Abstract

This work explores multilingual speech synthesis. We compare three models based on Tacotron that utilize various levels of parameter sharing. Two of them follow recent multilingual text-to-speech systems. The first one makes use of a fully-shared encoder and an adversarial classifier that removes speaker-dependent information from the encoder. The other uses language-specific encoders. We introduce a new approach that combines the best of both previous methods. It enables effective parameter sharing using a meta-learning technique, preserves encoder's flexibility, and actively removes speaker-specific information in the encoder. We compare the three models on two tasks. The first one aims at joint multilingual training on ten languages and reveals their knowledge-sharing abilities. The second concerns code-switching. We show that our model effectively shares information across languages, and according to a subjective evaluation test, it produces more natural and accurate code-switching speech.