CS482/682 Final Project Report Group 01

Title In Progress: Deep Fakes for Good

Jessica Soong (jsoong1), Cindy Huang (chuang95), Steve Luo (sluo15)

1 Introduction

Background People with amyotrophic lateral sclerosis (ALS) have increasing difficulty for speech and benefit from speech brain-computer interfaces to communicate with others [1]. However, deep neural network (DNN) models translating neural signals into speech can only be trained with generic speech for patients who have already lost speaking ability when they receive device implantation. One way to address the problem could be to develop a DNN to generate a more nature-sounding custom voice for the subjects.

Related Work Several deep learning models already exist for voice generation. Some of these include WaveNet [2]; DeepVoice [3]; LPCNet [4]; WaveRNN [5]; WaveGlow, a flow-based network that combines aspects of WaveNet and Glow [6]; Deep Convolutional TTS (DC-TTS), based on a fully convolutional sequence-to-sequence learning model [7]; and SampleRNN, a network that uses RNNs at different scales to model longer term dependencies [8].

2 Methods

Dataset Depending on the pre-trained model implemented, different datasets will be used. The two selected are the *TSP Speech Database* [9] and the *LJ Speech Dataset* [10]. The TSP speech database contains over 1400 utterances spoken by 24 speakers [9]. The database includes the original samples, as well as down-sampled versions of the data. We will only select samples from one speaker. The LJ dataset contains 13,100 short audio clips and transcriptions of a single speaker reading passages from 7 non-fiction books [10]. Once the transfer learning is confirmed to work with the aforementioned datasets, audio and text will be stripped from the Deep Learning lectures to train a model to read in Dr. Unberath's voice.

For preprocessing, the audio data will be transformed into a normalized Mel Spectrogram, which is a visual representation of an audio signal. Depending on the performance of the network, different ways to process the input text will be applied.Underused words can be removed, and the transcripts can be modified to use phonetic spelling.

Setup, Training and Evaluation Of the models mentioned so far, WaveGlow, SampleRNN, and DC-TTS show the most promise. They all have released/replicated pre-trained models, as well as GitHub repositories with code [11] [12] [13]. WaveGlow's benefits are that is parallelized and simpler than WaveNet [6]. SampleRNN is promising since in real-time deployment the speed of TTS is incredibly important, and SampleRNN has been shown to be 6 times faster than WaveRNN [14]. DC-TTS is needs minimal hardware to train (1080 Ti) and has a lower training time, but produces lower quality samples [7]. For simplicity, we will first take on DC-TTS due to the time and resource constraints.

For evaluation we will use short-term intelligibility score (STOI) to evaluate speech coherence, which has been shown to perform well when compared with subjective tests [15]. We will also use mean opinion score (MOS), which is a subjective test, where clips of generated and real audio are scored on a scale from 0 to 5, and averaging the results, with the real audio as a control [16].

References

[1] L. M. McCane, E. W. Sellers, D. J. McFarland, J. N. Mak, C. S. Carmack, D. Zeitlin, J. R. Wolpaw, and T. M. Vaughan, "Brain-computer interface (bci) evaluation in people with amyotrophic lateral sclerosis," *Amyotrophic lateral sclerosis and frontotemporal degeneration*, vol. 15, no. 3-4, pp. 207–215, 2014.

[2] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[3] S. Ö. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, J. Raiman, S. Sengupta, and M. Shoeybi, "Deep voice: Real-time neural text-to-speech," *CoRR*, vol. abs/1702.07825, 2017.

[4] J.-M. Valin and J. Skoglund, "Lpcnet: Improving neural speech synthesis through linear prediction," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5891–5895, IEEE, 2019.

[5] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. van den Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," *CoRR*, vol. abs/1802.08435, 2018.

[6] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3617–3621, IEEE, 2019.

[7] H. Tachibana, K. Uenoyama, and S. Aihara, "Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention," *CoRR*, vol. abs/1710.08969, 2017.

[8] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville, and Y. Bengio, "Samplernn: An unconditional end-to-end neural audio generation model," *arXiv preprint arXiv:1612.07837*, 2016.

[9] M. T. . S. P. Laboratory, "Speech database."

[10] K. Ito and L. Johnson, "The lj speech dataset." https://keithito.com/LJ-Speech-Dataset/, 2017.

[11] R. V. Ryan Prenger and B. Catanzaro, "Waveglow: a flow-based generative network for speech synthesis."

[12] J. R. Piotr Kozakowski, Katarzyna Kańska, "samplernn-pytorch."

[13] K. Park, "A tensorflow implementation of dc-tts: yet another text-to-speech model."

[14] P. Govalkar, J. Fischer, F. Zalkow, and C. Dittmar, "A comparison of recent neural vocoders for speech signal reconstruction," pp. 7–12, 09 2019.

[15] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE international conference on acoustics, speech and signal processing*, pp. 4214–4217, IEEE, 2010.

[16] I. T. Union, "Methods for subjective determination of transmission quality."