SynthSing: A Singing voice Synthesizer

Bunning, James
Mou, Shiyu
Murali, Sharada
Yang, Mu
Yang, Yixin

Introduction

- Most current singing voice synthesizers use concatenative methods
- Recent advances in TTS models achieve very realistic synthesis of the human voice, especially with models such as Wavenet[4]
- Use of machine learning methods with these advancements can result in quality far superior to concatenative methods[1]

Current research

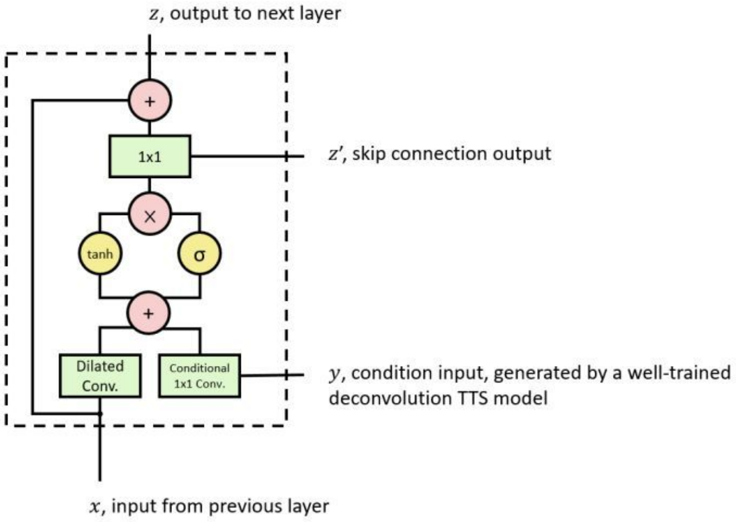
“WaveNet: A Generative Model for Raw Audio”, Aäron van den Oord, et al.

- Baseline model
- Idea is from TTS (Text-to-Speech).
- Like TTS, condition on linguistic features from input text (e.g. start and end timing of phonemes).
- Like Multi-speaker speech synthesis, also condition on singer identities.
- Additionally, condition on musical notes sequence.

Current research

Baseline Model: WaveNet.

- Basic structure
  - Dilated causal convolution
  - Gated activation units
  - Residual and skip connections
  - Conditional input variables
  - Global and local conditioning



$$\mathbf{z} = \tanh\left(W_{f,k} * \mathbf{x}\right) \odot \sigma\left(W_{g,k} * \mathbf{x}\right)$$

$$\mathbf{z} = \tanh\left(W_{f,k} * \mathbf{x} + V_{f,k}^{T}\mathbf{h}\right) \odot \sigma\left(W_{g,k} * \mathbf{x} + V_{g,k}^{T}\mathbf{h}\right)$$

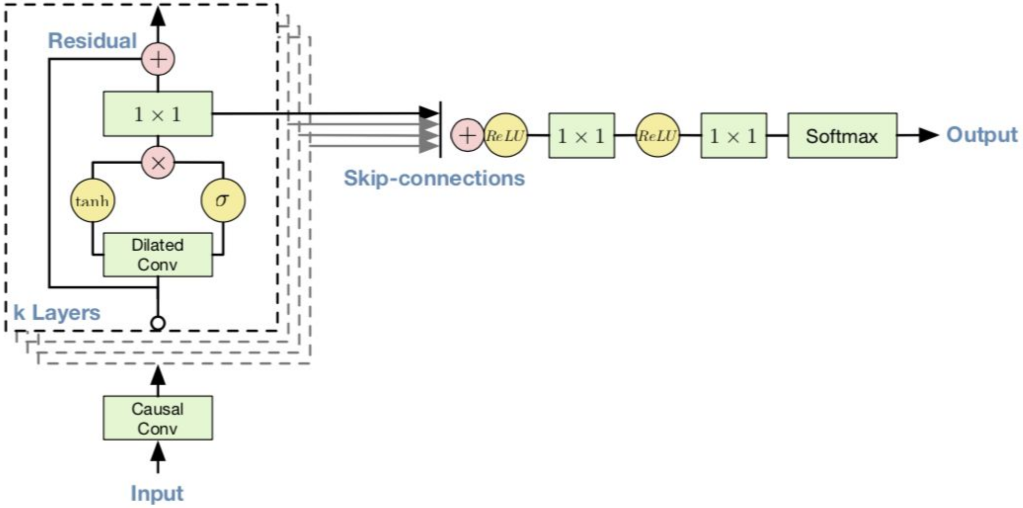
$$\mathbf{y} = f(\mathbf{h})$$

$$\mathbf{z} = \tanh\left(W_{f,k} * \mathbf{x} + V_{f,k} * \mathbf{y}\right) \odot \sigma\left(W_{g,k} * \mathbf{x} + V_{g,k} * \mathbf{y}\right)$$

Current research

Baseline Model: Wavenet.

- Overview of entire network

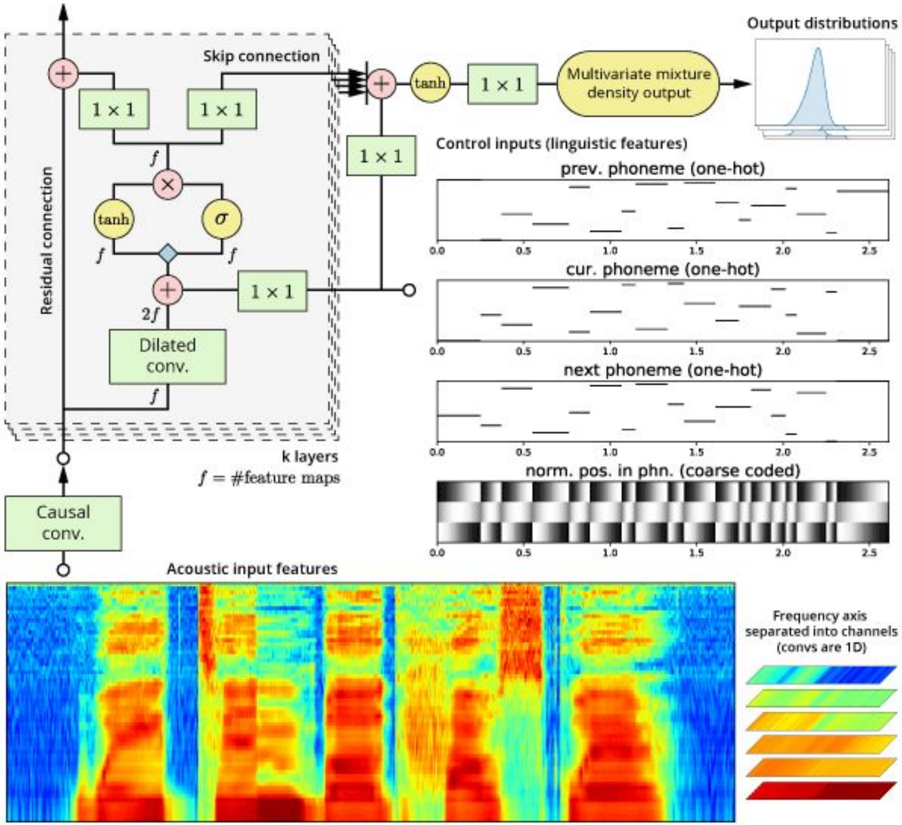


Current research

Blaauw, Merlijn, and Jordi Bonada, “A Neural Parametric Singing Synthesizer Modeling Timbre and Expression from Natural Songs,” Applied Sciences, 2017.

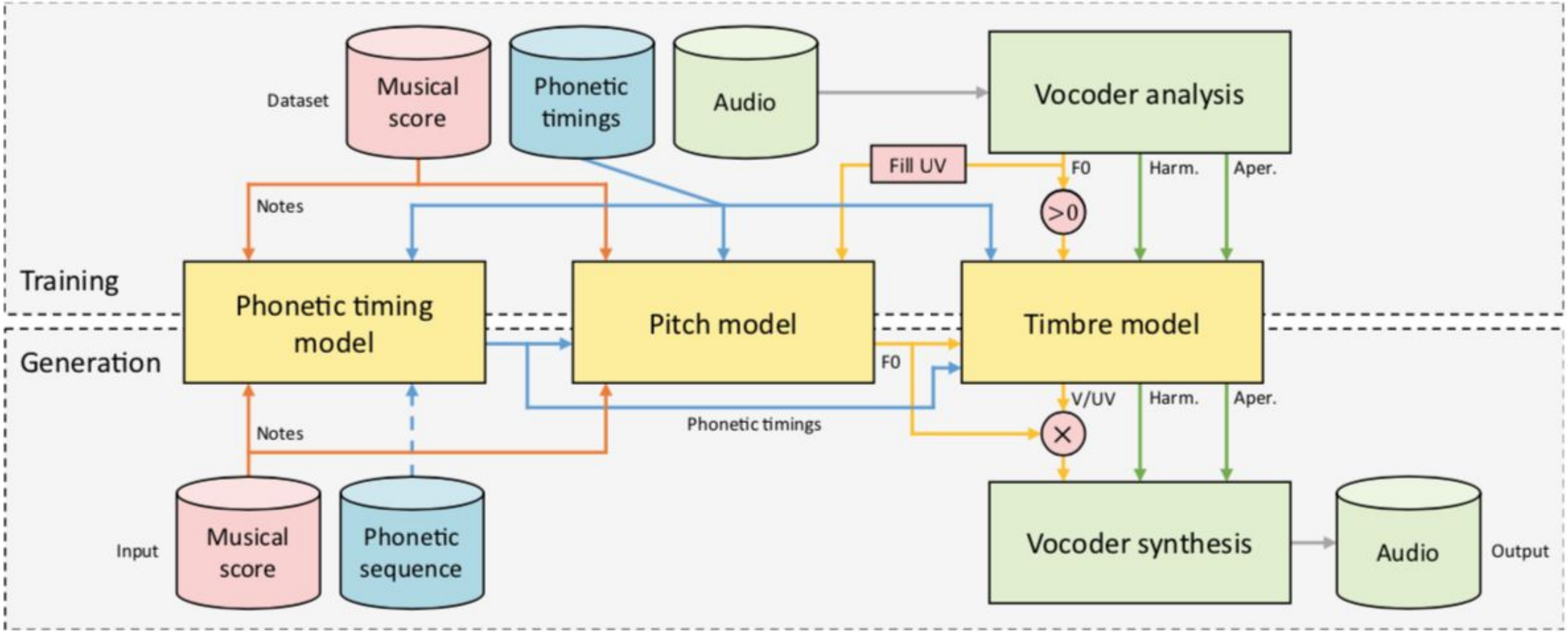
- Uses a modified wavenet architecture
- Models features of parametric vocoder instead of raw waveform
  - Separates influence of pitch and timbre
  - Fewer layers and model parameters
  - Allows training on smaller datasets (~30 minutes)
  - Reduces training and generation times (~8 hours of training)
- Generation is autoregressive, so errors may compound during synthesis
- Uses a constrained Gaussian Mixture output:
  - Mixture of 4 Gaussians with diagonal covariance
  - Using 4 free parameters - mean, variance, skewness, shape
  - Constrains possible output distributions

More details...



- Multi-stream network
  - Predicts harmonic spectral envelope, aperiodicity envelope, voiced/unvoiced
  - Modeled as independent networks, but takes one stream's output as additional input of another stream
- Acoustic features (from WORLD Vocoder[2])
  - 60-dimensional MFCCs
  - 4-dimensional band aperiodicity coefficients
- Control features
  - Previous, current and next phoneme identity (one-hot encoded)
  - Normalized position of frame within phoneme

Blaauw/Bonada model



Blaauw/Bonada model overview

- **Analysis part of vocoder**
  - Used to extract acoustic features.
- **Phonetic timing model**
  - Used to predict begin and end times of each phoneme.
    - During generation, have note begin and end times and phoneme sequence corresponding to each note (syllable), but don’t have access to begin and end times of each phoneme.
- **Pitch model**
  - Used to predict F0 from timed musical and phonetic information.
- **Timbre model**
  - Used to generate remaining acoustic features such as the harmonic spectral envelope, aperiodicity envelope, and voice/unvoiced (U/V) decision (from the predict phonetic timings and F0).
- **Synthesis part of vocoder**
  - Used to generate waveform signal from acoustic features.

Proposed improvements

- Learning to decorrelate speaker information from singing information
- This can be used to synthesize singing for new speakers with very little singer-specific data
- Changes to model architecture
  - Train an Auto-Encoder to learn embeddings for singers and/or styles, which are conditioned on during training of WaveNet framework.
  - This Auto-Encoder can be integrated with WaveNet and trained end-to-end.
- Generate short song in style of singer given new lyrics.

Data to be used

- Multiple datasets with clean singing and lyrics available:
  - MUSDB18
    - Studio-quality isolated drums, bass, vocals and other instruments
    - 150 full-length music tracks
  - Singing voice dataset
    - Singing musical scale recordings
    - Song recordings
    - Multiple recordings from 2 singers, 1 male and 1 female
- Use GENTLE for phoneme-level alignment of lyrics with the audio

Potential Performance Metrics

Quantitative Metrics:

- Mel-Cepstral Distortion (MCD)
- Band Aperiodcity Distortion (BAPD)
- Modulation Spectrum (MS) for Mel-Generalized Coefficients (MGC)
- Voiced/unvoiced decision metrics
- Timing metrics
- F0 metrics
- Modulation Spectrum (MS) for log F0

Listening Tests:

- Mean Opinion Score (MOS)
- Preference Test

Task Breakdown and assignments

1.	Collect and clean dataset	First 1-2 weeks
2.	Linguistic feature extraction	First 1-2 weeks
3.	Construct the baseline model	
4.	Train and test the baseline model	Within 4 weeks
5.	Transfer learning for new singers[5]	Within 8 weeks

References

1. Blaauw, Merlijn, and Jordi Bonada, “A Neural Parametric Singing Synthesizer Modeling Timbre and Expression from Natural Songs,” Applied Sciences, 2017.
2. WORLD Vocoder: https://github.com/mmorise/World
3. Gómez, Emilia, Blaauw, Merlijn, Bonada, Jordi, Chandna, Pritish, and Cuesta, Helena. “Deep Learning for Singing Processing: Achievements, Challenges and Impact on Singers and Listeners” (n.d.).
4. https://deepmind.com/blog/wavenet-generative-model-raw-audio/
5. Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis: https://google.github.io/tacotron/publications/speaker_adaptation/