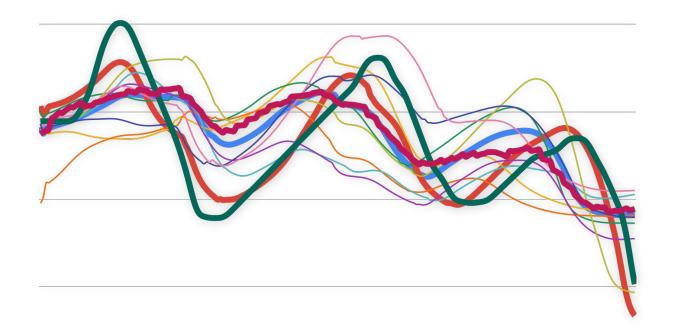


## CHIVE

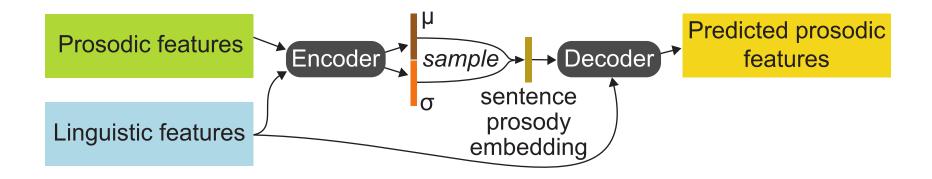
Varying Prosody in Speech Synthesis with a Linguistically Driven Dynamic Hierarchical Conditional Variational Network

Vincent Wan, Chun-an Chan, Tom Kenter, Jakub Vit, Rob Clark

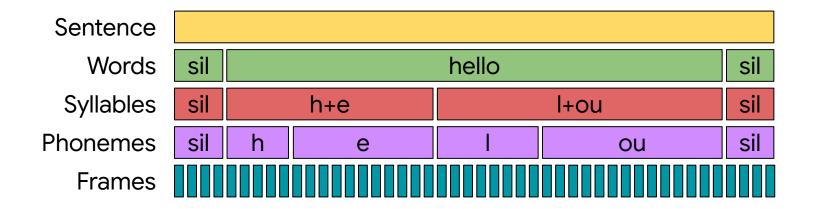
### Modelling intonation in prosody



# A conditional variational autoencoder captures the different intonations



#### Language has a hierarchical linguistic structure



#### Add linguistic knowledge to the network

predicted F frame values frame-rate F, RNN predicted co frame values frame-rate c RNN predicted phoneme durations duration prediction RNN phone-level RNN states phone-rate RNN phoneme-level features syllable-level features word-level features sentence-level features syllable-level RNN states syllable-rate RNN syllable-level features word-level features sentence-level features sentence prosody embedding

#### The structured model is better

