

Contents

| | |
|---|-------------|
| Abstract | v |
| Resumen | vii |
| Resum | ix |
| Contents | xiii |
| 1 Introduction | 1 |
| 1.1 Framework and motivation | 1 |
| 1.2 Scientific and technological goals | 3 |
| 1.3 Document structure | 4 |
| 2 Preliminaries | 5 |
| 2.1 Machine Learning | 5 |
| 2.2 Sequence-to-Sequence with Attention Mechanism | 7 |
| 2.3 Transformer | 10 |
| 2.4 Generative Adversarial Networks | 13 |
| 2.5 Automatic Speech Recognition | 14 |
| 2.6 Machine Translation | 15 |
| 2.7 Text-To-Speech | 16 |
| 2.7.1 Text-to-spectrogram | 18 |
| 2.7.2 Spectrogram-to-wave | 18 |
| 2.7.3 Evaluation metrics | 20 |
| 2.8 Speech-To-Speech Translation | 22 |
| 2.8.1 Streaming ASR | 22 |
| 2.8.2 Simultaneous MT | 23 |
| 2.8.3 Incremental TTS | 24 |

| | | |
|----------|--|-----------|
| 3 | Cross-lingual Voice Cloning with Tacotron 2 | 25 |
| 3.1 | Introduction | 25 |
| 3.2 | Tacotron 2 | 26 |
| 3.3 | Extending Tacotron 2 with cross-lingual voice cloning capabilities | 31 |
| 3.4 | Overcoming the exposure bias and attention failures | 32 |
| 3.5 | Improving stop token prediction | 34 |
| 3.6 | Proposed model and general training procedure | 35 |
| 3.7 | Conclusions | 36 |
| 4 | Cross-lingual Voice Cloning for UPV[Media] | 37 |
| 4.1 | Introduction | 37 |
| 4.2 | The UPV[Media] platform | 38 |
| 4.3 | The Docència en Xarxa multilingual TTS dataset | 41 |
| 4.4 | Model training | 43 |
| 4.5 | Evaluation | 46 |
| 4.5.1 | Naturalness | 47 |
| 4.5.2 | Speaker similarity | 48 |
| 4.5.3 | Real or synthetic | 49 |
| 4.5.4 | Questionnaire and comments | 50 |
| 4.6 | Conclusions | 51 |
| 5 | Robust, Efficient and Controllable Neural Text-To-Speech | 53 |
| 5.1 | Introduction | 53 |
| 5.2 | Non-autoregressive TTS with explicit duration modeling | 54 |
| 5.3 | GAN-based neural vocoders | 57 |
| 5.4 | The Blizzard Challenge 2021 | 58 |
| 5.4.1 | Introduction | 58 |
| 5.4.2 | Data processing | 59 |
| 5.4.3 | Forced-aligner autoencoder model | 60 |
| 5.4.4 | Acoustic model | 62 |
| 5.4.5 | Vocoder model | 64 |
| 5.4.6 | Subjective results | 64 |
| 5.5 | Conclusions | 68 |
| 6 | Simultaneous Speech-To-Speech Translation | 71 |
| 6.1 | Introduction | 71 |
| 6.2 | The Europarl-ST dataset | 72 |
| 6.3 | Streaming ASR | 73 |

| | | |
|----------|--|------------|
| 6.4 | Simultaneous Machine Translation | 73 |
| 6.5 | Incremental Multilingual Text-To-Speech | 75 |
| 6.5.1 | Adapted prefix-to-prefix framework | 75 |
| 6.5.2 | Model architecture | 76 |
| 6.5.3 | Experiments | 77 |
| 6.5.4 | Evaluation | 79 |
| 6.6 | S2S latency evaluation | 81 |
| 6.7 | Conclusions | 82 |
| 7 | Zero-Shot Speaker Adaptation | 85 |
| 7.1 | Introduction | 85 |
| 7.2 | Speaker conditioning via transfer learning | 86 |
| 7.3 | The LibriTTS multi-speaker English corpus | 87 |
| 7.4 | Proposed zero-shot multi-speaker architecture | 88 |
| 7.5 | Least Squares Generative Adversarial Networks for TTS acoustic modeling | 90 |
| 7.6 | Experiments | 93 |
| 7.7 | Evaluation | 95 |
| 7.8 | Integration into UPV[Media] transcription and translation pipeline | 97 |
| 7.9 | Conclusions | 99 |
| 8 | Conclusions and future work | 101 |
| 8.1 | Scientific and technological achievements | 101 |
| 8.2 | Publications | 102 |
| 8.3 | Future work | 104 |
| | List of figures | 105 |
| | List of tables | 107 |
| | Bibliography | 109 |