


Test slide

- Can you see my pointer?
- Can you hear this? 
- Can you see the video on the next slide?

Voice conversion and the geometry of self-supervised speech representations

Conversational AI Reading Group, June 2025

Herman Kamper

Electrical and Electronic Engineering, Stellenbosch University, South Africa

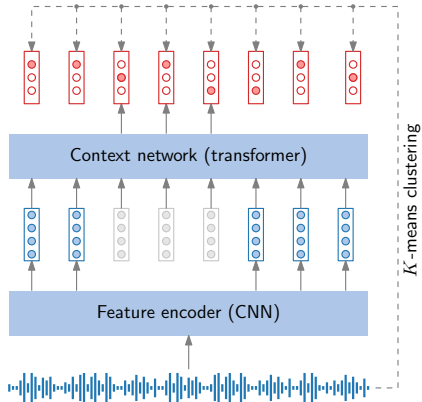
<http://www.kamperh.com/>



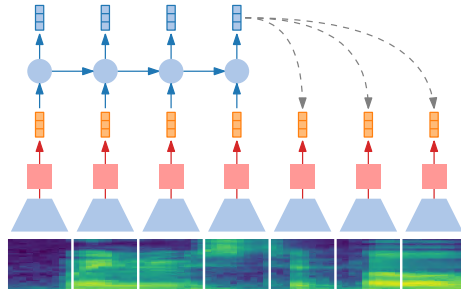


Self-supervised spoken language models

HuBERT / WavLM:

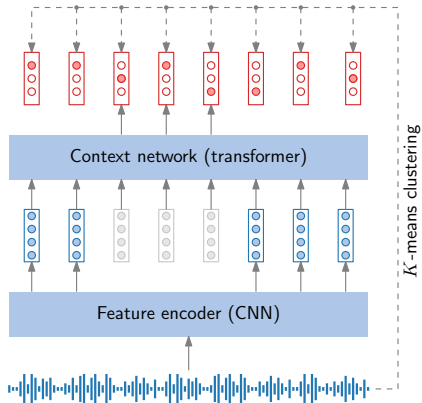


Contrastive predictive coding (CPC):

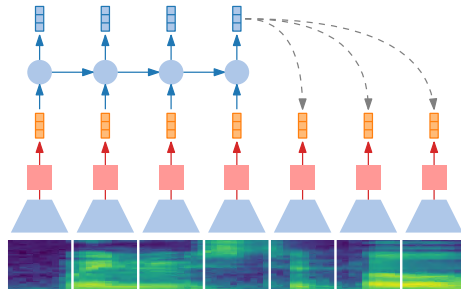


Self-supervised spoken language models

HuBERT / WavLM:



Contrastive predictive coding (CPC):



Caveat: SSL = WavLM layer six (1024 dimensional)

Voice conversion is useful for understanding SSL features

Agenda:

- Introduce two simple voice conversion approaches
- They give surprisingly good results, despite being dumb
- What does this tell us about the geometry of SSL features?

Voice conversion is useful for understanding SSL features

Agenda:

- Introduce two simple voice conversion approaches
- They give surprisingly good results, despite being dumb
- What does this tell us about the geometry of SSL features?

Main takeaways:

- The usefulness of voice conversion for probing
- Simpler methods are awesome

kNN-VC: Voice conversion with just nearest neighbours



Benjamin
van Niekirk

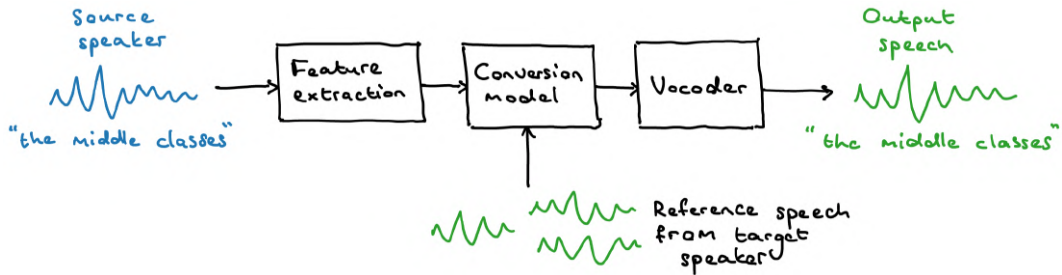


Matthew
Baas

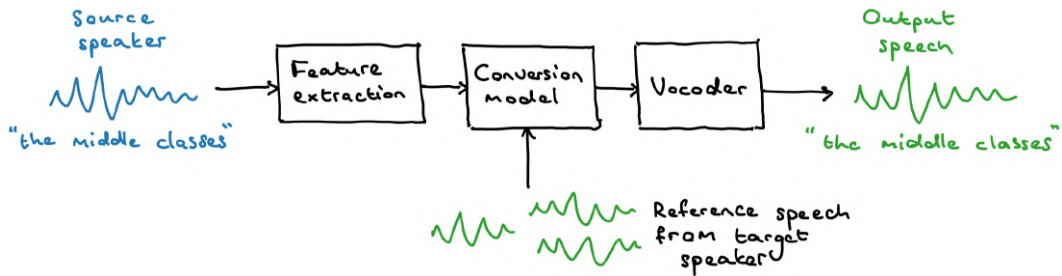
M. Baas, B. van Niekirk, and H. Kamper, "Voice conversion with just nearest neighbours," in *Interspeech*, 2023.

M. Baas and H. Kamper, "Voice conversion for stuttered speech, instruments, unseen languages and textually described voices," *Communications in Computer and Information Science*, 2023

Voice conversion



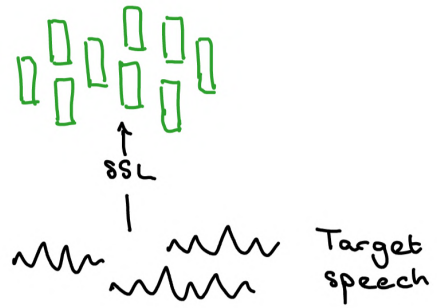
Voice conversion

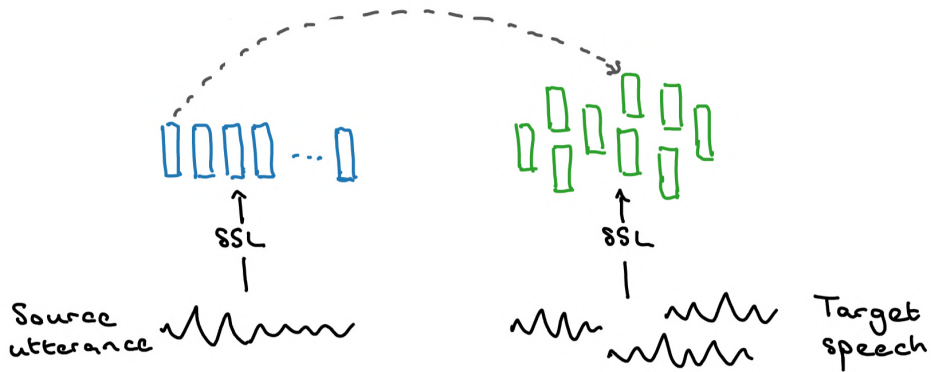


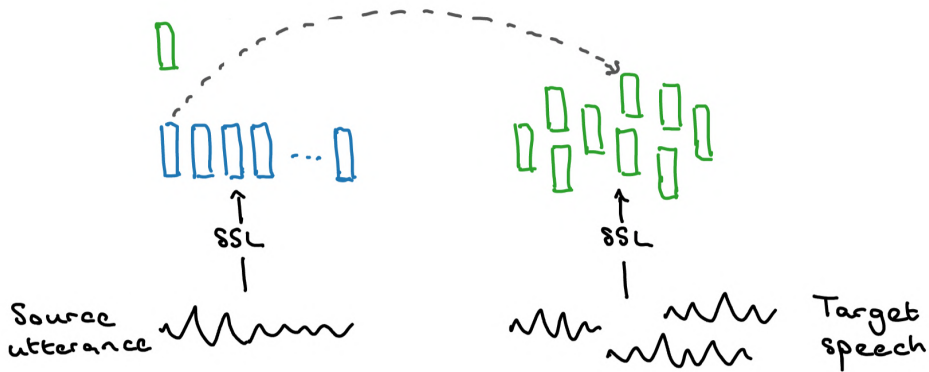
Source: [Play](#)

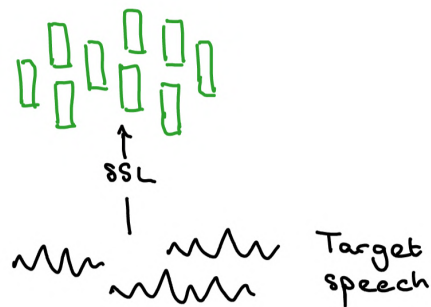
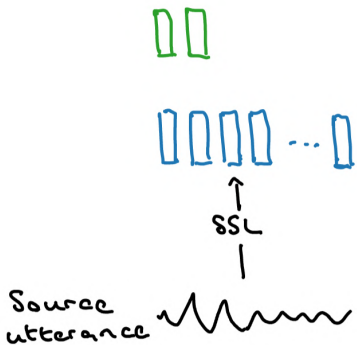
Reference: [Play](#)

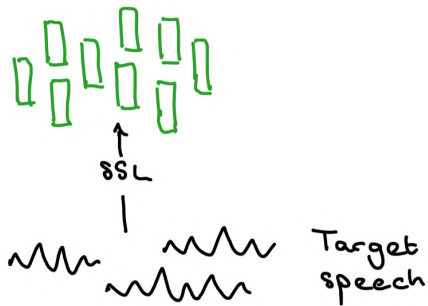
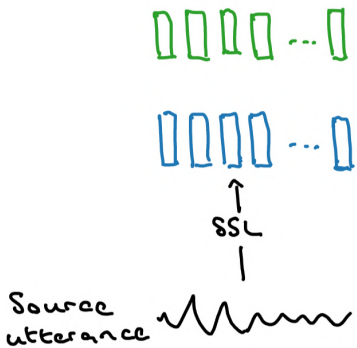
Output: [Play](#)

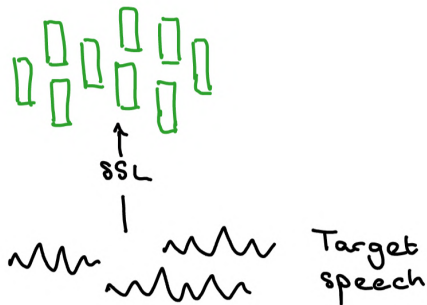
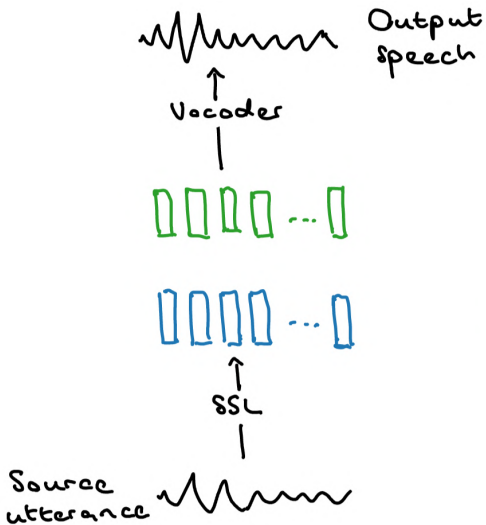




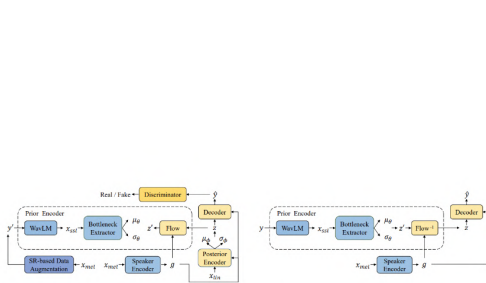








Existing voice conversion systems



FreeVC [2022]

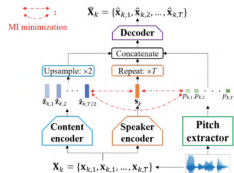
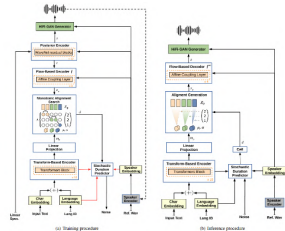


Figure 1: Diagram of the proposed VQMIVC system.

VQMIVC [2021]




YourTTS [2023]

Voice conversion results


Model	WER ↓	EER ↑	MOS ↑	SIM ↑
Ground truth	5.96	-	4.24	3.19
VQMIVC (Wang et al. 2021)	59.46	2.22	2.70	2.09
YourTTS (Casanova et al. 2022)	11.93	25.32	3.53	2.57
FreeVC (Li et al. 2022)	7.61	8.97	4.07	2.38
kNN-VC	7.36	37.15	4.03	2.91

Fun conversions

Cross-lingual conversion:

Source: 


Reference: 

Output: 

Whispered music conversion:

Source: 


Reference: 

Output: 

Human-to-animal conversion:





Source: 

Reference: 

Output: 

Applications of kNN-VC

- Stuttered reference speech (Baas and Kamper 2023):

Source:  Reference:  Output:  Baseline:  (TTS)

- Cross-lingual child voice conversion (Jacobs et al. 2025):

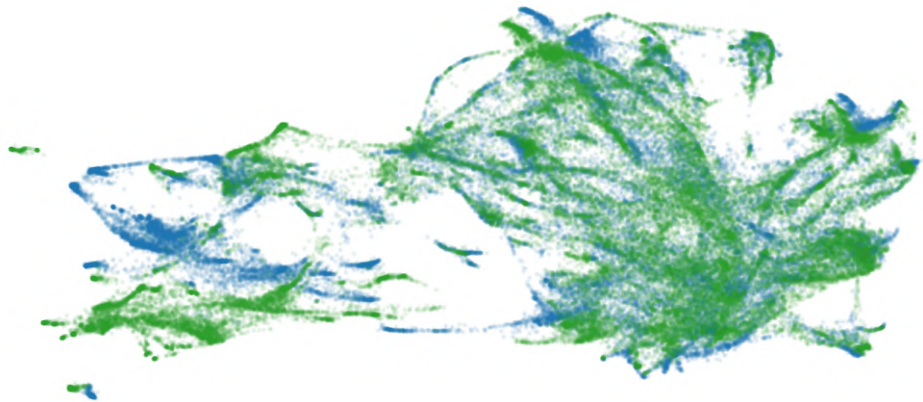
Source:  Reference:  Output:  (Afrikaans)

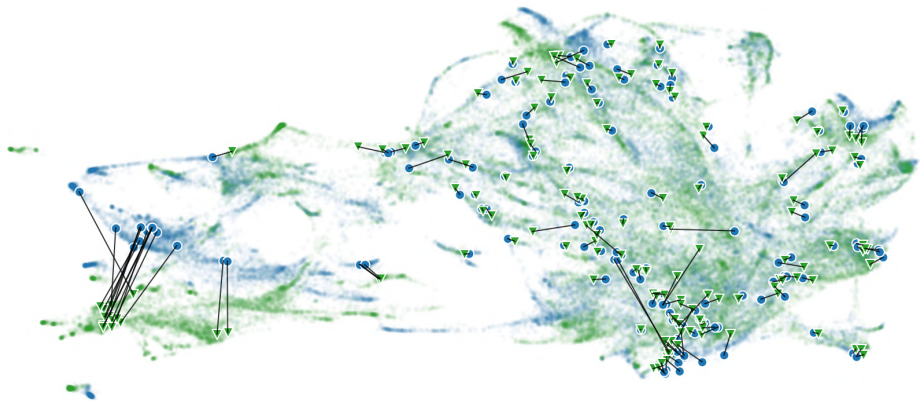
Source:  Reference:  Output:  (isiXhosa)

- Singing voice conversion (Shao et al. 2025)
- Dysarthric to healthy speech (El Hajal et al. 2025)
- Anonymisation (Franzreb et al. 2025)

What does this tell us about SSL representations?

- Phonetic content is matched through cosine distance
- But speaker characteristics are also still strongly captured





LinearVC: Voice conversion with just linear regression



Benjamin
van Niekerc



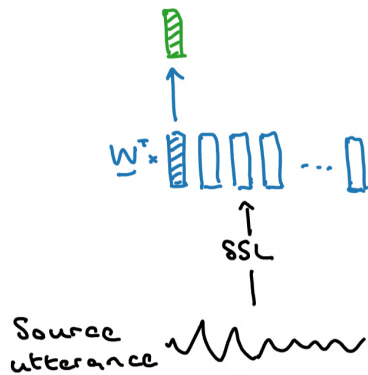
Julian
Zaïdi

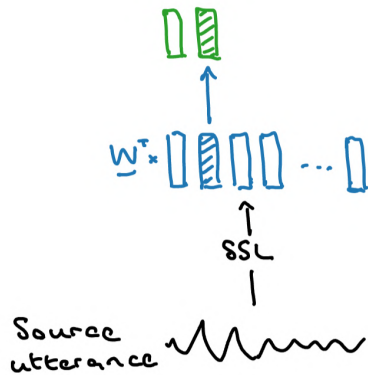


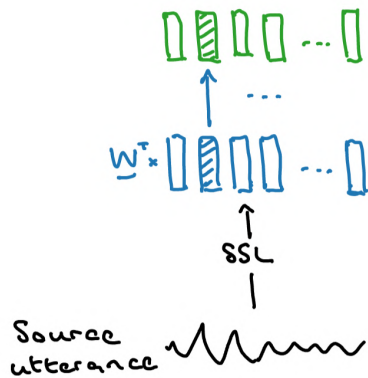
Marc-André
Carbonneau

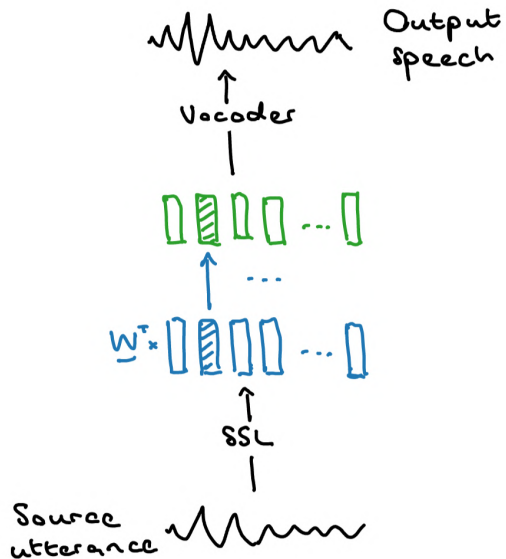
H. Kamper, B. van Niekerc, J. Zaïdi, and M-A. Carbonneau,
"LinearVC: Linear transformations of self-supervised features through the lens of voice conversion," in *Interspeech*, 2025.



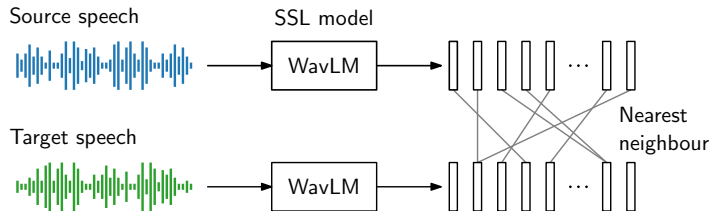




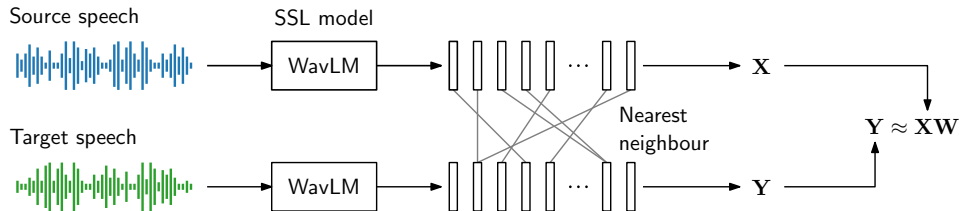




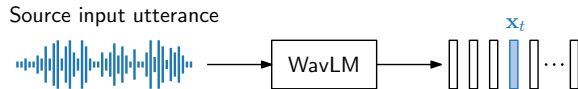
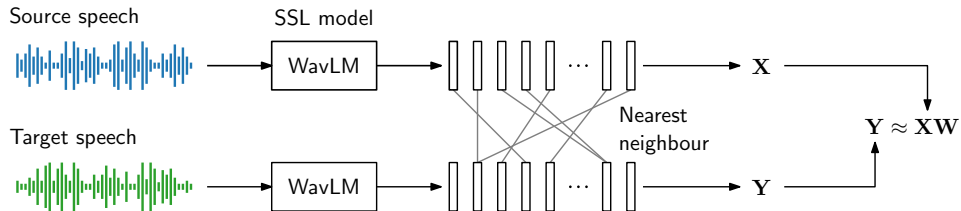
LinearVC



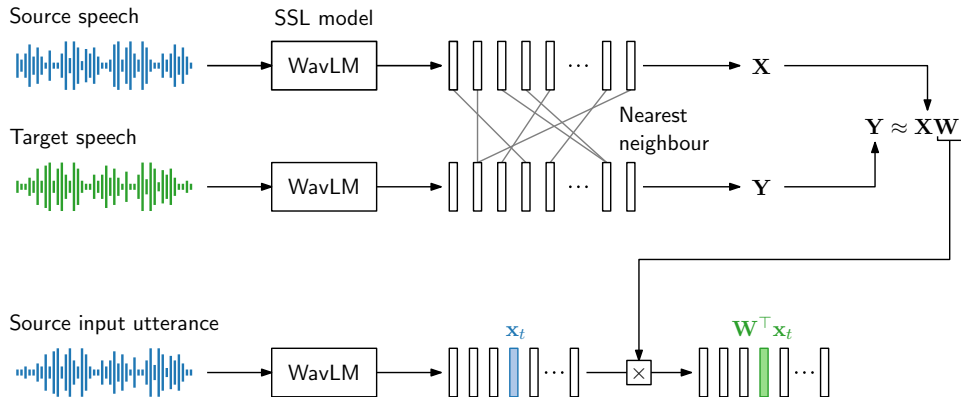
LinearVC



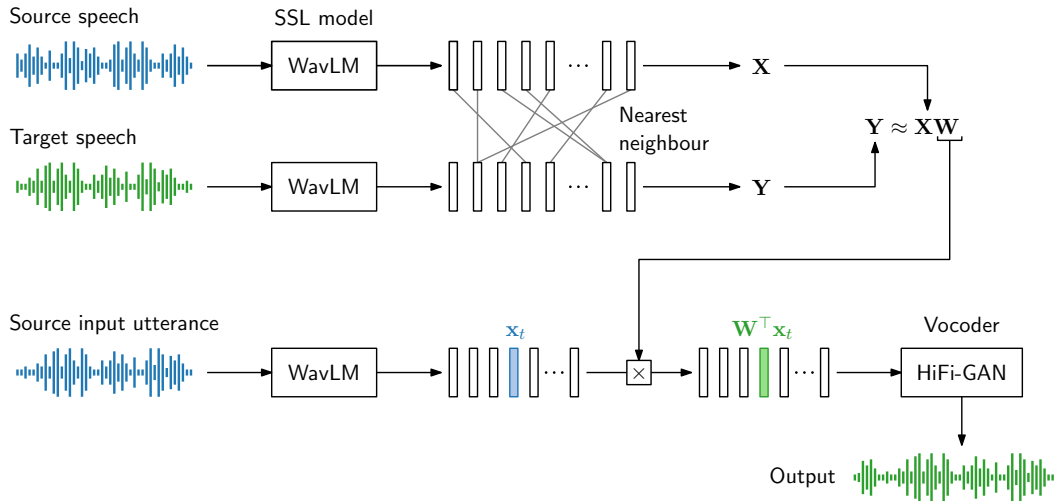
LinearVC



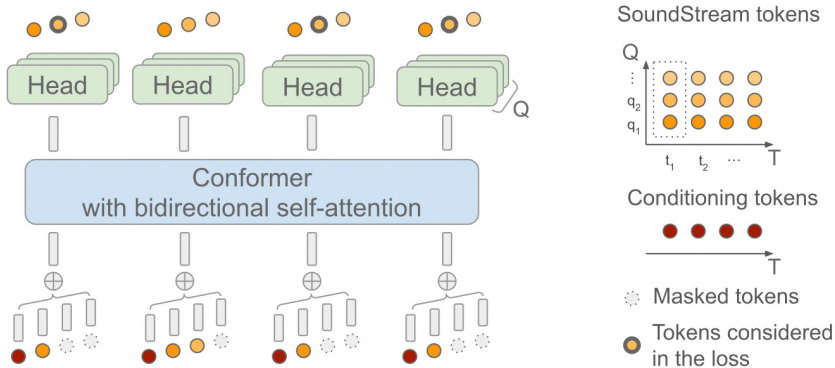
LinearVC



LinearVC



Codec-based spoken language model for voice conversion



SoundStorm (Borsos et al. 2023)

Voice conversion results


Model	WER ↓	EER ↑	Naturalness ↑	Similarity ↑
Ground truth	4.3	-	-	-
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

Samples

kNN-VC:

Source: 


Reference: 

Output: 

LinearVC:

Source: 


Reference: 

Output: 

SoundStorm:

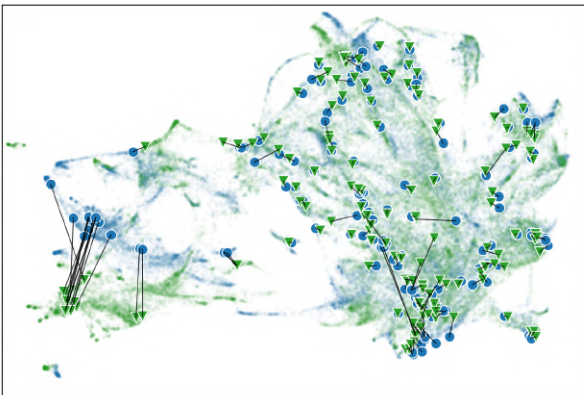
Source: 

Reference: 

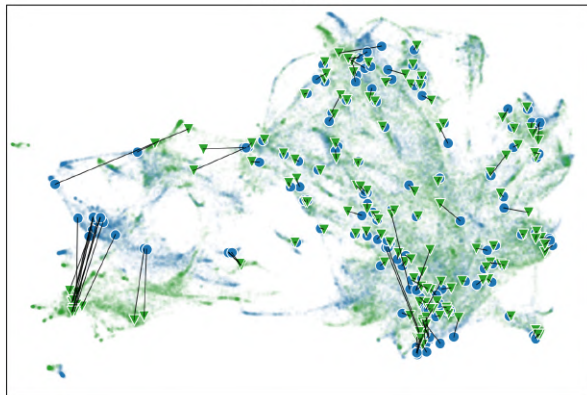
Output: 

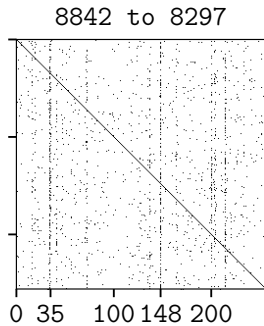
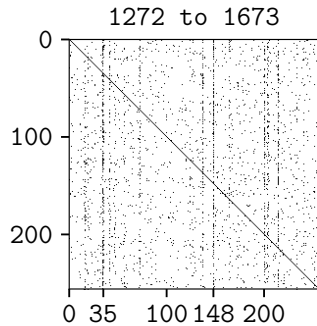
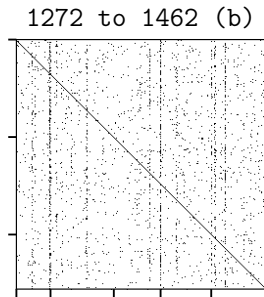
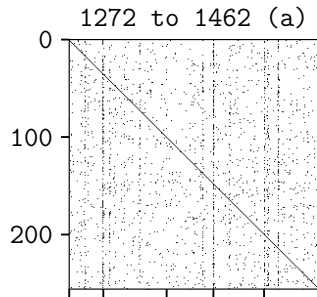
This should freak you out.
Let's try to make sense of this.

kNN-VC



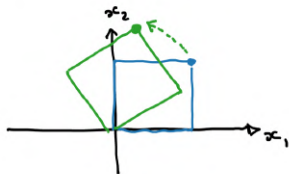
LinearVC





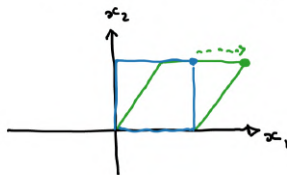
Rotation:

$$\underline{W} = \begin{bmatrix} \cos \theta & -\sin \theta & 0 \\ \sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{bmatrix}$$



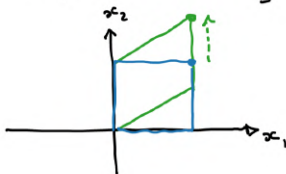
Shear in x_1 :

$$\underline{W} = \begin{bmatrix} 1 & \tan \theta & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$



Shear in x_2 :


$$\underline{W} = \begin{bmatrix} 1 & 0 & 0 \\ \tan \theta & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$



Voice conversion with just a bias vector

Source: 

Reference: 

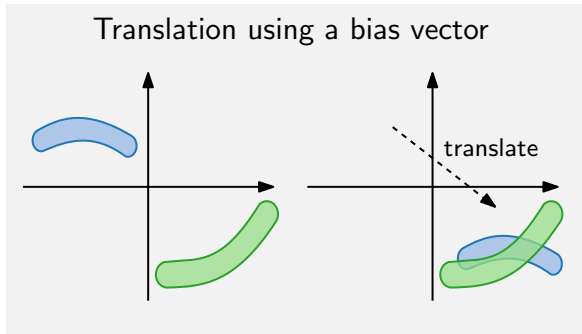
Output: 

Voice conversion with just a bias vector

Source: [Play](#)

Reference: [Play](#)

Output: [Play](#)

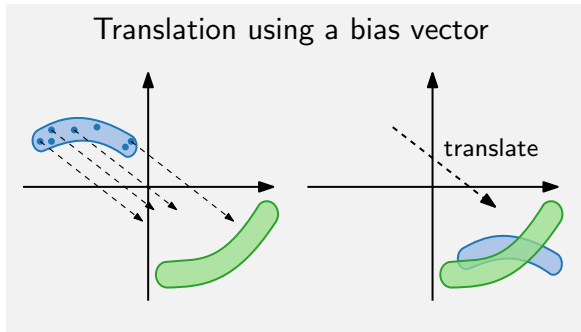


Voice conversion with just a bias vector

Source: [Play](#)

Reference: [Play](#)

Output: [Play](#)



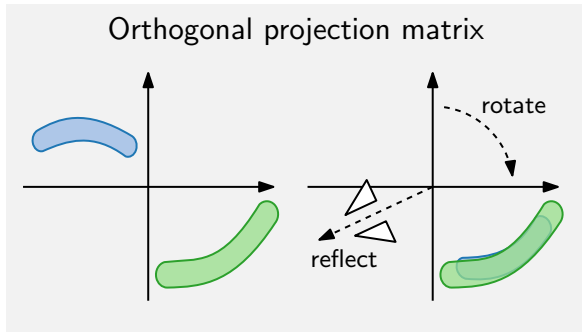
Voice conversion with just rotation

Source: [Play](#)

Reference: [Play](#)

Just orthogonal: [Play](#)

Full LinearVC: [Play](#)



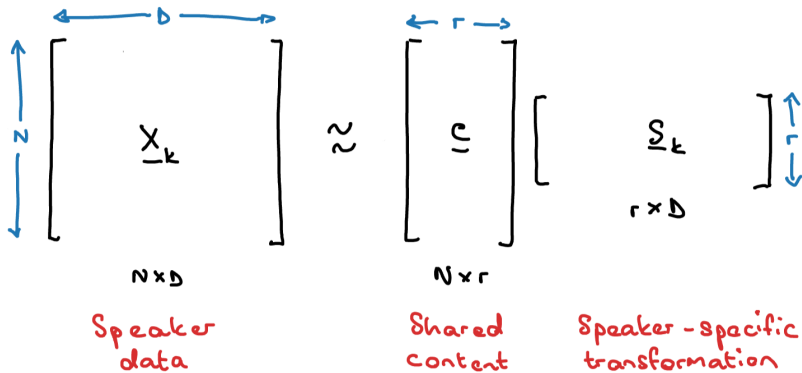
Let's try to visualise this

LinearVC with content factorisation

$$\begin{aligned} \min_{\mathbf{C}, \mathbf{S}_k} \quad & \sum_{k=1}^K \|\mathbf{X}_k - \mathbf{C}\mathbf{S}_k\|_F^2 \\ \text{subject to} \quad & \text{rank}(\mathbf{C}\mathbf{S}_k) \leq r \end{aligned}$$

LinearVC with content factorisation

$$\begin{aligned} \min_{\mathbf{C}, \mathbf{S}_k} \quad & \sum_{k=1}^K \|\mathbf{X}_k - \mathbf{C}\mathbf{S}_k\|_F^2 \\ \text{subject to} \quad & \text{rank}(\mathbf{C}\mathbf{S}_k) \leq r \end{aligned}$$



LinearVC with content factorisation

Source: [Play](#)

Reference: [Play](#)

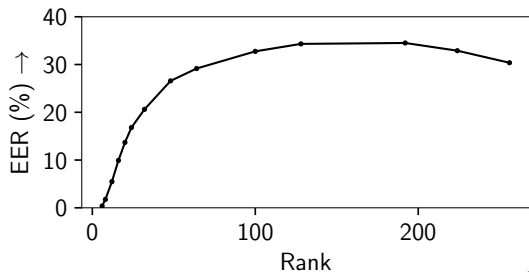
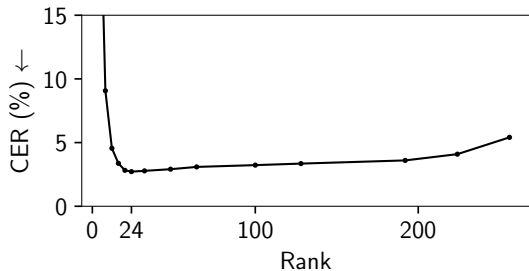
LinearVC cont. fact. $r = 6$: [Play](#)

LinearVC cont. fact. $r = 16$: [Play](#)

LinearVC cont. fact. $r = 100$: [Play](#)

LinearVC cont. fact. $r = 256$: [Play](#)

Standard LinearVC: [Play](#)

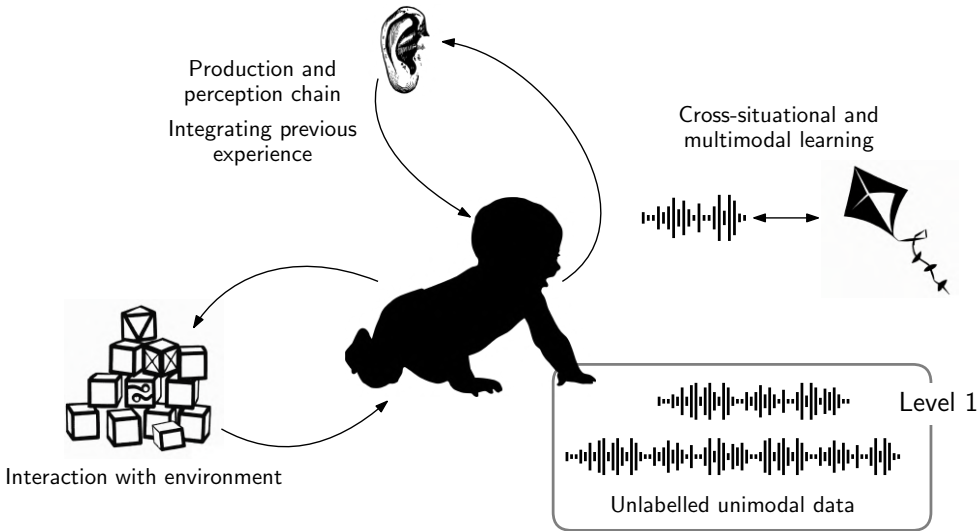


Voice conversion results

Model	WER ↓	EER ↑	Naturalness ↑	Similarity ↑
Ground truth	4.3	-	-	-
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
LinearVC content factorisation	4.7	35.2	62.3±3.7	64.2±3.1

Conclusion

- Simple approaches are very useful: Can do practical voice conversion!
- Probing experiments have their place, but ...
- Synthesis provides a unique perspective on SSL geometry
- Allows us to quickly see (hear, actually) salient effects
- Future work:
 - Formalise cartoon interpretations
 - Use content space in downstream applications



<https://bshall.github.io/knn-vc/>

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