Applications of deep learning: audio (speech)

This includes problems such as: speech denoising, speech recognition, text to speech, voice cloning.

5a. Text to speech
Citation: Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, Yonghui Wu. Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '18). [PDF] [Blog post] [Examples] [Github (via Mozilla, includes Colab notebook)] [Colab demo of Mozilla TTS (with pre-trained model)] [Also in ESPnet]
Citation: Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O. Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, John Miller. Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning. In Proceedings of the 2018 International Conference on Learning Representations (ICLR '18). [PDF] [Github (PyTorch implementation, includes notebooks)] [Examples]

5b. Speech to text
Citation: Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, Andrew Y. Ng. Deep Speech: Scaling up end-to-end speech recognition. arXiv:1412.5567 (2014). [PDF][Colab demo, pre-trained][Colab training demo] [Github (by Mozilla), documentation (includes information on training)][Discourse]
Citation: Sercan Arik, Gregory Diamos, Andrew Gibiansky, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, Yanqi Zhou. Deep Voice 2: Multi-Speaker Neural Text-to-Speech. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017). [PDF] [Blog post with code] [Github - TF] [Other pre-trained speech to text models: Wav2Letter+, Jasper, QuartzNet]

5c. Voice encoder (for speaker verification, speaker diarization, fake speech detection)
Citation: Li Wan, Quan Wang, Alan Papir, Ignacio Lopez Moreno. Generalized End-to-End Loss for Speaker Verification. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '18). [PDF] [Github: Resemblyzer. Trained via this repo]
[Also see pyannote.audio: Notebook, Github repo]

5d. Voice separation
Quan Wang, Hannah Muckenhirn, Kevin Wilson, Prashant Sridhar, Zelin Wu, John Hershey, Rif A. Saurous, Ron J. Weiss, Ye Jia, Ignacio Lopez Moreno. VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking. In Proceedings of the 20th Annual Conference of the International Speech Communication Association (INTERSPEECH '19). [PDF] [Video intro, video of live presentation] [Github (unofficial)][Github with pre-trained model and notebooks]

5e. Voice cloning

Citation: Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu. Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis. In Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018). [PDF] [Github (unofficial), including notebook using pretrained model. Notes on training, fine-tuning][Another notebook demo] (Warning: this project is not maintained) [Alternative approach in ESPnet]