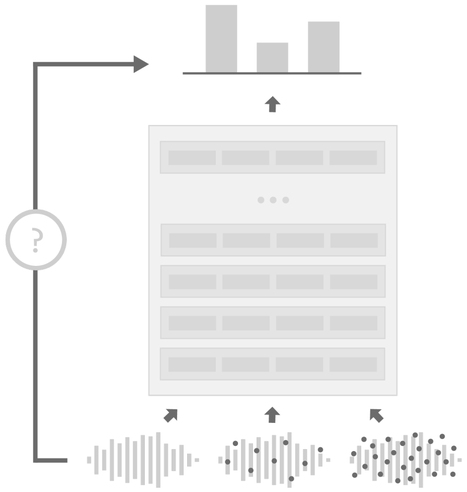
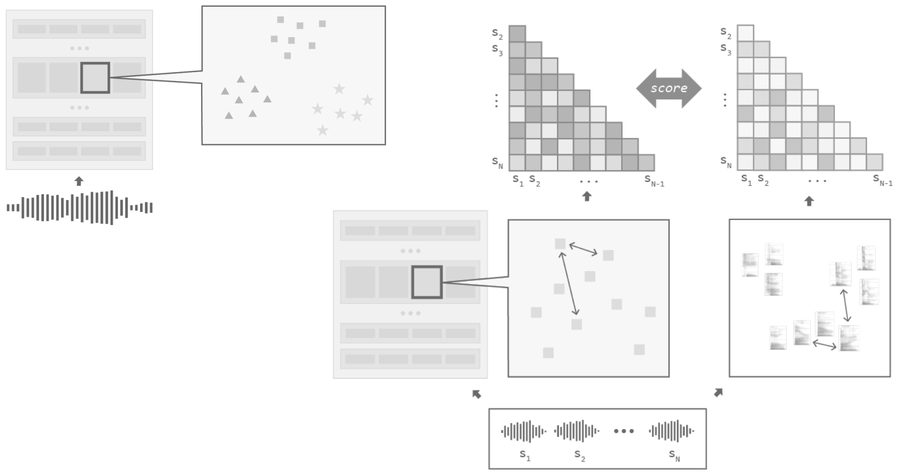
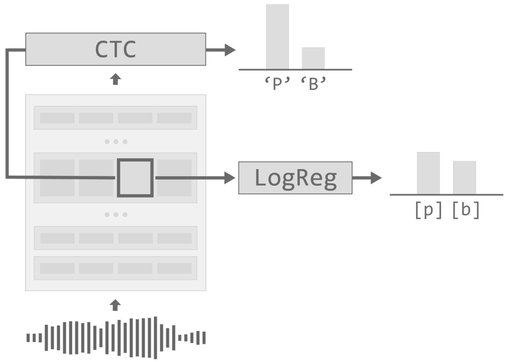
Interpretability Techniques for Speech Models

Conclusions & Outlook

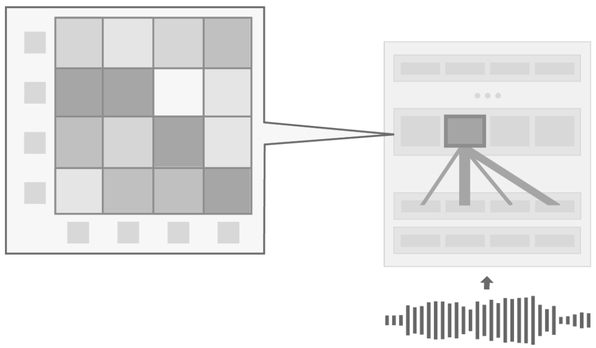
Marianne de Heer Kloots, 17-08-2025

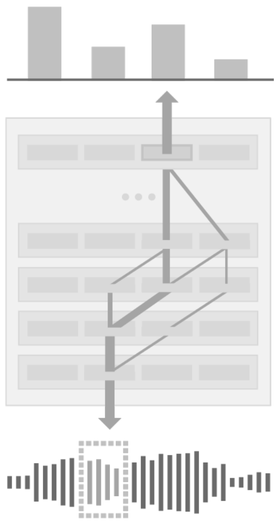


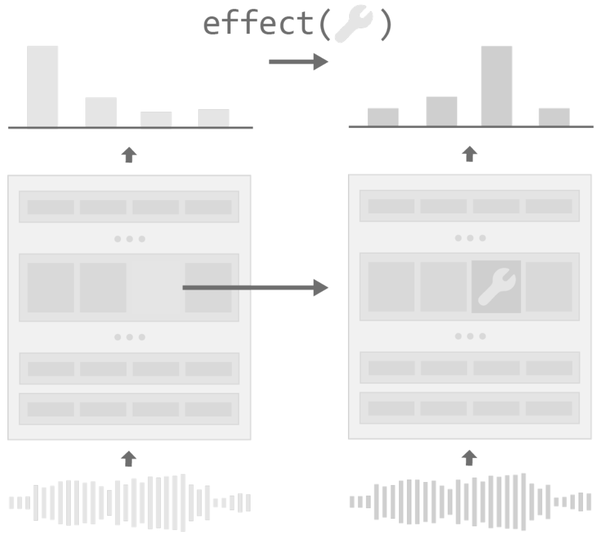


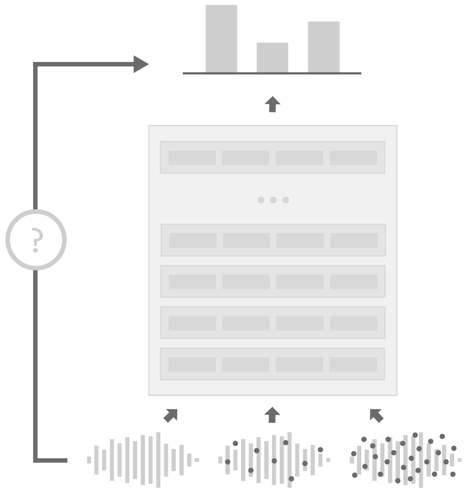


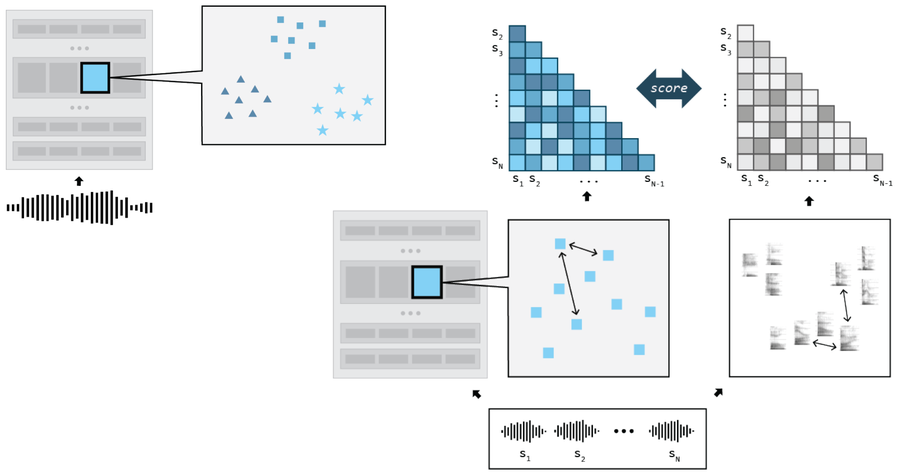
Interpretability Techniques for Speech Models

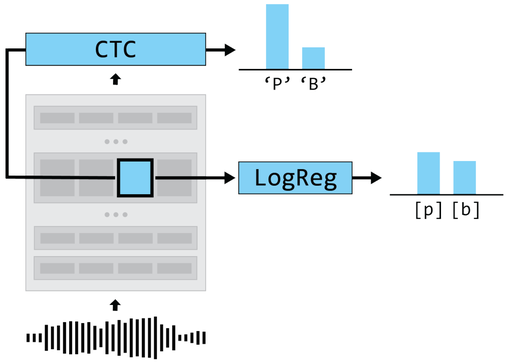




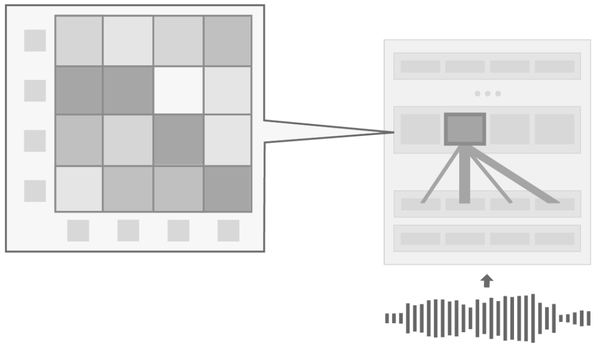


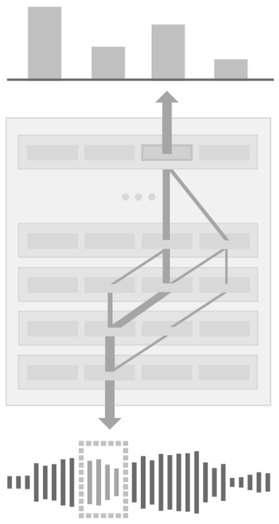


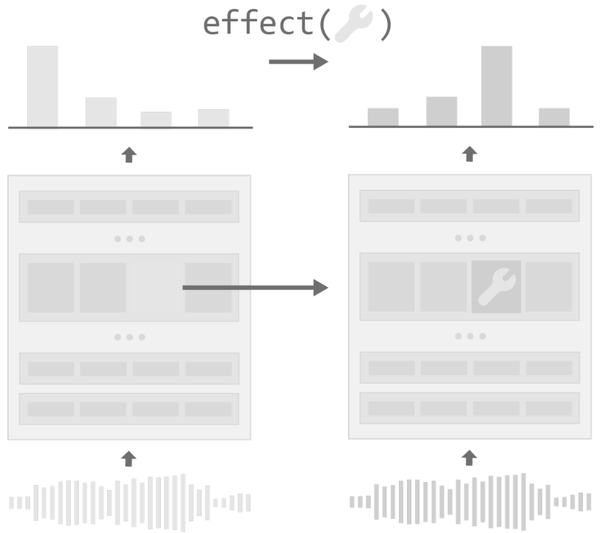


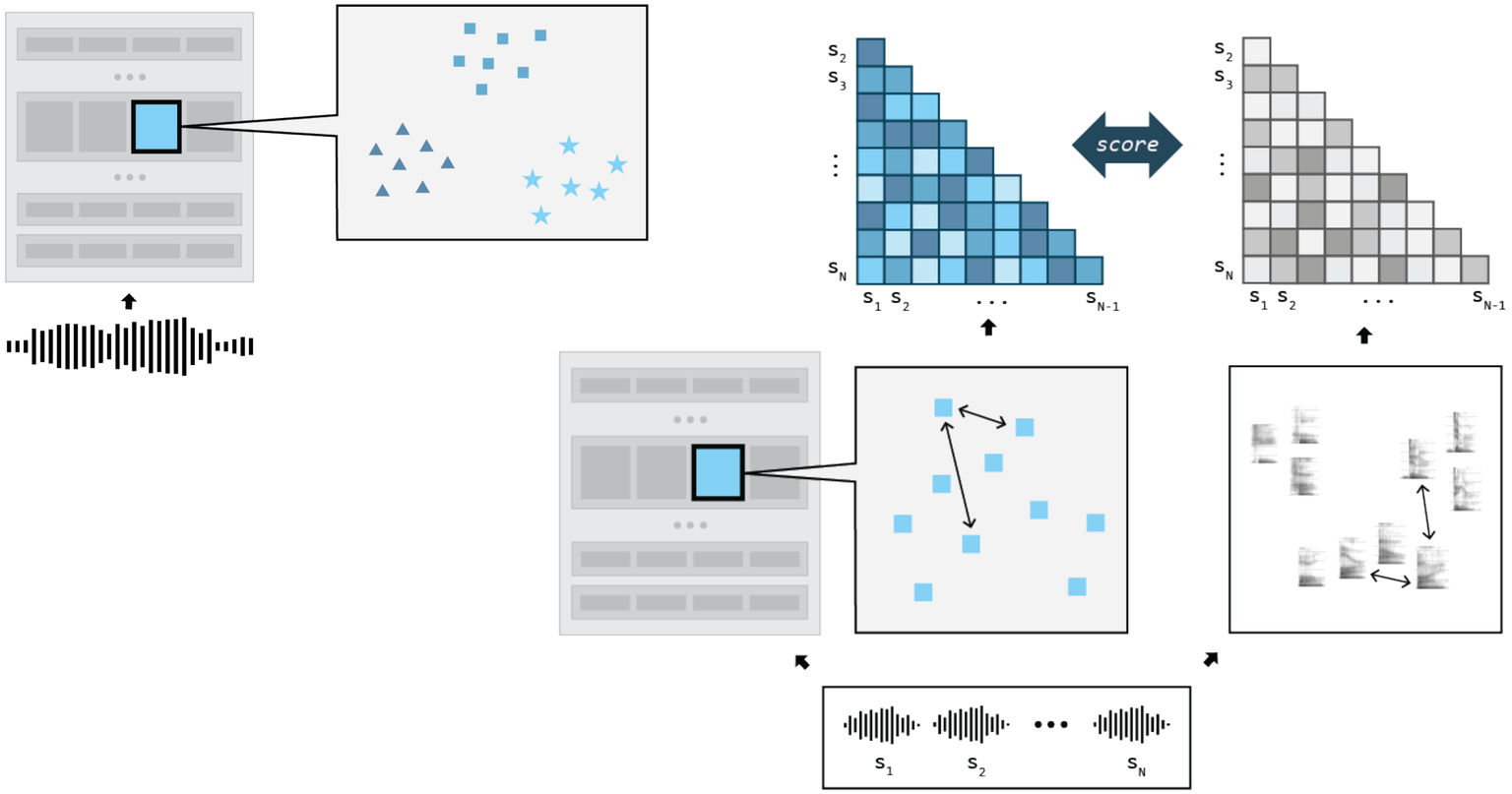


Interpretability Techniques for Speech Models



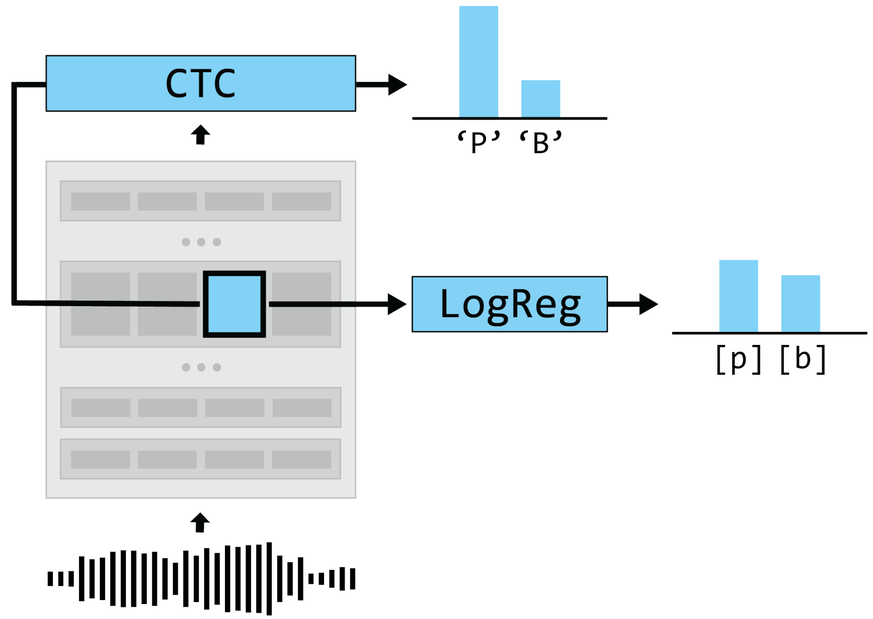




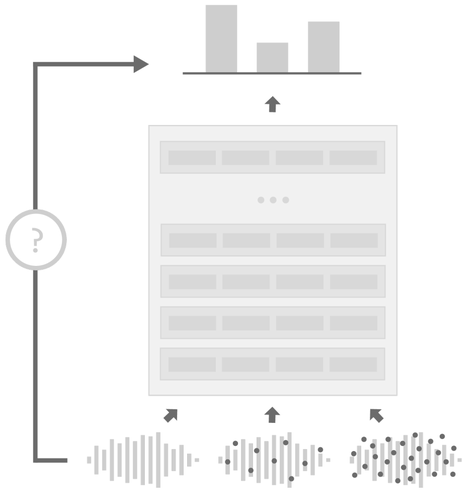


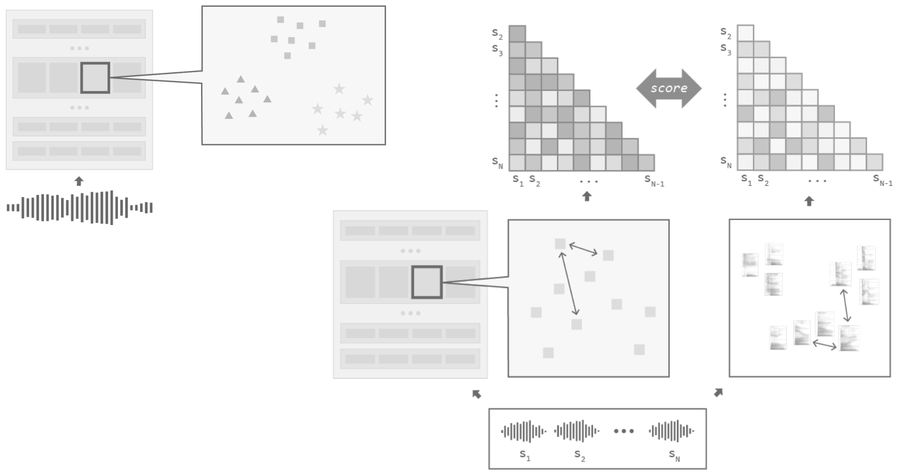
Representational analyses

- Dimensionality reduction
- Probing classifiers
- Representation space comparisons

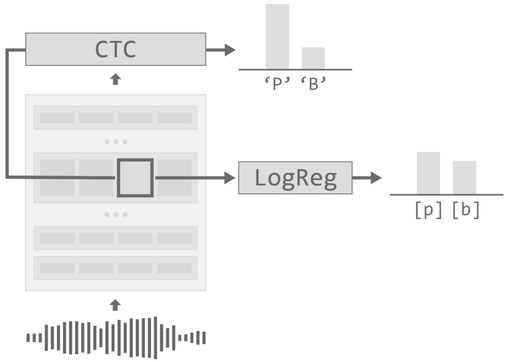


- Baselines and controls
- Zero-shot metrics vs. optimized diagnostic tools

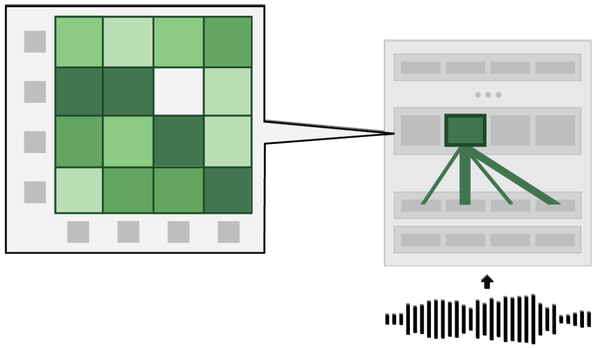


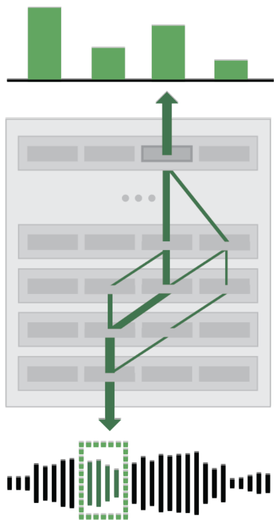


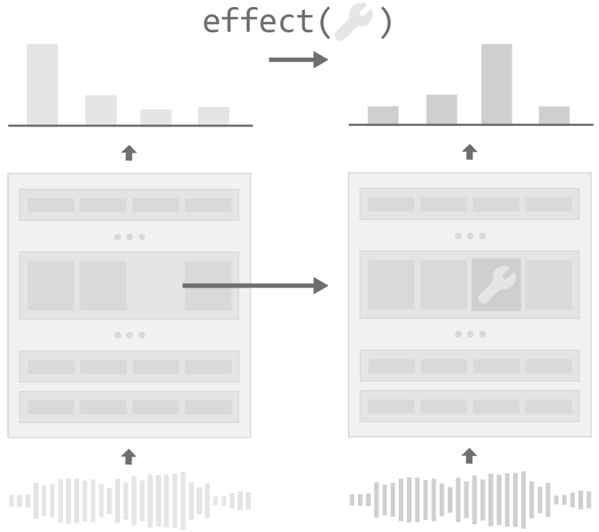
Representational analyses

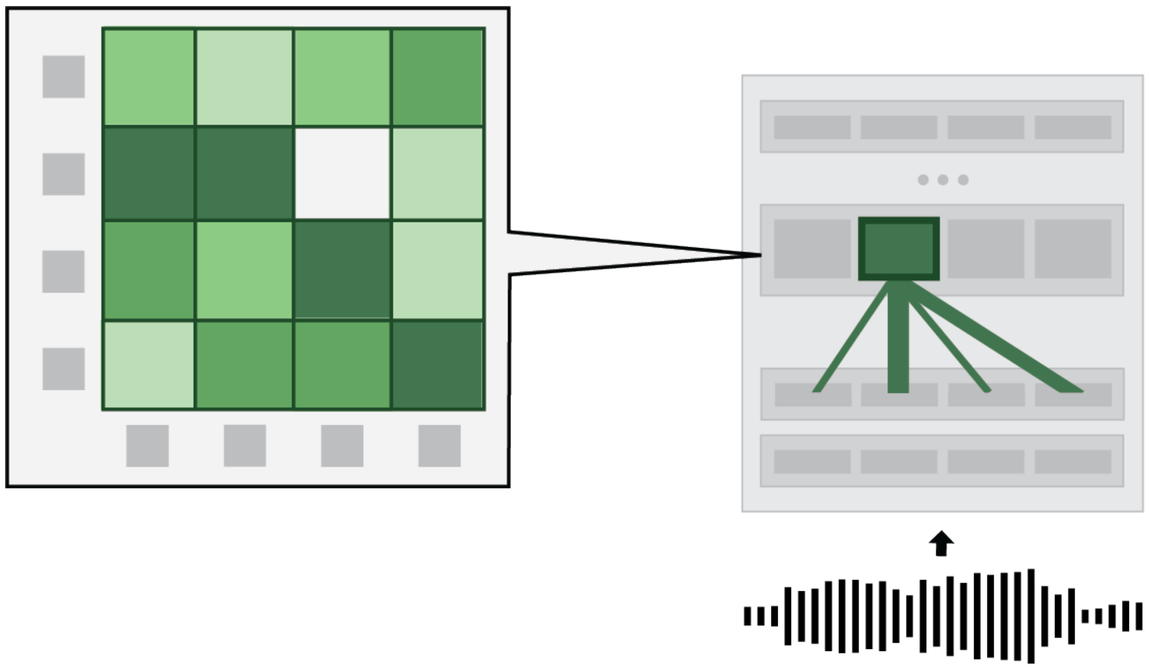


Interpretability Techniques for Speech Models

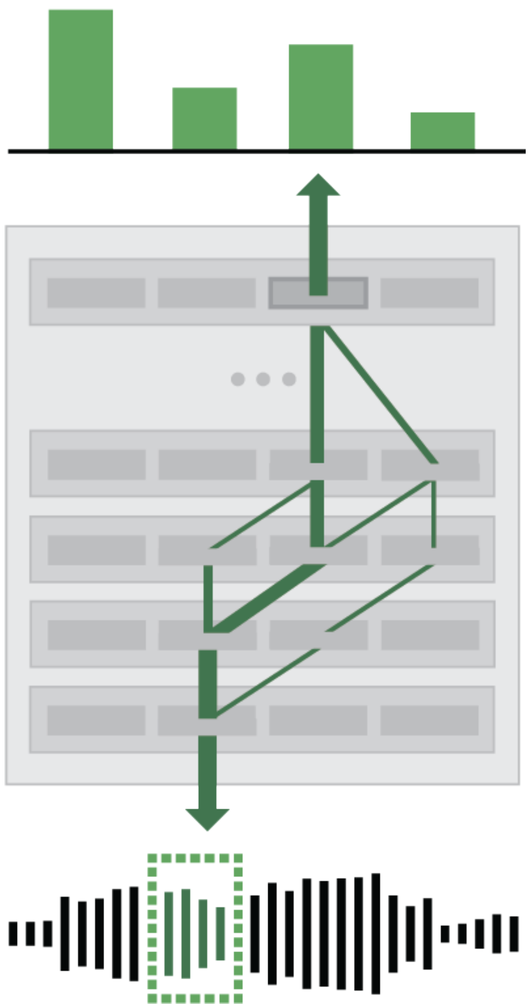






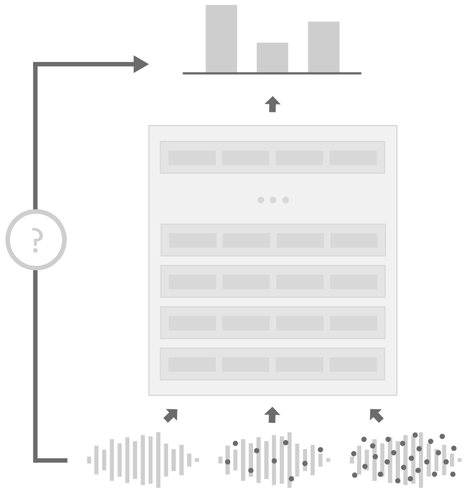


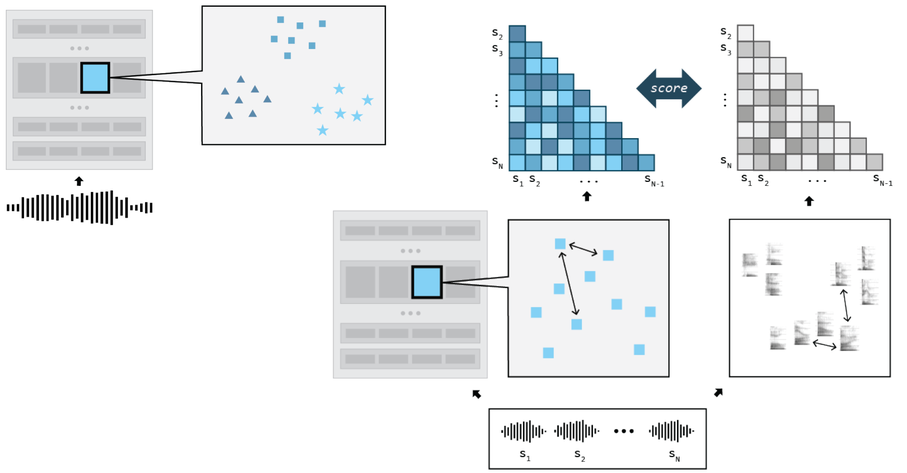
- Beyond attention weights
- Attribution reliability



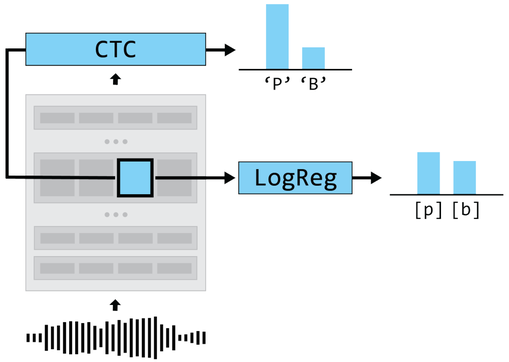
Feature Importance Scoring

- Context Mixing
- Feature Attribution

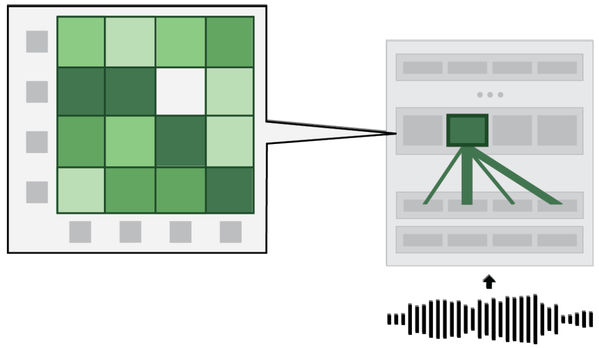




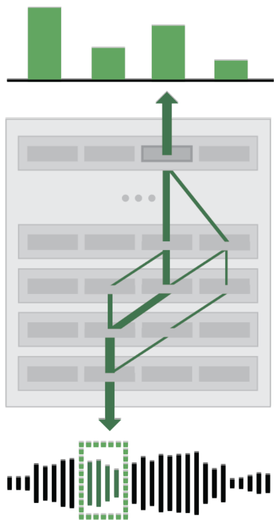
Representational analyses

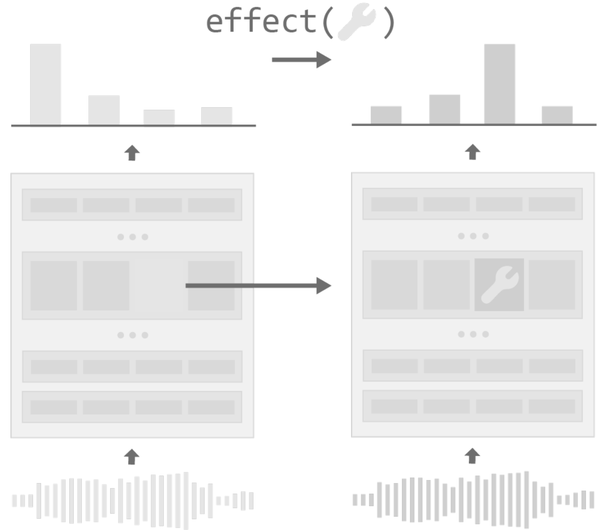


Interpretability Techniques for Speech Models



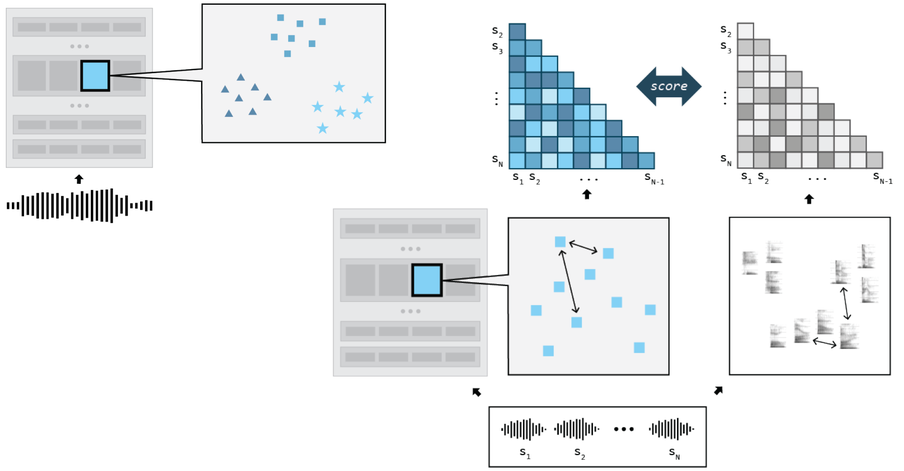
Feature Importance Scoring



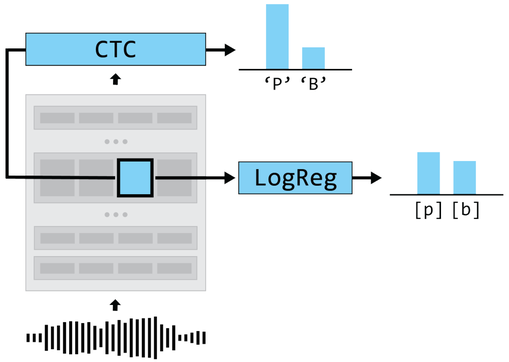


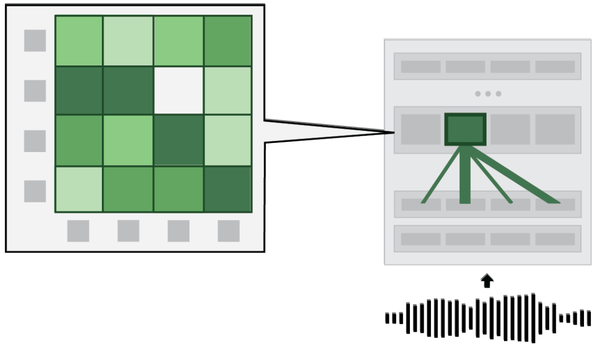
Interpretability Techniques for Speech Models

Behavioural analyses

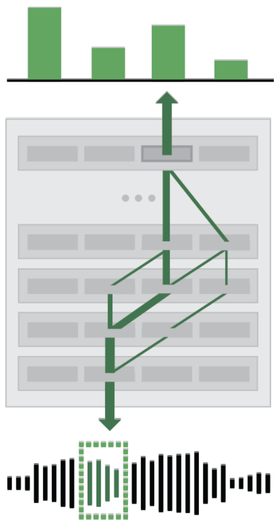


Representational analyses

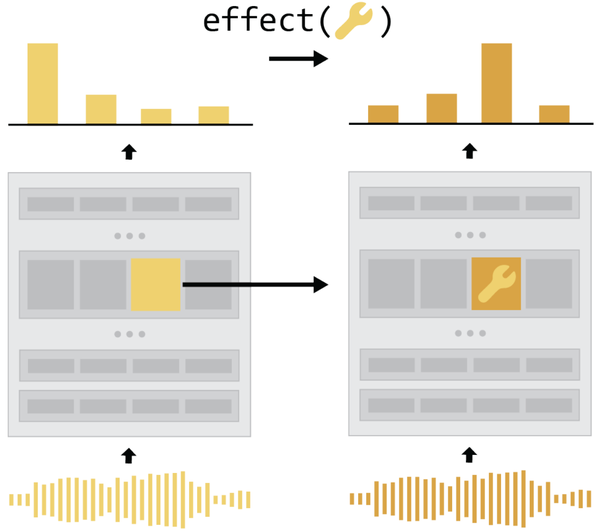




Feature Importance Scoring



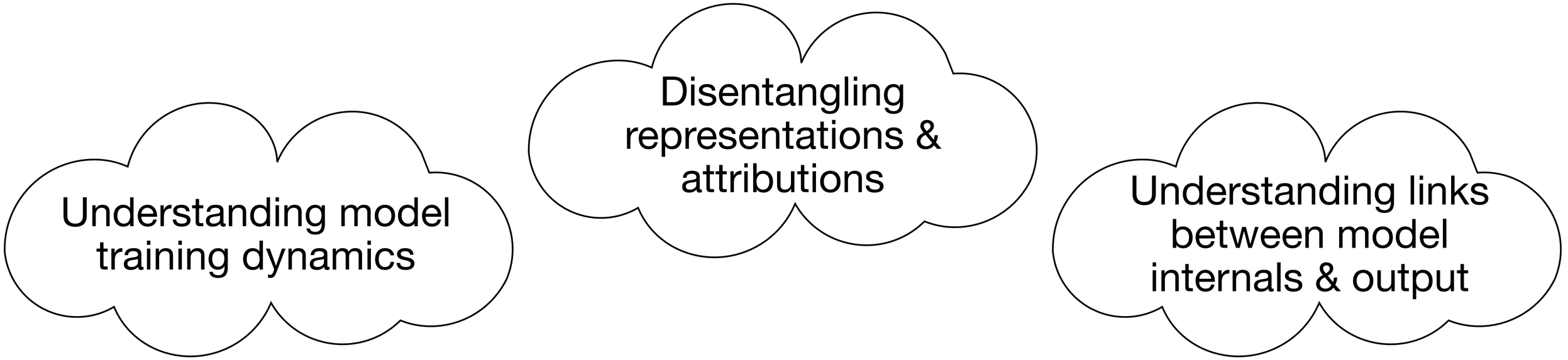
Survey bibliography



Causal analyses

Outlook:

Can interpretability be useful?



Outlook:

Can interpretability be useful?

Understanding model training dynamics

Disentangling representations & attributions

Understanding links between model internals & output

Understanding model training

Linguistic structure in speech models

Shen et al. (2024)

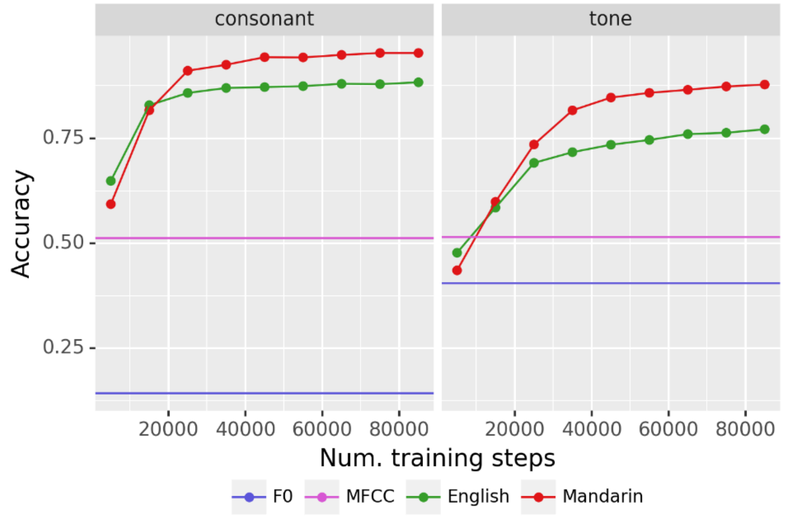


Figure 6: Classification accuracy of Mandarin lexical tones versus consonants for models pre-trained on English and Mandarin.

de Heer Kloots et al. (in prep)

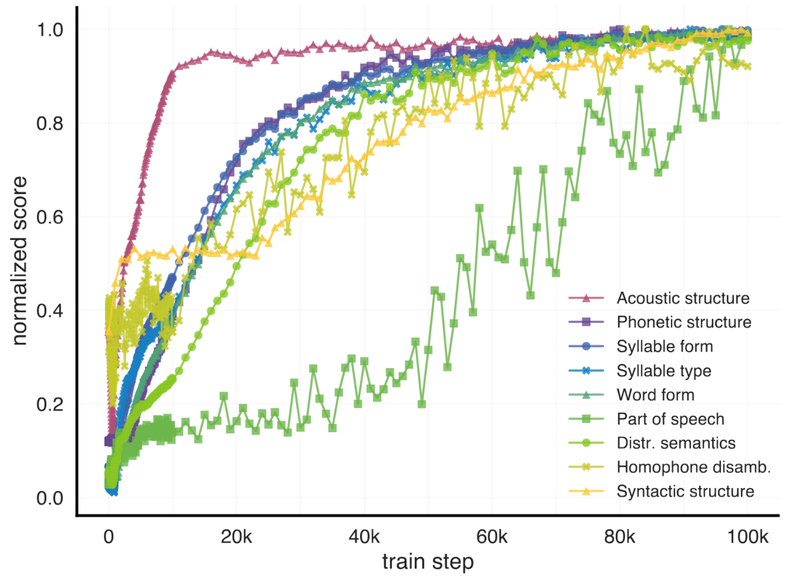


Figure 3: Development of linguistic encoding capacities across checkpoints in model training.

Shen, G., Watkins, M., Alishahi, A., Bisazza, A., & Chrupała, G. Encoding of lexical tone in self-supervised models of spoken language. (NAACL 2024)

de Heer Kloots, M., Mohebbi, H., Pouw, C., Shen, G., Bentum, M., Zuidema, W. Investigating language learning trajectories in self-supervised speech models. (in prep)

Understanding model training

Syntax acquisition in text models (Chen et al., 2025)

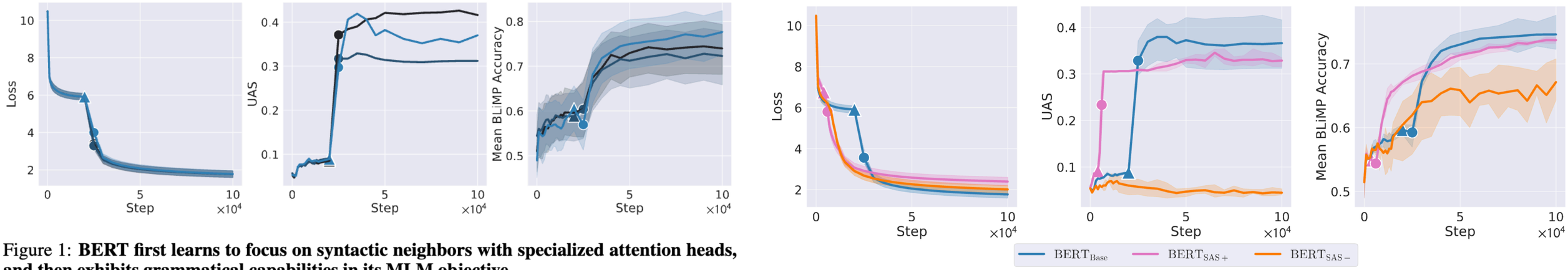


Figure 1: **BERT first learns to focus on syntactic neighbors with specialized attention heads, and then exhibits grammatical capabilities in its MLM objective.**

Chen, A., Shwartz-Ziv, R., Cho, K., Leavitt, M. L., & Saphra, N. Sudden Drops in the Loss: Syntax Acquisition, Phase Transitions, and Simplicity Bias in MLMs. (ICLR 2025)

Outlook:

Can interpretability be useful?

Understanding model training dynamics

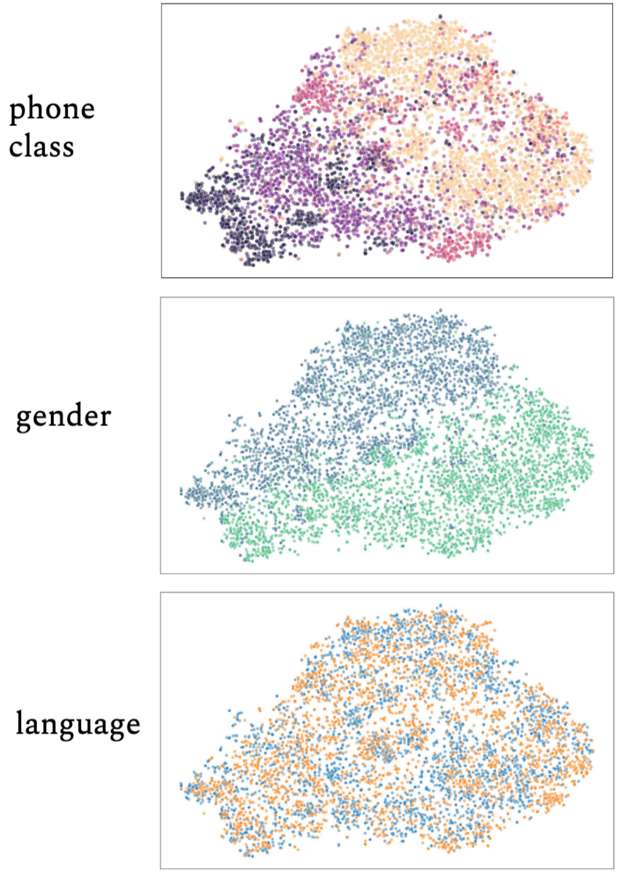
Disentangling representations & attributions

Understanding links between model internals & output

Disentangling representations & attribution

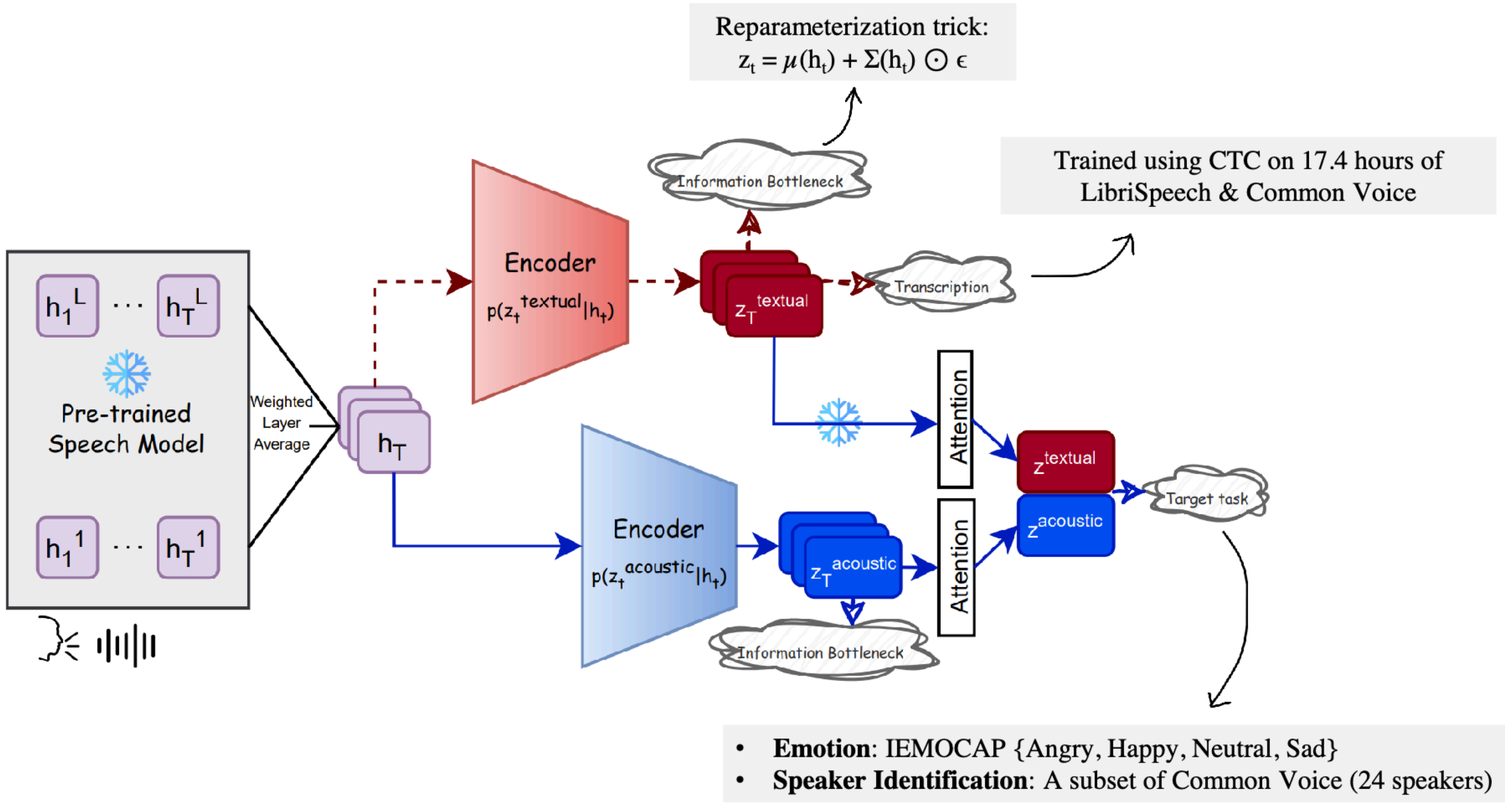
Speech model representations encode different kinds of interpretable information in an *entangled* way

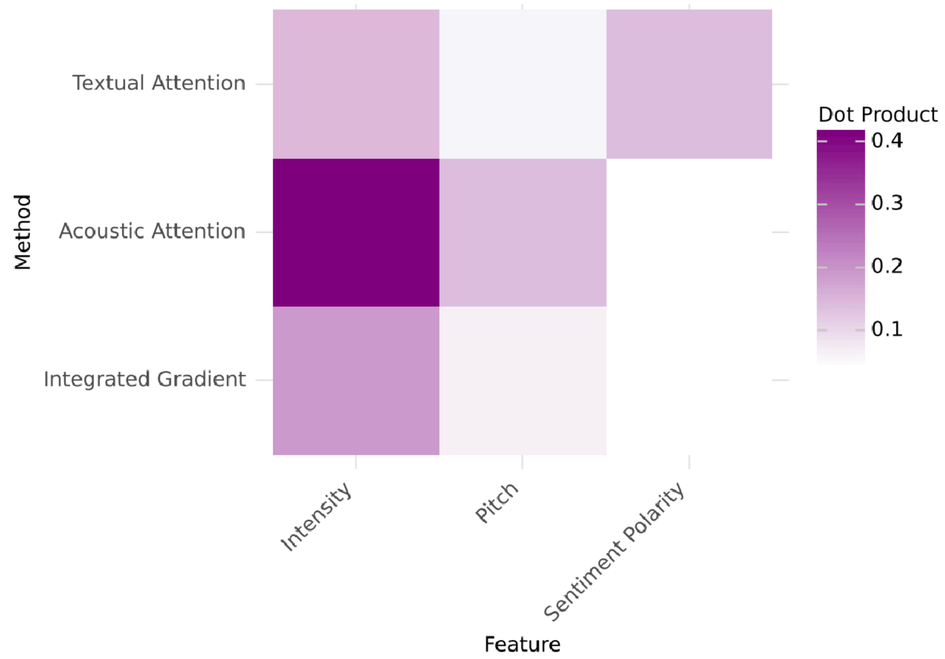
- Can we disentangle such dimensions for specific tasks?
- Can we attribute model predictions to such interpretable dimensions?



de Seyssel, M., Lavechin, M., Adi, Y., Dupoux, E., Wisniewski, G. Probing phoneme, language and speaker information in unsupervised speech representations. Proc. Interspeech 2022

Disentangling representations & attribution





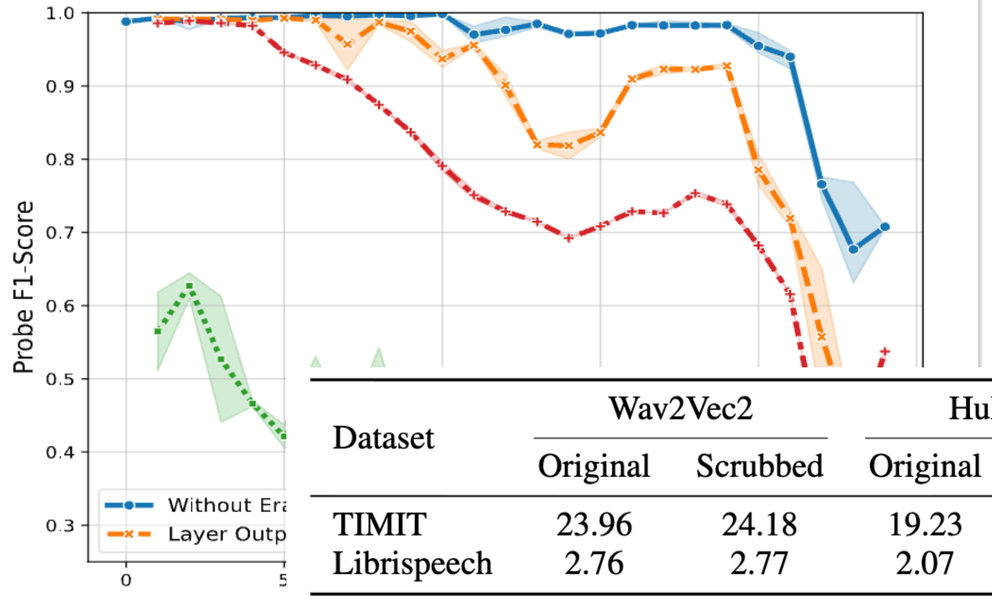
Mohebbi, H., Chrupała, G., Zuidema, W., Alishahi, A. Titov, I. (2025). Posthoc Disentanglement of Textual and Acoustic Features in Self-Supervised Speech Encoders. Actionable Interpretability@ICML 2025

Outlook:

Can interpretability be useful?

Understanding links between model internals and behaviour

Gender encoding & bias



Dataset	Wav2Vec2		HuBERT	
	Original	Scrubbed	Original	Scrubbed
TIMIT	23.96	24.18	19.23	21.57
Librispeech	2.76	2.77	2.07	2.90

Table 2: *WER before and after gender scrubbing*

Figure 1: *Gender scrubbing for the Wav2Vec2 ASR model.*

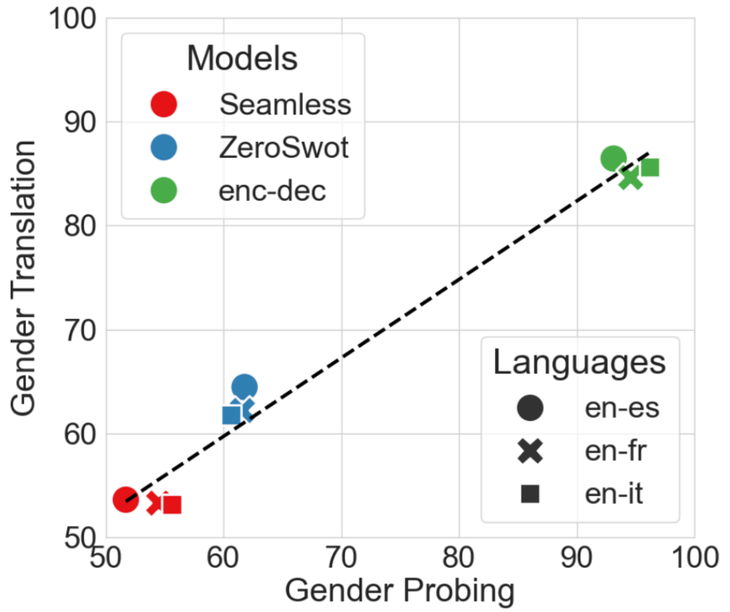


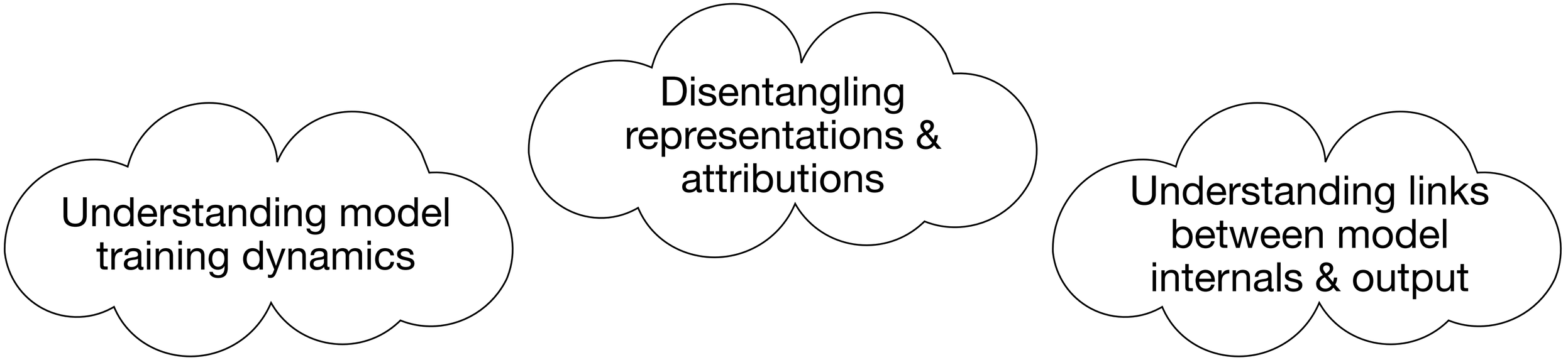
Figure 1: Correlation between overall gender probing performance (macro F1) and gender translation accuracy across models and languages on `test-speaker`.

Krishnan, A., Abdullah, B. M., & Klakow, D. On the encoding of gender in Transformer-based ASR representations. *Proc Interspeech 2024.*

Fucci, D., Gaido, M., Negri, M., Bentivogli, L., Martins, A., & Attanasio, G. Different Speech Translation Models Encode and Translate

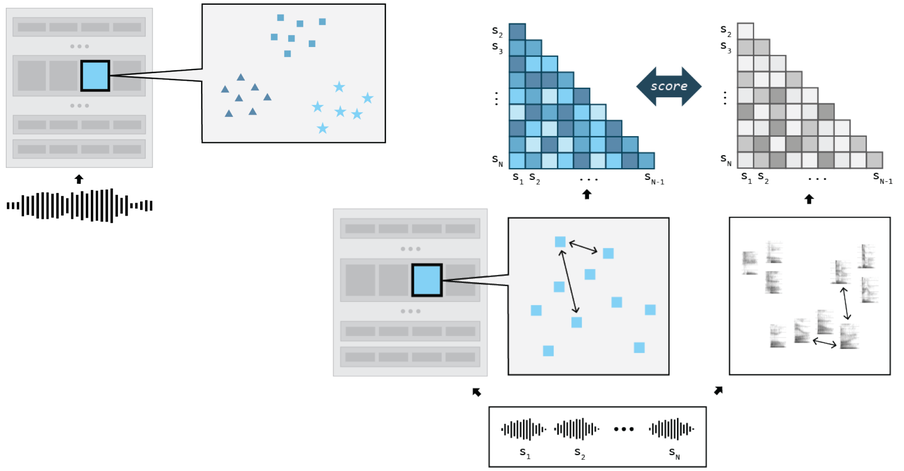
Outlook:

Can interpretability be useful?

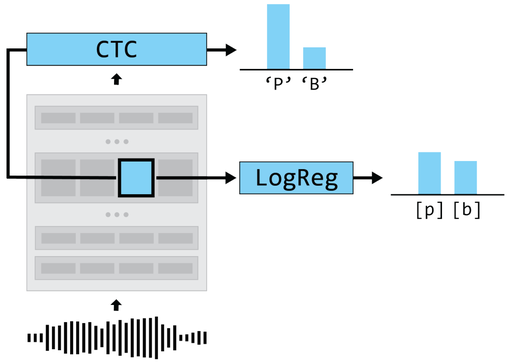


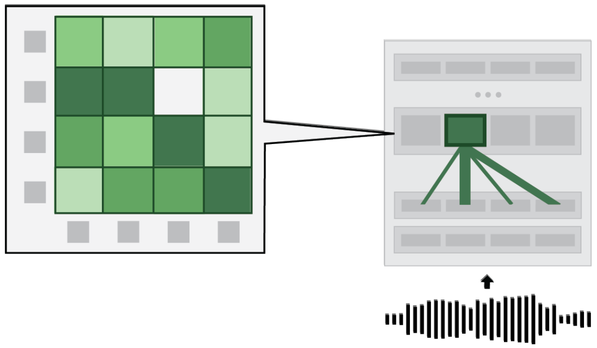
Interpretability Techniques for Speech Models

Behavioural analyses

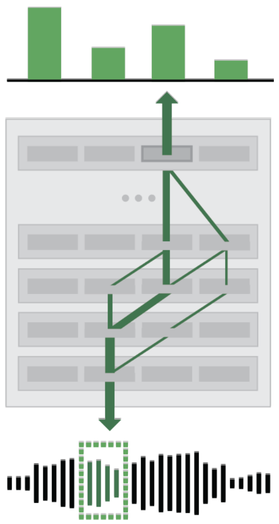


Representational analyses

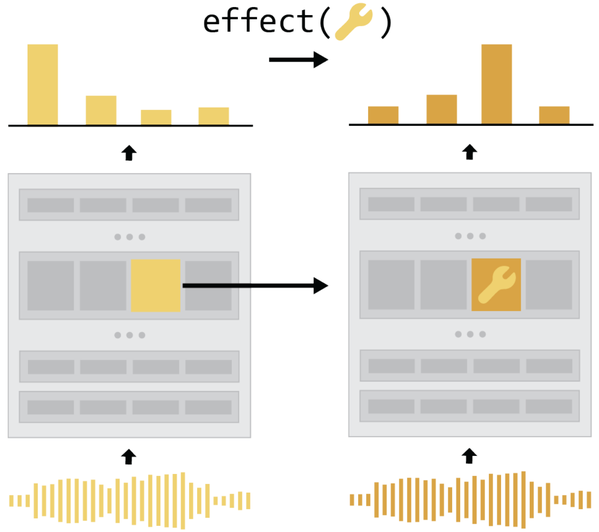




Feature Importance Scoring



Figures:
Marianne de Heer Kloots (2025)



Causal analyses