Concatenative Speech Synthesis
An Annotated Bibliography

Chase Adams (chasea2@illinois.edu)
University of Illinois at Urbana-Champaign

October 9, 2019

References

[1] S. Bansal, H. Kamper, K. Livescu, A. Lopez, and S. Goldwater, "Pre-training on high-resource speech recognition improves low-resource speech-to-text translation," *arXiv preprint arXiv:1809.01431*, 2018.

> The authors of this paper investigate direct speech-to-text translation when the source language islow-resource. They first modelled 300 hours of high-resource English ASR data and then applied transfer learning approaches to fine-tune those model parameters for the low-resource language at hand. They use BLEU machine translation scores to quantify the end-to-end system results. Much of this study focuses on the pre-training of the ASR encoder and how it yields the greatest improvement to the system as a whole.

[2] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

> This paper presents a concatenative synthesizer that achieves SOTA metrics on a number of tasks. Importantly the authors outline the Subjective 5-scale mean opinion score (MOS) which is an industry standard for detailing naturalness in speech. Interestingly the encoder-decoder framework focuses on translating text into a spectrogram which can then be interpreted into a waveform instead of the other way around. No hand-engineered HMM aligner is required to make this work.