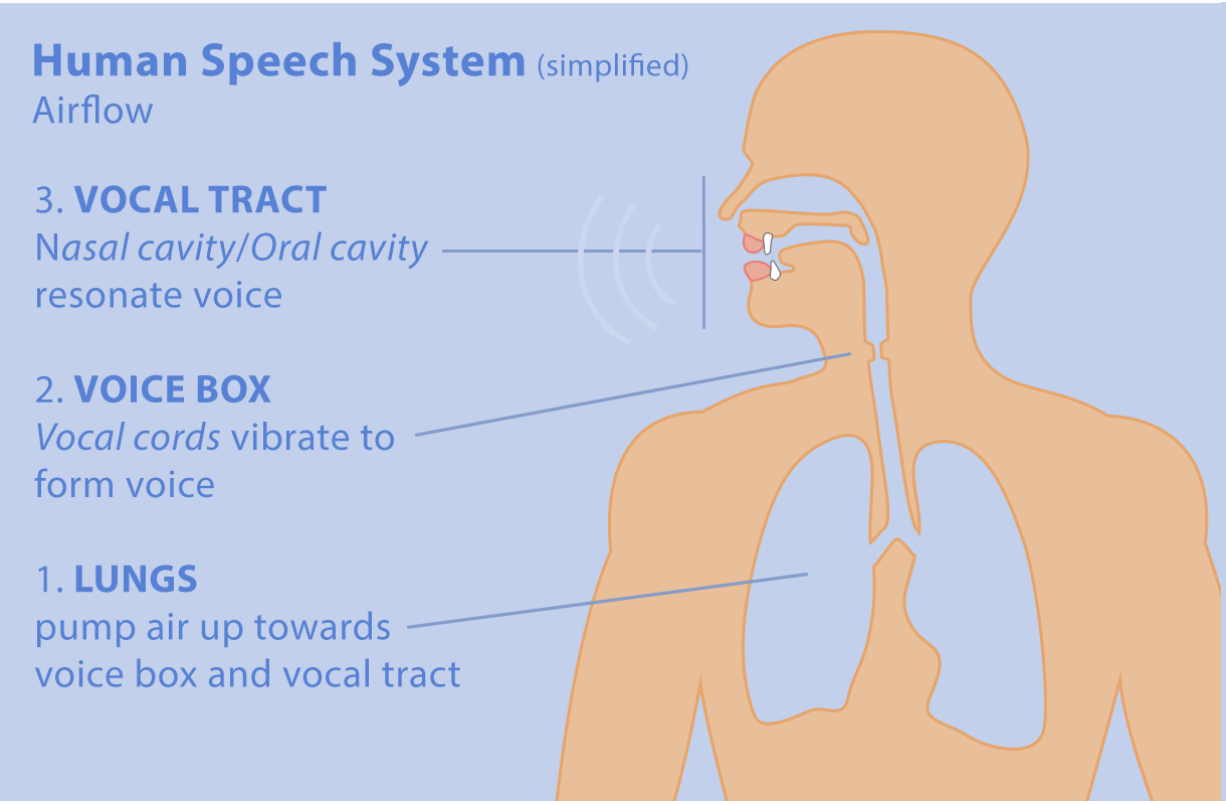
xuta@microsoft.com

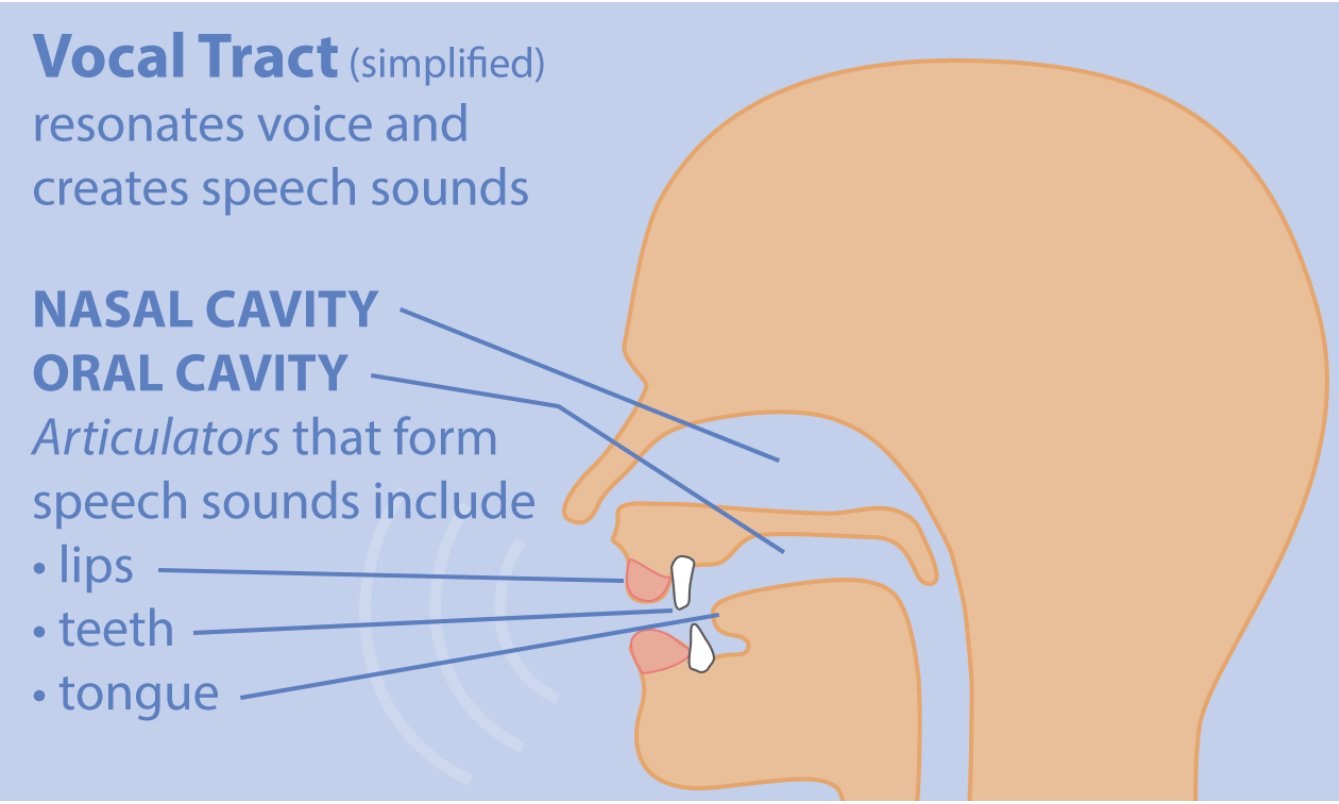
Self-introduction

- Xu Tan (谭旭)
- Senior Researcher @ Machine Learning Group, Microsoft Research Asia
- Research interests: deep learning and its applications on NLP/Speech/Music
  - Text to speech
  - Automatic speech recognition
  - Neural machine translation
  - Language/speech pre-training
  - Music understanding and generation
- Homepage: https://www.microsoft.com/en-us/research/people/xuta/
- Speech related research: https://speechresearch.github.io/

Text to speech synthesis

- The artificial production of human speech from text
  - Human speech system



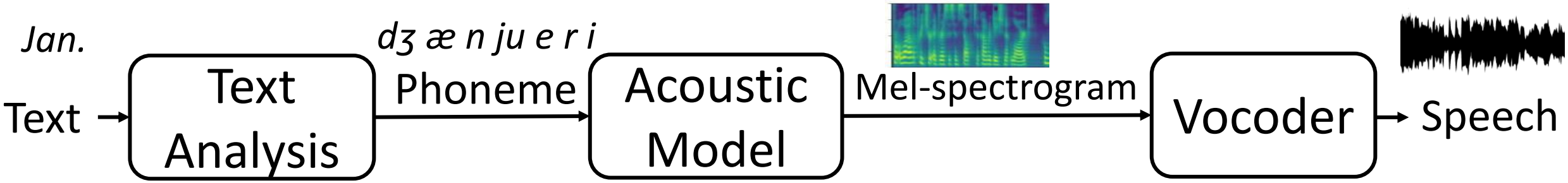


Overview of TTS technologies

- Articulatory Speech Synthesis, mid 1970s~
- Formant Speech Synthesis, late 1970s~
- Concatenative Speech Synthesis, 1990s~
- Statistical Parametric Speech Synthesis, 1995~

Overview of TTS technologies

- Articulatory Speech Synthesis, mid 1970s~
- Formant Speech Synthesis, late 1970s~
- Concatenative Speech Synthesis, 1990s~
- Statistical Parametric Speech Synthesis, 1995~
- Neural Speech Synthesis, 2013~
  - Neural based Statistical Parametric Speech Synthesis, 2013~
  - End-to-End Neural Speech Synthesis, 2016~



Overview of TTS technologies

- Articulatory Speech Synthesis, mid 1970s~
- Formant Speech Synthesis, late 1970s~
- Concatenative Speech Synthesis, 1990s~
- Statistical Parametric Speech Synthesis, 1995~
- Neural Speech Synthesis, 2013~
  - Neural based Statistical Parametric Speech Synthesis, 2013~
  - End-to-End Neural Speech Synthesis, 2016~

Concatenative

Statistical parametric (HMM)

Neural (Tacotron 2)

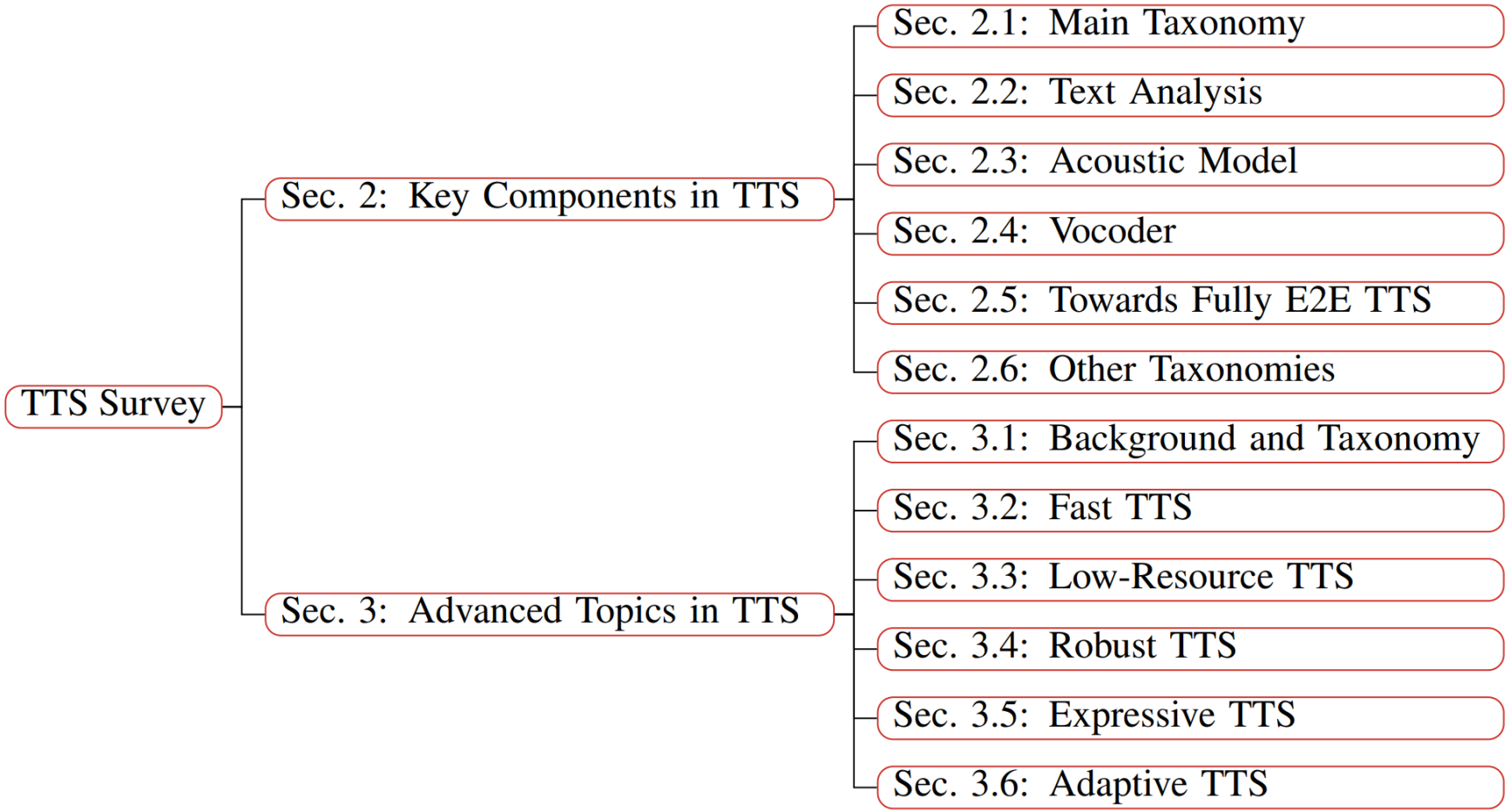
Neural (FastSpeech 2)

This survey paper

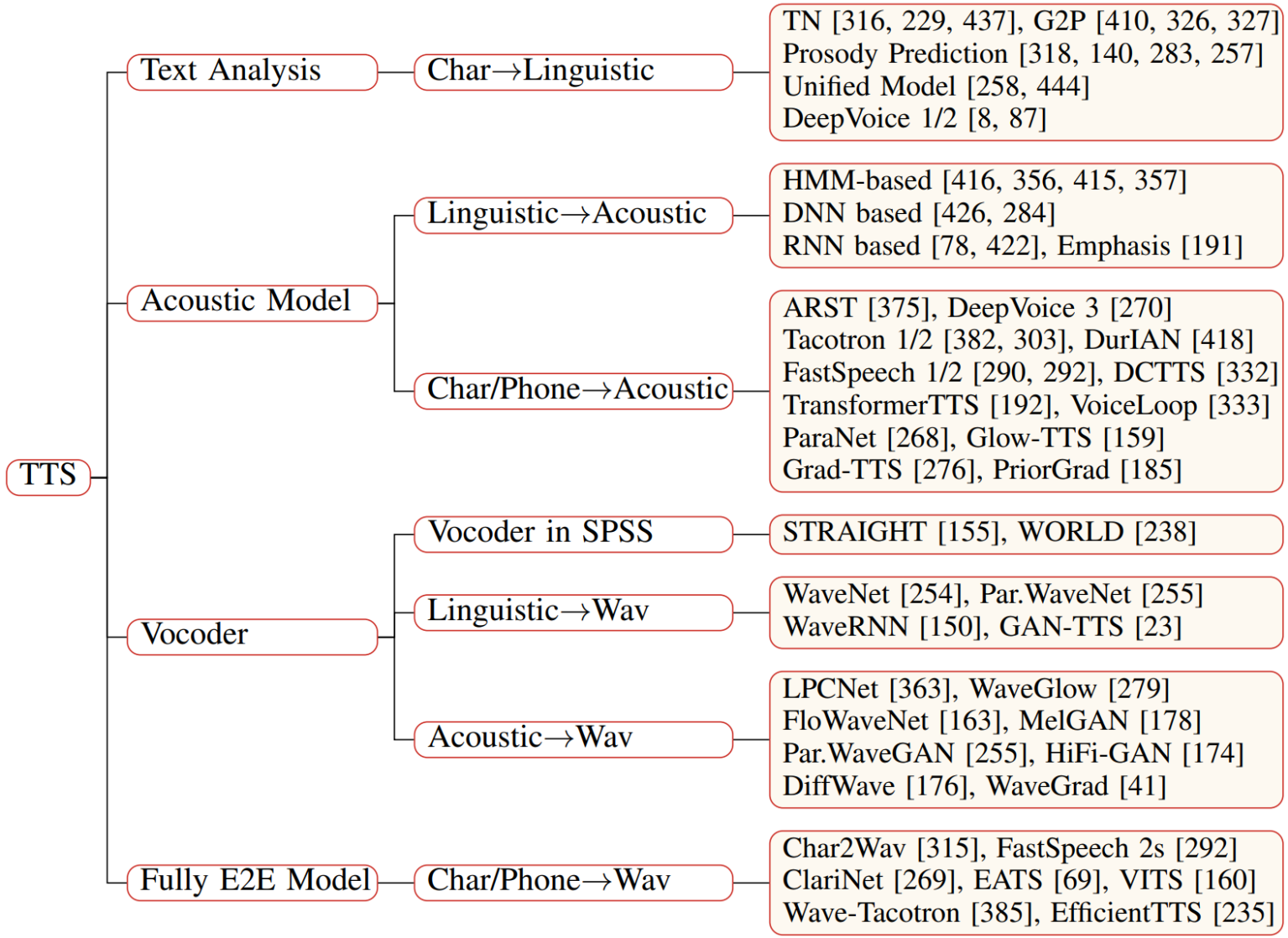
- Neural Speech Synthesis, 2013~
  - Neural based Statistical Parametric Speech Synthesis, 2013~
  - End-to-End Neural Speech Synthesis, 2016~

- A Survey on Neural Speech Synthesis
  - https://arxiv.org/pdf/2106.15561.pdf
  - 457 references, 63 pages

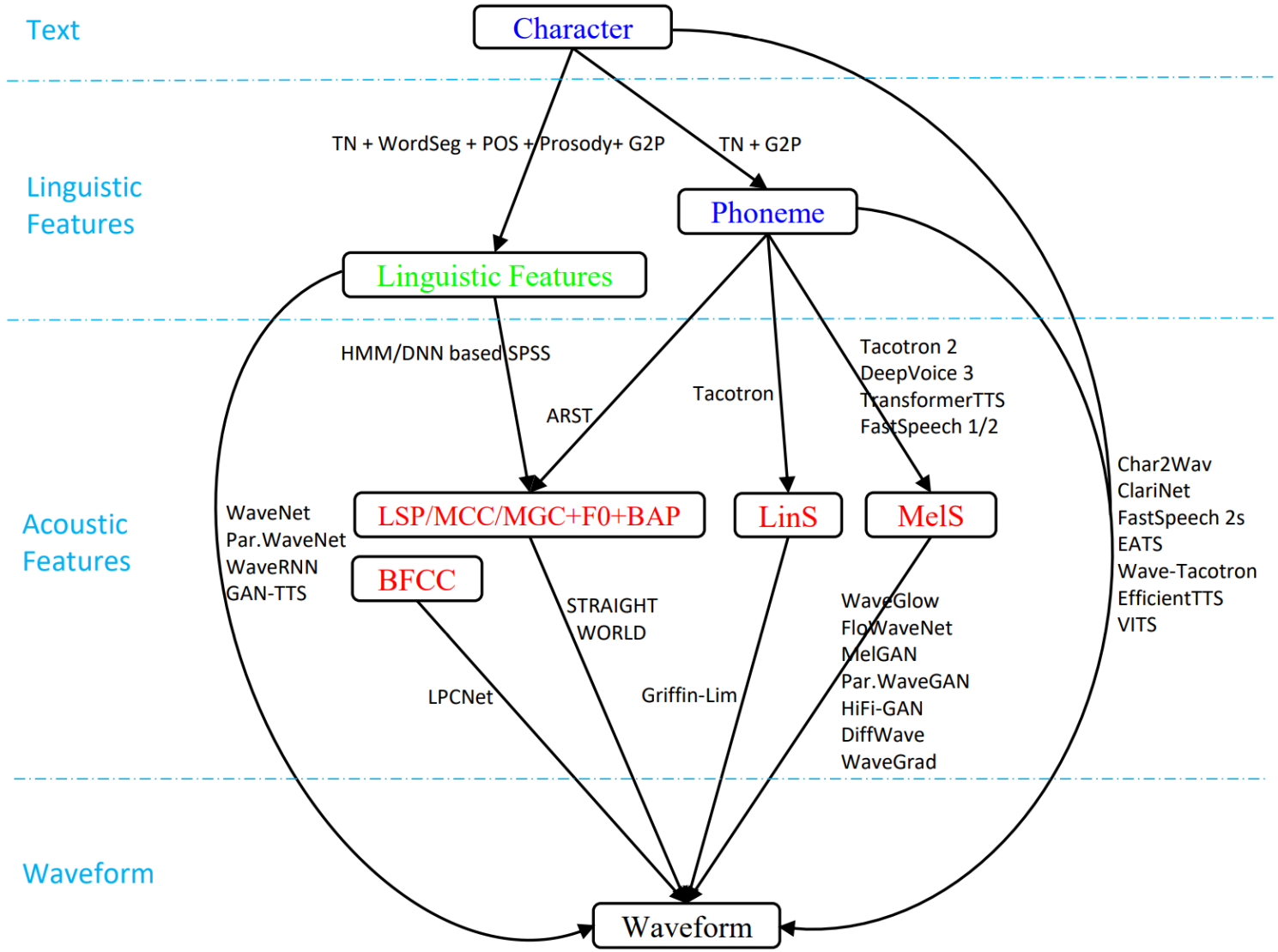
Organization of this survey paper



Key components in TTS



Key components in TTS



Text analysis

Task	Research Work
Text Normalization	Rule-based [317], Neural-based [316, 229, 413, 437], Hybrid [439]
Word Segmentation	[400, 451, 267]
POS Tagging	[298, 329, 227, 451, 138]
Prosody Prediction	[51, 412, 318, 190, 140, 328, 283, 64, 447, 216, 218, 3]
Grapheme to Phoneme	N-gram [42, 24], Neural-based [410, 289, 33, 326]
- - Polyphone Disambiguation	[448, 398, 230, 301, 327, 29, 263]

Text analysis

- Transforms input text into linguistic features, including
  - Text normalization
    - 1989 → nineteen eighty nine, *Jan. 24th → January twenty-fourth*
  - Phrase/word/syllable segmentation
    - synthesis → syn-the-sis
  - Part of speech (POS) tagging
    - Mary went to the store → noun, verb, prep, noun,
  - Prosody prediction
    - e.g., ToBI (Tones and Break Indices)
    - Mary went to the store ? → Mary’ store’ H%
  - Grapheme-to-phoneme conversion
    - *Speech → s p iy ch*
    - Polyphone disambiguation

Acoustic model

- Acoustic model in SPSS
- Acoustic models in end-to-end TTS
  - RNN-based (e.g., Tacotron series)
  - CNN-based (e.g., DeepVoice series)
  - Transformer-based (e.g., FastSpeech series)
  - Other (e.g., Flow, GAN, VAE, Diffusion)

Acoustic Model	Input→Output	AR/NAR	Modeling	Structure
HMM-based [416, 356]	Ling→MCC+F0	/	/	HMM
DNN-based [426]	Ling→MCC+BAP+F0	NAR	/	DNN
LSTM-based [78]	Ling→LSP+F0	AR	/	RNN
EMPHASIS [191]	Ling→LinS+CAP+F0	AR	/	Hybrid
ARST [375]	Ph→LSP+BAP+F0	AR	Seq2Seq	RNN
VoiceLoop [333]	Ph→MGC+BAP+F0	AR	/	hybrid
Tacotron [382]	Ch→LinS	AR	Seq2Seq	Hybrid/RNN
Tacotron 2 [303]	Ch→MelS	AR	Seq2Seq	RNN
DurIAN [418]	Ph→MelS	AR	Seq2Seq	RNN
Non-Att Tacotron [304]	Ph→MelS	AR	/	Hybrid/CNN/RNN
Para. Tacotron 1/2 [74, 75]	Ph→MelS	NAR	/	Hybrid/Self-Att/CNN
MelNet [367]	Ch→MelS	AR	/	RNN
DeepVoice [8]	Ch/Ph→MelS	AR	/	CNN
DeepVoice 2 [87]	Ch/Ph→MelS	AR	/	CNN
DeepVoice 3 [270]	Ch/Ph→MelS	AR	Seq2Seq	CNN
ParaNet [268]	Ph→MelS	NAR	Seq2Seq	CNN
DCTTS [332]	Ch→MelS	AR	Seq2Seq	CNN
SpeedySpeech [361]	Ph→MelS	NAR	/	CNN
TalkNet 1/2 [19, 18]	Ch→MelS	NAR	/	CNN
TransformerTTS [192]	Ph→MelS	AR	Seq2Seq	Self-Att
MultiSpeech [39]	Ph→MelS	AR	Seq2Seq	Self-Att
FastSpeech 1/2 [290, 292]	Ph→MelS	NAR	Seq2Seq	Self-Att
AlignTTS [429]	Ch/Ph→MelS	NAR	Seq2Seq	Self-Att
JDI-T [197]	Ph→MelS	NAR	Seq2Seq	Self-Att
FastPitch [181]	Ph→MelS	NAR	Seq2Seq	Self-Att
AdaSpeech 1/2/3 [40, 403, 404]	Ph→MelS	NAR	Seq2Seq	Self-Att
DenoiSpeech [434]	Ph→MelS	NAR	Seq2Seq	Self-Att
DeviceTTS [126]	Ph→MelS	NAR	/	Hybrid/DNN/RNN
LightSpeech [220]	Ph→MelS	NAR	/	Hybrid/Self-Att/CNN
Flow-TTS [234]	Ch/Ph→MelS	NAR*	Flow	Hybrid/CNN/RNN
Glow-TTS [159]	Ph→MelS	NAR	Flow	Hybrid/Self-Att/CNN
Flowtron [366]	Ph→MelS	AR	Flow	Hybrid/RNN
EfficientTTS [235]	Ch→MelS	NAR	Flow	Hybrid/CNN
GMVAE-Tacotron [119]	Ph→MelS	AR	VAE	Hybrid/RNN
VAE-TTS [443]	Ph→MelS	AR	VAE	Hybrid/RNN
BVAE-TTS [187]	Ph→MelS	NAR	VAE	CNN
GAN exposure [99]	Ph→MelS	AR	GAN	Hybrid/RNN
TTS-Stylization [224]	Ch→MelS	AR	GAN	Hybrid/RNN
Multi-SpectroGAN [186]	Ph→MelS	NAR	GAN	Hybrid/Self-Att/CNN
Diff-TTS [141]	Ph→MelS	NAR*	Diffusion	Hybrid/CNN
Grad-TTS [276]	Ph→MelS	NAR	Diffusion	Hybrid/Self-Att/CNN
PriorGrad [185]	Ph→MelS	NAR	Diffusion	Hybrid/Self-Att/CNN

Vocoder

- Autoregressive vocoder
- Flow-based vocoder

Flow		Evaluation $z = f^{-1}(x)$	Synthesis $x = f(z)$
AR	AF [261]	$z_t = x_t \cdot \sigma_t(x_{<t};\theta) + \mu_t(x_{<t};\theta)$	$x_t = \frac{z_t - u_t(x_{<t};\theta)}{\sigma_t(x_{<t};\theta)}$
	IAF [169]	$z_t = \frac{x_t - \mu_t(z_{<t};\theta)}{\sigma_t(z_{<t};\theta)}$	$x_t = z_t \cdot \sigma_t(z_{<t};\theta) + \mu_t(z_{<t};\theta)$
Bipartite	RealNVP [66]	$z_a = x_a,$	$x_a = z_a,$
	Glow [167]	$z_b = x_b \cdot \sigma_b(x_a;\theta) + \mu_b(x_a;\theta)$	$x_b = \frac{z_b - \mu_b(x_a;\theta)}{\sigma_b(x_a;\theta)}$

Vocoder	Input	AR/NAR	Modeling	Architecture
WaveNet [254]	Linguistic Feature	AR	/	CNN
SampleRNN [233]	/	AR	/	RNN
WaveRNN [150]	Linguistic Feature	AR	/	RNN
LPCNet [363]	BFCC	AR	/	RNN
Univ. WaveRNN [215]	Mel-Spectrogram	AR	/	RNN
SC-WaveRNN [265]	Mel-Spectrogram	AR	/	RNN
MB WaveRNN [418]	Mel-Spectrogram	AR	/	RNN
FFTNet [145]	Cepstrum	AR	/	CNN
Par. WaveNet [255]	Linguistic Feature	NAR	Flow	CNN
WaveGlow [279]	Mel-Spectrogram	NAR	Flow	Hybrid/CNN
FloWaveNet [163]	Mel-Spectrogram	NAR	Flow	Hybrid/CNN
WaveFlow [271]	Mel-Spectrogram	AR	Flow	Hybrid/CNN
SqueezeWave [433]	Mel-Spectrogram	NAR	Flow	CNN
WaveGAN [68]	/	NAR	GAN	CNN
GELP [149]	Mel-Spectrogram	NAR	GAN	CNN
GAN-TTS [23]	Linguistic Feature	NAR	GAN	CNN
MelGAN [178]	Mel-Spectrogram	NAR	GAN	CNN
Par. WaveGAN [402]	Mel-Spectrogram	NAR	GAN	CNN
HiFi-GAN [174]	Mel-Spectrogram	NAR	GAN	Hybrid/CNN
VocGAN [408]	Mel-Spectrogram	NAR	GAN	CNN
GED [96]	Linguistic Feature	NAR	GAN	CNN
Fre-GAN [161]	Mel-Spectrogram	NAR	GAN	CNN
Wave-VAE [268]	Mel-Spectrogram	NAR	VAE	CNN
WaveGrad [41]	Mel-Spectrogram	NAR	Diffusion	Hybrid/CNN
DiffWave [176]	Mel-Spectrogram	NAR	Diffusion	Hybrid/CNN
PriorGrad [185]	Mel-Spectrogram	NAR	Diffusion	Hybrid/CNN

Vocoder

- Autoregressive vocoder
- Flow-based vocoder
- GAN-based vocoder

GAN	Generator	Discriminator	Loss
WaveGAN [68]	DCGAN [287]	/	WGAN-GP [97]
GAN-TTS [23]	/	Random Window D	Hinge-Loss GAN [198]
MelGAN [178]	/	Multi-Scale D	LS-GAN [231] Feature Matching Loss [182]
Par.WaveGAN [402]	WaveNet [254]	/	LS-GAN, Multi-STFT Loss
HiFi-GAN [174]	Multi-Receptive Field Fusion	Multi-Period D, Multi-Scale D	LS-GAN, STFT Loss, Feature Matching Loss
VocGAN [408]	Multi-Scale G	Hierarchical D	LS-GAN, Multi-STFT Loss, Feature Matching Loss
GED [96]	/	Random Window D	Hinge-Loss GAN, Repulsive loss

Vocoder	Input	AR/NAR	Modeling	Architecture
WaveNet [254]	Linguistic Feature	AR	/	CNN
SampleRNN [233]	/	AR	/	RNN
WaveRNN [150]	Linguistic Feature	AR	/	RNN
LPCNet [363]	BFCC	AR	/	RNN
Univ. WaveRNN [215]	Mel-Spectrogram	AR	/	RNN
SC-WaveRNN [265]	Mel-Spectrogram	AR	/	RNN
MB WaveRNN [418]	Mel-Spectrogram	AR	/	RNN
FFTNet [145]	Cepstrum	AR	/	CNN
Par. WaveNet [255]	Linguistic Feature	NAR	Flow	CNN
WaveGlow [279]	Mel-Spectrogram	NAR	Flow	Hybrid/CNN
FloWaveNet [163]	Mel-Spectrogram	NAR	Flow	Hybrid/CNN
WaveFlow [271]	Mel-Spectrogram	AR	Flow	Hybrid/CNN
SqueezeWave [433]	Mel-Spectrogram	NAR	Flow	CNN
WaveGAN [68]	/	NAR	GAN	CNN
GELP [149]	Mel-Spectrogram	NAR	GAN	CNN
GAN-TTS [23]	Linguistic Feature	NAR	GAN	CNN
MelGAN [178]	Mel-Spectrogram	NAR	GAN	CNN
Par. WaveGAN [402]	Mel-Spectrogram	NAR	GAN	CNN
HiFi-GAN [174]	Mel-Spectrogram	NAR	GAN	Hybrid/CNN
VocGAN [408]	Mel-Spectrogram	NAR	GAN	CNN
GED [96]	Linguistic Feature	NAR	GAN	CNN
Fre-GAN [161]	Mel-Spectrogram	NAR	GAN	CNN
Wave-VAE [268]	Mel-Spectrogram	NAR	VAE	CNN
WaveGrad [41]	Mel-Spectrogram	NAR	Diffusion	Hybrid/CNN
DiffWave [176]	Mel-Spectrogram	NAR	Diffusion	Hybrid/CNN
PriorGrad [185]	Mel-Spectrogram	NAR	Diffusion	Hybrid/CNN

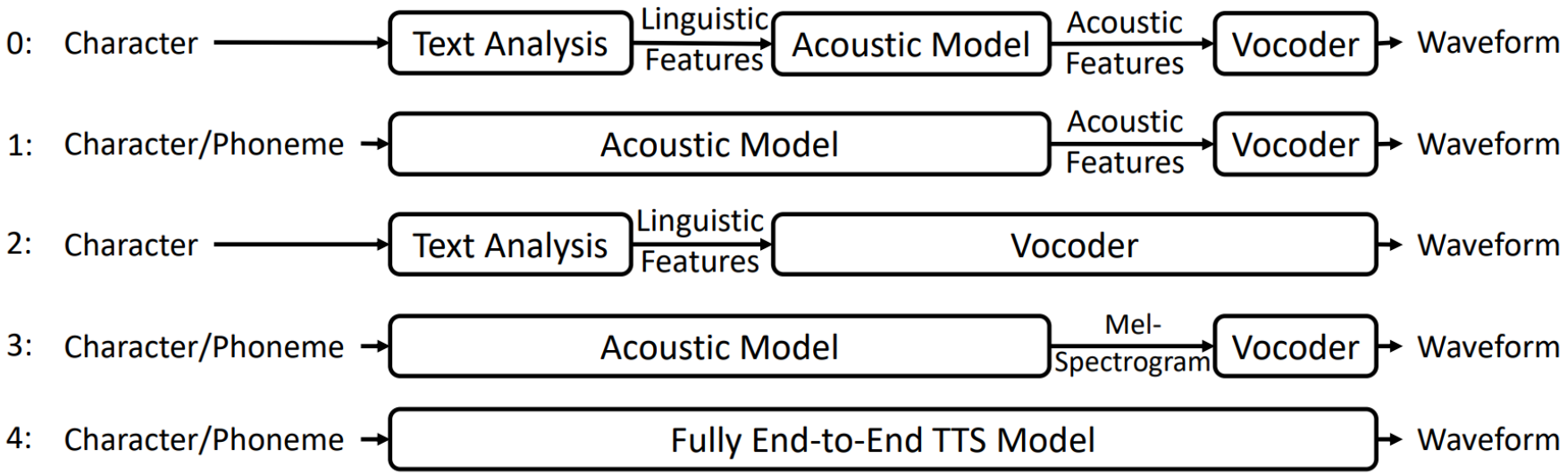
Vocoder

- Autoregressive vocoder
- Flow-based vocoder
- GAN-based vocoder
- VAE-based vocoder
- Diffusion-based vocoder

Generative Model	AR	VAE	Flow/AR	Flow/Bipartite	Diffusion	GAN
Vocoder (e.g.)	WaveNet	WaveVAE	Par.WaveNet	WaveGlow	DiffWave	MelGAN
Simple	Y	N	N	N	N	N
Parallel	N	Y	Y	Y	Y	Y
Latent Manipulate	N	Y	Y	Y	Y	Y*
Likelihood Estimate	Y	Y	Y	Y	Y	N

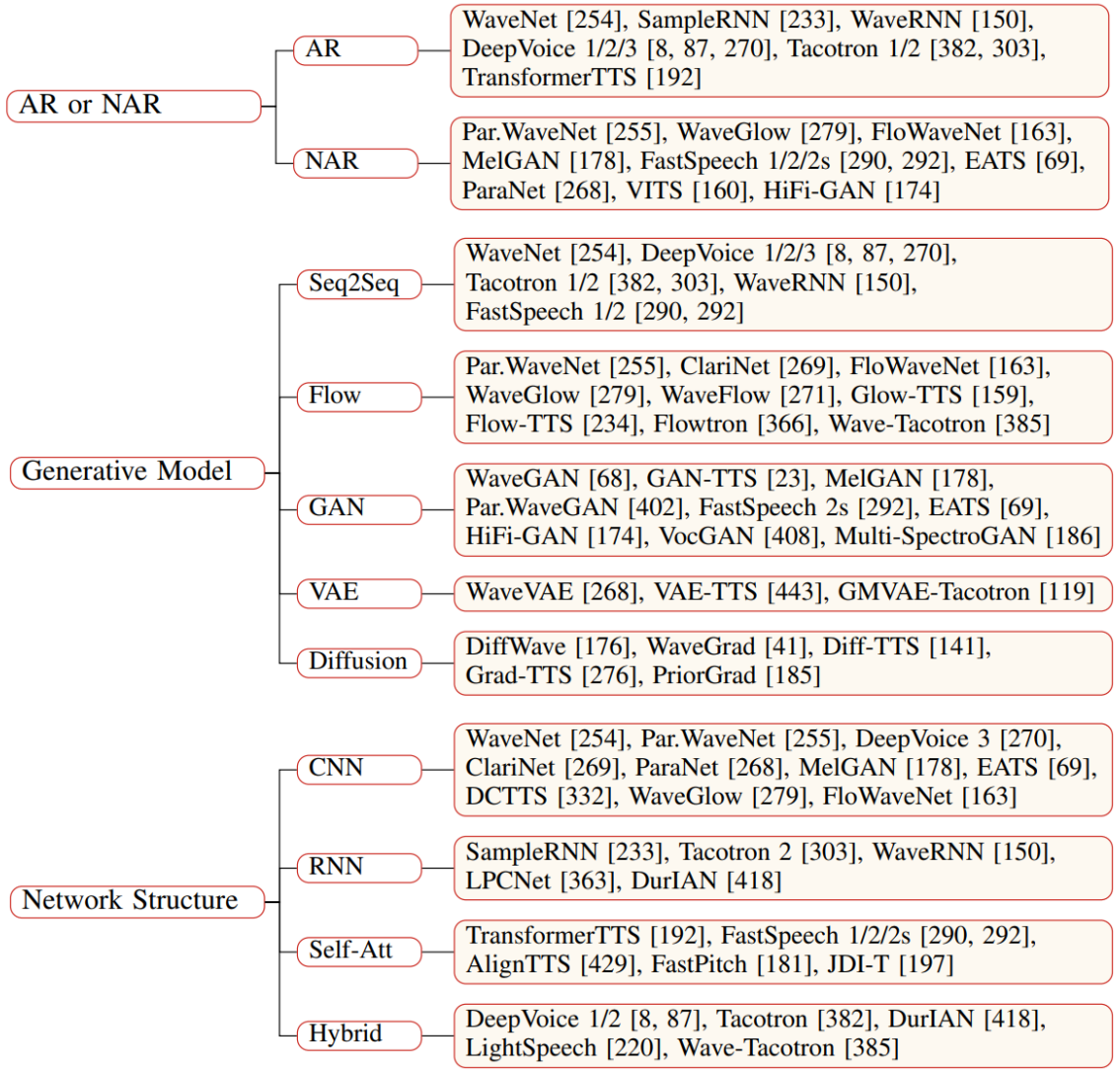
Vocoder	Input	AR/NAR	Modeling	Architecture
WaveNet [254]	Linguistic Feature	AR	/	CNN
SampleRNN [233]	/	AR	/	RNN
WaveRNN [150]	Linguistic Feature	AR	/	RNN
LPCNet [363]	BFCC	AR	/	RNN
Univ. WaveRNN [215]	Mel-Spectrogram	AR	/	RNN
SC-WaveRNN [265]	Mel-Spectrogram	AR	/	RNN
MB WaveRNN [418]	Mel-Spectrogram	AR	/	RNN
FFTNet [145]	Cepstrum	AR	/	CNN
Par. WaveNet [255]	Linguistic Feature	NAR	Flow	CNN
WaveGlow [279]	Mel-Spectrogram	NAR	Flow	Hybrid/CNN
FloWaveNet [163]	Mel-Spectrogram	NAR	Flow	Hybrid/CNN
WaveFlow [271]	Mel-Spectrogram	AR	Flow	Hybrid/CNN
SqueezeWave [433]	Mel-Spectrogram	NAR	Flow	CNN
WaveGAN [68]	/	NAR	GAN	CNN
GELP [149]	Mel-Spectrogram	NAR	GAN	CNN
GAN-TTS [23]	Linguistic Feature	NAR	GAN	CNN
MelGAN [178]	Mel-Spectrogram	NAR	GAN	CNN
Par. WaveGAN [402]	Mel-Spectrogram	NAR	GAN	CNN
HiFi-GAN [174]	Mel-Spectrogram	NAR	GAN	Hybrid/CNN
VocGAN [408]	Mel-Spectrogram	NAR	GAN	CNN
GED [96]	Linguistic Feature	NAR	GAN	CNN
Fre-GAN [161]	Mel-Spectrogram	NAR	GAN	CNN
Wave-VAE [268]	Mel-Spectrogram	NAR	VAE	CNN
WaveGrad [41]	Mel-Spectrogram	NAR	Diffusion	Hybrid/CNN
DiffWave [176]	Mel-Spectrogram	NAR	Diffusion	Hybrid/CNN
PriorGrad [185]	Mel-Spectrogram	NAR	Diffusion	Hybrid/CNN

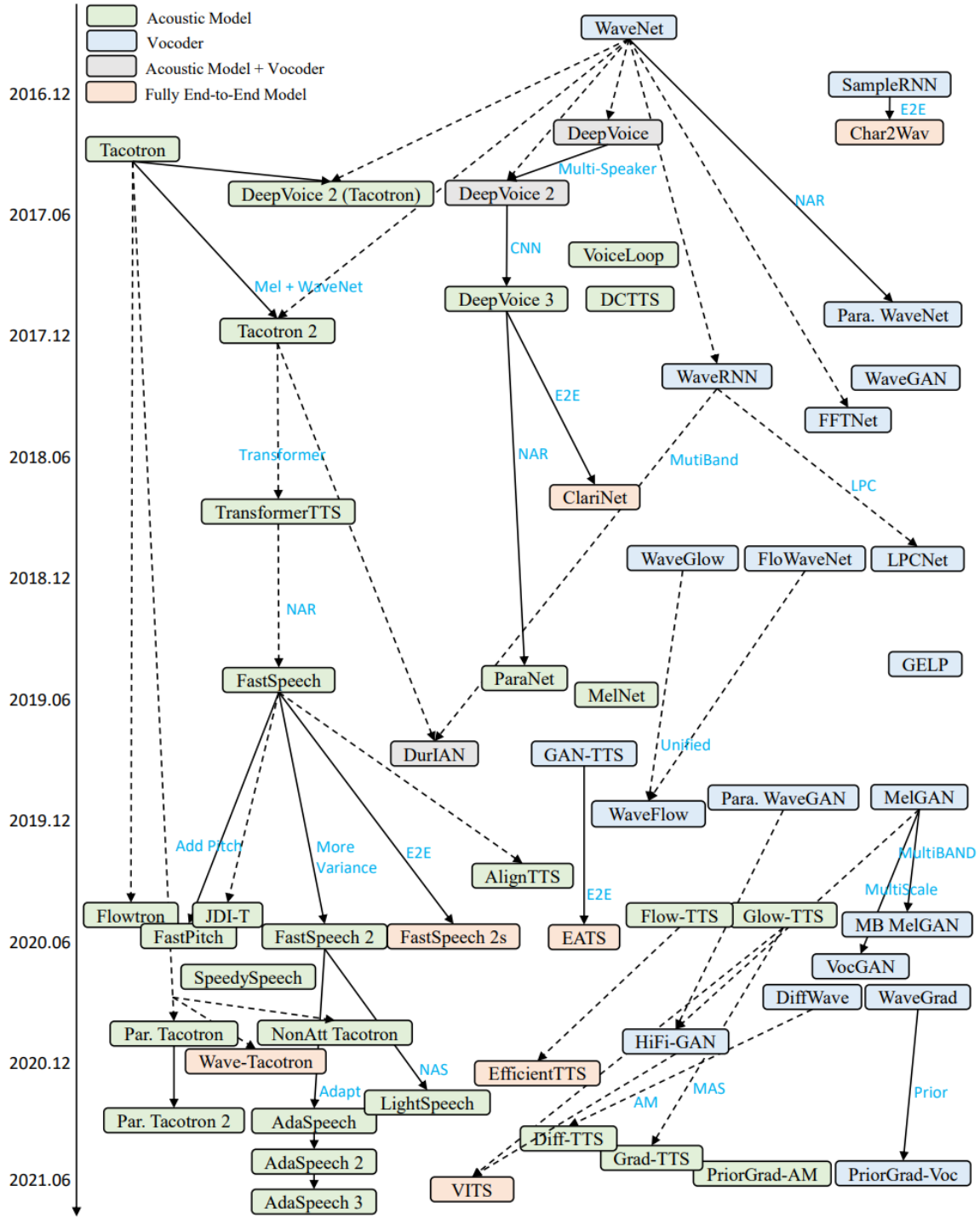
Towards Fully End-to-End TTS



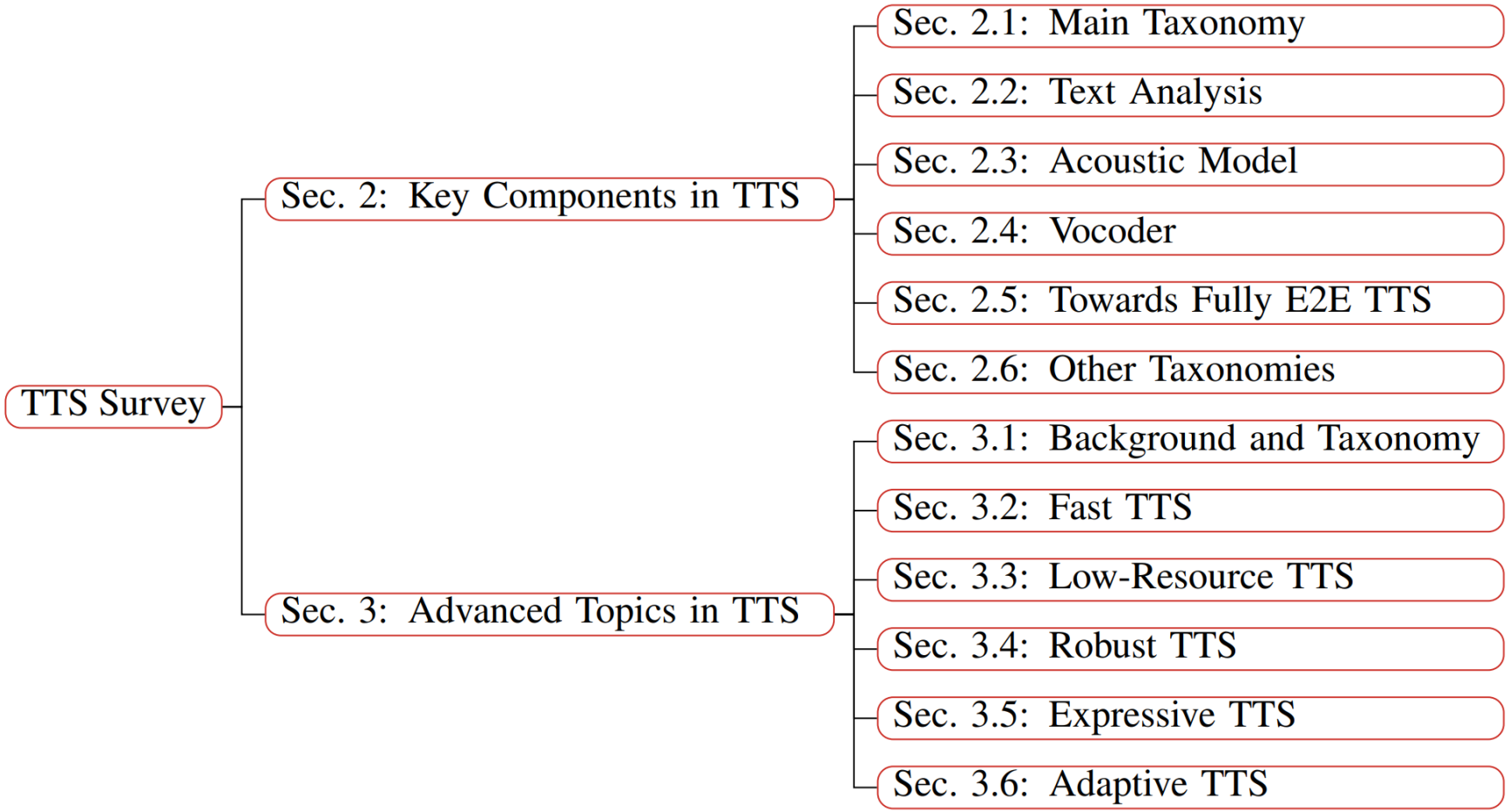
Stage	Models
0	SPSS [416, 356, 415, 425, 357]
1	ARST [375]
2	WaveNet [254], DeepVoice 1/2 [8, 87], Par. WaveNet [255], WaveRNN [150], HiFi-GAN [23]
3	DeepVoice 3 [270], Tacotron 2 [303], FastSpeech 1/2 [290, 292], WaveGlow [279], FloWaveNet [163]
4	Char2Wav [315], ClariNet [269], FastSpeech 2s [292], EATS [69], Wave-Tacotron [385], VITS [160]

Other taxonomies



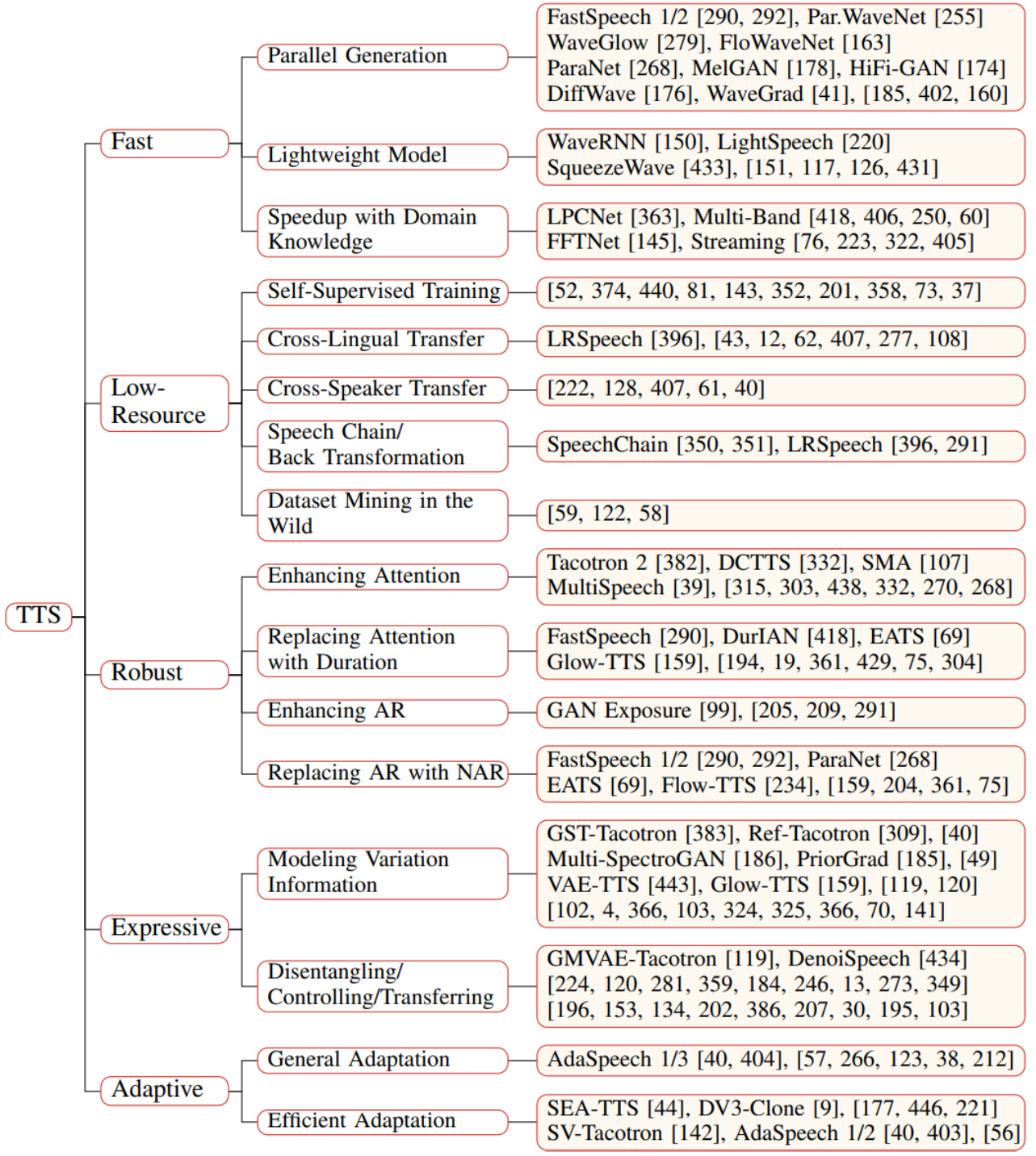

Acoustic Model
Vocoder
Acoustic Model + Vocoder
Fully End-to-End Model
2016.12
2017.06
2017.12
2018.06
2018.12
2019.06
2019.12
2020.06
2020.12
2021.06
WaveNet
SampleRNN
E2E
Char2Wav
DeepVoice
Multi-Speaker
Tacotron
DeepVoice 2 (Tacotron)
DeepVoice 2
NAR
CNN
VoiceLoop
Mel + WaveNet
DeepVoice 3
DCTTS
Para. WaveNet
Tacotron 2
WaveRNN
WaveGAN
E2E
FFTNet
Transformer
NAR
MutiBand
ClariNet
LPC
TransformerTTS
WaveGlow
FloWaveNet
LPCNet
NAR
GELP
FastSpeech
ParaNet
MelNet
DurIAN
GAN-TTS
Unified
WaveFlow
Para. WaveGAN
MelGAN
Add Pitch
More Variance
E2E
AlignTTS
E2E
MultiBAND
MultiScale
Flowtron
JDI-T
FastPitch
FastSpeech 2
FastSpeech 2s
EATS
Flow-TTS
Glow-TTS
MB MelGAN
VocGAN
SpeedySpeech
DiffWave
WaveGrad
Par. Tacotron
NonAtt Tacotron
HiFi-GAN
Wave-Tacotron
NAS
EfficientTTS
Adapt
LightSpeech
AM
MAS
Prior
Par. Tacotron 2
AdaSpeech
Diff-TTS
AdaSpeech 2
Grad-TTS
PriorGrad-AM
PriorGrad-Voc
VITS
AdaSpeech 3

Organization of this survey paper



Advanced topics in TTS

- Fast TTS
- Low-resource TTS
- Robust TTS
- Expressive TTS
- Adaptive TTS



Fast TTS

- Parallel generation

Modeling Paradigm	TTS Model	Training	Inference
AR (RNN)	Tacotron 1/2, SampleRNN, LPCNet	$\mathcal{O}(N)$	$\mathcal{O}(N)$
AR (CNN/Self-Att)	DeepVoice 3, TransformerTTS, WaveNet	$\mathcal{O}(1)$	$\mathcal{O}(N)$
NAR (CNN/Self-Att)	FastSpeech 1/2, ParaNet	$\mathcal{O}(1)$	$\mathcal{O}(1)$
NAR (GAN/VAE)	MelGAN, HiFi-GAN, FastSpeech 2s, EATS	$\mathcal{O}(1)$	$\mathcal{O}(1)$
Flow (AR)	Par. WaveNet, ClariNet, Flowtron	$\mathcal{O}(1)$	$\mathcal{O}(1)$
Flow (Bipartite)	WaveGlow, FloWaveNet, Glow-TTS	$\mathcal{O}(T)$	$\mathcal{O}(T)$
Diffusion	DiffWave, WaveGrad, Grad-TTS, PriorGrad	$\mathcal{O}(T)$	$\mathcal{O}(T)$

- Lightweight model
  - pruning, quantization, knowledge distillation, and neural architecture search
- Speedup with domain knowledge
  - linear prediction, multiband modeling, subscale prediction, multi-frame prediction, streaming synthesis

Low-resource TTS

Techniques	Data	Work
Self-supervised Training	Unpaired text or speech	[52, 374, 440, 81, 143, 352, 201, 358, 73]
Cross-lingual Transfer	Paired text and speech	[43, 396, 12, 407, 62, 277, 108]
Cross-speaker Transfer	Paired text and speech	[222, 128, 61, 407, 40]
Speech chain/Back transformation	Unpaired text or speech	[291, 396, 350, 351]
Dataset mining in the wild	Paired text and speech	[59, 122, 58]

- Self-supervised training
  - Text pre-training, speech pre-training, discrete token quantization
- Cross-lingual transfer
  - Languages share similarity, phoneme mapping/re-initialization/IPA/byte
- Cross-speaker transfer
  - Voice conversion, voice adaptation
- Speech chain/back transformation
  - TTS ←→ASR
- Dataset mining in the wild
  - Speech enhancement, denoising, disentangling

Robust TTS

- Alignment between characters/phonemes and mel-spectrograms
  - Enhance attention
  - Replace attention with duration prediction
- Exposure bias and error propagation in AR generation
  - Enhance AR
  - Replace AR with NAR

Robust TTS

Category	Technique	Work
Enhancing Attention	Content-based attention	[382, 192]
	Location-based attention	[315, 333, 367, 17]
	Content/Location hybrid attention	[303]
	Monotonic attention	[438, 107, 411]
	Windowing or off-diagonal penalty	[332, 438, 270, 39]
	Enhancing enc-dec connection	[382, 303, 270, 203, 39]
	Positional attention	[268, 234, 204]
Replacing Attention with Duration Prediction	Label from encoder-decoder attention	[290, 361, 197, 181]
	Label from CTC alignment	[19]
	Label from HMM alignment	[292, 418, 194, 252, 74, 304]
	Dynamic programming	[429, 193, 235]
	Monotonic alignment search	[159]
	Monotonic interpolation with soft DTW	[69, 75]
Enhancing AR	Professor forcing	[99, 205]
	Reducing training/inference gap	[361]
	Knowledge distillation	[209]
	Bidirectional regularization	[291, 452]
Replacing AR with NAR	Parallel generation	[290, 292, 268, 69]

Robust TTS

Perspective	Category	Work
External/Internal	External Internal	FastSpeech 1/2 [290, 292], DurIAN [418], TalkNet [19], [361, 74, 304] AlignTTS [429], Glow-TTS [159], EATS [69], [235, 75]
E2E Optimization	Not E2E E2E	[290, 361, 19, 292, 418, 194, 74, 304, 429, 197, 159] EATS [69], Parallel Tacotron 2 [75]

Attention? \ AR?	AR	Non-AR
Attention	Tacotron 2 [303], DeepVoice 3 [270]	ParaNet [268], Flow-TTS [234]
Non-Attention	DurIAN [418], Non-Att Tacotron [304]	FastSpeech [290, 292], EATS [69]

Expressive TTS

- Variation information
  - Category: Text, speaker/timbre, prosody/style/emotion, recording/environment
  - Modeling

Perspective	Category	Description	Work
Information Type	Explicit	Language/Style/Speaker ID	[445, 247, 195, 162, 39]
		Pitch/Duration/Energy	[290, 292, 181, 158, 239, 365]
	Implicit	Reference encoder	[309, 383, 224, 142, 9, 49, 37, 40]
		VAE	[119, 4, 443, 120, 324, 325, 74]
		GAN/Flow/Diffusion	[224, 186, 366, 234, 159, 141]
		Text pre-training	[81, 104, 393, 143]
Information Granularity	Language/Speaker Level	Multi-lingual/speaker TTS	[445, 247, 39]
	Paragraph Level	Long-form reading	[11, 395, 376]
	Utterance Level	Timbre/Prosody/Noise	[309, 383, 142, 321, 207, 40]
	Word/Syllable Level	Fine-grained information	[325, 116, 45, 335]
	Character/Phoneme Level		[188, 324, 430, 325, 45, 40, 189]
	Frame Level		[188, 158, 49, 434]

Expressive TTS

Technique	Description	Work
Disentangling with Adversarial Training	Disentanglement for control	[224, 120, 281, 434]
Cycle Consistency/Feedback for Control	Enhance style/timbre generation	[202, 386, 207, 30, 195]
Semi-Supervised Learning for Control	Use VAE and adversarial training	[103, 119, 120, 434, 302]
Changing Variance Information for Transfer	Different information in inference	[309, 383, 142, 443, 40]

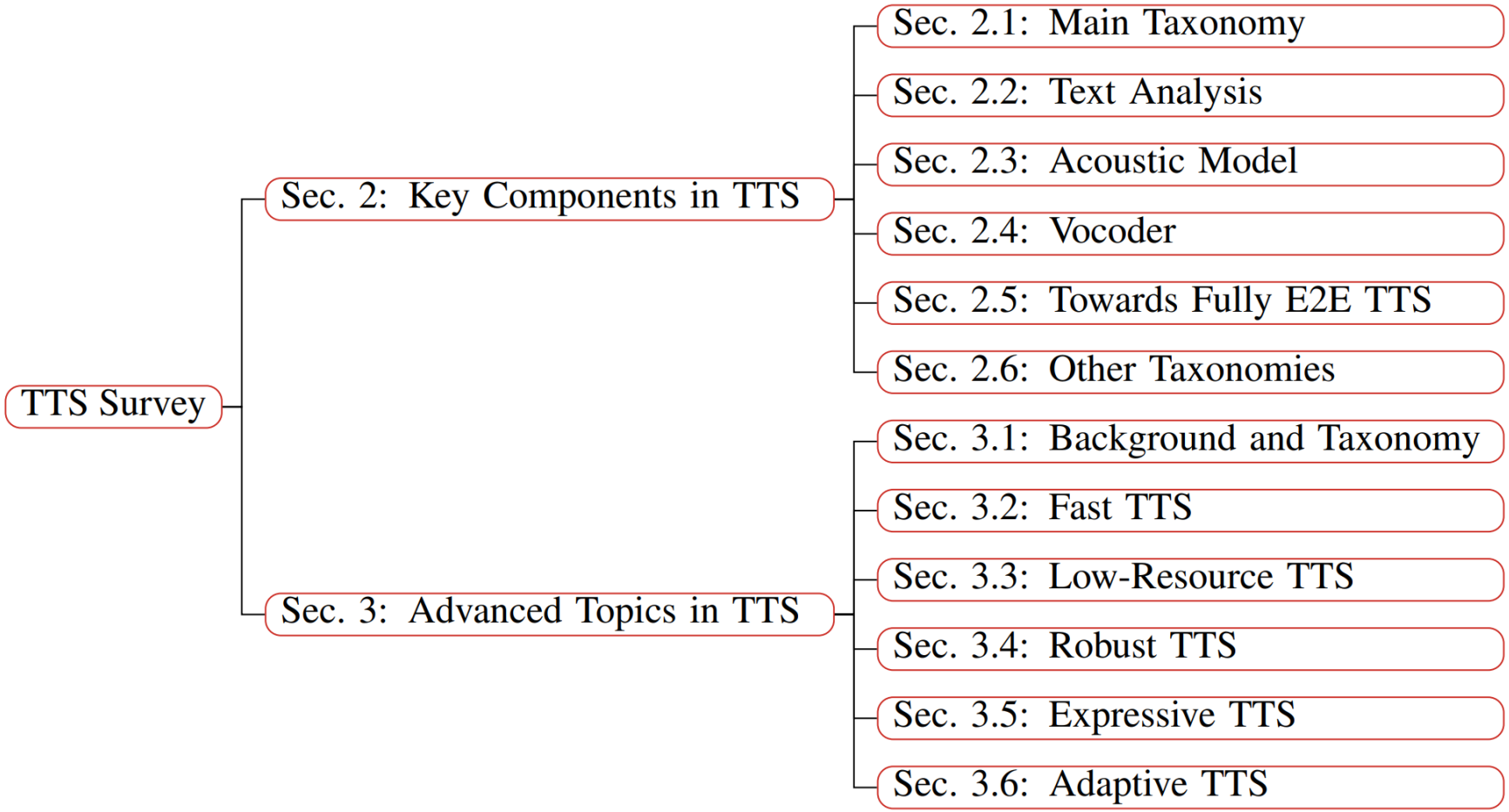
- Disentangling with Adversarial Training
- Cycle Consistency/Feedback Loss for Control
- Semi-Supervised Learning for Control
- Changing Variance Information for Transfer

Adaptive TTS

- Voice adaptation, voice cloning, custom voice

Category	Topic	Work
General Adaptation	Modeling Variation Information	[40]
	Increasing Data Coverage	[57, 407]
	Cross-Acoustic Adaptation	[40, 54]
	Cross-Style Adaptation	[404, 266, 123]
	Cross-Lingual Adaptation	[445, 38, 212]
Efficient Adaptation	Few-Data Adaptation	[44, 9, 177, 240, 446, 49, 40, 236]
	Untranscribed Data Adaptation	[403, 133, 221]
	Few-Parameter Adaptation	[9, 44, 40]
	Zero-Shot Adaptation	[9, 44, 142, 56]

Organization of this survey paper



Extension

- Voice conversion
- Singing voice synthesis
- Talking face synthesis

Resources

Open-Source Implementations	
ESPnet-TTS [105]	https://github.com/espnet/espnet
Mozilla-TTS	https://github.com/mozilla/TTS
TensorflowTTS	https://github.com/TensorSpeech/TensorflowTTS
Coqui-TTS	https://github.com/coqui-ai/TTS
Parakeet	https://github.com/PaddlePaddle/Parakeet
NeMo	https://github.com/NVIDIA/NeMo
WaveNet	https://github.com/ibab/tensorflow-wavenet
WaveNet	https://github.com/r9y9/wavenet_vocoder
WaveNet	https://github.com/basveeling/wavenet
SampleRNN	https://github.com/soroushmehr/sampleRNN_ICLR2017
Char2Wav	https://github.com/sotelo/parrot
Tacotron	https://github.com/keithito/tacotron
Tacotron	https://github.com/Kyubyong/tacotron
Tacotron 2	https://github.com/Rayhane-mamah/Tacotron-2
Tacotron 2	https://github.com/NVIDIA/tacotron2
DeepVoice 3	https://github.com/r9y9/deepvoice3_pytorch
TransformerTTS	https://github.com/as-ideas/TransformerTTS
FastSpeech	https://github.com/xcmyz/FastSpeech
FastSpeech 2	https://github.com/ming024/FastSpeech2
MelGAN	https://github.com/descriptinc/melgan-neurips
MelGAN	https://github.com/seungwonpark/melgan
WaveRNN	https://github.com/fatchord/WaveRNN
LPCNet	https://github.com/mozilla/LPCNet
WaveGlow	https://github.com/NVIDIA/WaveGlow
FloWaveNet	https://github.com/ksw0306/FloWaveNet
WaveGAN	https://github.com/chrisdonahue/wavegan
GAN-TTS	https://github.com/r9y9/gantts
Parallel WaveGAN	https://github.com/kan-bayashi/ParallelWaveGAN
HiFi-GAN	https://github.com/jik876/hifi-gan
Glow-TTS	https://github.com/jaywalnut310/glow-tts
Flowtron	https://github.com/NVIDIA/flowtron
DiffWave	https://github.com/lmnt-com/diffwave
WaveGrad	https://github.com/ivanvovk/WaveGrad
VITS	https://github.com/jaywalnut310/vits
TTS Samples	https://github.com/seungwonpark/awesome-tts-samples
Software/Tool for Audio	https://github.com/faroit/awesome-python-scientific-audio

TTS Tutorials & Keynotes	
TTS Tutorial at ISCSLP 2014 [282]	https://www.superlectures.com/iscslp2014/tutorial-4-deep-learning-for-speech-generation-and-synthesis
TTS Tutorial at ISCSLP 2016 [200]	http://staff.ustc.edu.cn/~zhling/download/ISCSLP16_tutorial_DLSPSS.pdf
TTS Tutorial at IEICE [378]	https://www.slideshare.net/jyamagis/tutorial-on-endtoend-texttospeech-synthesis-part-1-neural-waveform-modeling
Generative Models for Speech [21]	https://www.youtube.com/watch?v=vEAq_sBf1CA
Generative Model-Based TTS [423]	https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/45882.pdf
Keynote at INTERSPEECH [354]	http://www.sp.nitech.ac.jp/~tokuda/INTERSPEECH2019.pdf
TTS Tutorial at ISCSLP 2021 [339]	https://www.microsoft.com/en-us/research/uploads/prod/2021/02/ISCSLP2021-TTS-Tutorial.pdf
TTS Webinar [338]	https://www.youtube.com/watch?v=MA8PCvmr8B0
TTS Tutorial at IJCAI 2021 [340]	https://tts-tutorial.github.io/ijcai2021/

TTS Challenges	
Blizzard Challenge	http://www.festvox.org/blizzard/
Zero Resource Speech Challenge	https://www.zerospeech.com/
ICASSP2021 M2VoC	http://challenge.ai.iqiyi.com/detail?raceId=5fb2688224954e0b48431fe0
Voice Conversion Challenge	http://www.vc-challenge.org/

TTS Corpora

Corpus	#Hours	#Speakers	Sampling Rate (kHz)	Language
ARCTIC [173]	7	7	16	English
VCTK [369]	44	109	48	English
Blizzard-2011 [165]	16.6	1	16	English
Blizzard-2013 [166]	319	1	44.1	English
LJSpeech [136]	25	1	22.05	English
LibriSpeech [259]	982	2484	16	English
LibriTTS [428]	586	2456	24	English
VCC 2018 [214]	1	12	22.05	English
HiFi-TTS [16]	300	11	44.1	English
TED-LIUM [295]	118	666	/	English
CALLHOME [31]	60	120	8	English
RyanSpeech [421]	10	1	44.1	English
CSMSC [15]	12	1	48	Mandarin
HKUST [211]	200	2100	8	Mandarin
AISHELL-1 [28]	170	400	16	Mandarin
AISHELL-2 [71]	1000	1991	44.1	Mandarin
AISHELL-3 [305]	85	218	44.1	Mandarin
DiDiSpeech-1 [100]	572	4500	48	Mandarin
DiDiSpeech-2 [100]	227	1500	48	Mandarin
JSUT [314]	10	1	48	Japanese
KazakhTTS [243]	93	2	44.1/48	Kazakh
Ruslan [83]	31	1	44.1	Russian
HUI-Audio-Corpus [280]	326	122	44.1	German
India Corpus [106]	39	253	48	Multilingual
M-AILABS [88]	1000	/	16	Multilingual
MLS [278]	51K	6K	16	Multilingual
CSS10 [264]	140	1	22.05	Multilingual
CommonVoice [7]	2.5K	50K	48	Multilingual

Future directions

- High-quality speech synthesis
  - Powerful generative models
  - Better representation learning
  - Robust speech synthesis
  - Expressive/controllable/transferrable speech synthesis
  - More human-like speech synthesis
- Efficient speech synthesis
  - Data-efficient TTS
  - Parameter-efficient TTS
  - Energy-efficient TTS

Thank You!

Xu Tan
Microsoft Research Asia
xuta@microsoft.com

https://www.microsoft.com/en-us/research/people/xuta/
https://speechresearch.github.io/

Other information

- TTS tutorial @ IJCAI 2021
  - https://tts-tutorial.github.io/ijcai2021/
  - https://ijcai-21.org/tutorials/
  - August 19th
- We (Machine Learning Group, MSRA) are hiring!
  - FTE (social/campus hire): Speech, NLP, Machine Learning, Deep Learning
  - Intern: TTS, ASR, healthcare
  - xuta@microsoft.com