Eva Székely (KTH Stockholm)

Breathing and speech planning in spontaneous speech synthesis

Breathing and speech planning in spontaneous speech are coordinated processes, often exhibiting disfluent patterns. While synthetic speech is not subject to respiratory needs, integrating breath into synthesis has advantages for naturalness and recall. At the same time, a synthetic voice reproducing disfluent breathing patterns learned from the data can be problematic. To address this, we first propose training stochastic TTS on a corpus of overlapping breath-group bigrams, to take context into account. Next, we introduce an unsupervised automatic annotation of likely-disfluent breath events, through a product-of-experts model that combines the output of two breath-event predictors, each using complementary information and operating in opposite directions. This annotation enables creating an automatically-breathing spontaneous speech synthesiser with a more fluent breathing style. A subjective evaluation on two spoken genres (impromptu and rehearsed) found the proposed system to be preferred over the baseline approach treating all breath events the same.