A closer look at latent representations of end-to-end TTS models

Martin Lenglet, Olivier Perrotin, Gérard Bailly, *Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab*

How are acoustic parameters encoded in a model?

Context

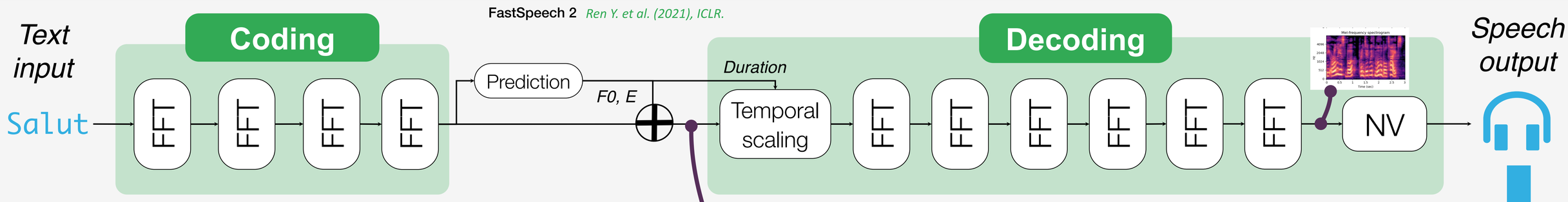
Hypothesis on neural audio modelling:

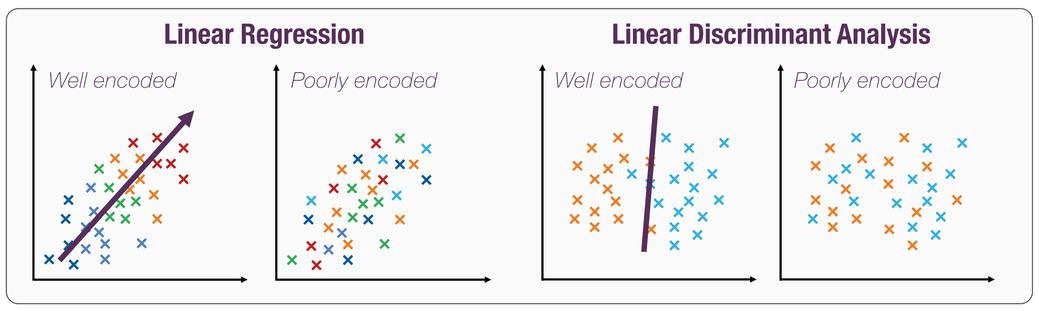
- Given the high-quality output of audio modelling in general, most speech information should be encoded in the model.

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

$$\begin{bmatrix} -0.11 & -0.98 & 0.41 & 0.48 & 0.09 \\ 0.13 & 0.33 & 0.76 & -0.05 & -0.67 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -0.35 & -0.62 & -0.82 & -0.14 & 0.17 \\ 0.63 & -0.25 & 0.93 & -0.16 & -0.62 \end{bmatrix}$$

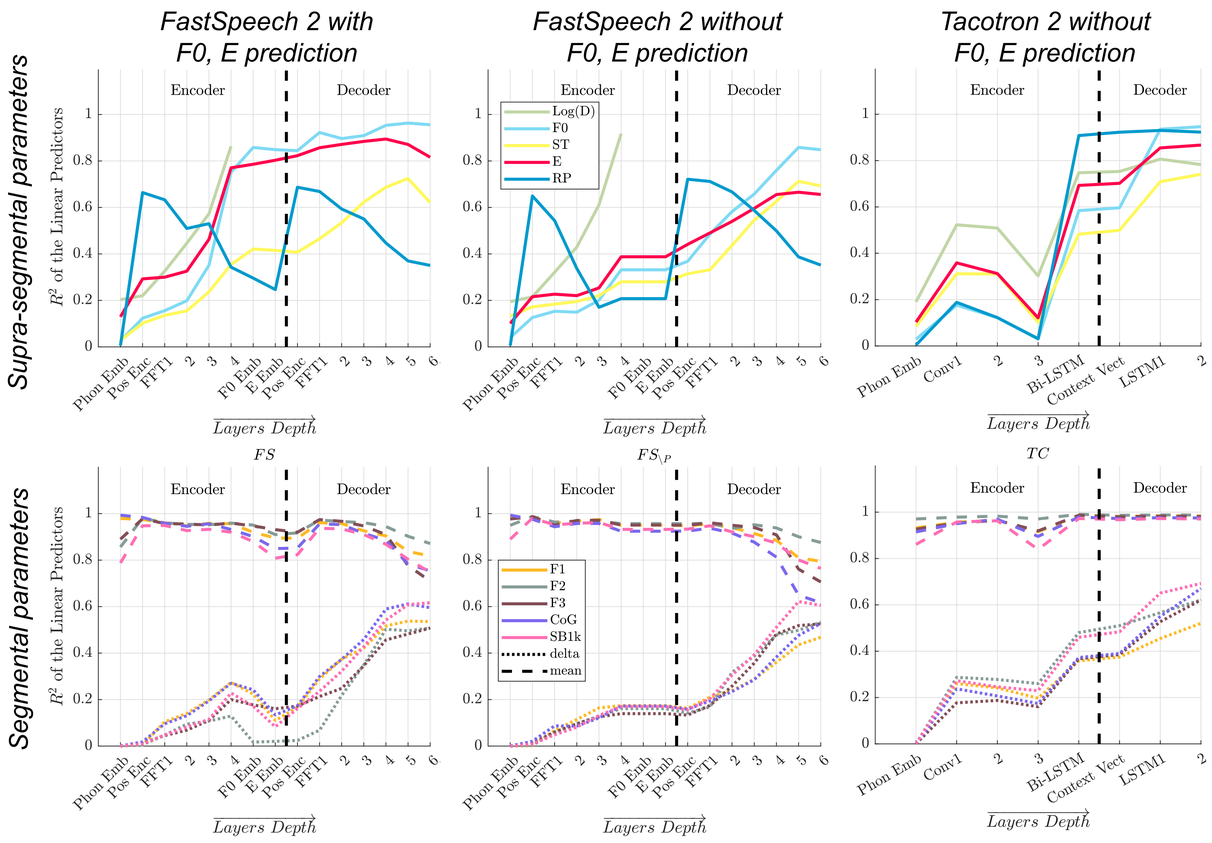
Linear Prediction

$$\begin{bmatrix} / & 150 & 145 & 140 & 120 \end{bmatrix}$$

Features / character

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

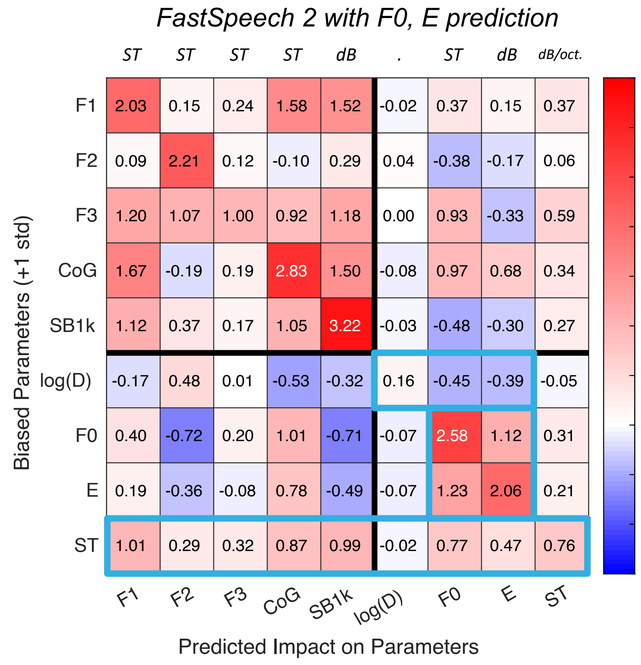
Results



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

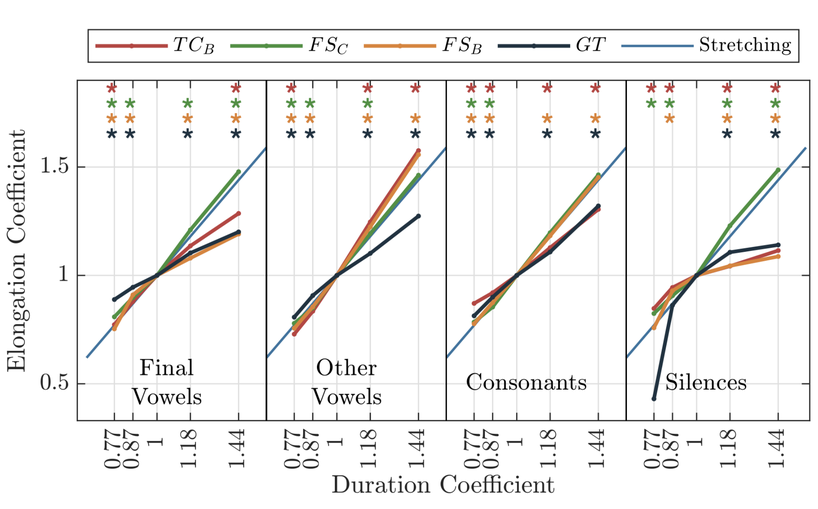
- Progressive **linear coding of acoustic parameters**
- Phonetic first, then prosodic
- Helped with forced predictions

- 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?**

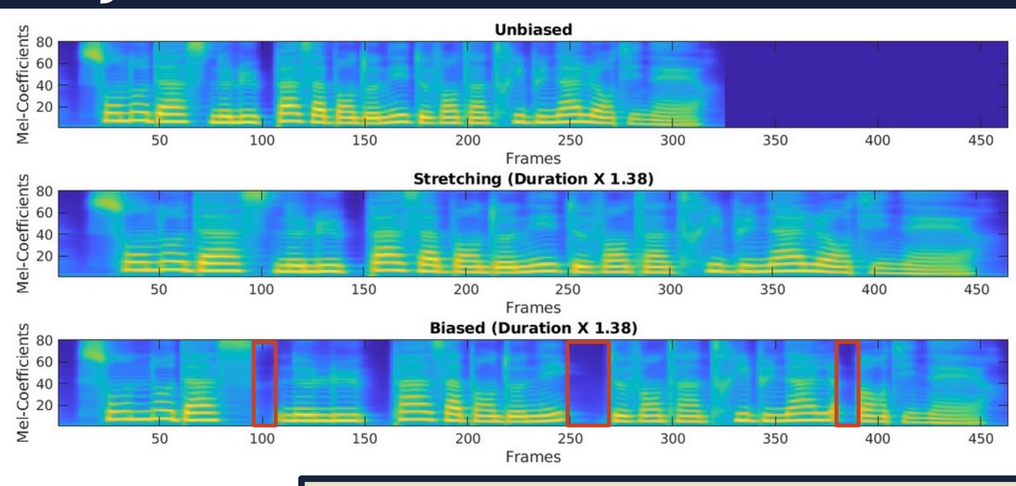


➤ Powerful speech analysis tool learnt on massive data

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



- 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

Grenoble Image Parole Signal Automatique
UMR CNRS 5216 - Grenoble Campus
38400 Saint Martin d'Hères - FRANCE