Attention, I'm Trying to Speak! Speech Synthesis

Akash Mahajan (CS224n Final Project)

akashmjn@stanford.edu

Stanford | ENGINEERING

I. Problem Formulation

- We implement and end to end text-to-speech (TTS) model from Tachibana et. al.
- We show that it is possible to build a decent quality TTS model in \$50-\$100
- Attention is a bottleneck - experimented with two modifications for improvement
- Interesting observations made about intermediate modules / embeddings

II. Dataset

- 13k unaligned 1-10s audio-sentence pairs from female English speaker. Total size ~24h
- Text preprocessed to 32 char tokens in `a-z'.?<space>PE` ($\mathbf{L}_{Nx32}$)
- Audio converted to freq-domain as normalized spectrograms ($\mathbf{Y}_{Tx80}$) ($\mathbf{Z}_{4Tx513}$) - [0,1]

III. Model

- Text2Mel - convolutional seq2seq model (Gehring et. al): characters to $\mathbf{Y}_{Tx80}$ frames
- SSRN - upscale $\mathbf{Y}_{Tx80}$ frames to $\mathbf{Z}_{4Tx513}$ frames
- Text2Mel during training:

$$\mathbf{K}, \mathbf{V}_{N\times d} = TextEnc(\mathbf{L}_{N\times e})$$

$$\mathbf{Q}_{T\times d} = AudioEnc(\mathbf{S}_{T\times F})$$

$$\mathbf{S}_{T\times F} = 0 \oplus \mathbf{Y}_{0:F-1} \quad \text{(shifted left)}$$

$$\mathbf{R}_{T\times d} = \mathbf{AV} = softmax\left(\frac{\mathbf{QK}^T}{\sqrt{d}}\right)\mathbf{V}$$

$$\hat{\mathbf{Y}}_{T\times F} = AudioDec(\mathbf{R} \oplus \mathbf{Q})$$

$$\hat{\mathbf{Z}}_{4T\times F_o} = SSRN(\mathbf{Y})$$

$$\mathbf{J}_{L1} = \mathbb{E}|\mathbf{Y} - \hat{\mathbf{Y}}|$$

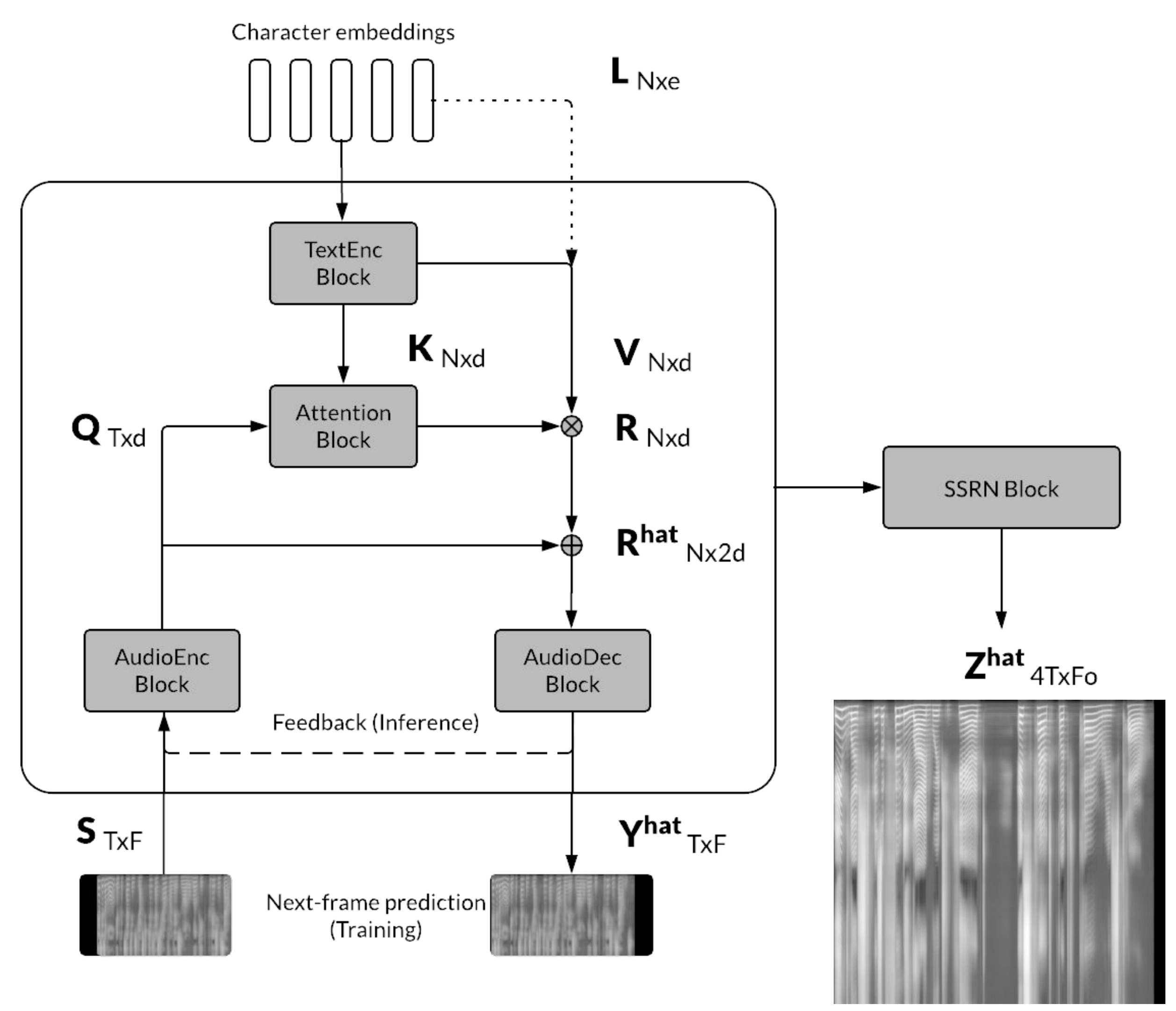
$$\mathbf{J}_{CE} = -\mathbb{E}[\mathbf{Y}\log\hat{\mathbf{Y}} + (1-\mathbf{Y})\log(\mathbf{1}-\hat{\mathbf{Y}})]$$

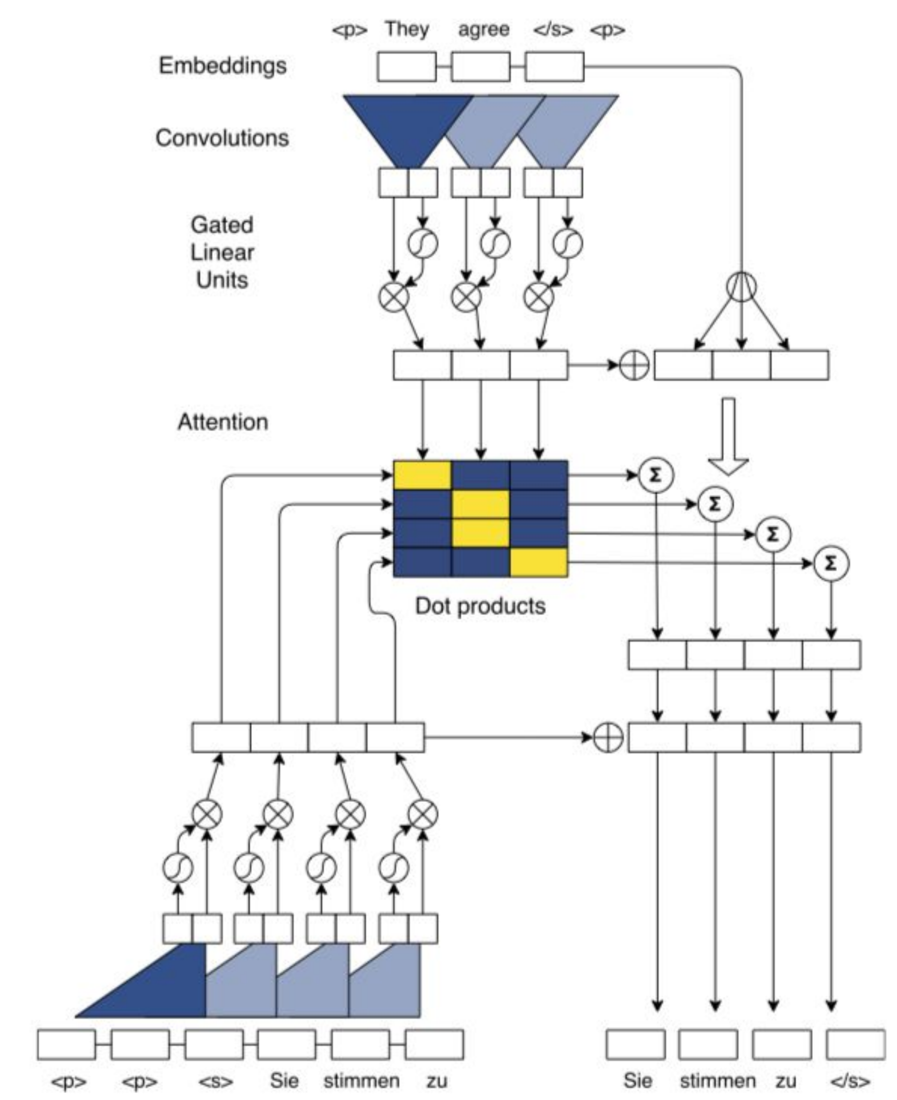
- Text2Mel during inference:

$$\mathbf{S}_{1:t+1,F} = \mathbf{S}_{1:t,F} \oplus \hat{\mathbf{Y}}_{t,F} \quad \text{(feedback)}$$

$$\hat{\mathbf{Z}}_{4T\times F_o} = SSRN(\hat{\mathbf{Y}})$$

IV. Approach & Experiments





Guided Attention Loss

- Standard attention does not converge
- L1 loss keep decreasing (AudioEnc, AudioDec)
- Generating from this model produces gibberish sounds that sound like speech!
- An additional loss that penalizes terms in **A** that are far from the diagonal speeds up convergence (Tachibana)

$$W_{n,t} = (1 - \exp(-n/N) - t/T)^2/2g^2$$

$$\mathbf{J}_{att} = \mathbb{E}(A \circ W)$$

Positional, Local Encodings

- Combine character-level embeddings $\mathbf{L}_{Nx32}$ with value encoding **V**.
- Attention needs to be initialized properly - use a prior that enforces a linear monotonicity
- A value h_p added to both **K** and **Q**. Produces best scores, and greatly improves quality of char embeddings

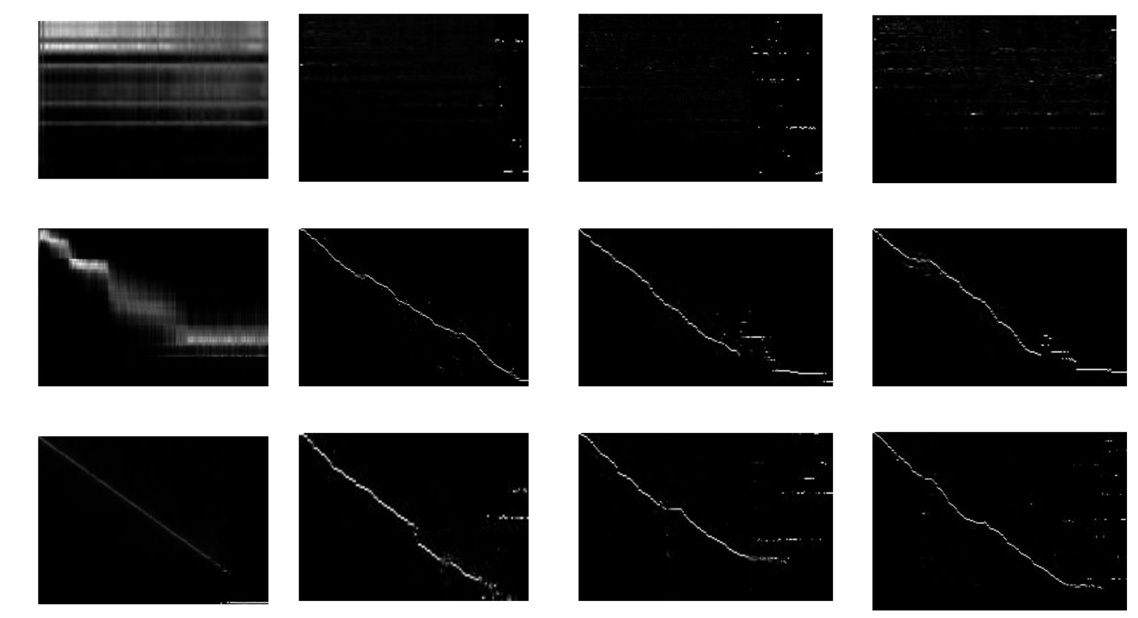
$$h_p(i) = sin(\omega_s i/10000^{k/d}) \quad \text{(even i)}$$

$$= cos(\omega_s i/10000^{k/d}) \quad \text{(odd i)}$$

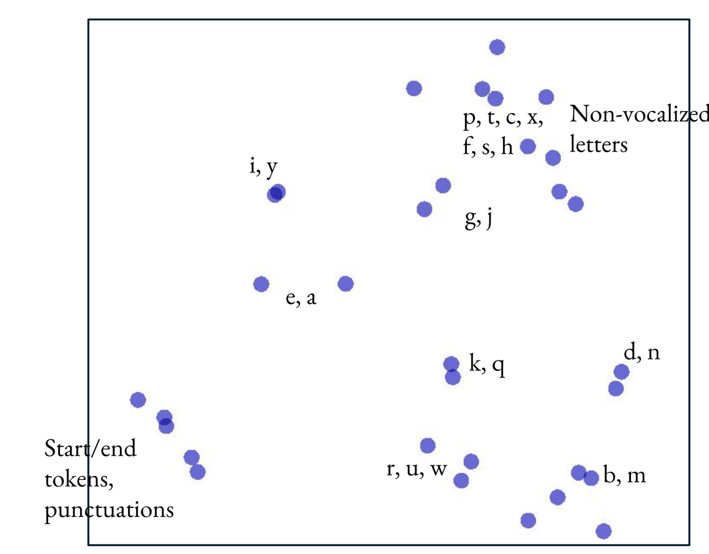
Results of Model Variation Experiments (best and lowest highlighted)

Model	Variation	L1 (Train)	L1 (Validation)	Guided Attention
M1	Standard attention, CE loss	0.0288	**0.0611**	$\mathbf{27.5 \times 10^{-4}}$
M2	M1 + guided attention	0.0249	0.0484	3.99×10^{-4}
M3	M2 + local char encodings	0.0245	0.0485	4.19×10^{-4}
M4	M1 + positional encodings	**0.0230**	0.0490	8.42×10^{-4}
M5	M4 without CE loss	0.0235	0.0490	17×10^{-4}

V. Observations



- Learned attention alignments converge faster
- Common attention failure modes: discontinuities/jumps, sparse outputs at silences
- Learned character embedding capture sound/semantic relationships (below)



Figures:
(above) attention alignments during training at 5k, 20k, 40k, 60k steps L-R for M1 (top), M2 (middle), M5 (bottom)

(left) T-SNE plot of character embeddings learned by M5 model.

VI. Discussion and Future Work

- The model works surprisingly well compared to ones like Tacotron that train for over a week
- Attention can be constrained during inference to avoid common failure modes
- Can explore semi-supervised learning by first training just AudioEnc, AudioDec like a "audio language model" without labels
- Transfer to new language with lesser data
- MOS-like subjective evaluation scores

VII. Selected References

- H. Tachibana, K. Uenoyama, and S. Aihara. Efficiently Trainable Text-to-Speech System Based on Deep Convolutional Networks with Guided Attention. 10 2017.
- J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin. Convolutional Sequence to Sequence Learning. 5 2017
- W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller. Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning. 10 2017.