INTERNATIONAL AUDIO LABORATORIES ERLANGEN
A joint institution of Fraunhofer IIS and Universität Erlangen-Nürnberg

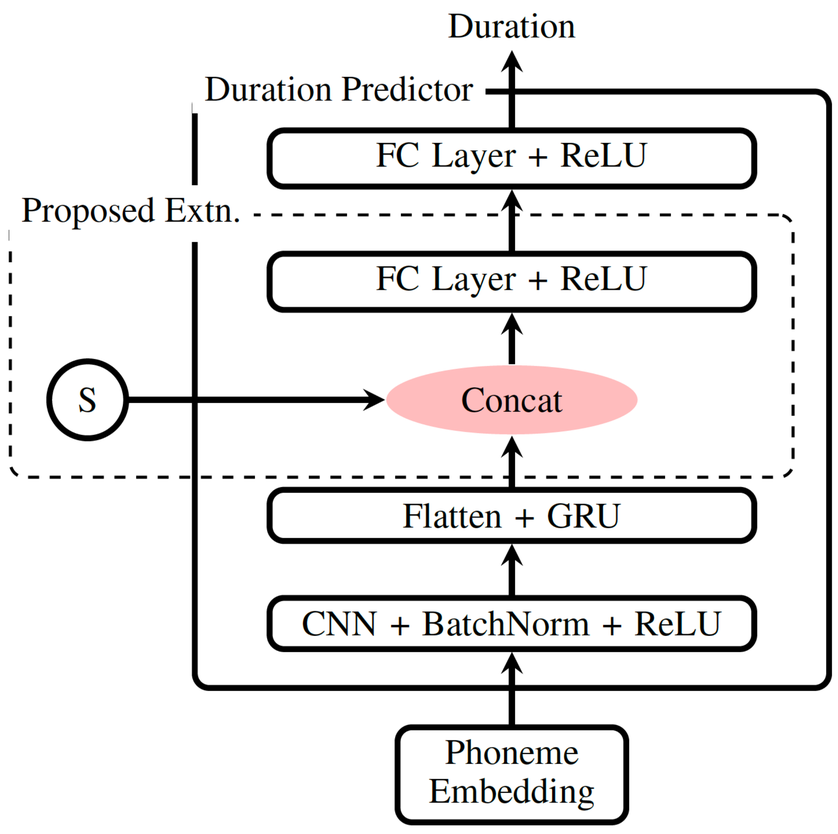
Multi-Speaker Text-to-Speech Using ForwardTacotron with Improved Duration Prediction

Kishor Kayyar Lakshminarayana, Christian Dittmar, Nicola Pia, Emanuël Habets

1. Introduction

- Many neural models such as Tacotron-1/2 can produce nearly natural speech from text.
- Non-autoregressive text-to-speech (TTS) models like FastSpeech and ForwardTacotron, which have low inference time, use a duration predictor to predict the phoneme durations.
- Typically, the duration predictor in such models is speaker independent.
- We propose to extend the duration-predictor for speaker-dependent durations with effective and efficient use of speaker embeddings.
- We show that the extended model is able to learn speaker specific rhythm.
- We also demonstrate that a large variation in duration does not impact naturalness ratings in subjective tests.

2. Proposed Method

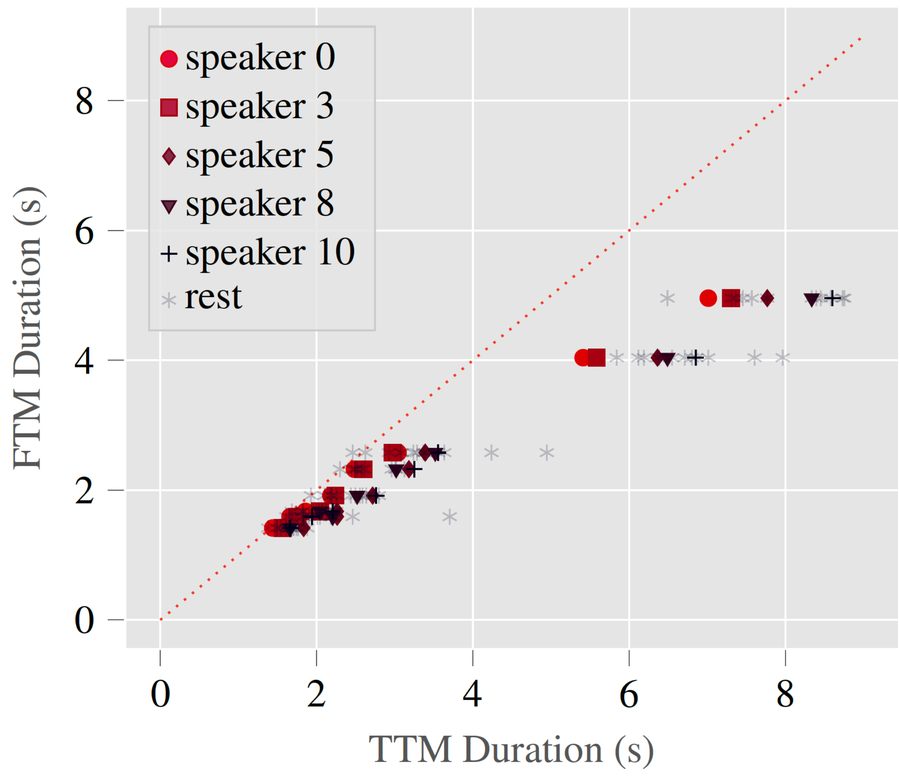


Block diagram of the duration predictor with the proposed extension, ‘S’ stands for speaker embedding.

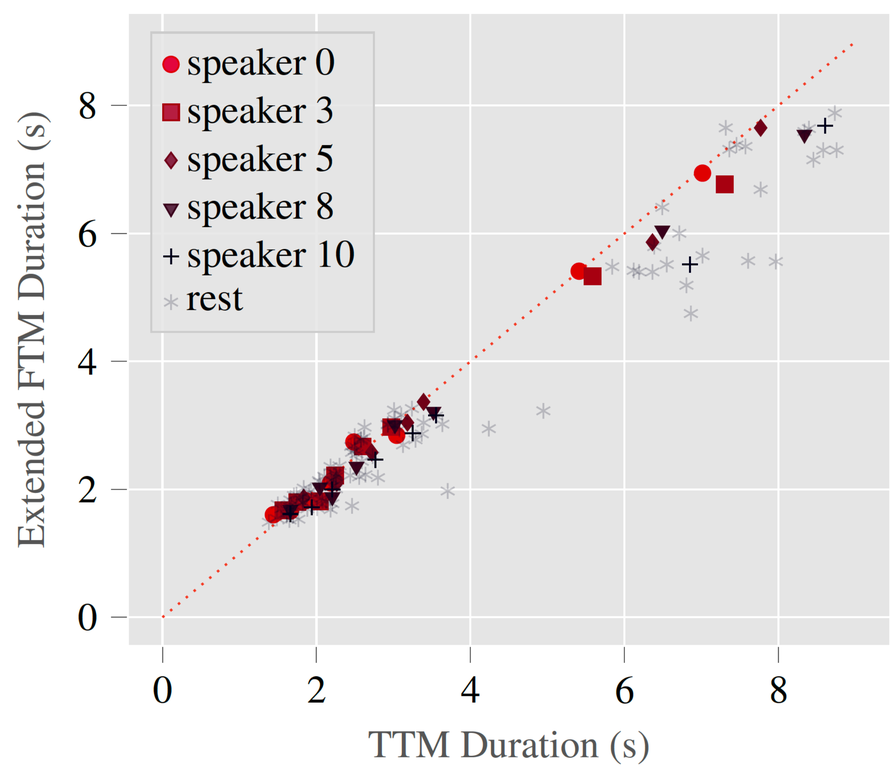
3. Performance Evaluation

- Automatic Speech Recognition done on 100 Semantically Unpredictable Sentences (SUS) showed a marginal improvement in the Word Error Rate from 29.6% to 28.9% with the proposed method.
- The paired comparison listening test of the baseline and proposed methods had no statistically significant differences in terms of overall quality.

- Scatter plot of the teacher Tacotron model (TTM) ground-truth durations against (a) the baseline ForwardTacotron model (FTM) and (b) proposed extended FTM durations for different speakers. A few speakers have been highlighted for better legibility.



(a)



(b)

4. Conclusions

- Made the duration prediction in a multi-speaker non-autoregressive TTS system speaker dependent with minimal computational overhead.
- The synthesized speech pace with the proposed method is shown to be speaker dependent and closely matched to the auto-regressive baseline.
- Subjective tests show that the perceived quality is not impacted by the presence or absence of a speaker’s characteristic synthesis pace.

Parts of this work have been supported by the SPEAKER project (FKZ 01MK20011A), funded by the German Federal Ministry for Economic Affairs and Climate Action. In addition, this work was supported by the Free State of Bavaria in the DSAI project.

Friedrich-Alexander-Universität Erlangen-Nürnberg
Fraunhofer IIS