Using generative modelling to produce varied intonation for speech synthesis

Zack Hodari, Oliver Watts, Simon King

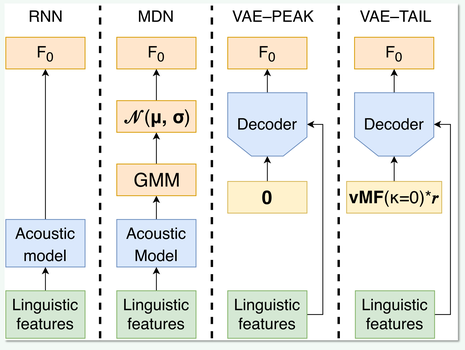
Centre for Speech Technology Research,
University of Edinburgh

Correspondence: zack.hodari@ed.ac.uk

*Our model can produce more varied intonation **without** sacrificing naturalness*

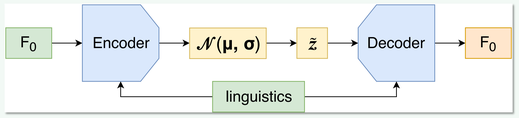
speech samples

Models

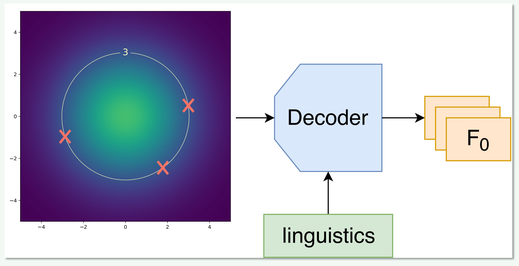


1. RNN and MDN are similar to standard SPSS-based TTS. VAE–PEAK should produce average prosody similar to RNN

VAE–TAIL should produce more varied intonation

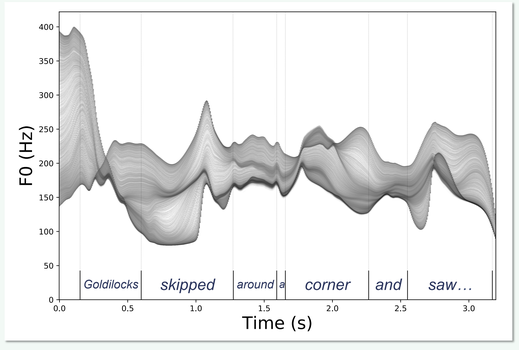


2. Variational autoencoder



3. VAE–TAIL

Sampling 3 F0 contours from VAE prior



4. Density plot of 10,000 F0 contours sampled from VAE prior

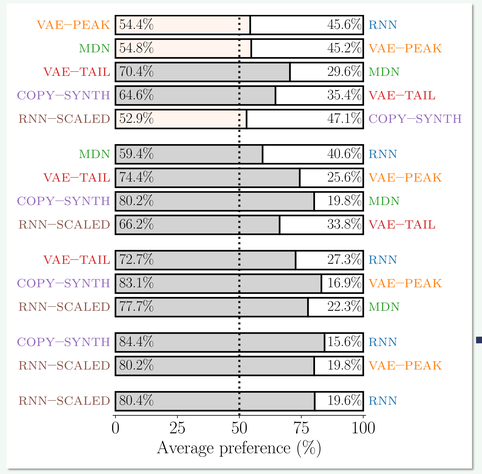
Overview

Normal speech synthesis voices produce average prosody

Most methods to alleviate this reduce the naturalness of the voice

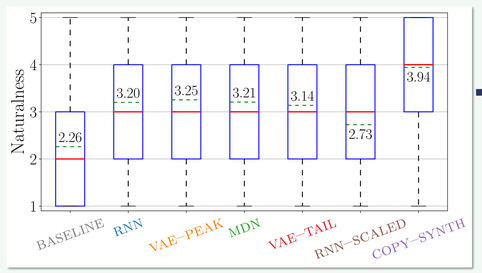
Our method can produce multiple renditions of a sentence

We demonstrate that our model's output is significantly more varied but not at the expense of naturalness



6. Pairwise preference

Q: *"Choose which clip has more varied intonation"*



7. MOS

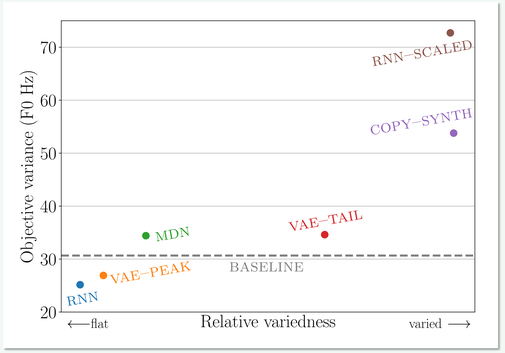
Q: *"Rate the naturalness of each clip"*

Listening tests

We evaluate naturalness and variedness, allowing us to compare the trade-off between these factors

We predict F0 and use parameters from natural speech for synthesis

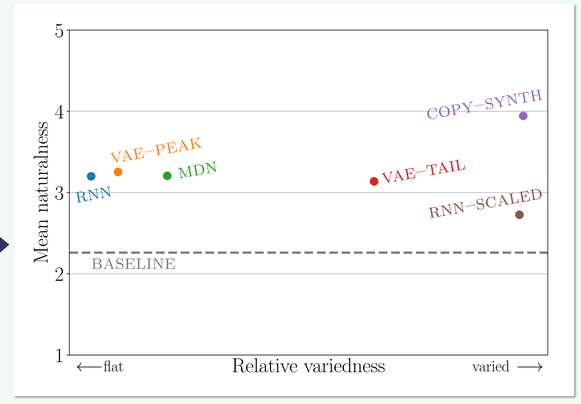
Subjective and objective F0 variation do not directly correspond, therefore VAE–TAIL and RNN–SCALED were calibrated by ear to match the level of variation in COPY–SYNTH



5. Subjective vs. objective intonation variation

8. Relative variedness

(derived from pairwise preference results)



9. Naturalness–Variedness tradeoff

VAE–TAIL has same naturalness as other TTS voices, but is much more varied