In this work we focus on modeling the (harmonic) spectral envelope only.

**Dataset**
- Professional Spanish female singer
- Single pitch, constant cadence and no pauses
- 123 short sentences, total duration 10min, 120k frames

**Acoustic features**
- WORLD VOCODER [6], 5ms frame shift, 32kHz sample rate
- Reduce dimensionality to 35 mel-frequency scale filter bank features

**Linguistic features**
- Previous, current and next phoneme identity (one-hot encoded vectors)
- Position of frame within current phoneme (3-state coarse-coded vectors)
- No duration features (constant cadence dataset)

**Avoiding translations in frequency**
- Standard 2D convolutions are translation invariant in both spatial dimensions
- For spectrogram data we want to allow translation invariance only in the time dimension, not in the frequency dimension
- To encourage this behavior we condition the frequency dimension on its one-hot encoded bin index

**Handling varying length sequences**
- Spectrogram sequences can be arbitrarily long and vary in length
- To facilitate training we split sequences into fixed length sub-sequences
- We use a leading overlap that covers at least the receptive field
- Out of bounds data is zero-padded, and the loss is masked accordingly

**Improving generation robustness**
- Small, simple dataset and high capacity model causes over-fitting
- Over-fits to training objective (ground truth context), poor autoregressive generation (predicted context)
- Reduce over-fitting using denoising PixelCNN objective

- \[ \mathcal{L} = - \mathbb{E}_x \mathbb{E}_{x_{<i}} \log p(x_i | x_{<i}) + \log p(x_i | x_{<i}, h) \]
- \[ p(x | x) = \mathcal{N}(\mathbf{x}; \mathbf{x}, \Lambda) \]

**References**