



HAL
open science

A Closer Look at Latent Representations of End-to-end TTS Models

Martin Lenglet, Olivier Perrotin, Gérard Bailly

► **To cite this version:**

Martin Lenglet, Olivier Perrotin, Gérard Bailly. A Closer Look at Latent Representations of End-to-end TTS Models. Journée commune AFIA-TLH / AFCP – “Extraction de connaissances interprétables pour l’étude de la communication parlée”, Dec 2023, Avignon, France. . hal-04269953

HAL Id: hal-04269953

<https://hal.science/hal-04269953>

Submitted on 23 Jan 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

How are acoustic parameters encoded in a model?

Context

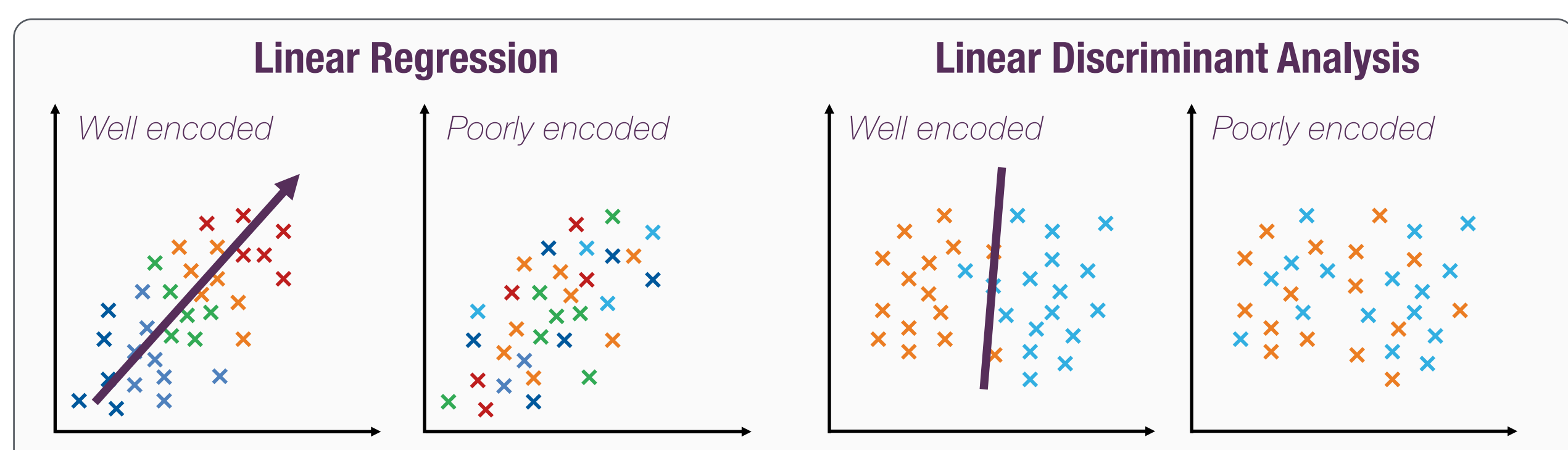
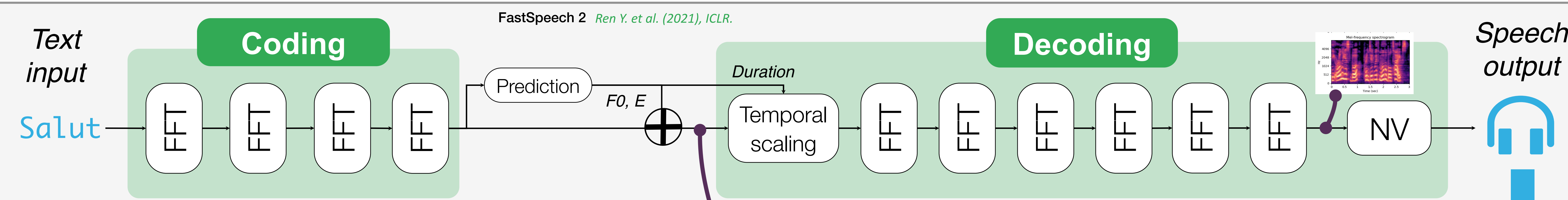
Hypothesis on neural audio modelling:

- Given the high-quality output of audio modelling in general, most speech information should be encoded in the model.

Previous studies on speech model probing:

- Phonetic on TTS: *Perquin et al. (2020), arXiv*
- Acoustic on TTS: *Tits et al. (2021), Informatics 8(4)*
- Language in SSL: *Vaidya et al. (2022), ICML*
- Articulatory in SSL: *Cho et al. (2023), ICASSP*

Method



- Training material: 33h audiobook – Single French speaker
- Probing material: 2000 utterances from training set
- Test material: 2000 utterances NOT from training set

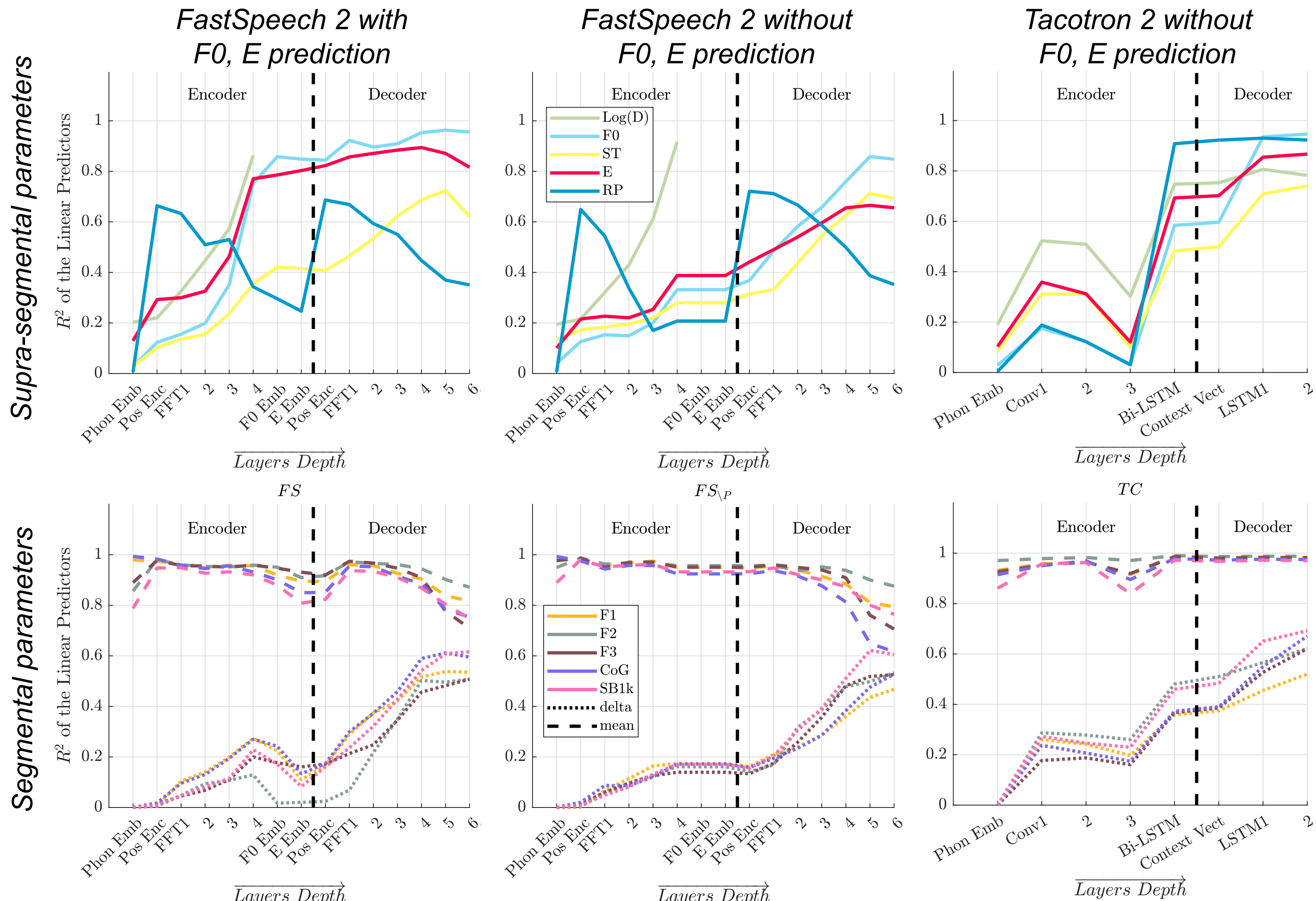
-0.11	-0.98	0.41	0.48	0.09
0.13	0.33	0.76	-0.05	-0.67
⋮	⋮	⋮	⋮	⋮
-0.35	-0.62	-0.82	-0.14	0.17
0.63	-0.25	0.93	-0.16	-0.62

Linear Prediction

Features / character

- Formants
- F0, Energy, Duration
- Phonetic class
- Pauses
- Etc.

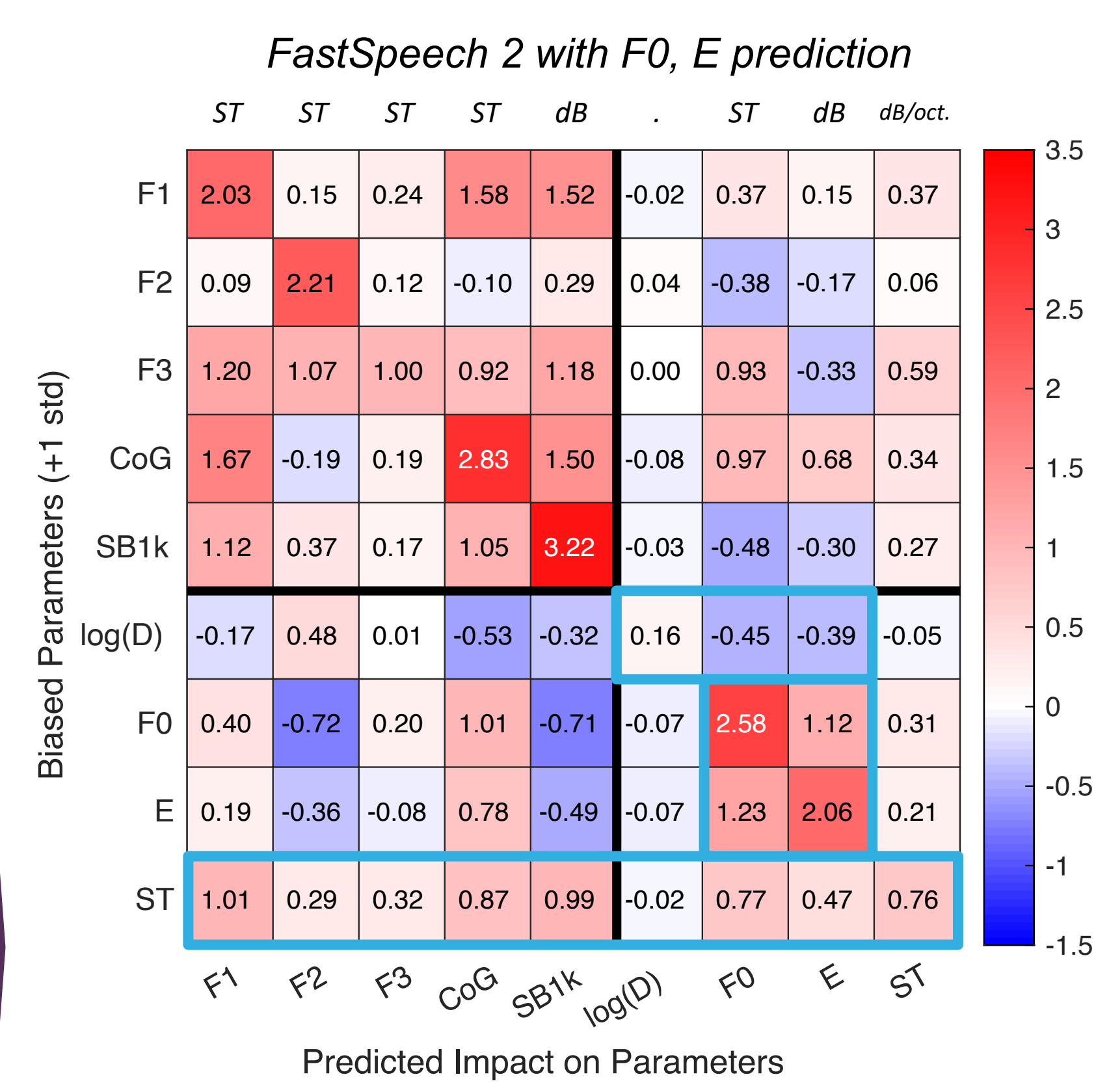
Results



Goodness of fit expressed as the R^2 of the multi-linear regression.

- Progressive linear coding of acoustic parameters
- Phonetic first, then prosodic
- Helped with forced predictions

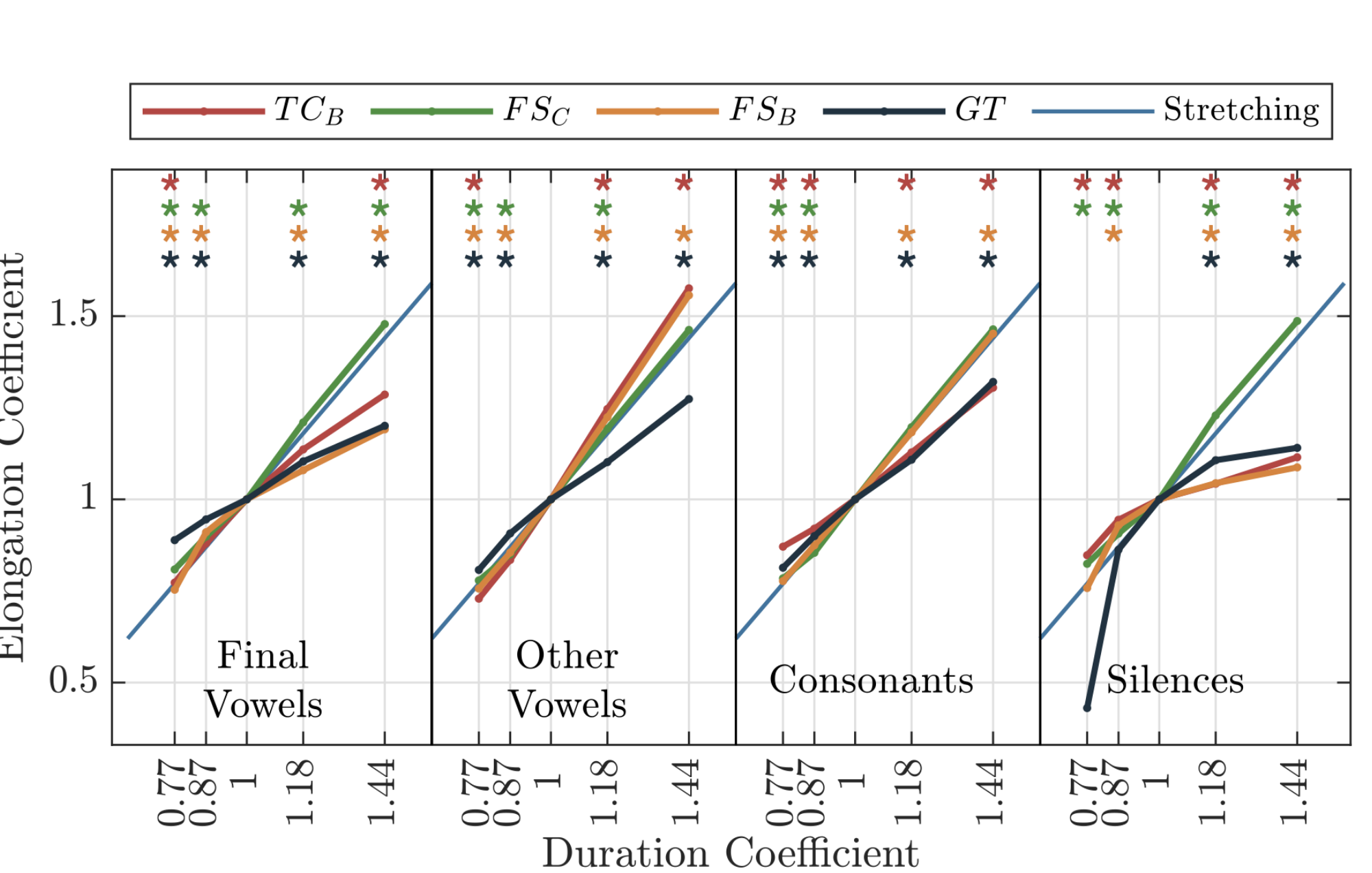
- Observation of covariations
- Consistent with literature
 - F0 and Energy
 - Duration, F0 and Energy
 - Vocal effort correlates (F0, Energy, F1, Spectral tilt)
- New ones to find?



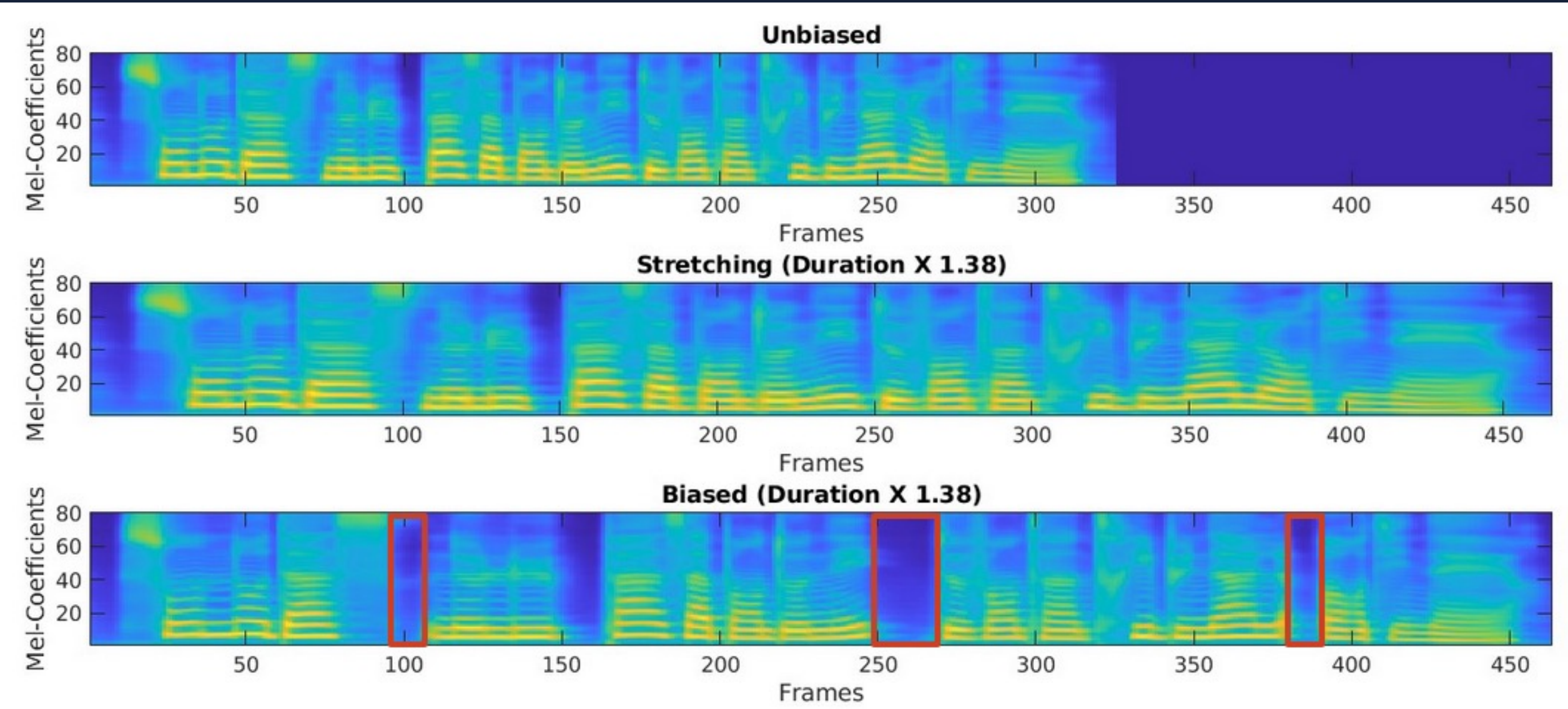
➤ Powerful speech analysis tool learnt on massive data

Can we control them, and how does the system behave?

Results



- Control of duration
 - Single modification for all phones
 - Different behaviour depending on segment
- Control of percentage of pauses
 - Position of pauses left to the model
- Evaluation of both
 - Significant preference (resp. worst score) when pauses are well (resp. wrongly) placed



➤ Single linear control, multiple non-linear effects