Context aware modelling of prosody (CAMP)

A two-stage approach to modelling prosody in context

Zack Hodari

ICASSP 2021

Motivation

– Speech has two channels of information: lexical and prosodic

– TTS models the lexical information well, but not the prosodic information

– Humans use **context** information to plan their prosody

To improve prosody we need more **context**

Capturing prosody

– Prosody has no orthography

– ***stage-1***: Learn a disentangled prosody representation

Synthesising prosody

– Prosody is determined by context

– ***stage-2***: Use additional context information for prosody prediction

CAMP
Context aware model of prosody

– ***stage-1***: Prosody representation learning

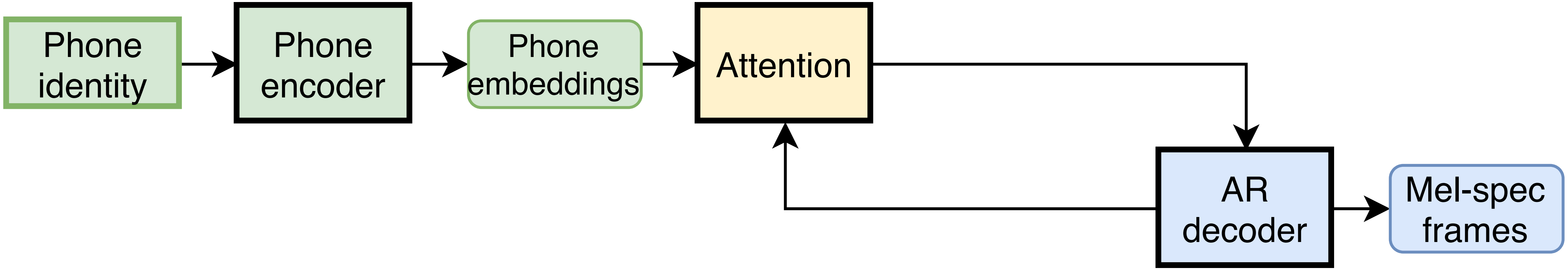
– ***stage-2***: Prosody prediction using context

Models and experiments

- DurIAN+ Tacotron-2 with an explicit duration model
- ORA Autoencoder using oracle prosody
- CAMP ORA using predicted prosody

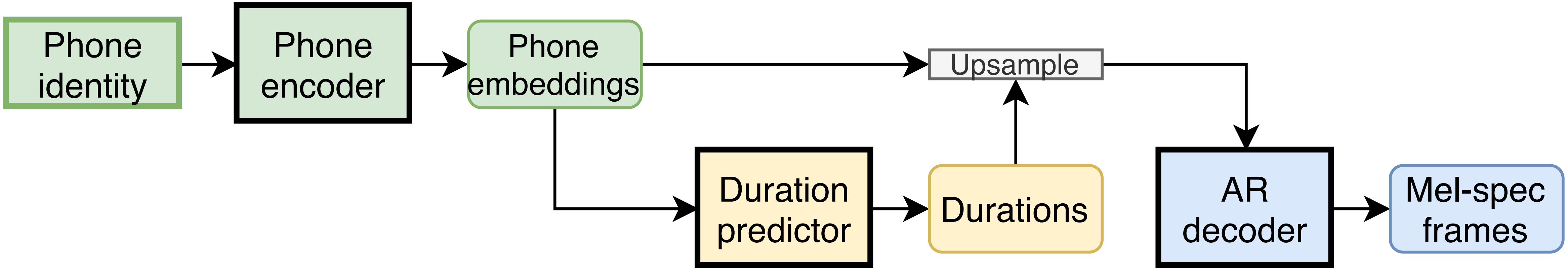
S2S

– Tacotron-2 like model



DurIAN+

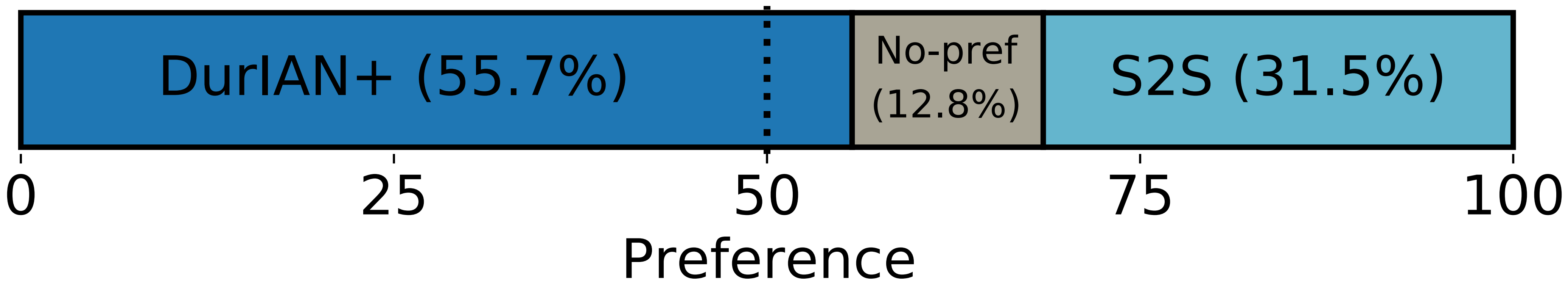
– Tacotron-2 with jointly-trained duration model



Baseline preference test

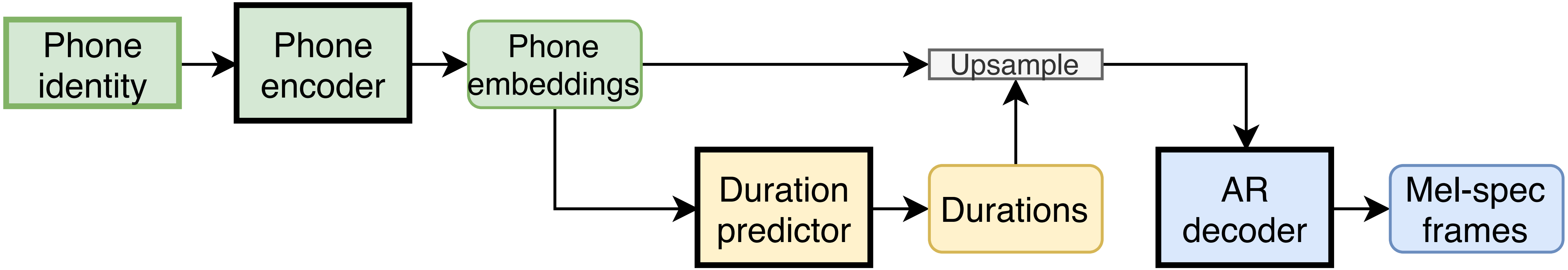
– S2S uses implicit duration modelling (i.e. attention)

– DurIAN+ uses explicit duration modelling (i.e. a duration model)



DurIAN+

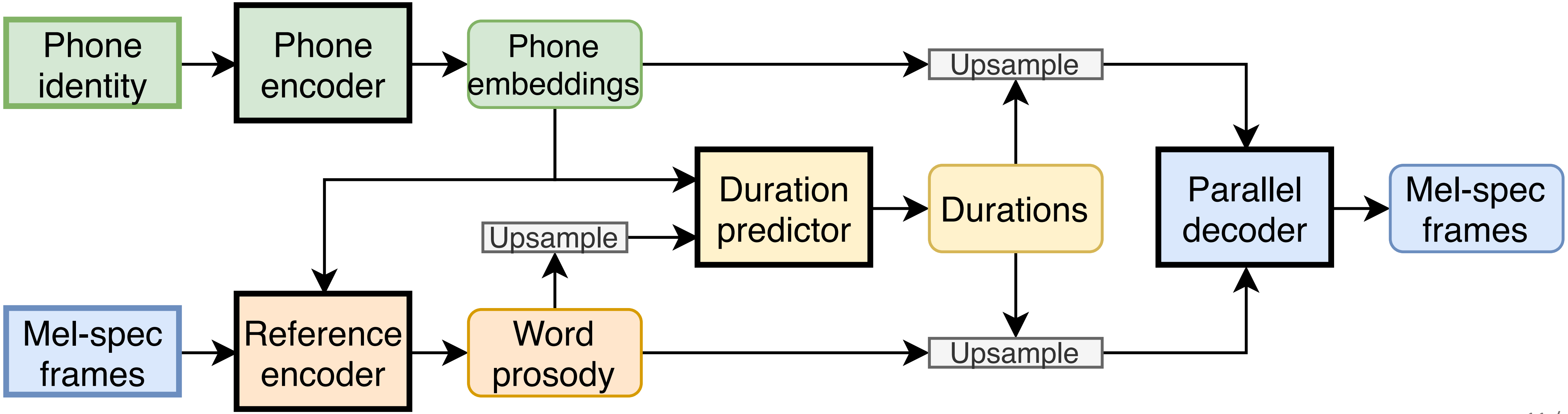
– Tacotron-2 with jointly-trained duration model



ORA – Oracle prosody

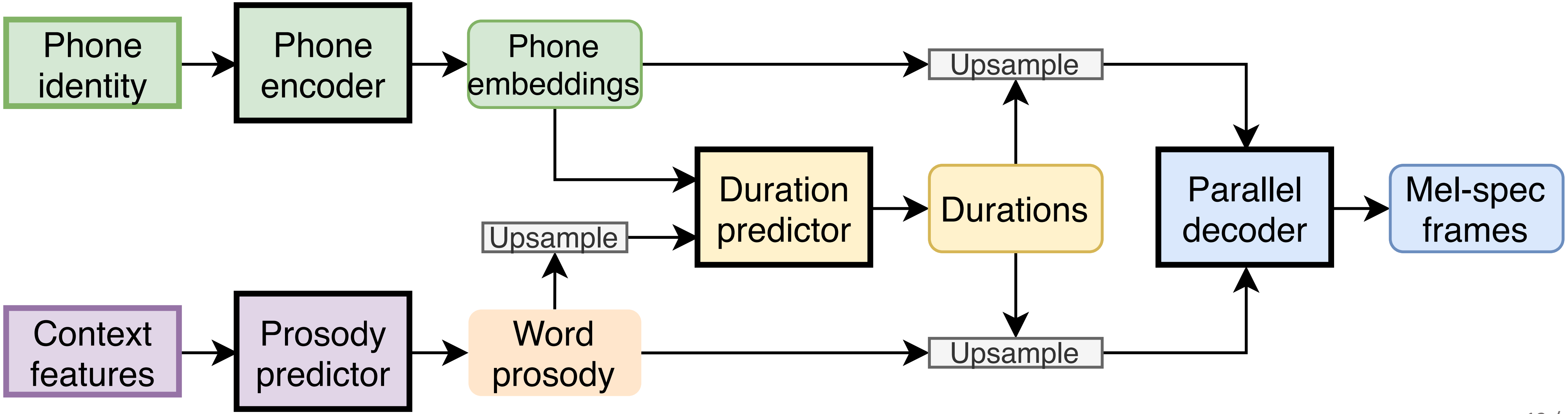
TRAINING STAGE 1

– Autoencoder model for representation learning



CAMP – Predicted prosody

- Context-based prediction of prosody
- **Proposed model using two-stage training**



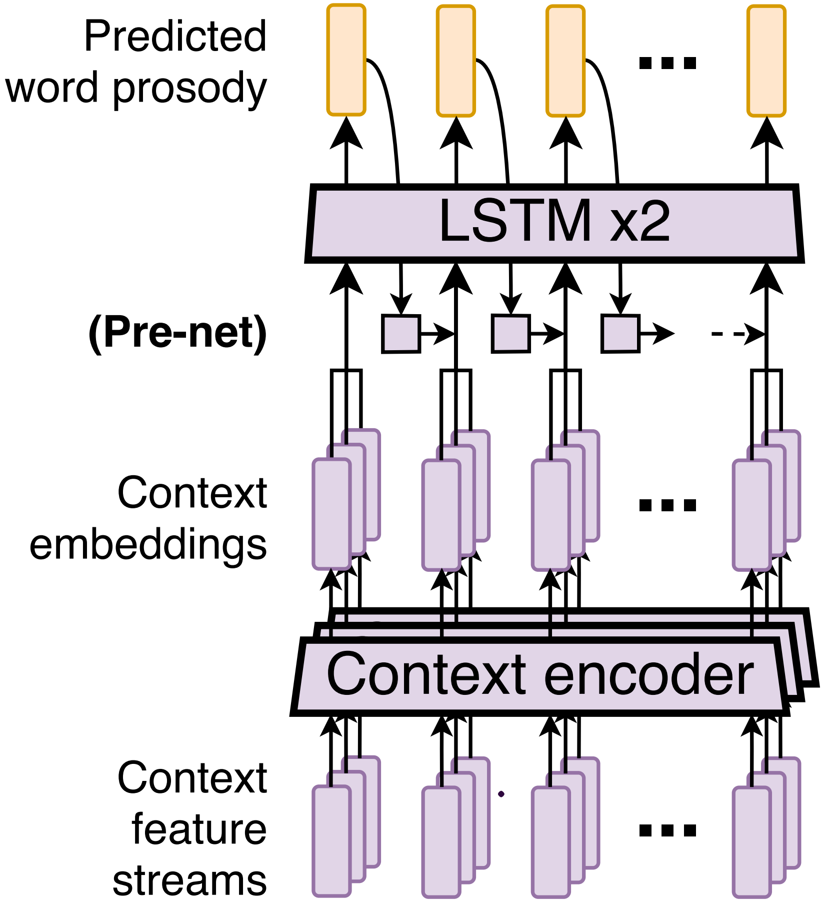
Prosody predictor

TRAINING STAGE 2

– Predicts prosody representation

– Replaces reference encoder

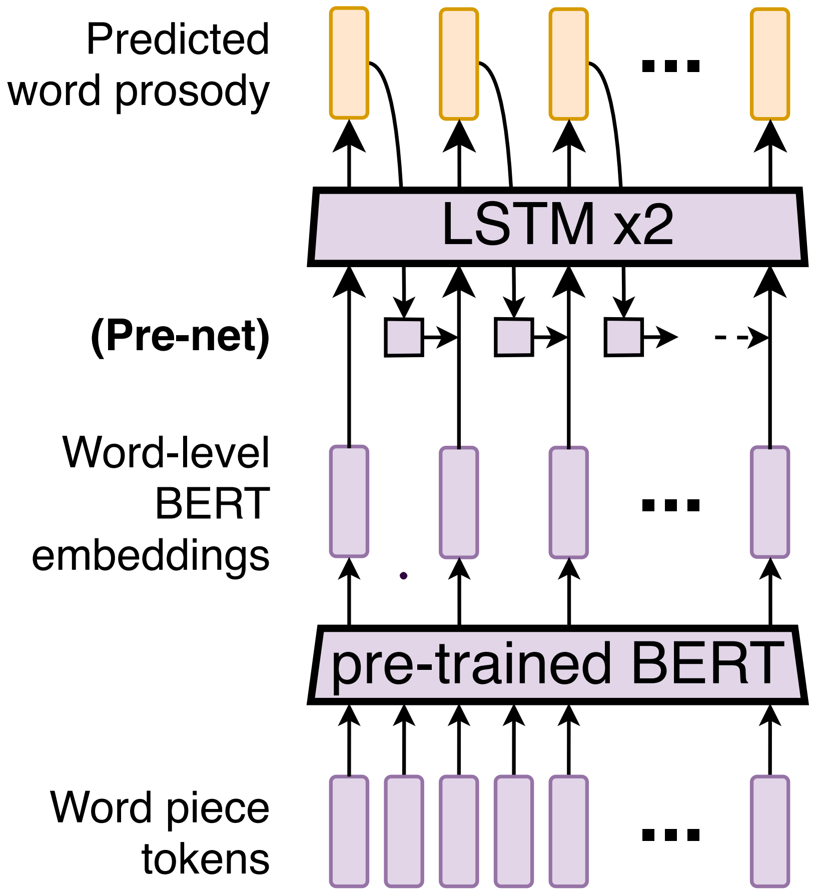
– Uses 1 or more context encoders



Prosody predictor

TRAINING STAGE 2

- – Predicts prosody representation
- – Replaces reference encoder
- – Uses fine-tuned BERT$_{\text{BASE}}$

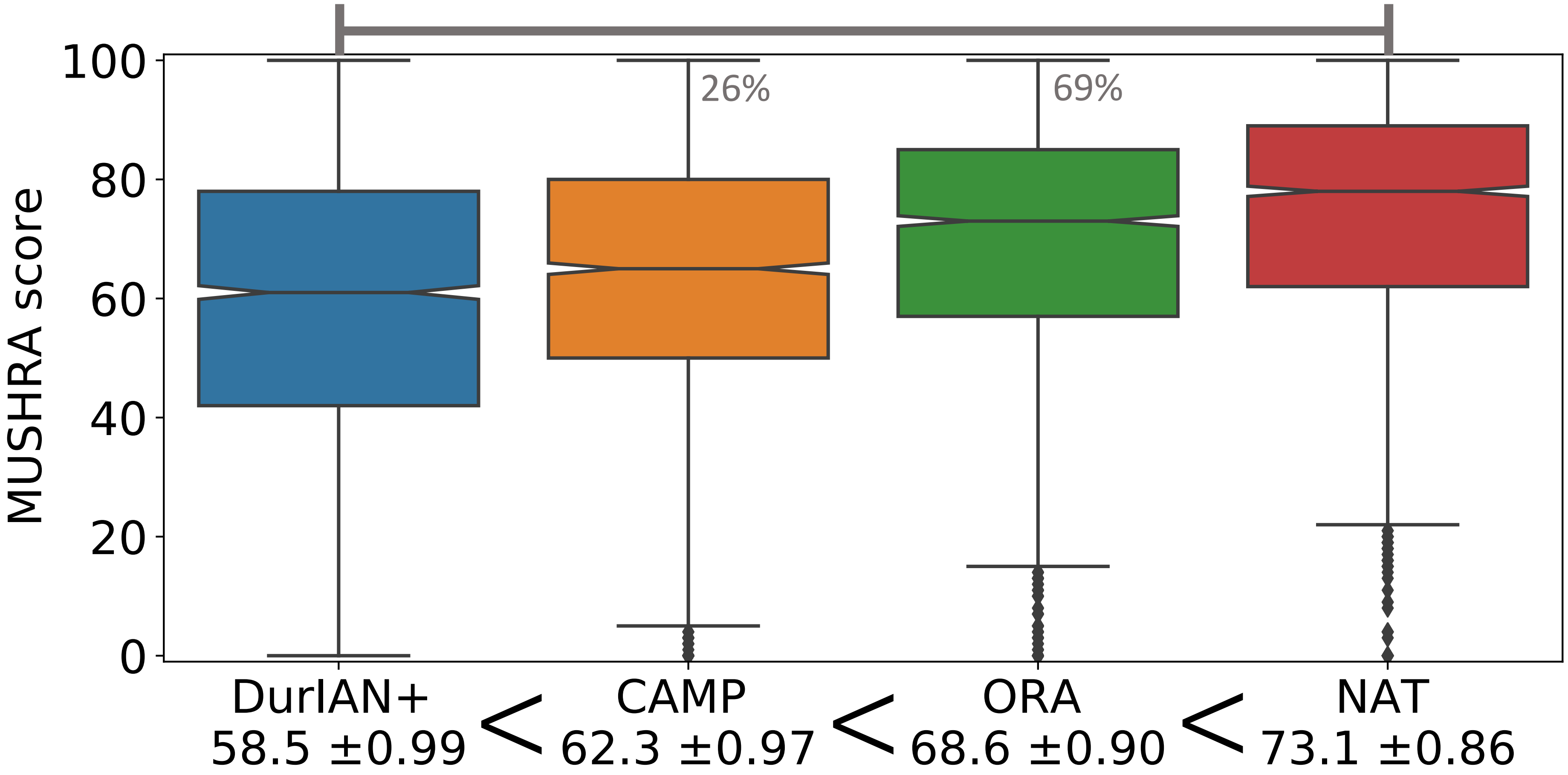


Evaluation of CAMP

– MUSHRA evaluation of our proposed model – CAMP

DurIAN+	Lower-bound	Duration-based Tacotron-2
CAMP	***Proposed***	Predicted prosody using BERT
ORA	Top-line	Oracle prosody
NAT	Upper-bound	Natural speech (no vocoding)

CAMP


MUSHRA score
100
80
60
40
20
0
26%
69%
DurIAN+
58.5 ±0.99
<
CAMP
62.3 ±0.97
<
ORA
68.6 ±0.90
<
NAT
73.1 ±0.86

Conclusion

– Train duration model jointly

– Use a prosodically-relevant loss

– Incorporate additional context

Thanks!