CAMP (Context aware model of prosody)

A two-stage approach to modelling prosody in context

Zack Hodari, Alexis Moinet, Sri Karlapati, Jaime Lorenzo-Trueba, Thomas Merritt, Arnaud Joly, Ammar Abbas, Penny Karanasou, Thomas Drugman

Correspondence: zack.hodari@ed.ac.uk

Paper 2639

Motivation

Prosody is hard to define, so we learn a disentangled representation of prosody (**Stage-1**).

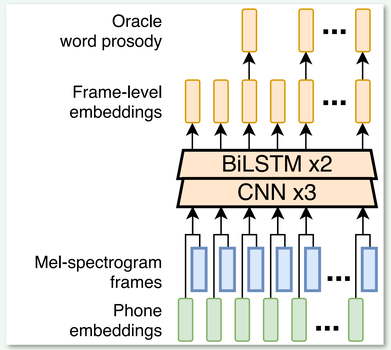
Prosody is determined by context information. This missing context must be incorporated into TTS models when predicting prosody, we achieved this using additional features (**Stage-2**).

Use more context to directly predict prosody

Stage-1

The reference encoder in ORA learns a prosodic representation. Disentanglement is encouraged using:

1. An information bottleneck in time using a word-level representation, and
2. Conditioning on phonetic information already available to the model through the phone encoder.

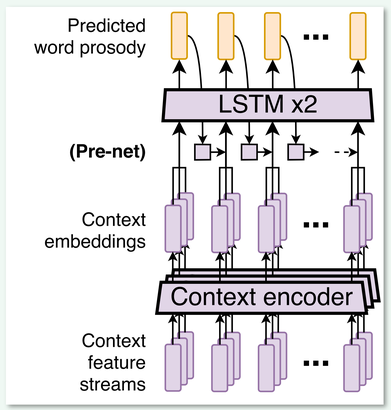


Stage-2

A prosody predictor is trained to mimic the reference encoder using one or more context features.

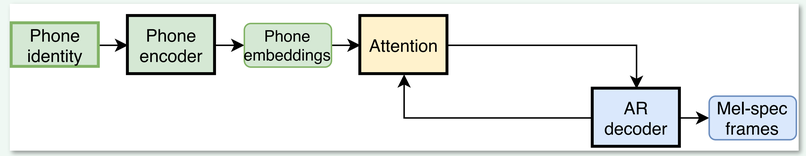
Our best prosody model used one context encoder, a pre-trained BERT, this was fine-tuned on the task of prosody representation prediction.

More context can be added through additional features, or by training on longer input sequences.



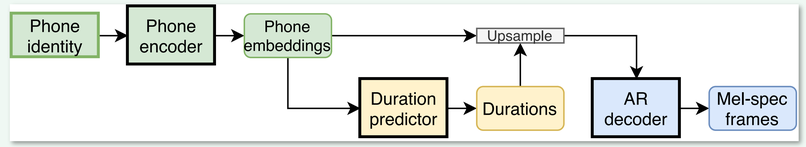
S2S

Tacotron-2 like model using attention



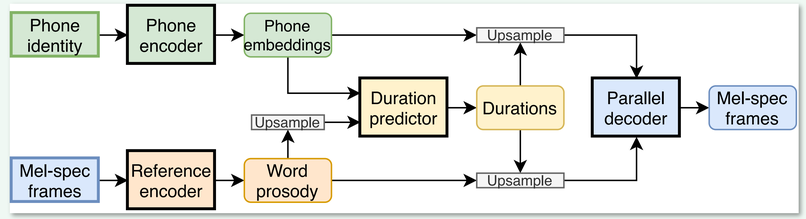
DurIAN+

S2S using a duration model instead of attention



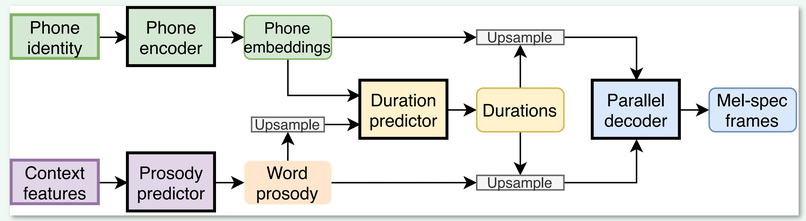
ORA

Autoencoder used to learn a disentangled word-level prosody representation



CAMP

TTS model using context to predict prosody. Best context features were fine-tuned BERT embeddings



MUSHRA test

- **DurIAN+** baseline
- Proposed **CAMP** w/BERT
- Top-line **ORA**cle prosody
- **NAT**ural speech (no vocoding)

CAMP$_{BERT}$ closed the gap between state-of-the-art and human speech by **26%**