Sainath, Tara N., Weiss, Ron J., Senior, Andrew, Wilson, Kevin W., and Vinyals, Oriol. Learning the speech front-end with raw waveform CLDNNs. In *Interspeech*, pp. 1–5, 2015.

Takaki, Shinji and Yamagishi, Junichi. A deep auto-encoder based low-dimensional feature extraction from FFT spectral envelopes for statistical parametric speech synthesis. In *ICASSP*, pp. 5535–5539, 2016.

Takamichi, Shinnosuke, Toda, Tomoki, Black, Alan W., Neubig, Graham, Sakriani, Sakti, and Nakamura, Satoshi. Postfilters to modify the modulation spectrum for statistical parametric speech synthesis. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 24(4):755–767, 2016.

Theis, Lucas and Bethge, Matthias. Generative image modeling using spatial LSTMs. In *NIPS*, pp. 1927–1935, 2015.

Toda, Tomoki and Tokuda, Keiichi. A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. *IEICE Trans. Inf. Syst.*, E90-D(5):816–824, 2007.

Toda, Tomoki and Tokuda, Keiichi. Statistical approach to vocal tract transfer function estimation based on factor analyzed trajectory hmm. In *ICASSP*, pp. 3925–3928, 2008.

Tokuda, Keiichi. Speech synthesis as a statistical machine learning problem. `http://www.sp.nitech.ac.jp/~tokuda/tokuda_asru2011_for_pdf.pdf`, 2011. Invited talk given at ASRU.

Tokuda, Keiichi and Zen, Heiga. Directly modeling speech waveforms by neural networks for statistical parametric speech synthesis. In *ICASSP*, pp. 4215–4219, 2015.

Tokuda, Keiichi and Zen, Heiga. Directly modeling voiced and unvoiced components in speech waveforms by neural networks. In *ICASSP*, pp. 5640–5644, 2016.

Tuerk, Christine and Robinson, Tony. Speech synthesis using artificial neural networks trained on cepstral coefficients. In *Proc. Eurospeech*, pp. 1713–1716, 1993.

Tüske, Zoltán, Golik, Pavel, Schlüter, Ralf, and Ney, Hermann. Acoustic modeling with deep neural networks using raw time signal for LVCSR. In *Interspeech*, pp. 890–894, 2014.

Uria, Benigno, Murray, Iain, Renals, Steve, Valentini-Botinhao, Cassia, and Bridle, John. Modelling acoustic feature dependencies with artificial neural networks: Trajectory-RNADE. In *ICASSP*, pp. 4465–4469, 2015.

van den Oord, Aäron, Kalchbrenner, Nal, and Kavukcuoglu, Koray. Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759*, 2016a.

van den Oord, Aäron, Kalchbrenner, Nal, Vinyals, Oriol, Espeholt, Lasse, Graves, Alex, and Kavukcuoglu, Koray. Conditional image generation with PixelCNN decoders. *CoRR*, abs/1606.05328, 2016b. URL `http://arxiv.org/abs/1606.05328`.

Wu, Yi-Jian and Tokuda, Keiichi. Minimum generation error training with direct log spectral distortion on LSPs for HMM-based speech synthesis. In *Interspeech*, pp. 577–580, 2008.

Yamagishi, Junichi. English multi-speaker corpus for CSTR voice cloning toolkit, 2012. URL `http://homepages.inf.ed.ac.uk/jyamagis/page3/page58/page58.html`.

Yoshimura, Takayoshi. *Simultaneous modeling of phonetic and prosodic parameters, and characteristic conversion for HMM-based text-to-speech systems*. PhD thesis, Nagoya Institute of Technology, 2002.

Yu, Fisher and Koltun, Vladlen. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016. URL `http://arxiv.org/abs/1511.07122`.

Zen, Heiga. An example of context-dependent label format for HMM-based speech synthesis in English, 2006. URL `http://hts.sp.nitech.ac.jp/?Download`.