Advanced Methods in Machine Learning 07.06.2018

Singing Voice Generation

Nir Raviv 200683548 *Roi Tabach 203022983* *Andrey Leshenko 322026527*

nirraviv@mail.tau.ac.il *roi.tabach@gmail.com* *andrey.leshenko@gmail.com*

1 Introduction

(Andrey)
The problem of generating singing voices that sound as close to human as possible. Given text + music notes (MIDI), output a singing voice of the text.

2 Related Works

During the recent years many neural networks architecture based on fully-visible probabilistic autoregressive generative models are applied in many fields, such as natural images (PixelCNN) [13], raw audio waveform (WaveNet) [12] and video (Video Pixel Networks) [5]. These methods predict the distribution for each sample conditioned on all previous ones while remain efficiently trained.

Since WaveNet was published in 2016, there has been several major attempts at the problem of synthesizing speech with neural networks, including Deep Voice 1 [1], 2 [2], and 3 [10], Tacotron [14], Char2Wav [11], and others. Deep Voice 1 and 2 are built as TTS pipelines, separating grapheme-to-phoneme conversion, duration and frequency prediction, and waveform synthesis. Deep Voice 3, Tacotron and Char2Wav propose sequence-to-sequence models for neural TTS. Griffin-Lim *et al.* [4] text-to-spectrogram conversion model, Tacotron, used spectrogram-to-waveform synthesis. Char2Wav predicts the parameters of the WORLD vocoder [7] and uses a SampleRNN [6] conditioned upon WORLD parameters for waveform generation. These works achieve state of the art performance, with human listeners rating as more natural sounding than the previous state of the art methods of concatenative methods.

Different singing generators are based on statistical parametric methods centered around Hidden Markov Models (HMM) which allow joint modeling of timbre and musical expression but perform less than the previous methods by causing “buzzy” sound [9]. This work was extended to feedforward DNNs [8].

Merlijin *et al.* [3] presented synthesizer which can generate synthetic singing voice given musical score with phonetic lyrics using different models to learn phonetic timing, pitch and timbre combining autoregressive generative model with mixture density output instead of softmax output allow skewness or truncated distribution and multiple modes.

3 Proposed Method

(Andrey)
We will train on one language and generate on a different language. Moreover, we will try and generate singing voice from text, not from a phoneme sequence what is your approach and why it will improve things.
o; ***note: our main difference is using of non-phonetic lyrics as far as I understand [3] using phonetic lyrics. see section 2.6!***

4 Evaluation Method

We will evaluate both TTS and singing generator separately and combined using mean opinion score (MOS) tests on the following datasets:

- NIT-SONG070-F001 dataset published by the Nagoya Institute of Technology (Nitech)
- MIR1K dataset containing 1000 sentences, extracted from karaoke of chinese pop songs
- Stanford's DAMP dataset that contains 10,000s of English songs recorded by the Sing! Android karaoke app.

We will compare our final results to the following publicly accessible systems:

- Sinsy-HMM [9]
- Sinsy-DNN [8]
- The Original songs

5 References

[1] Sercan O Arik, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, et al. Deep voice: Real-time neural text-to-speech. *arXiv preprint arXiv:1702.07825*, 2017. 1-1

[2] Sercan O Arık, Gregory Diamos, Andrew Gibiansky, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou. Deep voice 2: Multi-speaker neural text-to-speech. *arXiv preprint arXiv:1705.08947*, 2017. 1-1

[3] Merlijn Blaauw and Jordi Bonada. A neural parametric singing synthesizer. *arXiv preprint arXiv:1704.03809*, 2017. 1-1, 1-2

[4] Daniel Griffin and Jae Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2):236–243, 1984. 1-1

[5] Nal Kalchbrenner, Aaron van den Oord, Karen Simonyan, Ivo Danihelka, Oriol Vinyals, Alex Graves, and Koray Kavukcuoglu. Video pixel networks. *arXiv preprint arXiv:1610.00527*, 2016. 1-1

[6] Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, and Yoshua Bengio. Samplernn: An unconditional end-to-end neural audio generation model. *arXiv preprint arXiv:1612.07837*, 2016. 1-1

[7] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa. World: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE TRANSACTIONS on Information and Systems*, 99(7):1877–1884, 2016. 1-1

[8] Masanari Nishimura, Kei Hashimoto, Keiichiro Oura, Yoshihiko Nankaku, and Keiichi Tokuda. Singing voice synthesis based on deep neural networks. In *INTERSPEECH*, pages 2478–2482, 2016. 1-1, 1-2

[9] Keiichiro Oura, Ayami Mase, Yoshihiko Nankaku, and Keiichi Tokuda. Pitch adaptive training for hmm-based singing voice synthesis. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 5377–5380. IEEE, 2012. 1-1, 1-2

[10] Wei Ping, Kainan Peng, Andrew Gibiansky, S Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. Deep voice 3: Scaling text-to-speech with convolutional sequence learning. In *Proc. 6th International Conference on Learning Representations*, 2018. 1-1

[11] Jose Sotelo, Soroush Mehri, Kundan Kumar, Joao Felipe Santos, Kyle Kastner, Aaron Courville, and Yoshua Bengio. Char2wav: End-to-end speech synthesis. 2017. 1-1

[12] Aaron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016. 1-1

[13] Aaron van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. In *Advances in Neural Information Processing Systems*, pages 4790–4798, 2016. 1-1

[14] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*, 2017. 1-1