Evaluating Speech–Phoneme Alignment and Its Impact on Neural Text-To-Speech Synthesis

Frank Zalkow*, Prachi Govalkar*, Meinard Müller*,†, Emanuël A. P. Habets*,†, Christian Dittmar*

* Fraunhofer IIS, Erlangen, Germany, † International Audio Laboratories Erlangen, Germany

ICASSP 2023
4 - 10 JUNE, RHODES ISLAND, GREECE

Background

- Parallel TTS architectures provide excellent synthesis quality at fast inference.
- Training requires speech recordings, corresponding phoneme-level transcripts, and the temporal alignment of each phoneme to the utterances.
- Alignments are usually estimated using automatic speech–phoneme alignment methods.

Problem

- In the literature, either the alignment methods' accuracy (objective evaluation) or their impact on the TTS system's synthesis quality is evaluated (subjective evaluation).
- The relationship between the objective and subjective evaluation is usually unclear.

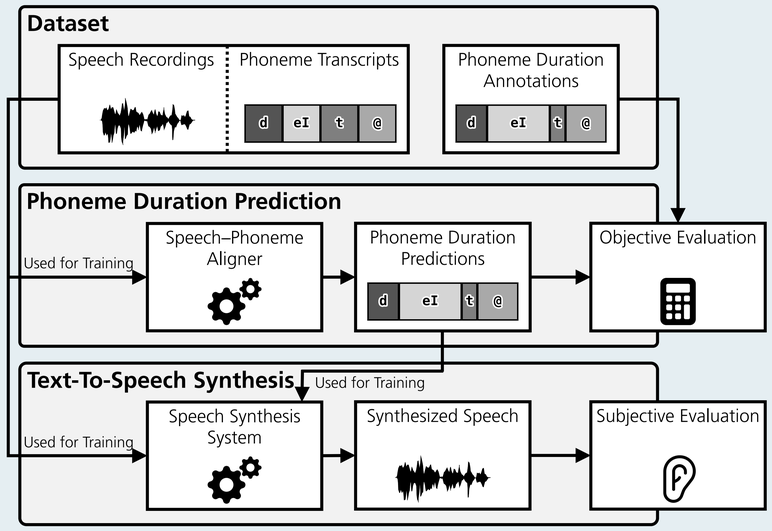
Contribution

- We performed experiments with five state-of-the-art speech–phoneme aligners and evaluated their benefit with objective and subjective measures.

Insight

- Small alignment errors do not decrease the synthesis quality, which implies that the alignment error may not be the crucial factor when choosing an aligner for training.

Overview



Methods

- Speech–phoneme alignment methods

	Description	Train Time
LOW	Assigning equal phoneme durations (lower anchor baseline)	0.0 h
MFA [3]	Montreal Forced Aligner, three-stage GMM-HMM model	0.4 h
TAC [4]	Tacotron 2, autoregressive TTS system	89.8 h
RAD [5,1]	Alignment framework using cost matrix	1.5 h
CTC [6]	CTC-based system with phonetic attention & spectral decoder	4.5 h
CTC∗	Custom simplification of previous system	1.8 h

- Datasets

	TIMIT	TC-STAR
Total duration	5.5 h	9.75 h
Avg. utterance	3.0 s	6.4 s
Speakers	630 (mixed gender)	1 (male)

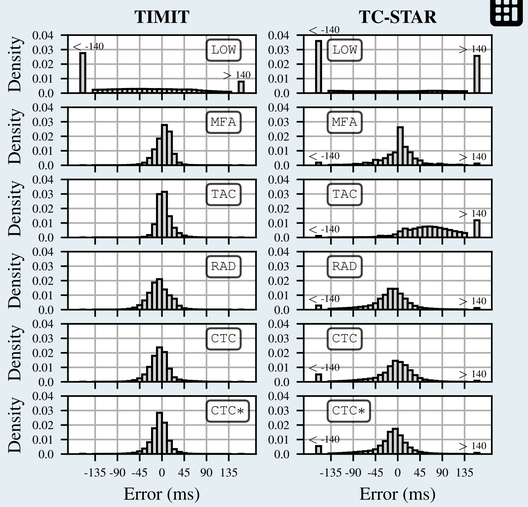
- Using alignment methods to estimate phoneme durations of datasets
- Objective evaluation by comparing estimates with ground truth
- Using estimates to train TTS system [2], which is then evaluated subjectively

Objective Evaluation

- Comparison of predicted phoneme start times with ground-truth from datasets
- Mean/median absolute difference in ms

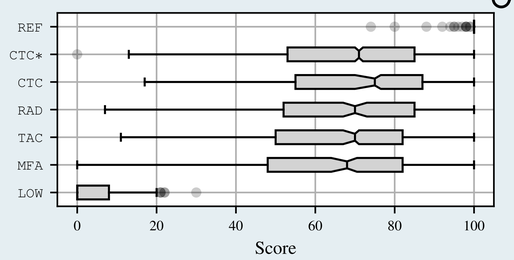
	TIMIT		TC-STAR	
	Mean	Median	Mean	Median
LOW	137.8	98.7	222.0	181.5
MFA	13.5	10.9	31.1	20.0
TAC	11.5	7.5	82.3	70.1
RAD	18.3	14.9	38.7	21.4
CTC	17.3	11.6	45.3	19.7
CTC∗	15.2	10.0	45.8	18.4

- Histograms of the signed differences in ms (on the right)



Subjective Evaluation

- Listening test according to MUSHRA-like methodology (reference: copy synthesis, 34 participants) using TC-STAR



- Pairwise comparisons using the Wilcoxon signed-rank test & Bonferroni correction
- Value below 0.05 means significantly different score distributions

	MFA	TAC	RAD	CTC	CTC∗
MFA	1.000	0.509	0.017	0.000	0.000
TAC		1.000	1.000	0.000	0.027
RAD			1.000	0.302	1.000
CTC				1.000	1.000
CTC∗					1.000

Demo

https://www.audiolabs-erlangen.de/resources/NLUI/2023-ICASSP-eval-alignment-tts

Acknowledgements

We thank all participants of our listening test. Furthermore, we thank Alexander Adami for fruitful discussions on the listening test design and its evaluation. 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. The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits IIS. The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the FAU.

References

[1] Badlani, Łancucki, Shih, Valle, Ping, and Catanzaro, *One TTS alignment to rule them all*. ICASSP, 2022.

[2] Govalkar, Mustafa, Pia, Bauer, Yurt, Özer, and Dittmar, *A lightweight neural TTS system for high-quality German speech synthesis*. ITG Conf. on Speech Communication, 2021.

[3] McAuliffe, Socolof, Mihuc, Wagner, and Sonderegger, *Montreal forced aligner: Trainable text-speech alignment using Kaldi*. Interspeech, 2017.

[4] Shen, Pang, Weiss, Schuster, Jaitly, Yang, Chen, Zhang, Wang, Ryan, Saurous, Agiomyrgiannakis, and Wu, *Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions*. ICASSP, 2018.

[5] Shih, Valle, Badlani, Łancucki, Ping, and Catanzaro, RAD-TTS: *Parallel flow-based TTS with robust alignment learning and diverse synthesis*. ICML Workshop, 2021.

[6] Teytaut and Roebel, *Phoneme-to-audio alignment with recurrent neural networks for speaking and singing voice*, Interspeech, 2021.