A Neural Parametric Singing Synthesizer

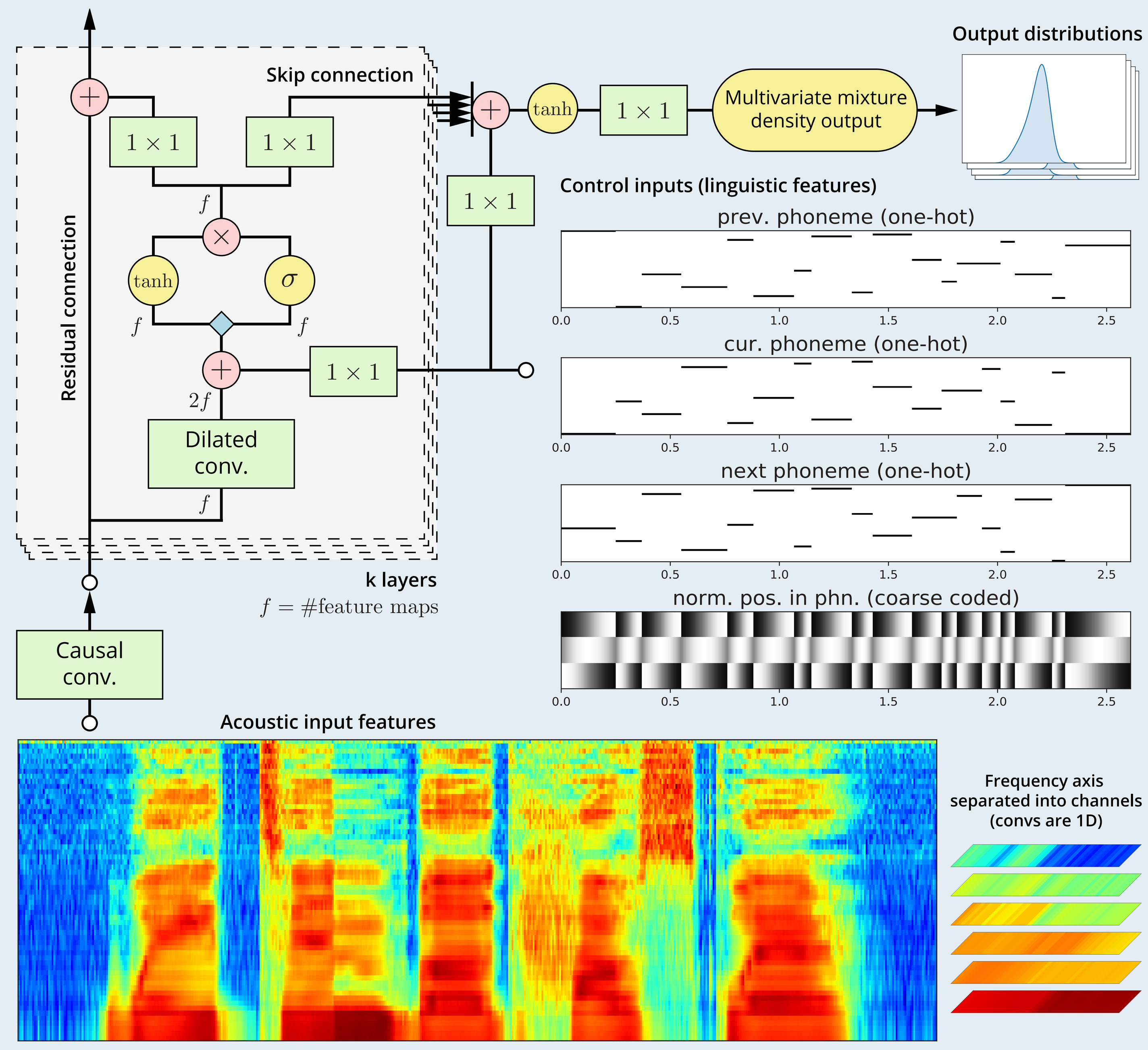
Merlijn Blaauw, Jordi Bonada
Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain

Key points

- Singing synthesizer based on WaveNet
- Models vocoder features rather than raw waveform
- Motivation
  - » Using a vocoder, the quality of resynthesis exceeds that of generative models; close the gap by improving model
  - » The large timbre-pitch space of singing voice can be reproduced with a relatively small amount of training data (e.g. 30 min.)
  - » Allows for faster synthesis, making application more practical
- Improved flexibility compared to sample-based approaches

Model and architecture

- Autoregressive probabilistic model, like WaveNet, with similar network architecture
- Uses dilated convolutions, gated activations, residual connections and skip outputs
- Scaled down to significantly less layers, while maintaining a similar receptive field
- Conditioned on a set of control inputs
- Input is 2D time-frequency data, rather than 1D waveform data
- The 2D input is processed using 1D convolutions, the input channels correspond to different frequency bins

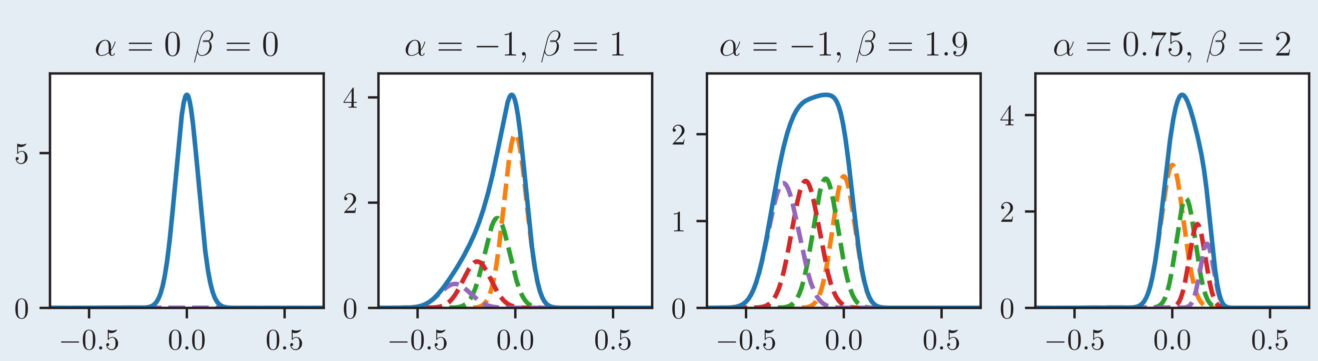


Front-end

- Acoustic features
  WORLD vocoder, 5 ms hop time, 32 kHz, reduced dimensionality
  - » Mel-frequency spectral coefficients, 60 dimensional
  - » Band aperiodicity coefficients, 4 dimensional
- Control features
  - » Previous, current, next phoneme identity (one-hot encoded)
  - » Normalized position of frame within phoneme (3-state coarse coded)

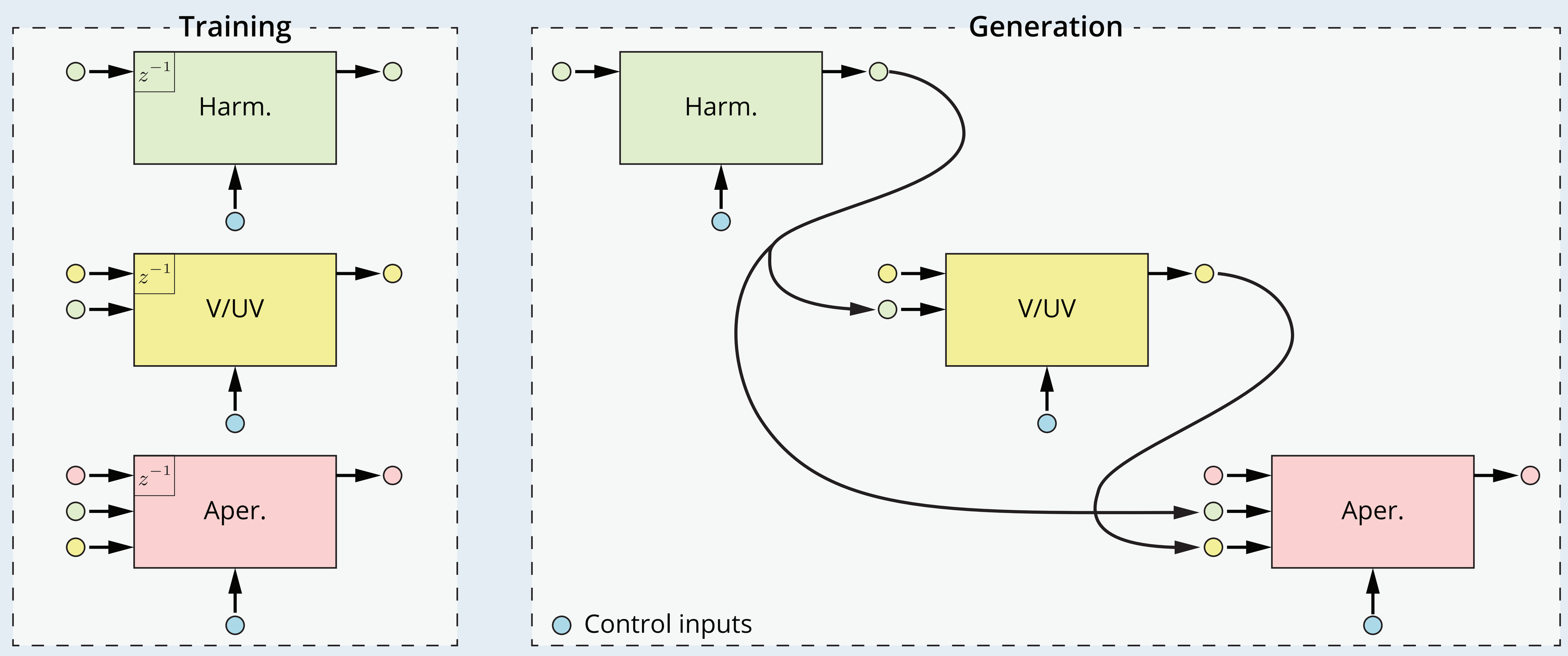
Constrained Gaussian mixture output

- This class of model typically predicts a categorical distribution over binned data
- A 256-way softmax per output feature requires too many parameters
- Instead, we use a mixture of 4 Gaussians, with diagonal covariance
- The 12 parameters of the mixture are obtained by mapping 4 free parameters: mean μ, variance σ^2, skewness α, shape β
- This mapping also constrains the possible output distributions; in particular to avoid distributions with multiple modes or very small variances



Multi-stream network

- Our model predicts several feature streams
  - » Harmonic spectral envelope, aperiodicity envelope and voiced/unvoiced decision
  - » Pitch and phoneme durations are not predicted in this work, but are obtained from an auxiliary model or target recording
- Streams are modeled as independent networks
- However, one stream's network may take other streams as additional input



Regularization

- Training is parallelized by using ground-truth past, but generation is autoregressive
- Even with good validation loss, errors may compound during synthesis
- An unregularized model often relies too much on past inputs and too little on control inputs, which can cause synthesized lyrics to change arbitrarily
- We propose a denoising objective; noise is added to all (non-control) inputs, but the clean signal is predicted

$$\mathcal{L} = -\log p(\mathbf{x}_t|\tilde{\mathbf{x}}_{<t}) \quad \text{with} \quad \tilde{\mathbf{x}}_{<t} \sim \mathcal{N}(\tilde{\mathbf{x}}_{<t}; \mathbf{x}_{<t}, \lambda I)$$

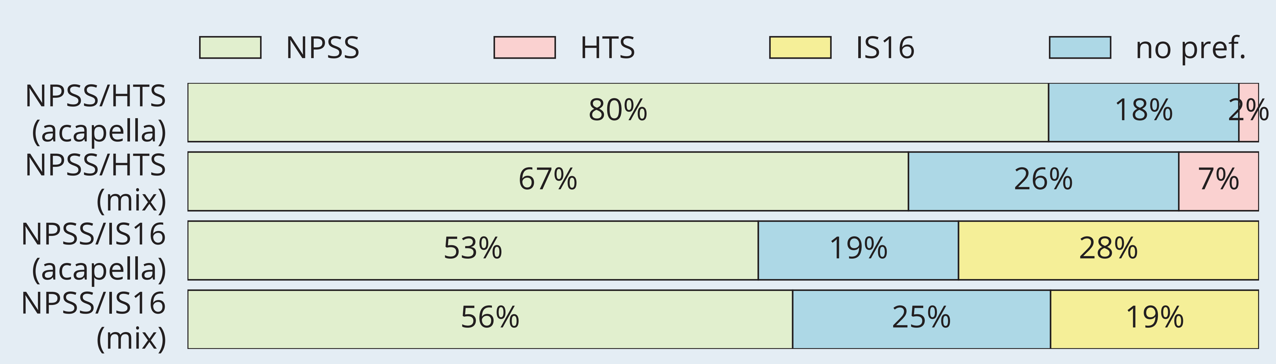
- Increased output noise can be alleviated by sampling from a corresponding lower temperature distribution at synthesis

Fast generation

- Autoregressive generation is generally slow because it cannot be parallelized
- Advantages of our model compared to modeling raw waveform
  - » Much lower sample rate (e.g. 200 vs. 16000 samples per second)
  - » Fewer layers and model parameters (e.g. 5 vs. 30 layers, 1.3M vs. 47M params.)
- Additionally, we use a fast generation algorithm based on efficient caching of computations, implemented on CPU (rather than GPU)
- We are able to achieve generation speeds of 10-20x real-time

Experiments, results and demos

- Two English voices, male and female (35 min., single pitch)
- One Spanish voice, female (16 min., single pitch)
- A/B preference listening tests for our system ("NPSS") vs. two baseline systems: HMM-based ("HTS") and concatenation-based ("IS16")



Conclusions

- Notably improves quality compared to conventional HMM-based approach
- Less reliant on "perfect" phonetic segmentation than sample-based methods
- Many practical applications thanks to fast generation and low memory footprint
- Very flexible approach, with many future directions; jointly modeling timbre and expression, multi-speaker training, model adaptation, ...

upf. Universitat Pompeu Fabra Barcelona
MTG Music Technology Group