Text-To-Speech Synthesis

Current trends

Sisamaki Eirini

University of Crete

Outline

1. **Overview of TTS**
   - **Introduction - History**
   - **Basic components of TTS**

2. Neural text to speech

3. Previous Work
   - Tacotron 2 - Greek

4. Current work
   - Stable Training of Parallel WaveNet

5. Challenging Research Topics

6. Summary

Introduction: Speech Synthesis

Artificial production of human speech

Applications

- Screen readers for people with visual impairment or dyslexia
- Speech synthesizers
- Games, animations entertainment production
- Educational tools for foreign languages
- Natural language processing interfaces
- Personal Assistants

Introduction: Text-To-Speech

Text-To-Speech (TTS) synthesis is called the automatic conversion of written to spoken language.

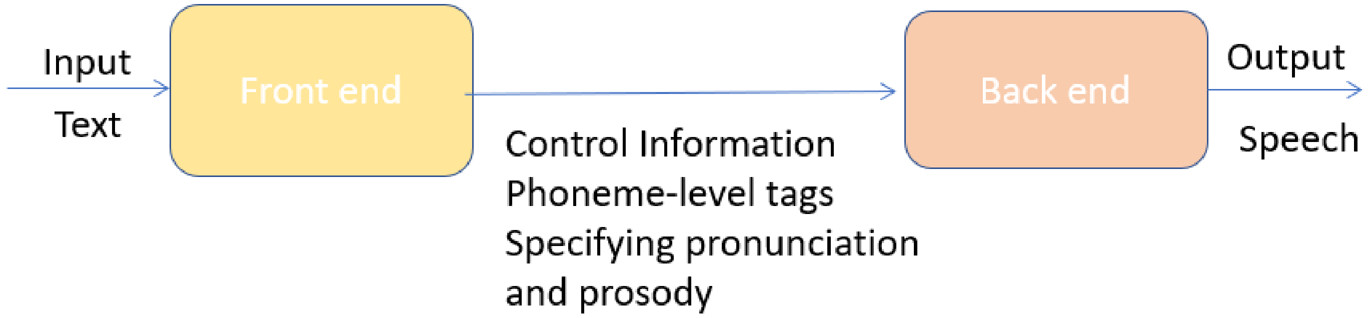
Disciplines: acoustics, linguistics, digital signal processing,

Statistics, deep learning

A TTS system is divided into two parts.

- the **front end** converts text in to a linguistic specification and
- the **back end** uses the specification to generate a waveform.

One-to-many mapping problem



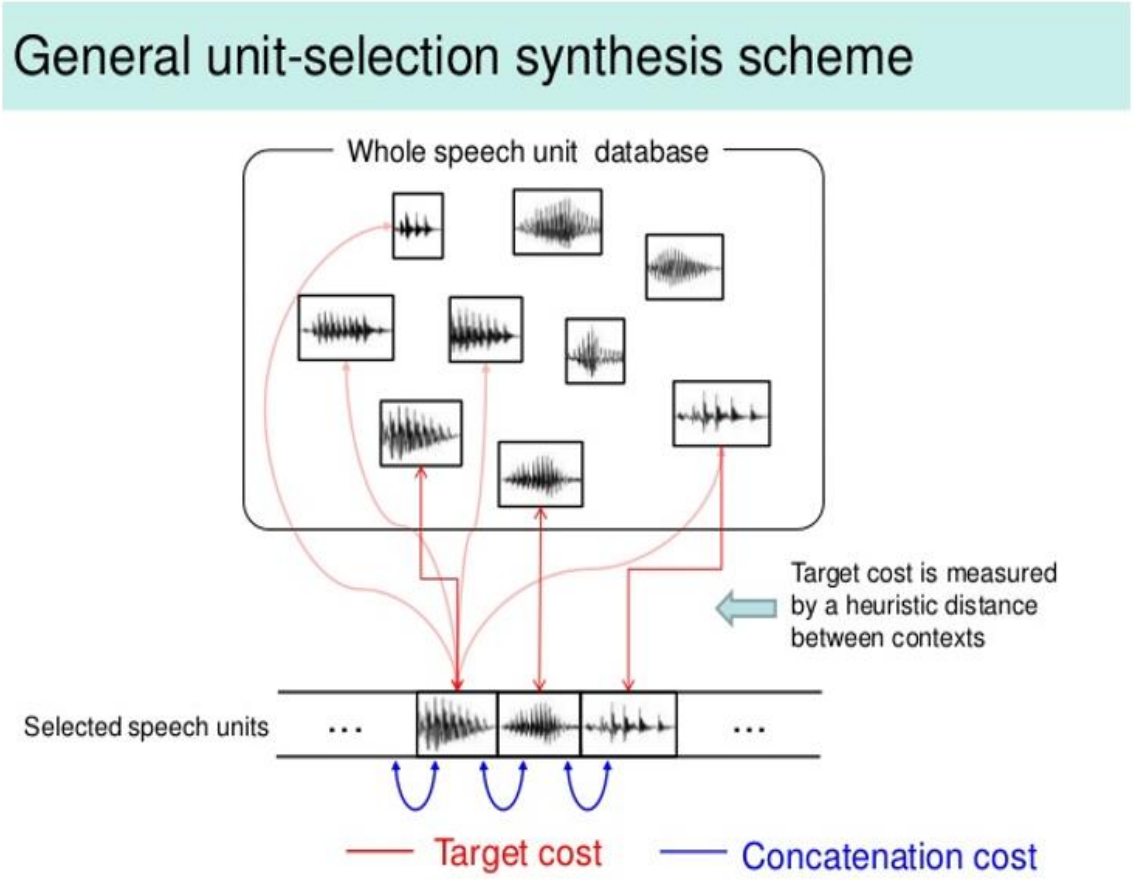
The quality of the synthesized speech is measured by Intelligibility and naturalness

History:

Concatenative Speech Synthesis

Recordings of short audio sequences are combined to create speech.

Advantages
• Clean and clear speech
Problems
• Very large data requirements.
• Unnatural speech (No emotional and intonation components)
• Long development time.
• Language specific

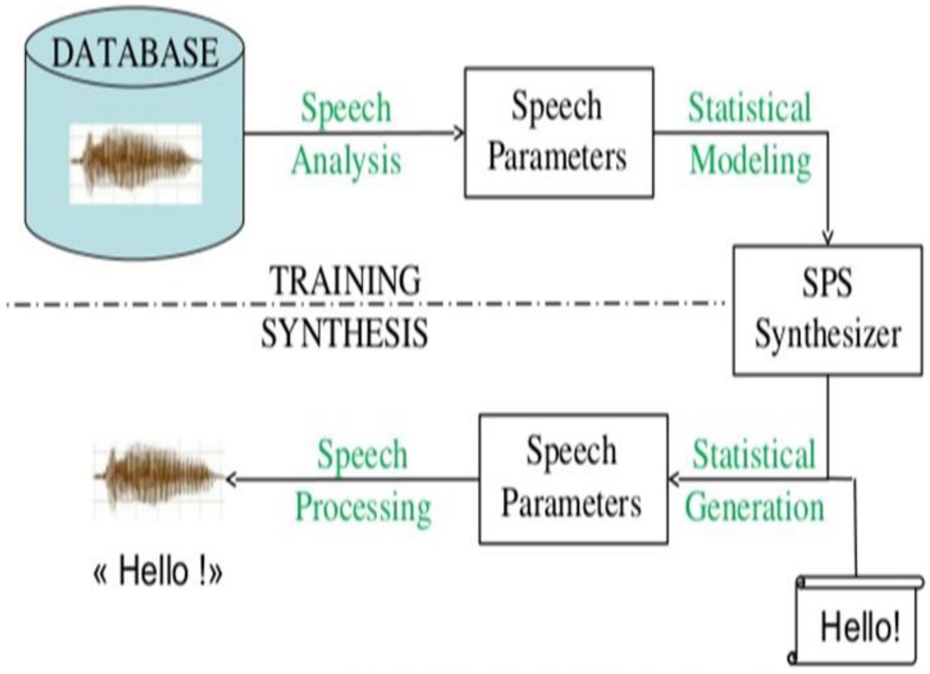


History:

Statistical Parametric Speech Synthesis

- **Linguistic features** (phonemes, duration, etc) are extracted from the text.
- **Speech parameters** (cepstrum, frequency, Mel Spectrogram, linear spectrogram) are extracted from the corresponding speech signal.
- A **vocoder** (a system that generates wave-form) encodes them.

Advantages	Problems
Transforming voice characteristics, speaking styles and emotions	Artifacts lead to production of muffled speech and buziness.
Less hard work than concatenative	
Less data requirements.	
Multilingual support	

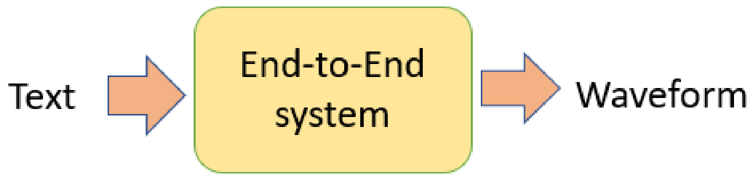


2015 Bachelor Thesis: “Parameter Estimation of LDMs for speech synthesis”

TTS in the era of Deep Learning

End-to-end neural networks have dramatically improved the quality of synthetic speech.

An **end-to-end TTS** system can be trained on **<text, audio>** pairs.



Advantages
Rich conditioning on various attributes such as speaker, language, sentiment, etc.
More robust than a multi-component system.
Potential for transfer learning and it can adapt to new data.
It may be trained on huge amount of often noisy data found in the real world.

TTS in the era of Deep Learning

- Deep Voice 1 (Arik et al., 2017)
- Deep Voice2 (Ariketal.,2017)
- DeepVoice3 (Pingetal.,2018)
- Tacotron (Wang et al., 2017)
- Tacotron 2 (Shen et al., 2018)
- Char2Wav (Sotelo et al., 2017)
- VoiceLoop (Taigman et al., 2018)
- ClariNet (Ping et al., 2018)
- FastSpeech (2020), FastSpeech 2 (2021)
- ParaNet (2020)
- Glow-TTS (2020)
- WaveGrad 2 (2021)

Vocoders:

- Parallel WaveNet (2017)
- WaveRNN (2018)
- MelGAN (2019)
- Parallel WaveGAN (2019), WaveGAN(2020)
- GAN-TTS (2019)
- LPCNet (2019)
- HiFi-GAN (2020)
- WaveGlow (2018), WaveFlow (2019)
- FloWaveNet (2020)
- WaveGrad (2020), DiffWave (2020)

Basic Components in TTS

a text analysis module,

converts a text sequence into linguistic features

an acoustic model

generates acoustic features from linguistic features

a vocoder

synthesizes waveform from acoustic features.

Text Decoding (Word Tokenization and Normalization)
Texts include non-standard word sequences from a variety of different semiotic classes.

- F0, V/UV, energy, Mel-scale Frequency Cepstral Coefficients (MFCC), Bark-Frequency Cepstral Coefficients (BFCC), Linear prediction coefficient (LPC),
- Linear-spectrogram, Mel-spectrogram

Outline

1. Overview of TTS
   - Introduction - History
   - Basic components of TTS

2. **Neural text to speech**

3. Previous Work
   - Tacotron 2 - Greek

4. Current work
   - Stable Training of Parallel WaveNet

5. Challenging Research Topics

6. Summary

VOCODERS (neural)

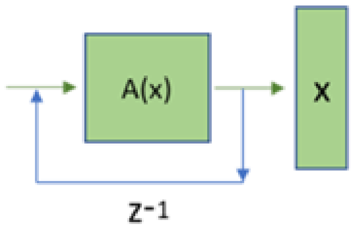
- Sequential generation of samples
    - Autoregressive

- Parallel generation of samples
    - Flow-based
    - GAN-based
    - VAE-based
    - Diffusion-based

More applications: Speech enhancement, Denoising,

Voice conversion, Source separation

Autoregressive



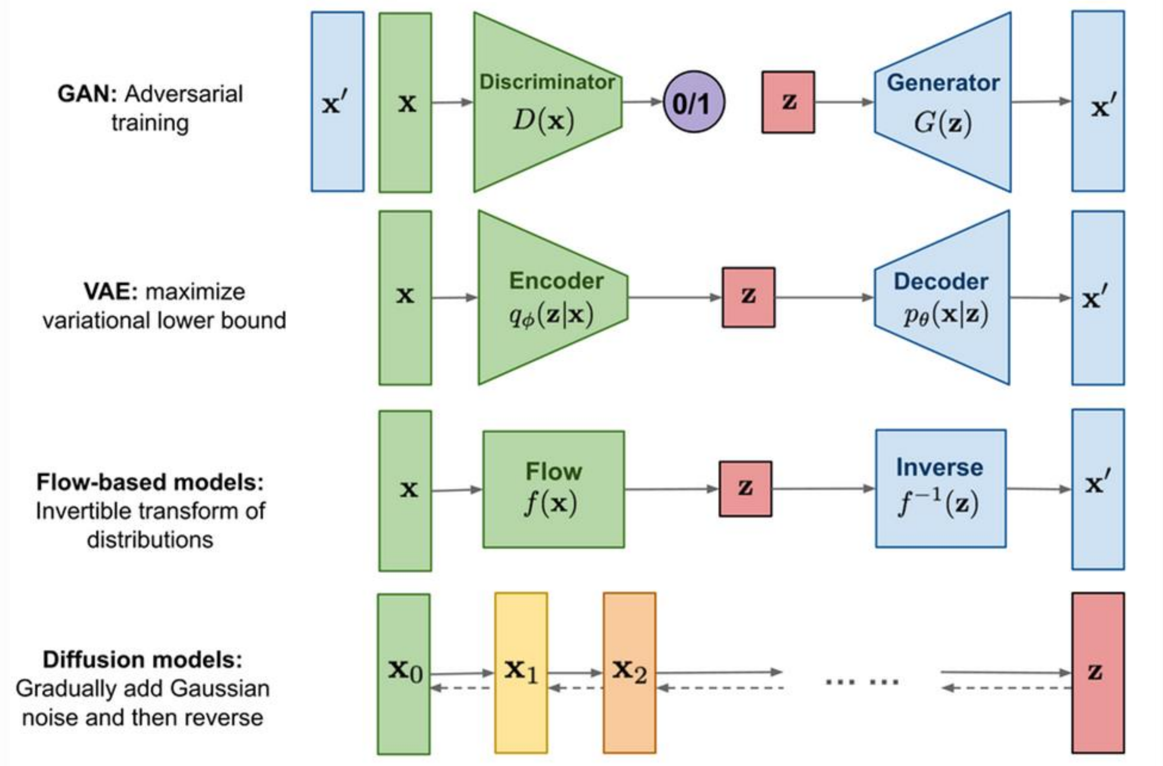


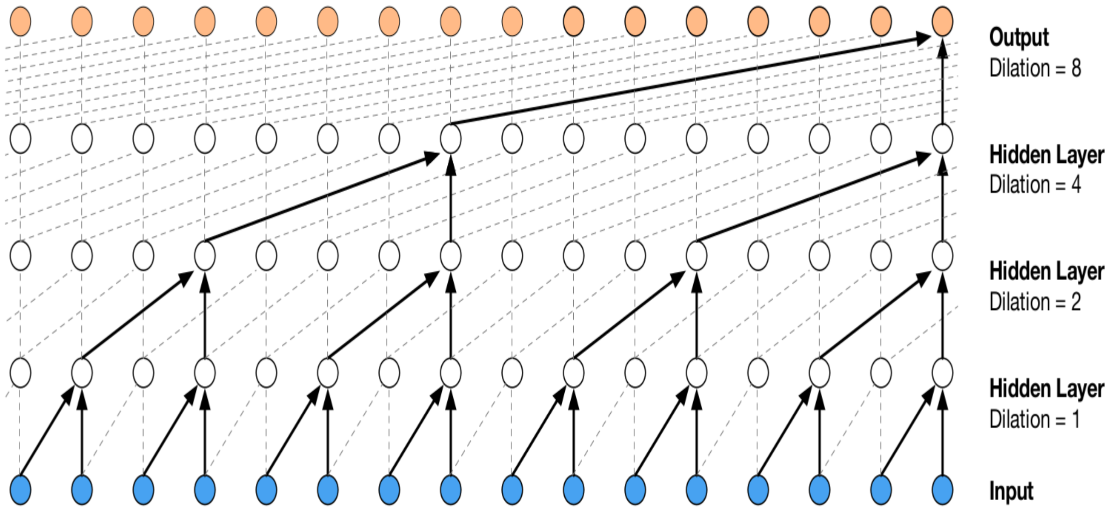
Fig. 1. Overview of different types of generative models.

Autoregressive Models: WaveNet

$$p(\mathbf{x}) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1})$$

Directly predict waveform instead of mel-spectrogram

WaveNet : linguistic features, F0, duration →waveform



Visualization of a stack of dilated causal convolutional layers

Key Components

- Causal dilated convolution
- Gated activation + residual + skip: powerful non-linearity
- Softmax at output: classification rather than regression. Also
  - MoL(Mixture of Logistics) in Parallel WaveNet
  - A single Gaussian in ClariNet

Achievements

- High-fidelity speech (SOTA)
- Efficient training on parallel hardware
- Trained maximizing likelihood

Limitations

- Generating waveform one sample at a time
- Ill-suited for development on GPU, TPU

Flow-based Models

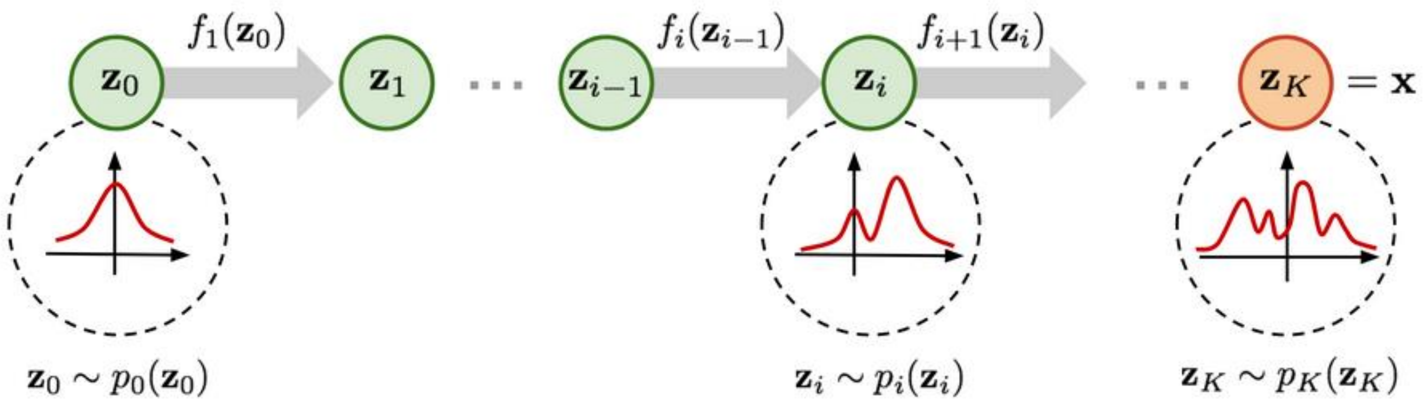


Fig. 2. Illustration of a normalizing flow model, transforming a simple distribution $p_0(\mathbf{z}_0)$ *to a complex one* $p_K(\mathbf{z}_K)$ *step by step.*

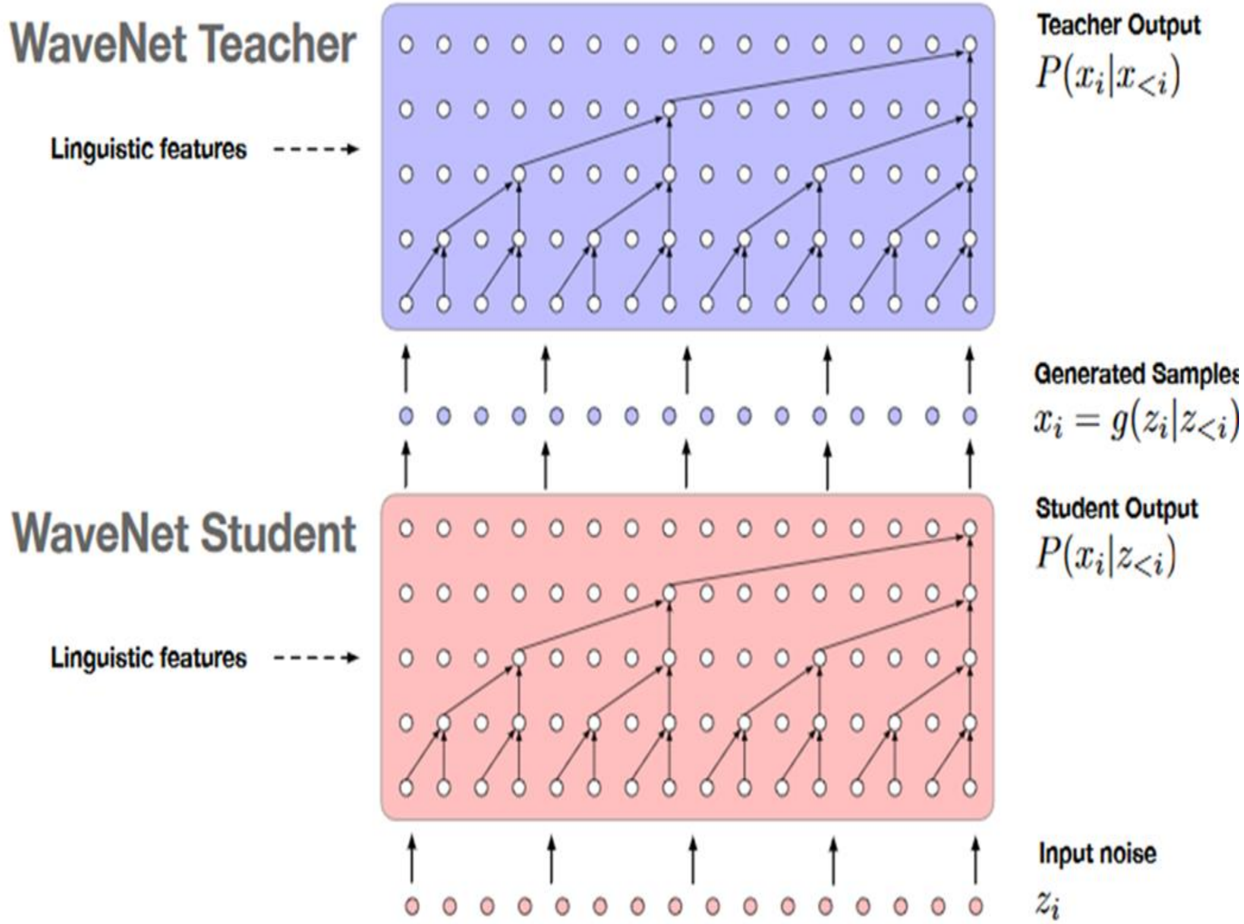
$$\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{z}; 0, \boldsymbol{I}) \tag{1}$$

$$\boldsymbol{x} = \boldsymbol{f}_0 \circ \boldsymbol{f}_1 \circ \dots \boldsymbol{f}_k(\boldsymbol{z}) \tag{2}$$

$$\log p_\theta(\boldsymbol{x}) = \log p_\theta(\boldsymbol{z}) + \sum_{i=1}^{k} \log |\det(\boldsymbol{J}(\boldsymbol{f}_i^{-1}(\boldsymbol{x})))| \tag{3}$$

$$\boldsymbol{z} = \boldsymbol{f}_k^{-1} \circ \boldsymbol{f}_{k-1}^{-1} \circ \dots \boldsymbol{f}_0^{-1}(\boldsymbol{x}) \tag{4}$$

Parallel WaveNet



$$x_{1:N}, \mu, s = F(z_{1:N}, c_{1:M})$$

F Parallel Feed-forward Neural Network
x_n waveform samples
$z_{1:N}$ white noise
c_i conditional features
μ location, s scale

Characteristics

- Non-autoregressive (parallel generation)
- Minimize the KL divergence
  Cauchy Schwarz divergence or other divergence between student and teacher
- Or train with Generalized Energy Distance

Limitations

- Regularization and
- Auxiliary losses for the student models to converge

Probability Density Distillation loss

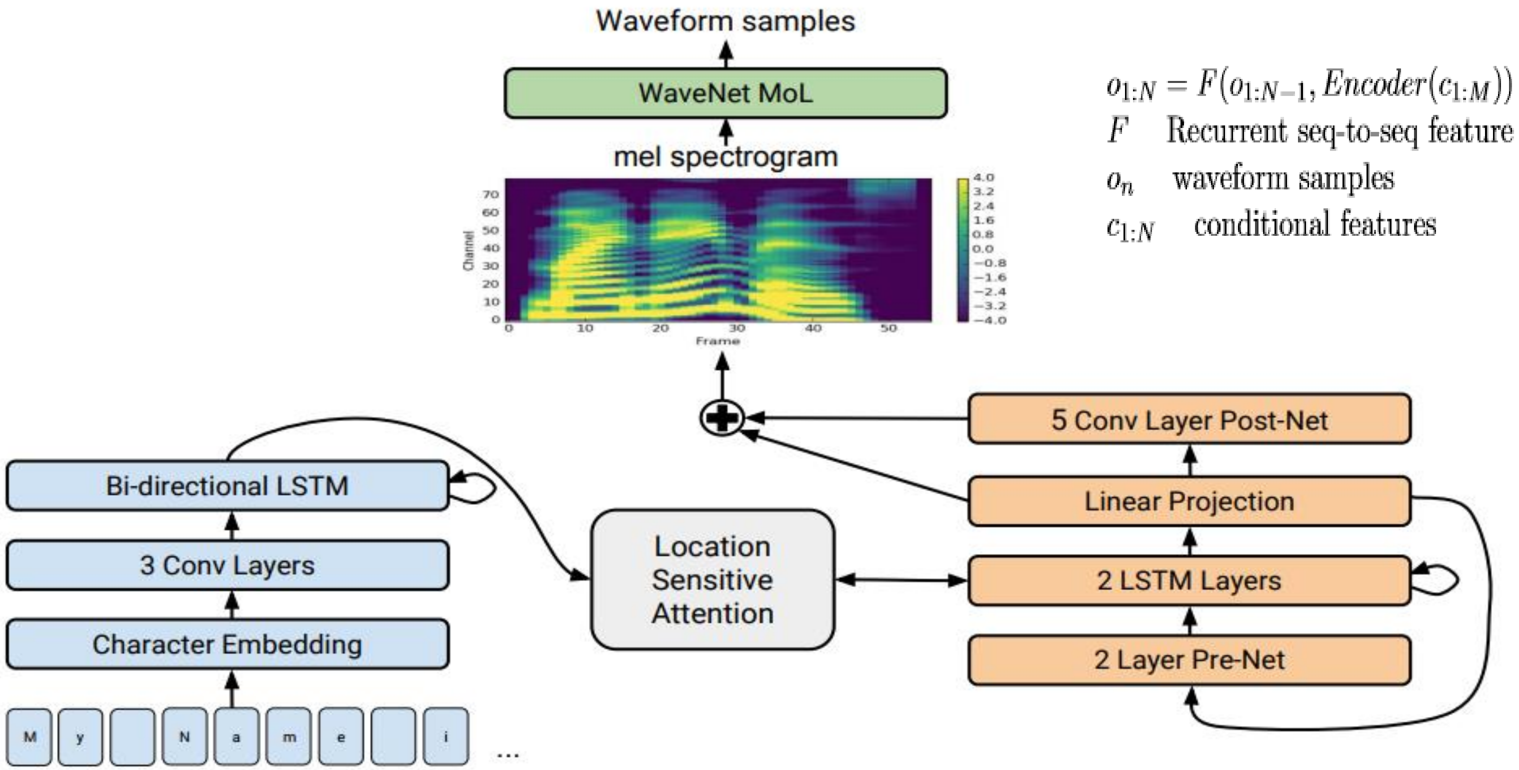
$$D_{\text{KL}}(P_S||P_T) = H(P_S, P_T) - H(P_S)$$

Towards end-to-end TTS

Simplify/remove text analysis, and simplify acoustic features

Tacotron 2

(2018)



$$o_{1:N} = F(o_{1:N-1}, Encoder(c_{1:M}))$$

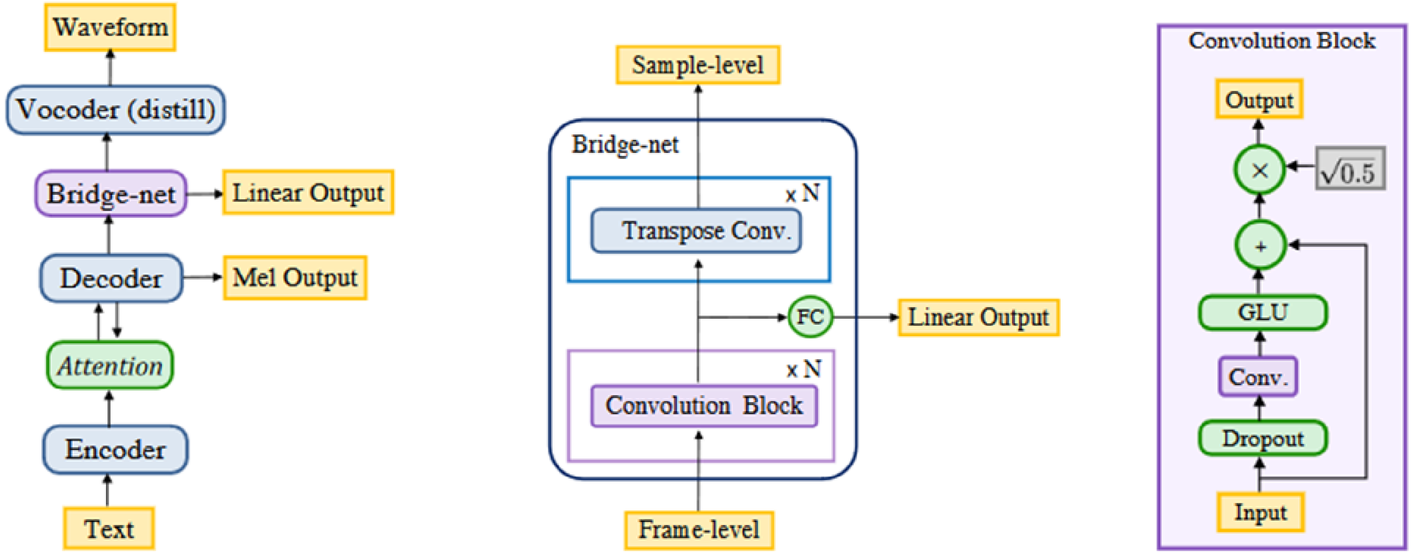
F Recurrent seq-to-seq feature prediction NN

o_n waveform samples

$c_{1:N}$ conditional features

Fully end-to-end, direct text to waveform synthesis

ClariNet (2018): autoregressive acoustic model and non-autoregressive vocoder



(a) Text-to-wave architecture (b) Bridge-net (c) Convolution block

$$o_{1:N} = F(Encoder(c_{1:M}))$$

F Text-to-Wave Fully Convolutional NN

o_n waveform samples

$c_{1:N}$ conditional features

(a) Text-to-wave model converts textual features into waveform. All components feed their hidden representation to others directly. (b) Bridge-net maps frame-level hidden representation to sample-level through several convolution blocks and transposed convolution layers interleaved with soft sign non-linearities. (c) Convolution block is based on gated linear unit

Fully end-to-end

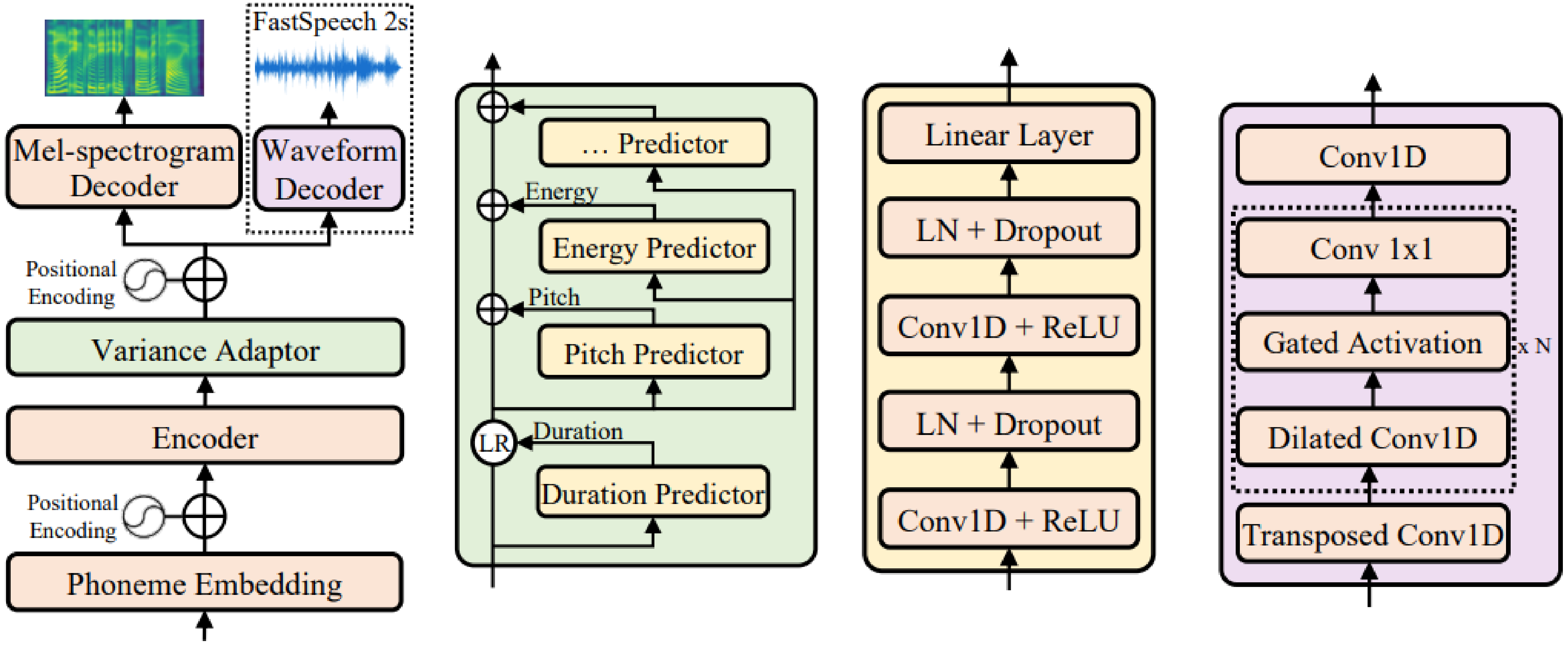
FastSpeech 2s (2021): fully parallel text to wave model

$$o_{1:N} = F(Duration, Pitch, Energy, Encoder(c_{1:M}))$$

F Fully-end-to-end parallel model

o_n waveform samples

$c_{1:N}$ phonemes



(a) FastSpeech 2 (b) Variance adaptor (c) Duration/pitch/energy predictor (d) Waveform decoder

Outline

1. Overview of TTS
   - Introduction - History
   - Basic components of TTS

2. Neural text to speech

3. **Previous Work**
   - **Tacotron 2 - Greek**

4. Current work
   - Stable Training of Parallel WaveNet

5. Challenging Research Topics

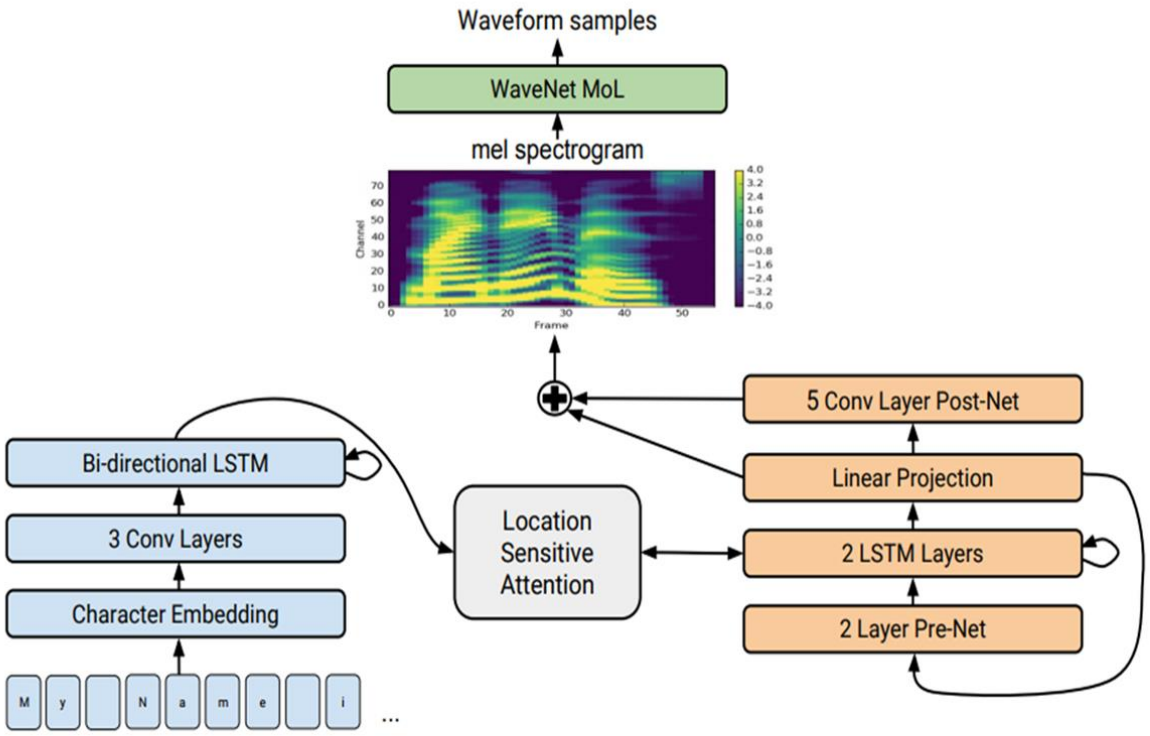
6. Summary

Tacotron 2

The proposed system consists of two components

1. a recurrent sequence-to-sequence feature prediction network with attention which predicts a sequence of mel spectrogram frames from an input character sequence, and

2. a modified version of WaveNet which generates time-domain waveform samples conditioned on the predicted mel-spectrogram frames.



Experiments:
Spanish Greek language adaptation

Tacotron	WaveNet	utterances	time(hours)	Success
Greek+Spanish	Greek	2154 gr+11133sp	3.0+18.0	✓
Greek+Spanish	Greek+Spanish	2154gr+11133sp	3.0+18.0	✓ improved

The Spanish Dataset was obtained from **M-Ailabs**, book=**'don_quijote'**, multi-speaker.

Synthesized

Real recording

Experiments: Listening Test

Mean Opinion Scores Evaluation

- 30 volunteers
- 16 sentences from the test set of our internal dataset as the evaluation set
- each sentence appears in two samples (Original recording and synthesized)
- Naturality, intelligibility (and quality of sound, artifacts)
- scale from 1 to 5 (Perfect, very good, good, bad, very bad)
- Two speakers (male – female), s027, s023

System	MOS (Mean Opinion Score)
Original Recordings	3.82±0.19
Tacotron-2 Spanish/Greek	3.15±0.20

Experiments: Speaker Adaptation

Training with Greek Harvard Database.
From step 145000 to 150000 the training needs almost 3h.

George Anna

Outline

1. Overview of TTS
   - Introduction - History
   - Basic components of TTS

2. Neural text to speech

3. Previous Work
   - Tacotron 2 - Greek

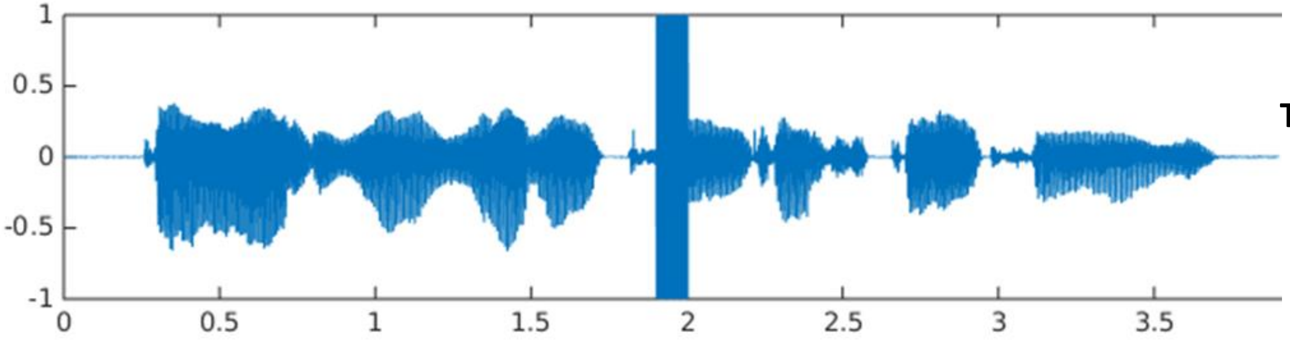
4. **Current work**
   - **Stable Training of Parallel WaveNet**

5. Challenging Research Topics

6. Summary

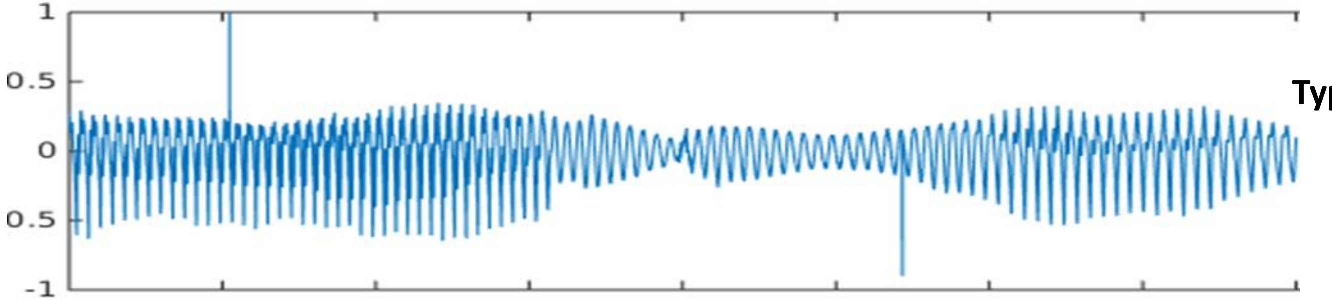
Stable Training of Parallel WaveNet

- Problem definition and analysis
  - Type I and II artifacts of WaveNet
  - The loss function as the source of instabilities



Type I artifacts

Type I artifacts usually occur in the WaveNet that predicts the parameters of a distribution.



Type II artifacts

Noise at the end (CS-Divergence)

Noisy (CS-Divergence)

Noisy (KL-Divergence)

Related Work

ClariNet 2018, Wu et al.,

FloWaveNet 2019,

Parallel WaveGAN 2019

have trained Parallel WaveNet

with a set of losses

Stable Training of Parallel WaveNet

The **Kullback-Leibler** divergence Loss function

$$D_{KL}(q||p) = \log \frac{\sigma_p}{\sigma_q} + \frac{\sigma_q^2 - \sigma_p^2 + (\mu_p - \mu_q)^2}{2\sigma_p^2}$$

The **Cauchy-Schwarz** divergence Loss function

$$D_{CS} = \frac{1}{2}\left(\log(\sigma_p^2 + \sigma_q^2) + \frac{(\mu_p + \mu_q)^2}{\sigma_p^2 + \sigma_q^2} - \log(2\sigma_p\sigma_q)\right)$$

Student distribution

$q(x) = \mathcal{N}(\mu_q, \sigma_q)$

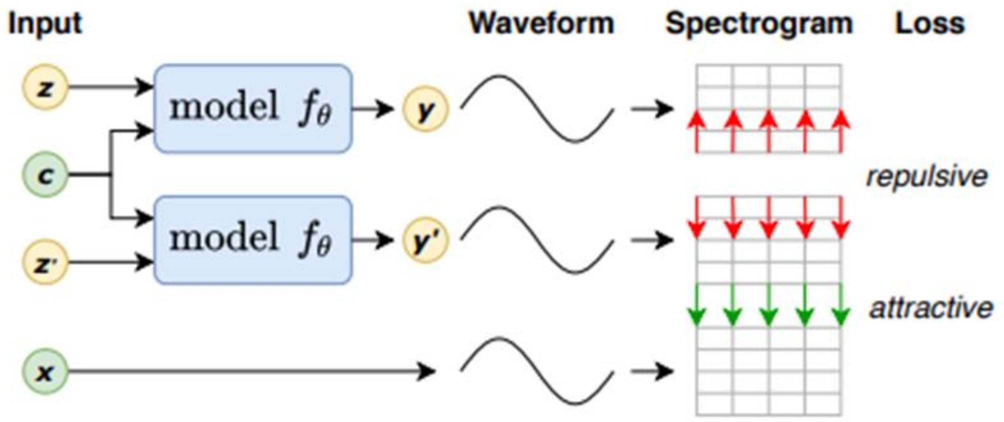
Teacher distribution

$p(x) = \mathcal{N}(\mu_p, \sigma_p)$

Training with Generalized Energy Distance

Training procedure

- For each training example we generate two independent batches of audio samples from our model, conditioned on the same features, which are then used to compute our training loss.



A. Gritsenko, 2020, "A spectral energy distance for parallel speech synthesis"

A Generalized Energy Distance based on spectrograms

- The **minibatch loss**: **unbiased** estimator of GED

$$L^*_{\text{GED}}(q) = \sum_{i=1}^{M} 2d(\mathbf{x}_i, \mathbf{y}_i) - d(\mathbf{y}_i, \mathbf{y}'_i).$$

where $y_i = f_\theta(c_i, z_i), \quad y'_i = f_\theta(c_i, z'_i)$

- The performance of the energy score strongly depends on the choice of metric d(·, ·).
- We thus have to select a distance function that emphasizes those features of the generated audio that are most important to the human ear.

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k \in [2^6, \ldots, 2^{11}]} \sum_t ||\mathbf{s}^k_t(\mathbf{x}_i) - \mathbf{s}^k_t(\mathbf{x}_j)||_1 + \alpha_k || \log \mathbf{s}^k_t(\mathbf{x}_i) - \log \mathbf{s}^k_t(\mathbf{x}_j)||_2,$$

- We combine multiple frame-lengths k into a single multi-scale spectrogram loss (k frame-length, t time-slice).

Experiments – Results (Stable Training of Parallel WaveNet)

KL_Divergence+ GED **CS_Divergence+ GED** **GED** **real**

	Experiment	MOS
1.	GED + KL-Divergence	3.04 ± 0.16
2.	GED + CS-Divergence	3.04 ± 0.16
3.	CS - Divergence	2.21 ± 0.14
4.	KL- Divergence	3.00 ± 0.15
5.	**GED**	$\mathbf{4.03 \pm 0.19}$
6.	Original Speech	4.99 ± 0.01

Table 1. Mean Opinion Score (MOS) obtained by Listening Test with 10 participants.

KL : Kullback-Leibler Divergence
CS : Cauchy Schwarz Divergence
GED: Generalized Energy Distance

Outline

1. Overview of TTS
   - Introduction - History
   - Basic components of TTS

2. Neural text to speech

3. Previous Work
   - Tacotron 2 - Greek

4. Current work
   - Stable Training of Parallel WaveNet

5. **Challenging Research Topics**

6. Summary

Challenging Research Topics

Robustness
Expressiveness
Controllability
Techniques for low-resource TTS
Reducing the cost of TTS

Future Goals
Speech Synthesis for the Greek language and other low resource languages, in combination with: 1. Work on specific problems of the existing vocoders (stability, quality, speed). 2. Use of improved database (having cleaned the previous noisy database, and added Greek Harvard database with two speakers, 1 male and 1 female)

Outline

1. Overview of TTS
    - Introduction - History
    - Basic components of TTS
2. Neural text to speech
3. Previous Work
    - Tacotron 2 - Greek
4. Current work
    - Stable Training of Parallel WaveNet
5. Challenging Research Topics
6. **Summary**

Summary

TTS technology evolves from concatenative synthesis, statistical parametric synthesis, and neural based end-to-end synthesis.

Mainstream TTS model uses separate acoustic model and vocoder, but fully end-to-end TTS model is on the way.

Improving the quality while reducing the cost is always the goal of TTS

- Quality: Intelligibility, naturalness, robustness, expressiveness and controllability
- Cost: Engineering cost (end-to-end), serving cost (inference speedup), data cost (low resource)

Thank you for your time!

References

[1] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J. Weiss, Mohammad Norouzi, and William Chan. Wavegrad: Estimating gradients for waveform generation. *ArXiv*, abs/2009.00713, 2021.

[2] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J. Weiss, Mohammad Norouzi, Najim Dehak, and William Chan. Wavegrad 2: Iterative refinement for text-to-speech synthesis. *ArXiv*, abs/2106.09660, 2021.

[3] Griffin D. and Lim J. Signal estimation from modified short-time fourier transform, 1984.

[4] Chris Donahue, Julian McAuley, and Miller Puckette. Adversarial audio synthesis. In *ICLR*, 2019.

[5] Alexey A. Gritsenko, Tim Salimans, Rianne van den Berg, Jasper Snoek, and Nal Kalchbrenner. A spectral energy distance for parallel speech synthesis. *ArXiv*, abs/2008.01160, 2020.

[6] Nal Kalchbrenner, Erich Elsen, K. Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron van den Oord, S. Dieleman, and K. Kavukcuoglu. Efficient neural audio synthesis. *ArXiv*, abs/1802.08435, 2018.

[7] Sungwon Kim, Sang gil Lee, Jongyoon Song, Jaehyeon Kim, and Sungroh Yoon. Flowavenet : A generative flow for raw audio. *ArXiv*, abs/1811.02155, 2019.

[8] Simon King. A text-to-speech system: Festival. 2018.

[9] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

[10] Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. 2018.

[11] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *ArXiv*, abs/2010.05646, 2020.

[12] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *ArXiv*, abs/2009.09761, 2021.

[13] Kainan Peng, Wei Ping, Zhao Song, and Kexin Zhao. Parallel neural text-to-speech. *ArXiv*, abs/1905.08459, 2019.

[14] Wei Ping, Kainam Peng, and Jitong Chen. Clarinet: Parallel wave generation in end-to-end text-to-speech. 30 July 2018.

References

[15] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. *ArXiv*, abs/2006.04558, 2021.

[16] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech: Fast, robust and controllable text to speech. In *NeurIPS*, 2019.

[17] J. Shen1, R. Pang1, Ron J. Weiss1, M. Schuster1, N. Jaitly1, Z. Yang, Z. Chen1, and RJ Skerry-Ryan1 Y. Zhang1, and Y. Wang1, and Y. Agiomyrgiannakis1 R. A. Saurous1, and Y. Wu1. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. 2017.

[18] Jean-Marc Valin and Jan Skoglund. Lpcnet: Improving neural speech synthesis through linear prediction. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5891–5895, 2019.

[19] Aaron van den Oord, Yazhe Li, I. Babuschkin, K. Simonyan, Oriol Vinyals, K. Kavukcuoglu, G. V. D. Driessche, Edward Lockhart, Luis C. Cobo, Florian Stimberg, Norman Casagrande, Dominik Grewe, Seb Noury, S. Dieleman, Erich Elsen, Nal Kalchbrenner, H. Zen, A. Graves, Helen King, T. Walters, D. Belov, and D. Hassabis. Parallel wavenet: Fast high-fidelity speech synthesis. 2018.

[20] Yuxuan Wang, R. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Y. Xiao, Z. Chen, Samy Bengio, Quoc V. Le, Yannis Agiomyrgiannakis, R. Clark, and R. Saurous. Tacotron: Towards end-to-end speech synthesis. 2017.

[21] Yi-Chiao Wu, Kazuhiro Kobayashi, Tomoki Hayashi, Patrick Lumban Tobing, and Tomoki Toda. Collapsed speech segment detection and suppression for wavenet vocoder. In *Interspeech 2018*, pages 1988–1992.

[22] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. Probability density distillation with generative adversarial networks for high-quality parallel waveform generation. *ArXiv*, abs/1904.04472, 2019.

[23] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6199–6203. 2020.

[13] Jesse Engel, Lamtharn Hantrakul, Chenjie Gu, and Adam Roberts, "Ddsp: Differentiable digital signal processing," 2020.

[14] Xin Wang, Shinji Takaki, and Junichi Yamagishi, "Neural source-filter waveform models for statistical parametric speech synthesis," 2019.

[15] Shinji Takaki, Hirokazu Kameoka, and Junichi Yamagishi, "Training a neural speech waveform model using spectral losses of short-time fourier transform and continuous wavelet transform," 2019.