Simple, training-free methods for SSL-based speech processing

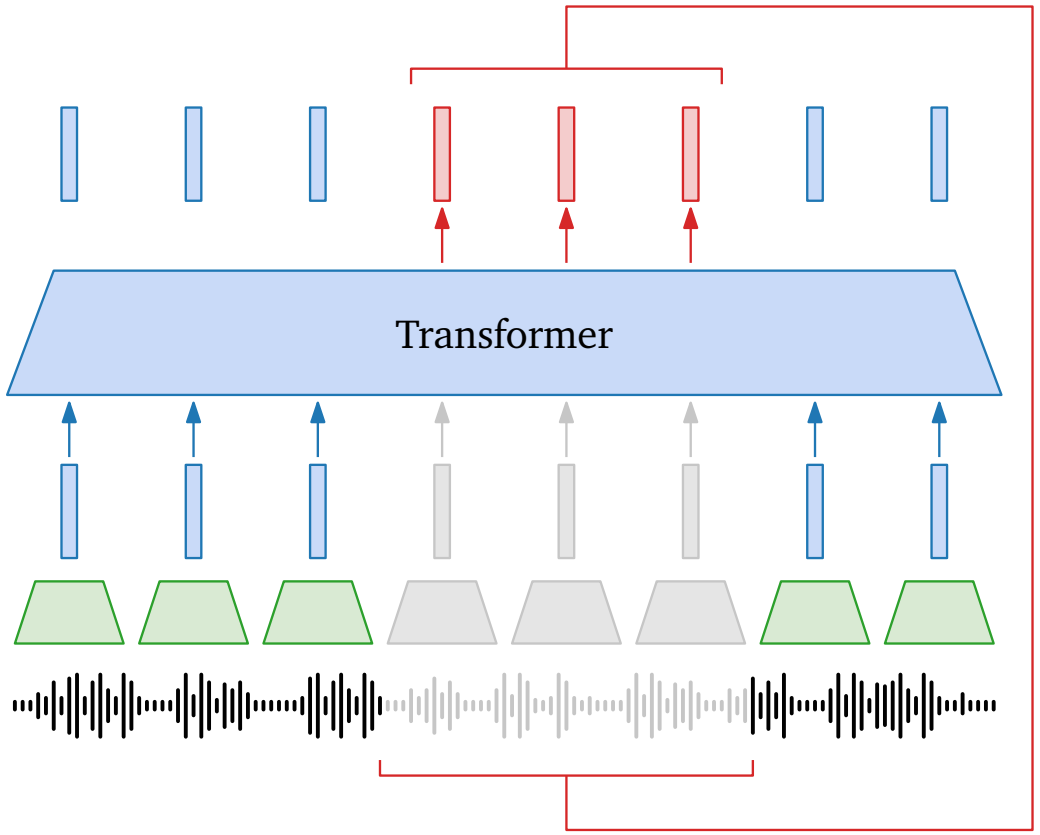
Herman Kamper

Electrical and Electronic Engineering, Stellenbosch University, South Africa

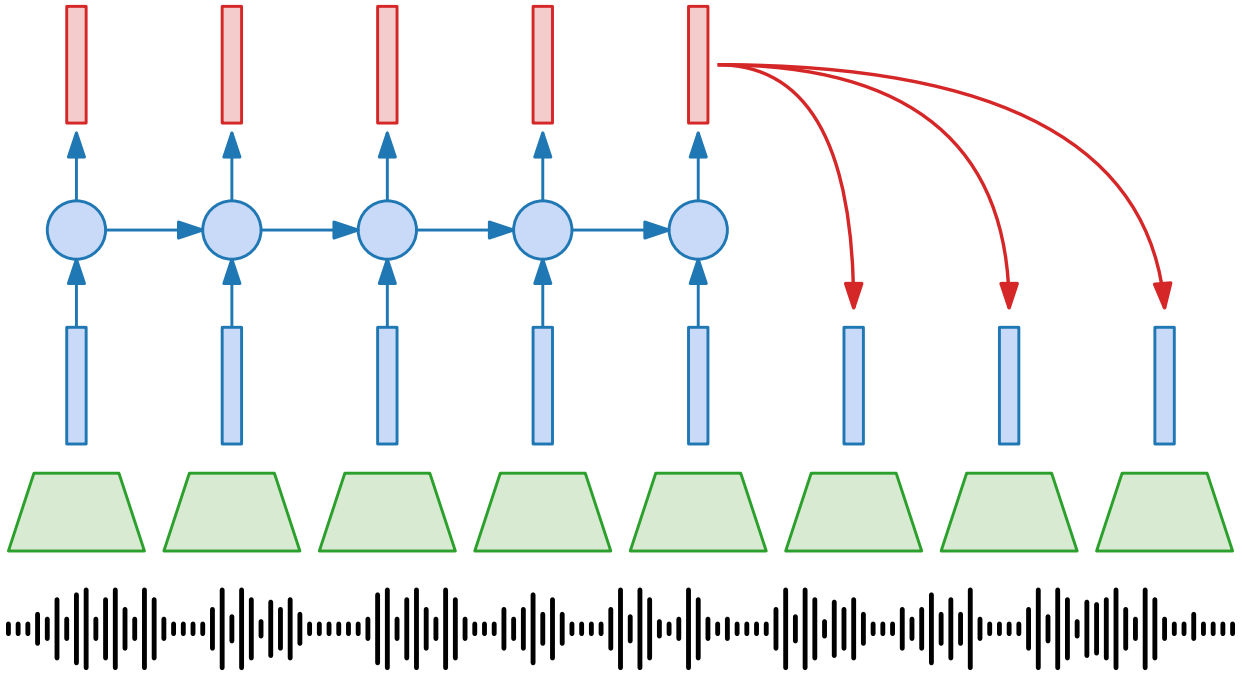
https://www.kamperh.com/

Self-supervised speech models

HuBERT / WavLM:



Contrastive predictive coding (CPC):



Takeaways for today

1. Synthesis can give unique insights into geometry of SSL features
2. Simple, training-free methods still have a place in speech processing

1. Through the lens of voice conversion:
Linear transformations of SSL features

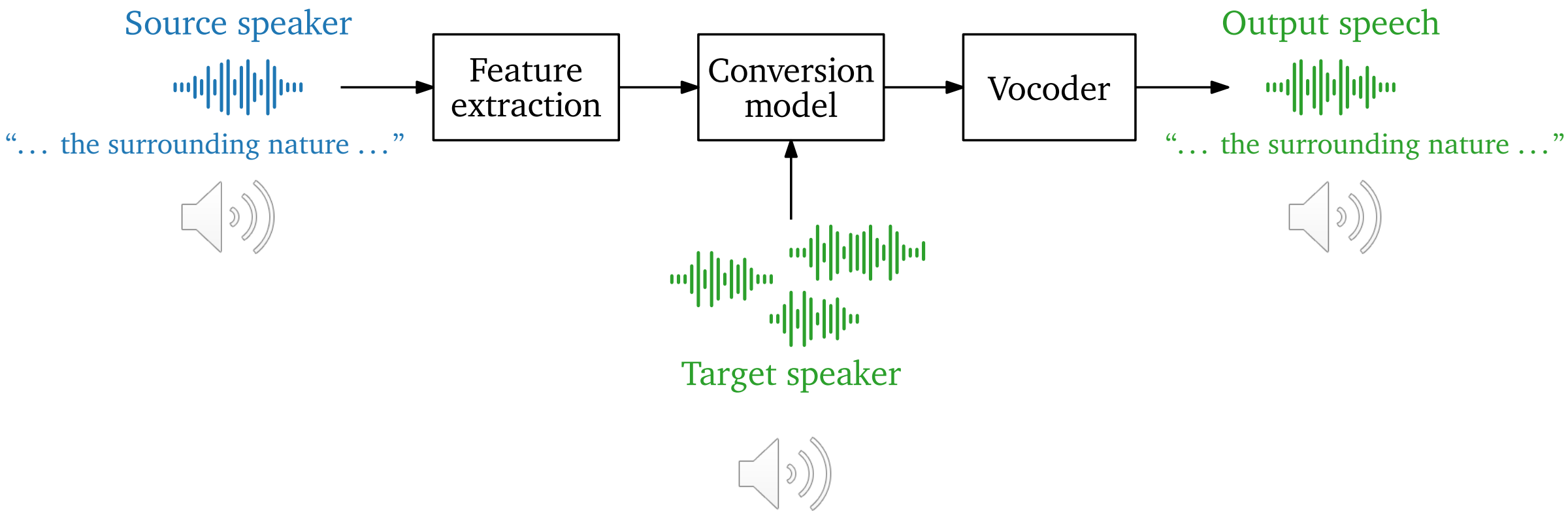
Benjamin van Niekerk

Julian Zaïdi

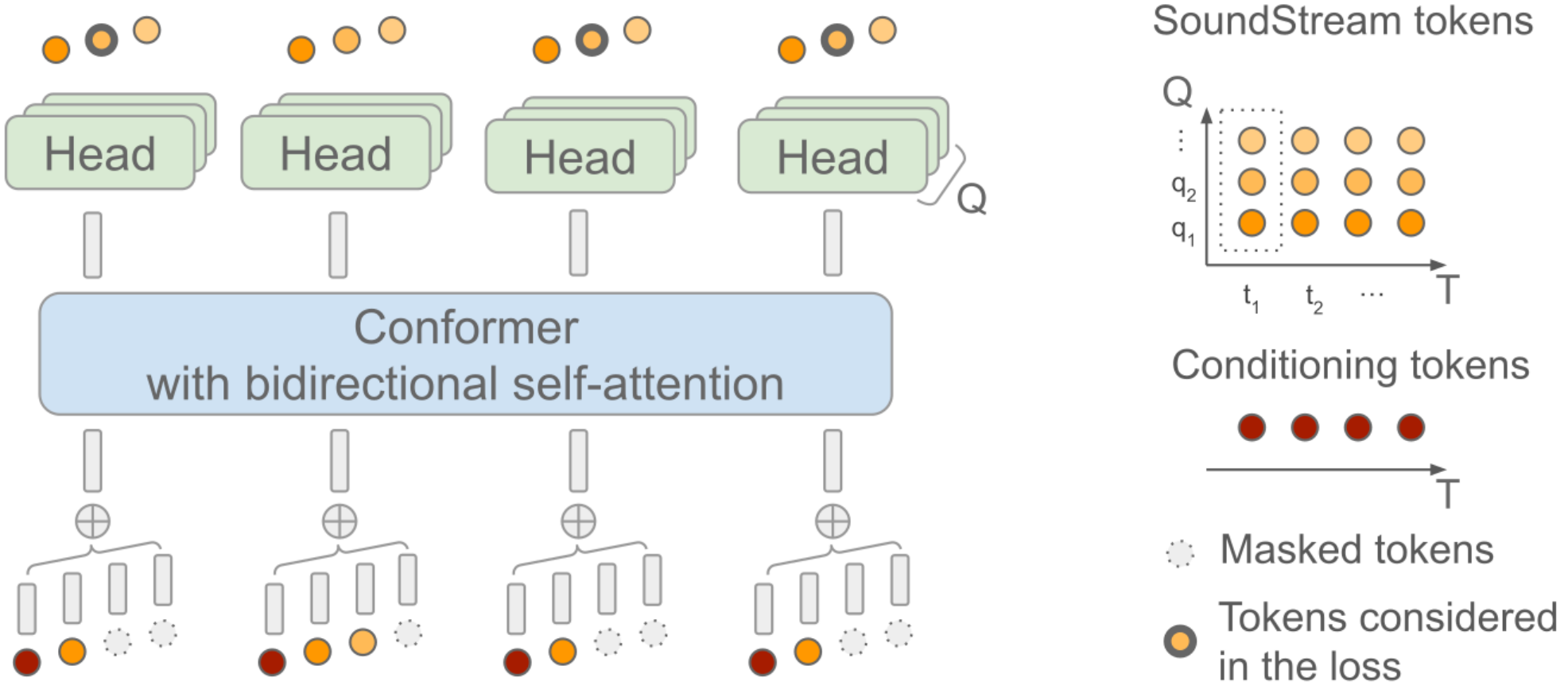
Marc-André Carbonneau

H. Kamper, B. van Niekerk, J. Zaïdi, and M-A. Carbonneau,“LinearVC: Linear transformations of self-supervised features through the lens of voice conversion,” in *Interspeech*, 2025.

Voice conversion



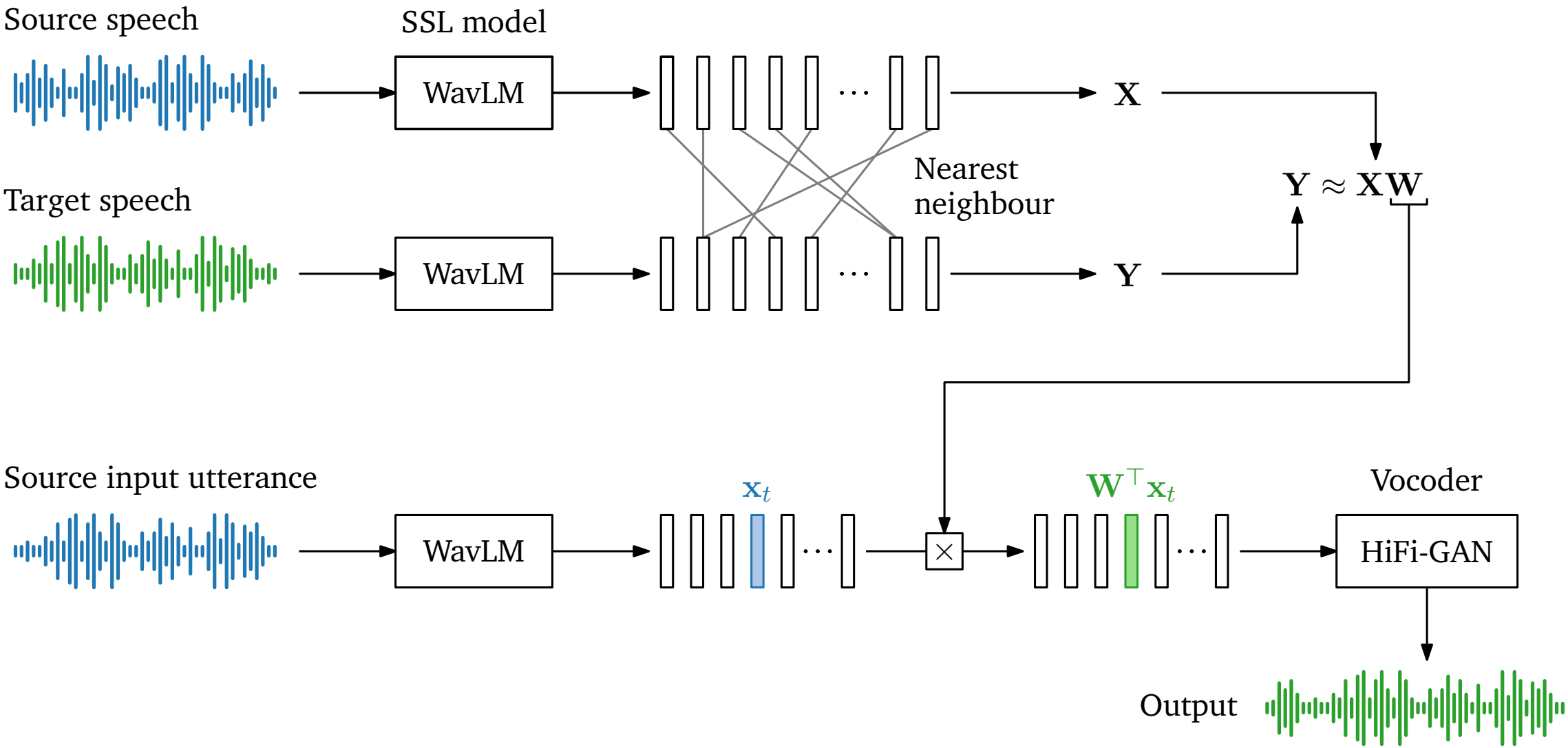
Codec-based spoken language model



Z. Borsos, M. Sharifi, D. Vincent, E. Kharitonov, N. Zeghidour, and M. Tagliasacchi, "SoundStorm: Efficient parallel audio generation," *arXiv preprint*, 2023.

LinearVC intuition

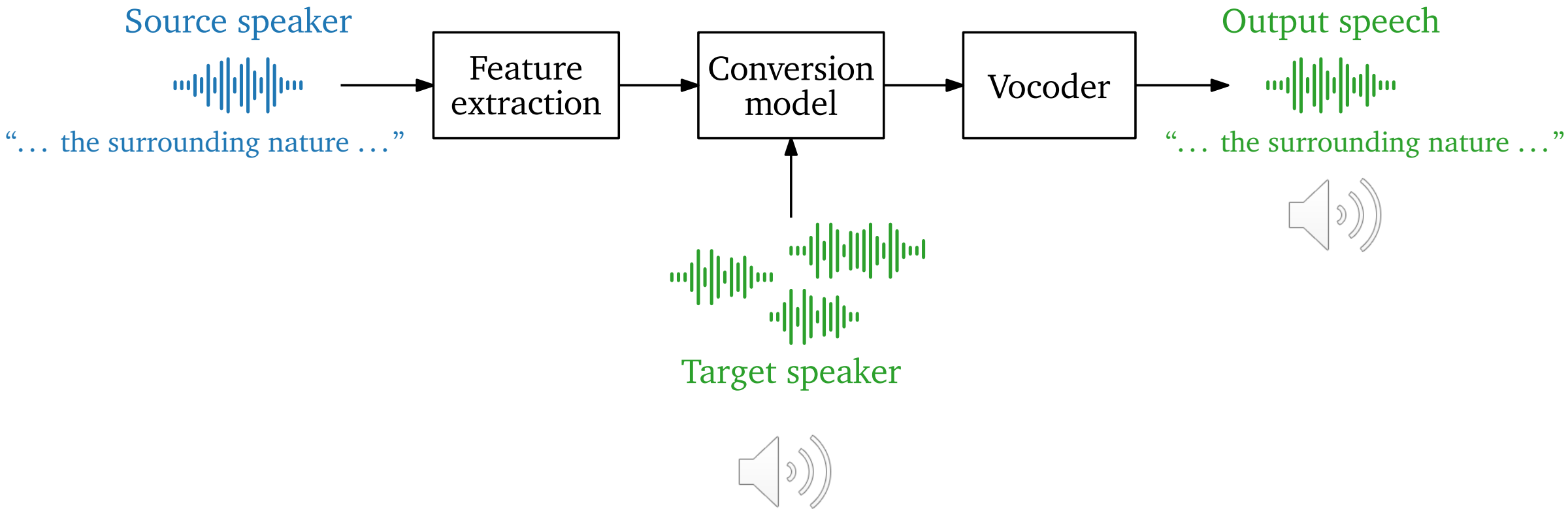
LinearVC


Source speech
SSL model
WavLM
X
Target speech
WavLM
Nearest neighbour
Y
Y ≈ XW
Source input utterance
WavLM
$\mathbf{x}_t$
$\mathbf{W}^\top \mathbf{x}_t$
Vocoder
HiFi-GAN
Output

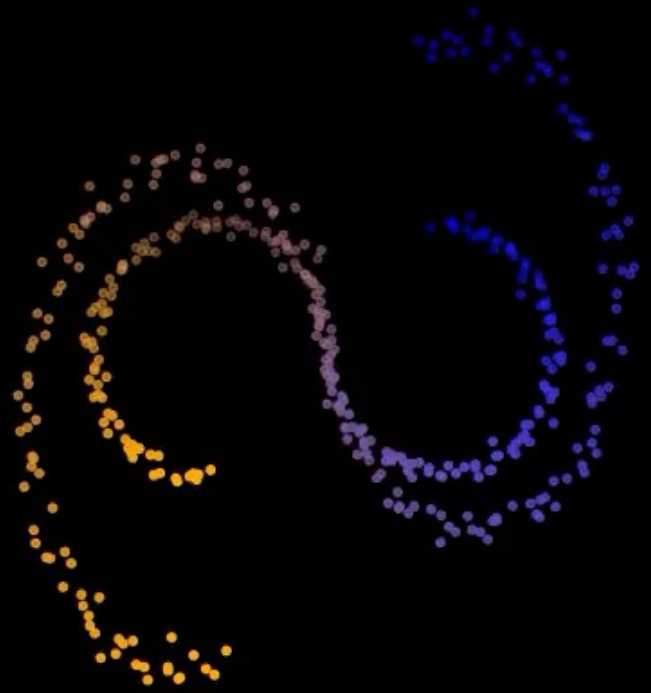
Voice conversion results (%)

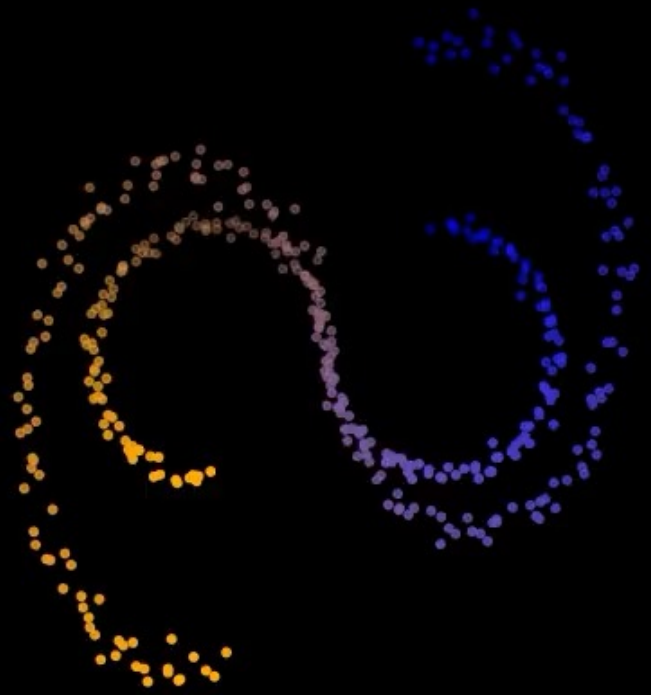
Model	WER↓	EER↑	Naturalness↑	Similarity↑
kNN-VC (Baas et al. 2023)	5.7	**38.9**	60.6±3.6	67.2±2.7
FreeVC (Li et al. 2022)	5.7	10.5	**71.1±3.6**	48.7±2.9
SoundStorm (Borsos et al. 2023)	**4.6**	30.2	58.6±4.0	**68.6±3.2**
LinearVC	4.9	33.6	62.5±3.5	67.5±2.6
Ground truth	4.3	50.0	-	-

LinearVC samples



https://www.kamperh.com/linearvc/

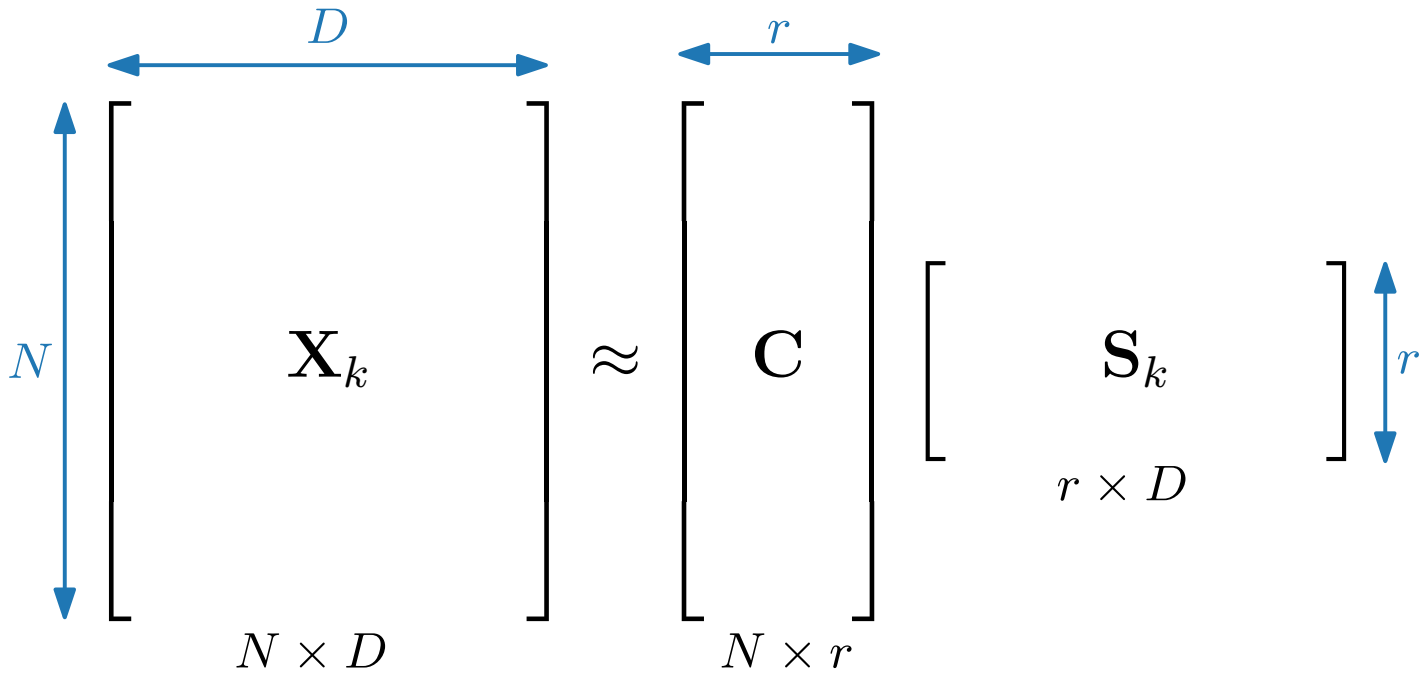




LinearVC with content factorisation

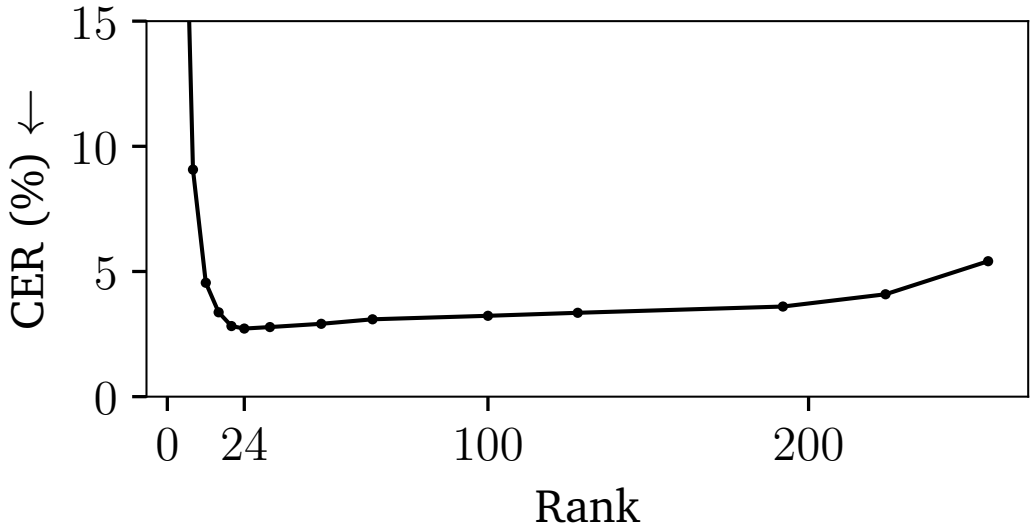
$$\min_{\mathbf{C},\mathbf{S}_k} \quad \sum_{k=1}^{K} \|\mathbf{X}_k - \mathbf{C}\mathbf{S}_k\|_F^2$$

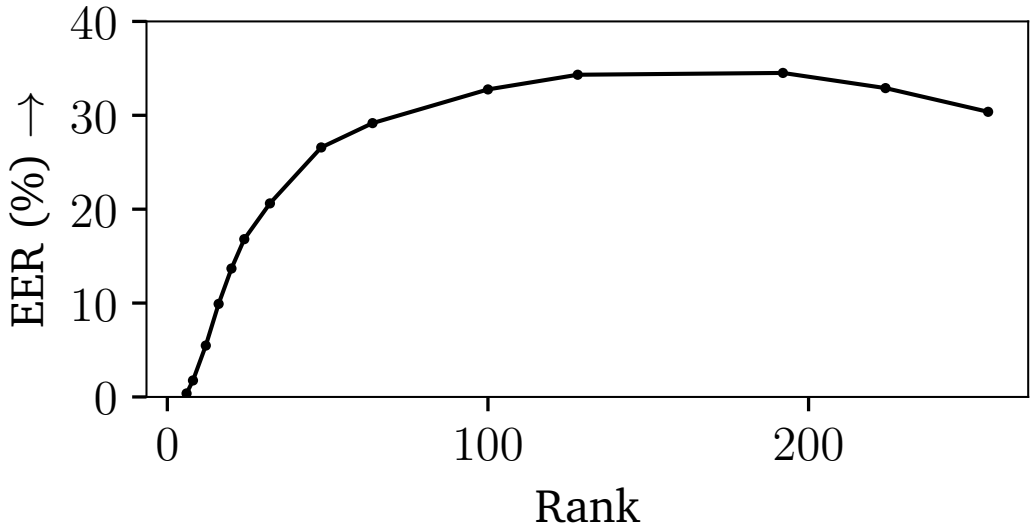
$$\text{subject to} \quad \operatorname{rank}(\mathbf{C}\mathbf{S}_k) \leq r$$

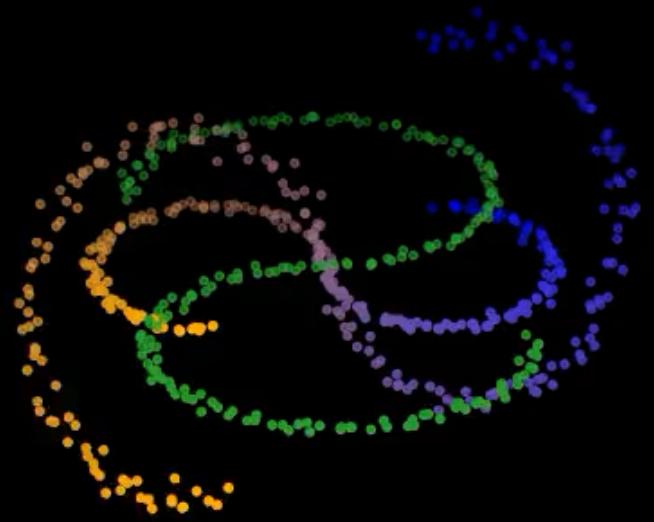


LinearVC with content factorisation

- Source:
- Target:
- Rank = 256:
- Rank = 100:
- Rank = 16:
- Rank = 6:





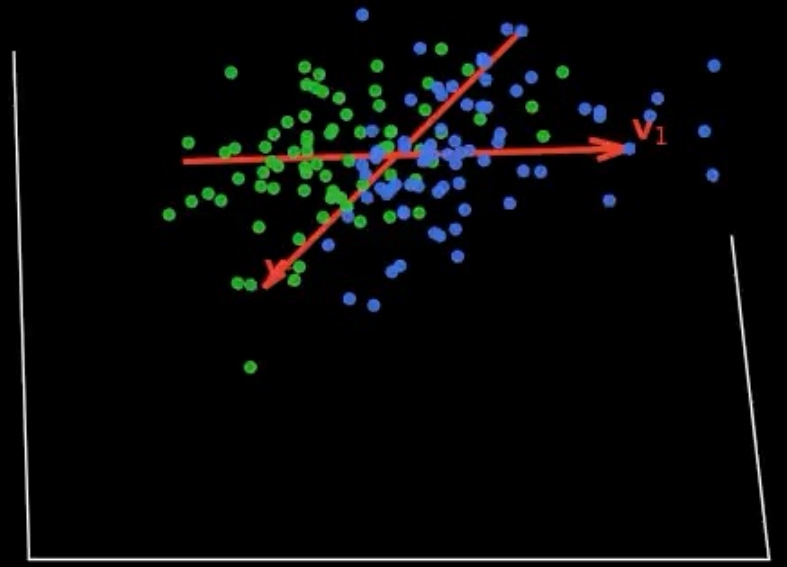


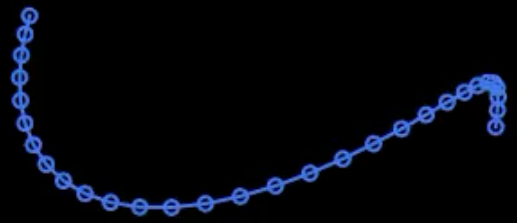
2. Through the lens of voice manipulation:
Interpreting the dimensions of SSL features

Benjamin van Niekerk

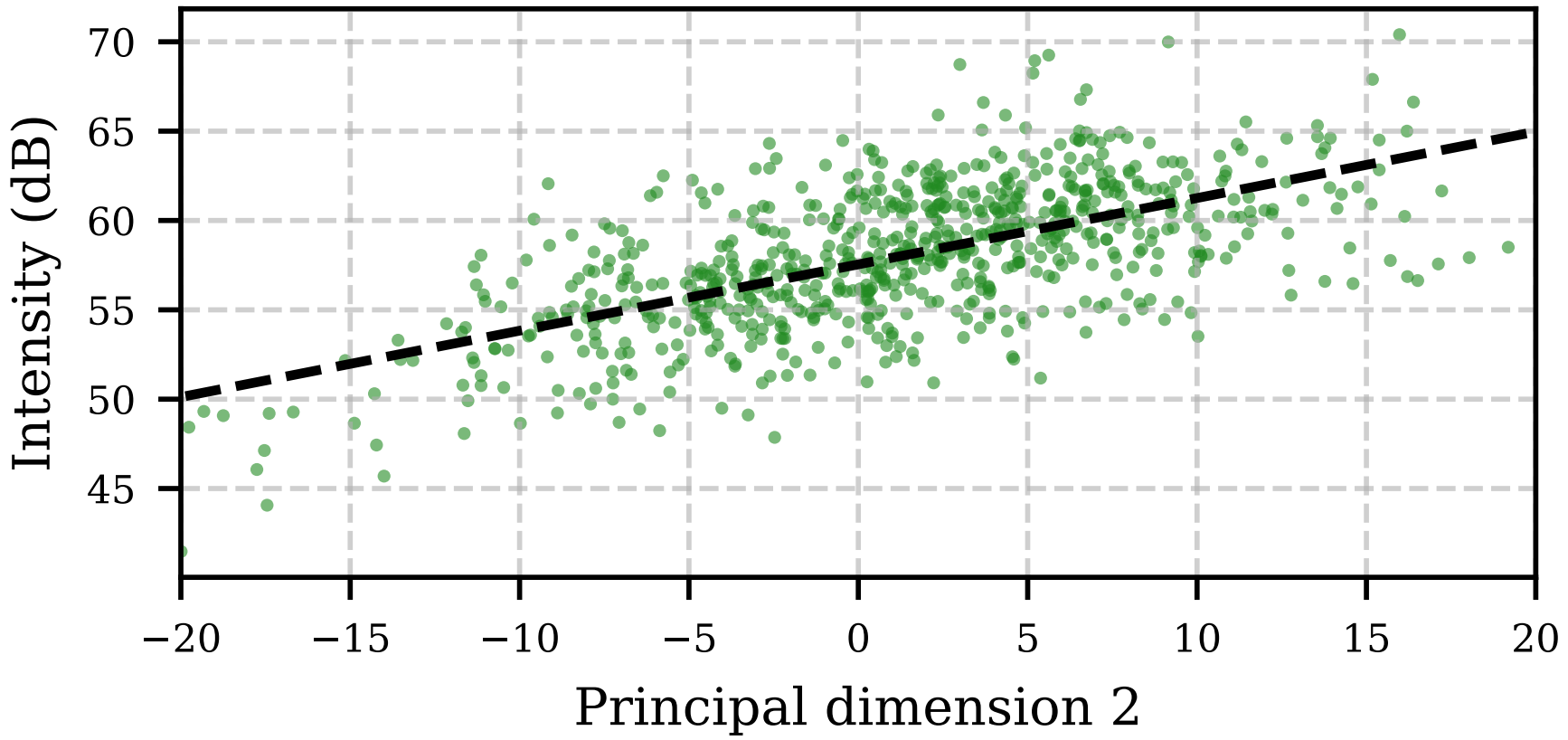
Kyle Janse van Rensburg

K. Janse van Rensburg, B. van Niekerk, and H. Kamper, “Interpreting speaker characteristics in the dimensions of self-supervised speech features,” *arXiv preprint*, 2026.

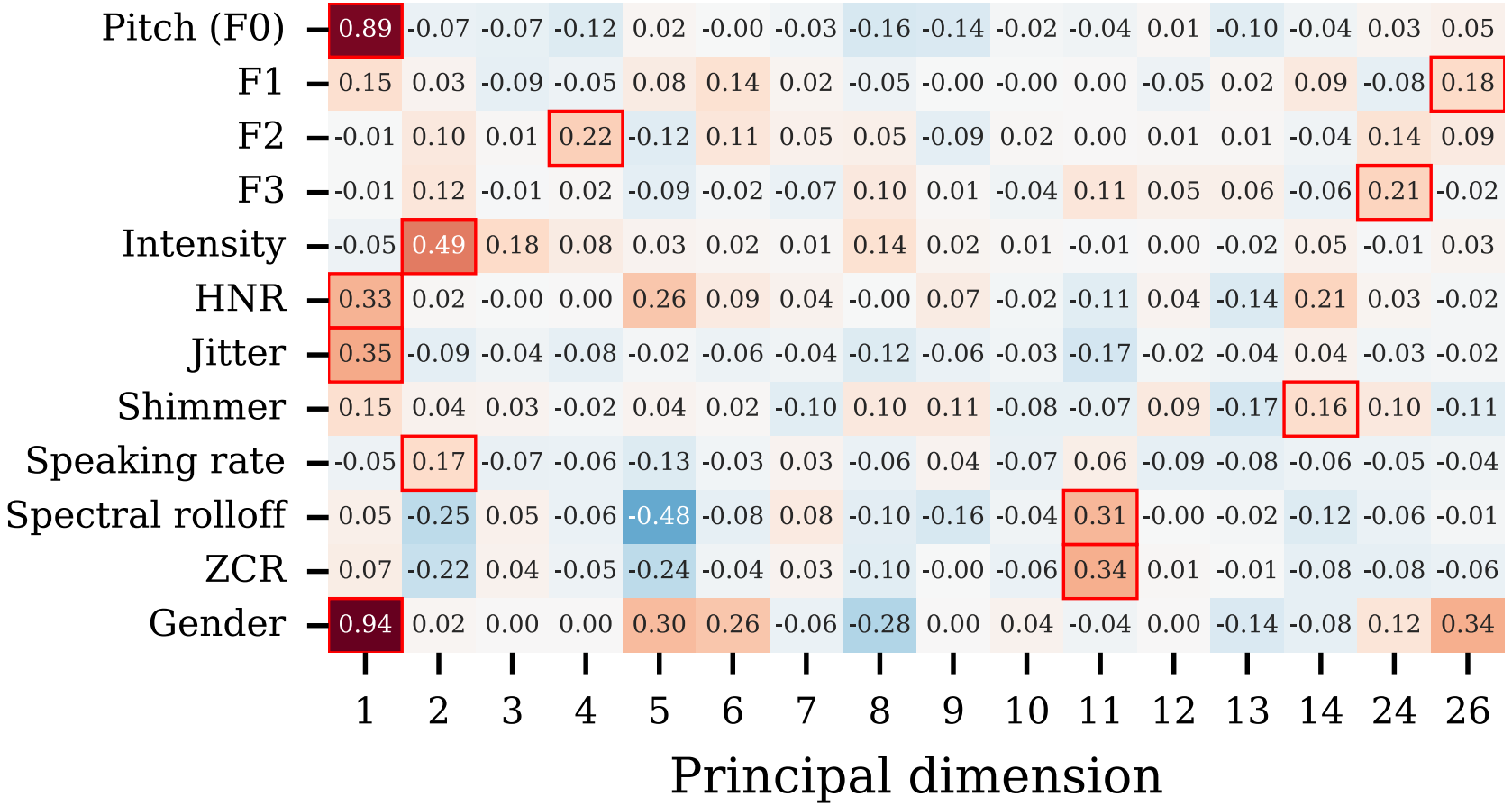

v_1



Correlation analysis



Correlation analysis



3. Through the lens of unsupervised segmentation: The norms of SSL features

Nicol Visser

Simon Malan

Danel Slabbert

N. Visser, S. Malan, D. Slabbert, and H. Kamper, “ZeroSyl: Simple zero-resource syllable tokenization for spoken language modeling,” *arXiv preprint*, 2026.

What Do Self-Supervised Speech Models Know About Words?

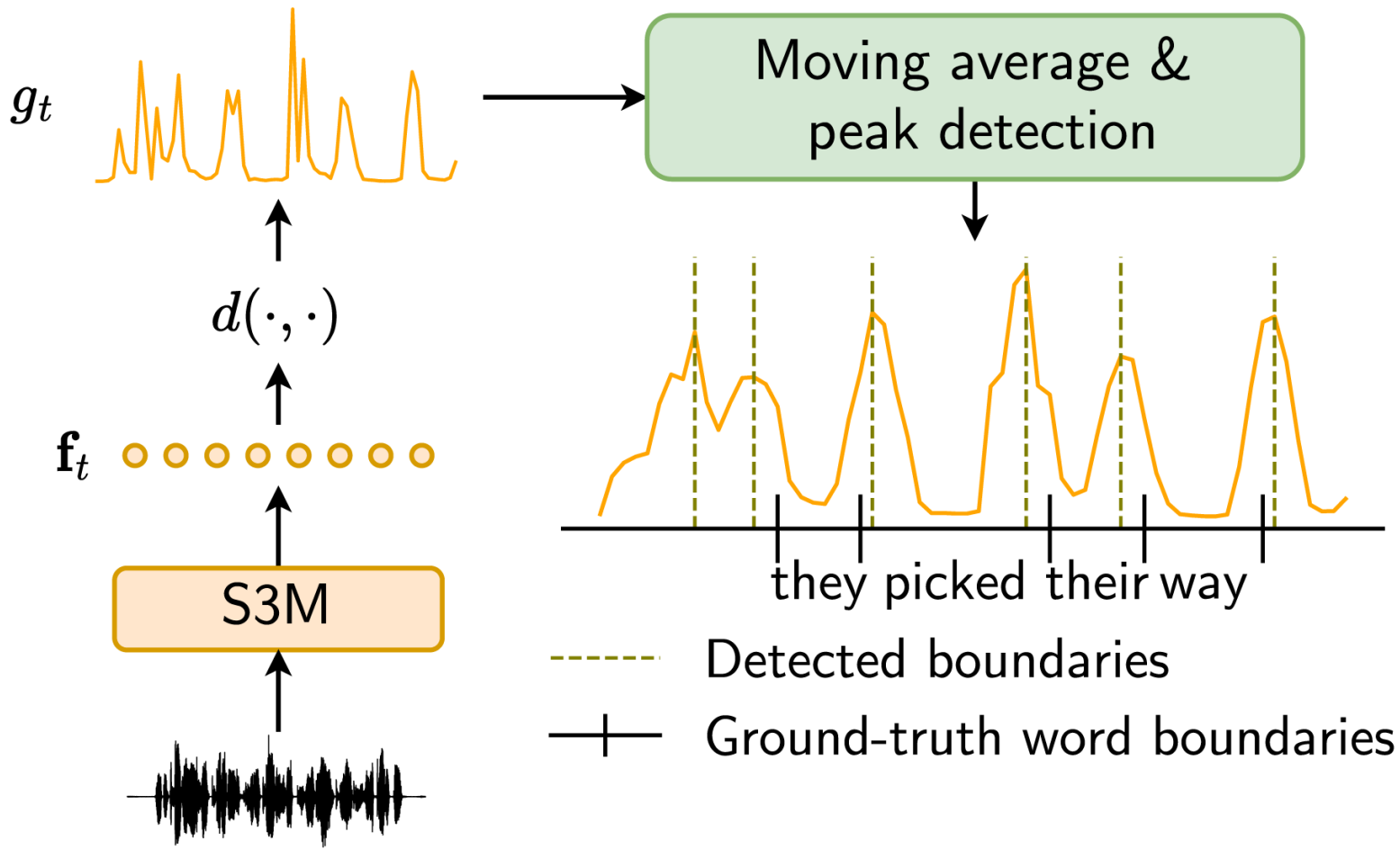
Ankita Pasad, Chung-Ming Chien, Shane Settle, Karen Livescu

Check for updates

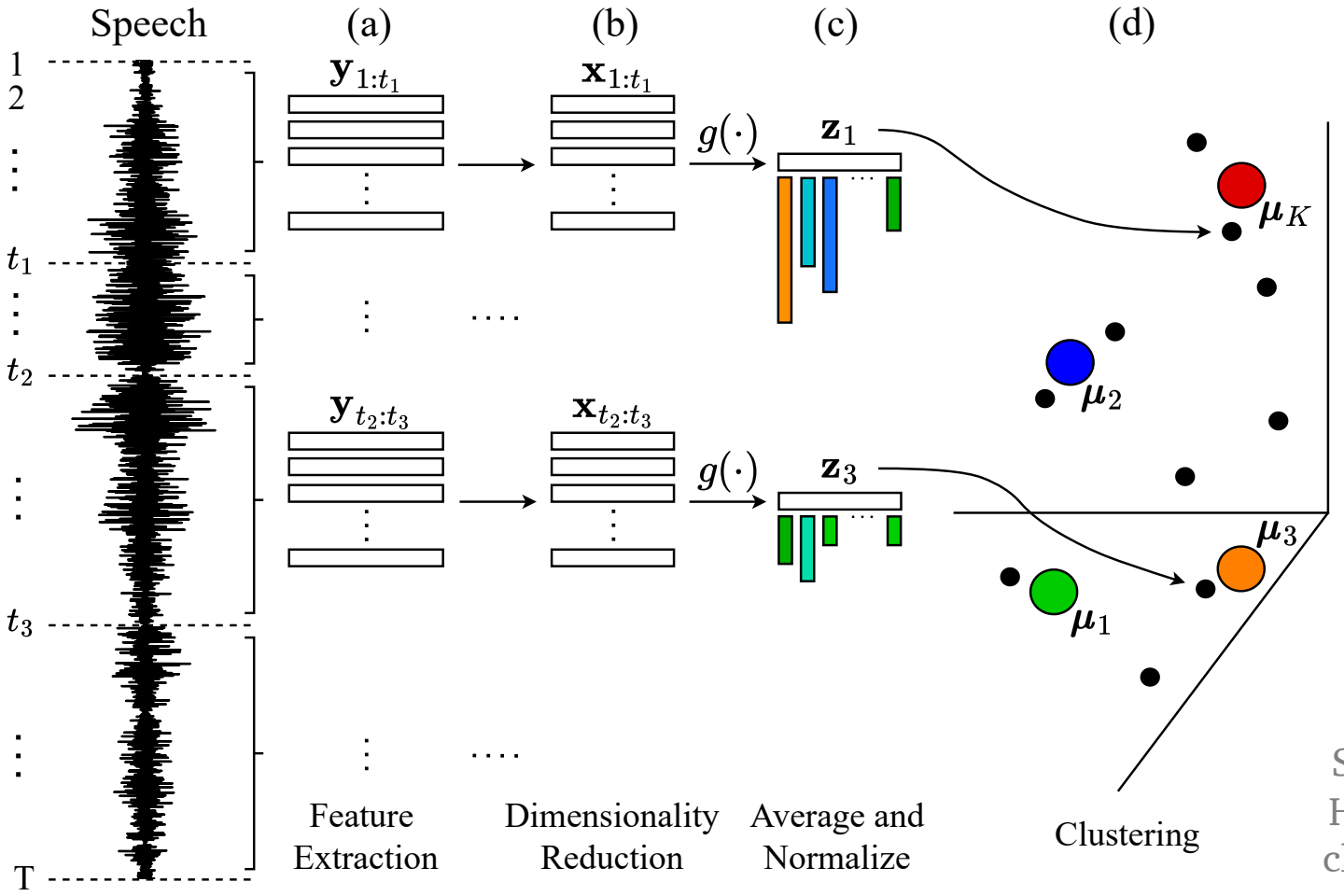
> Author and Article Information

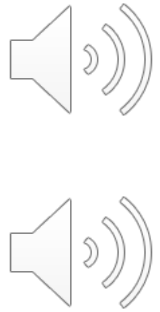
Transactions of the Association for Computational Linguistics (2024) 12: 372–391.

https://doi.org/10.1162/tacl_a_00656



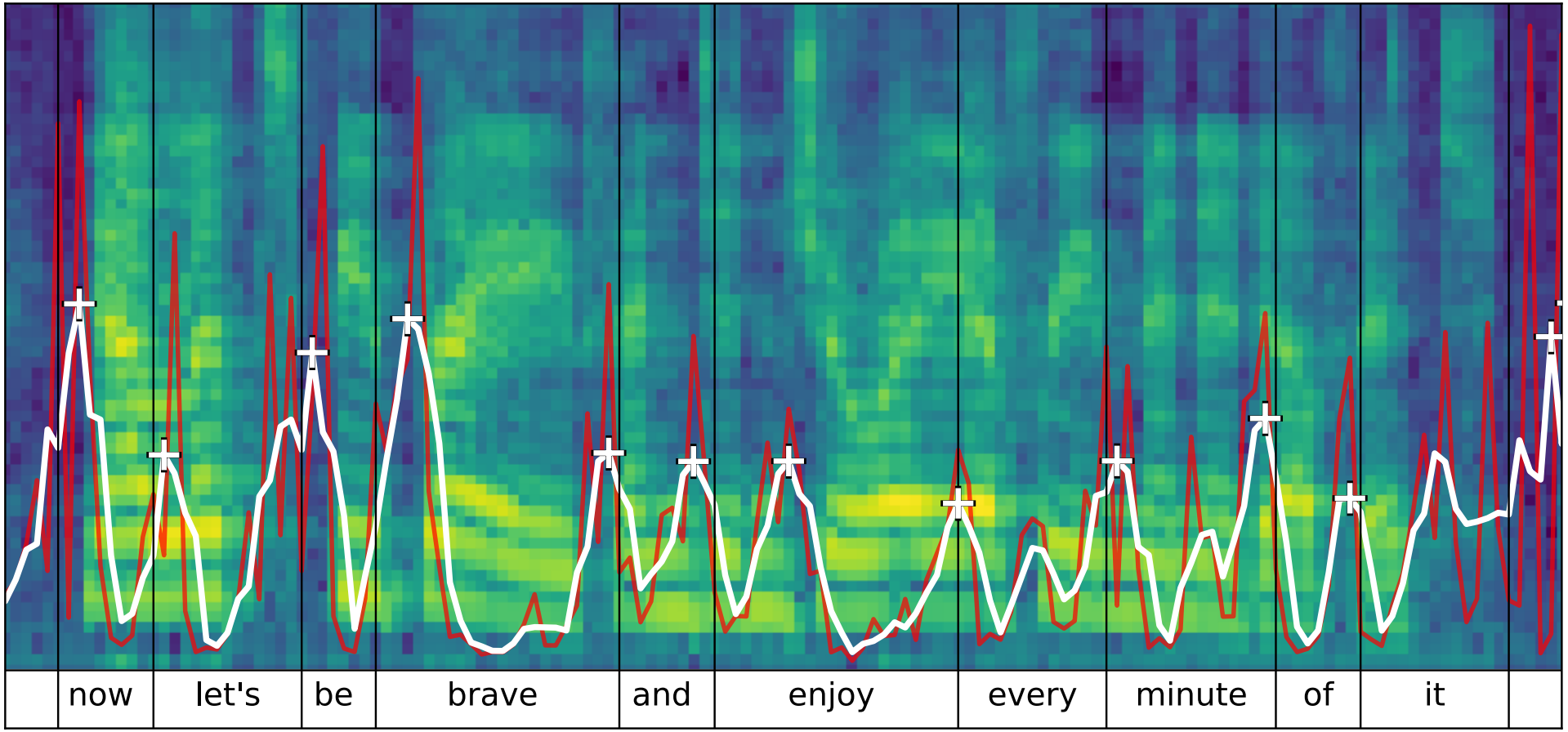
Prominence-based unsupervised word discovery



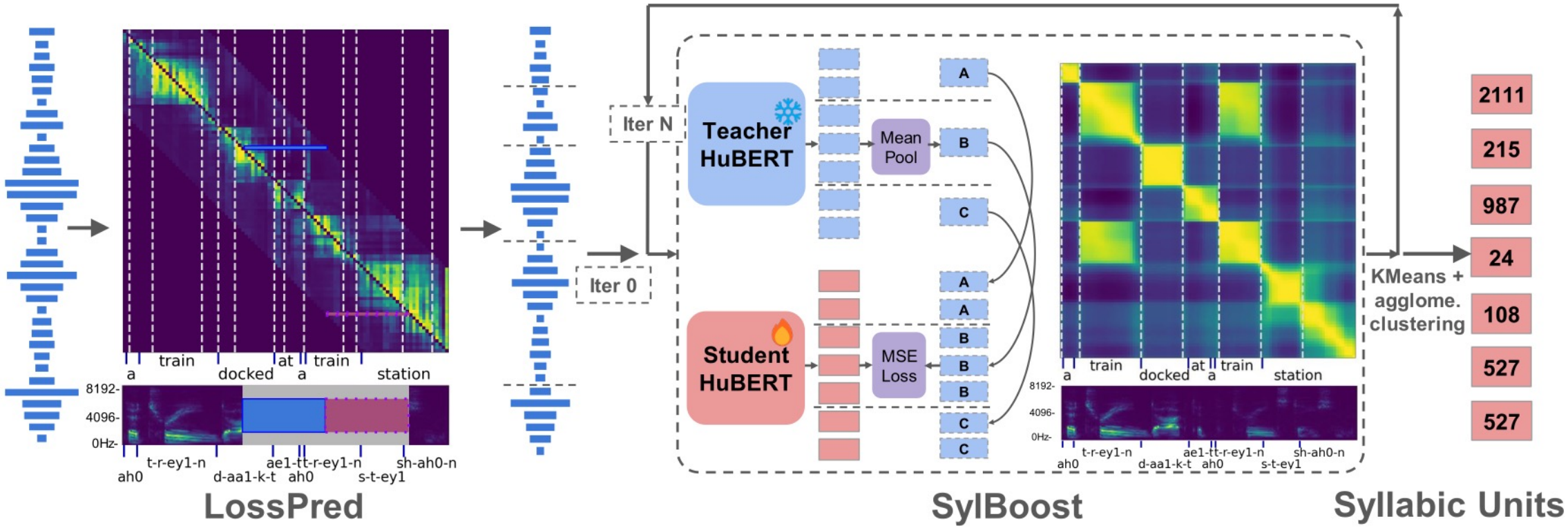


S. Malan, B. van Niekerk, and H. Kamper, "Should top-down clustering affect boundaries in unsupervised word discovery?," *TASLP*, 2026.

Prominence-based unsupervised speech segmentation



SyllableLM and Sylber

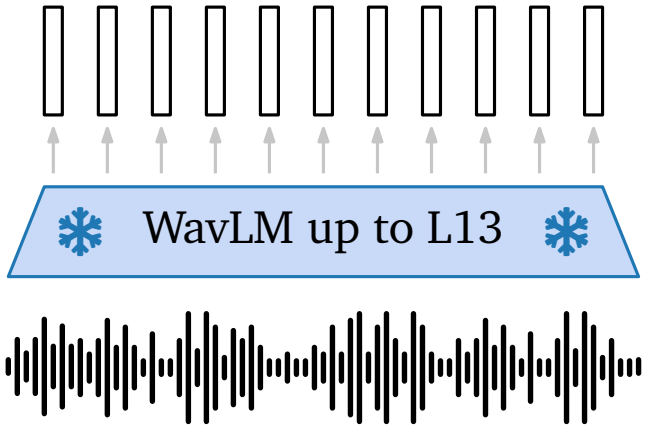


A. Baade, P. Peng, and D. Harwath, "SyllableLM: Learning coarse semantic units for speech language models," in *ICLR*, 2025.

Syllable boundary segmentation results

Model	Boundary Prec.	Boundary Recall	Boundary F1	Boundary R-score	Token Prec.	Token Recall	Token F1
SyllableLM 5.00 Hz	71	79	75	77	54	58	56
SyllableLM 6.25 Hz	62	87	72	59	44	59	50
SyllableLM 8.33 Hz	50	91	65	27	31	52	39
Sylber	65	74	69	71	48	54	51
PromSeg	62	65	64	69	40	41	41
ZeroSyl	69	75	72	75	52	56	54

ZeroSyl: Simple syllable tokenisation for spoken language modelling

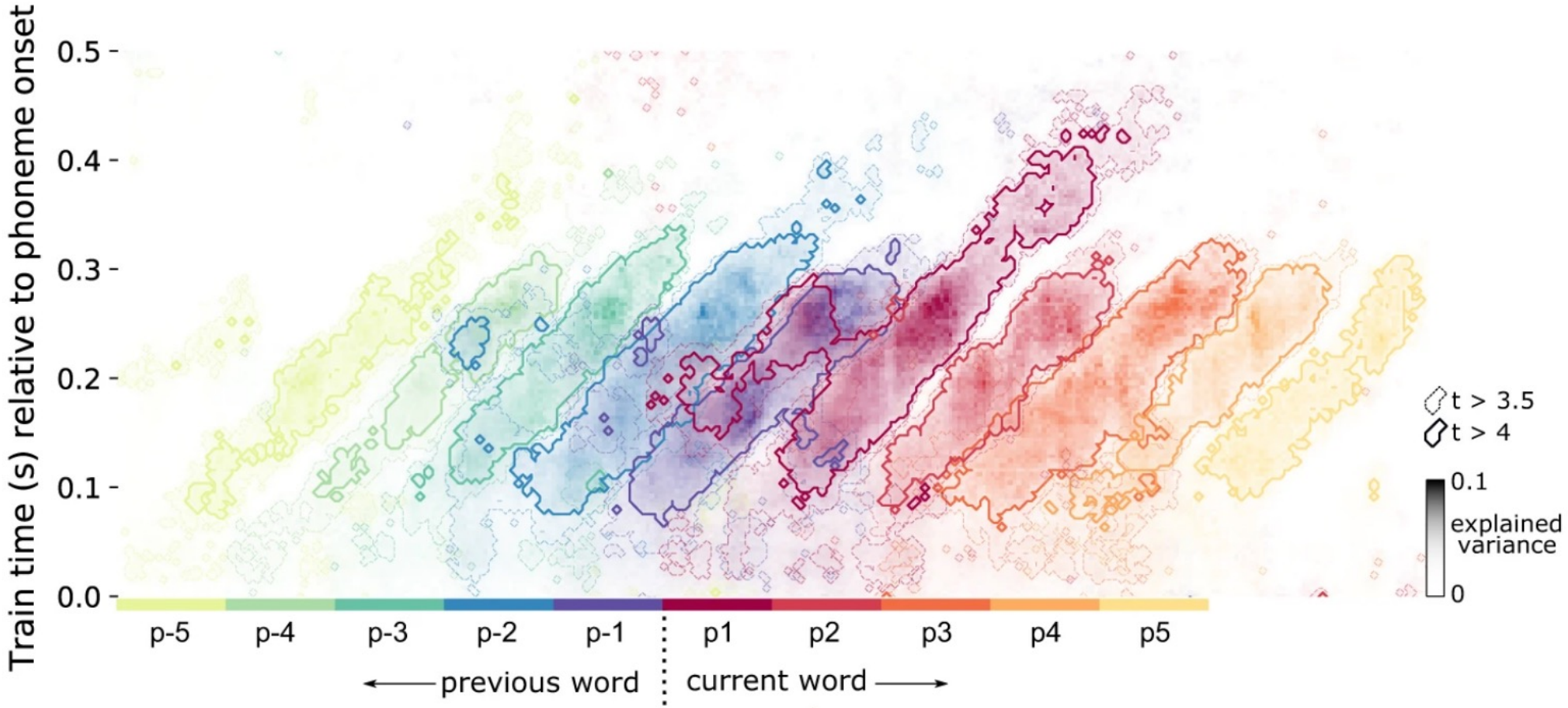


Spoken language modelling results

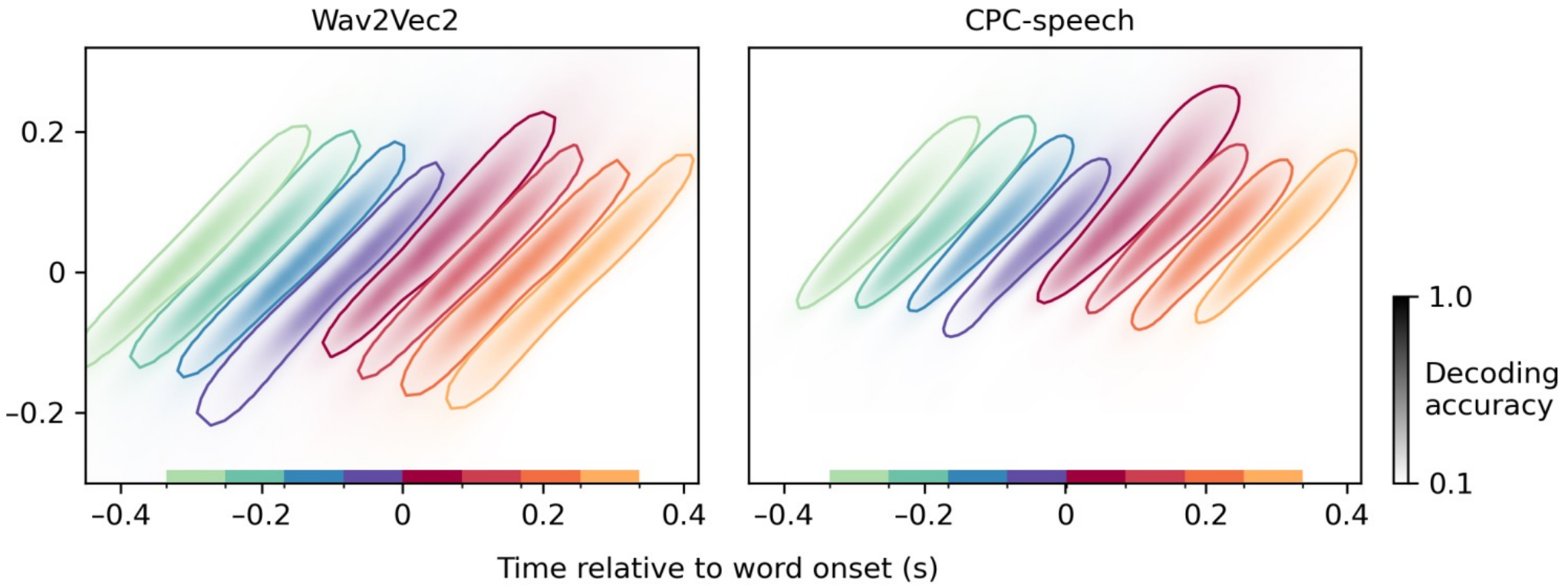
Model	Bitrate (bps)	Purity (%)	sWUGGY (%)	sBLIMP (%)	tSC (%)
SyllableLM 6.25Hz	73	67.5	66.1	56.4	67.6
Sylber	53	73.5	66.0	59.1	65.8
ZeroSyl	**52**	**79.9**	**68.0**	**60.5**	**68.1**
Text topline (BPE)			79.3	67.7	83.3

Spoken language modelling results

Model	Bitrate (bps)	Purity (%)	sWUGGY (%)	sBLIMP (%)	tSC (%)
SyllableLM 6.25Hz	73	67.5	66.1	56.4	67.6
Sylber	53	73.5	66.0	59.1	65.8
ZeroSyl	**52**	**79.9**	**68.0**	60.5	**68.1**
ZeroSyl – Unnormalised			66.8	**63.8**	57.2
Text topline (BPE)			79.3	67.7	83.3



L Gwilliams, J. R. King, A. Marantz, and D. Poeppel, "Neural dynamics of phoneme sequences reveal position-invariant code for content and order," *Nature Communications*, 2022.



O. D. Liu, H. Tang, N. H. Feldman, and S. Goldwater, “Brain-like dynamics in speech representations can emerge through self-supervised learning,” *bioRxiv preprint*, 2026.

Bringing it together

1. Synthesis can give unique insights into geometry of SSL features
2. Simple, training-free methods still have a place in speech processing
   - Looked at three diverse tasks as examples
   - Huge benefit: take advantage of big models as these become better