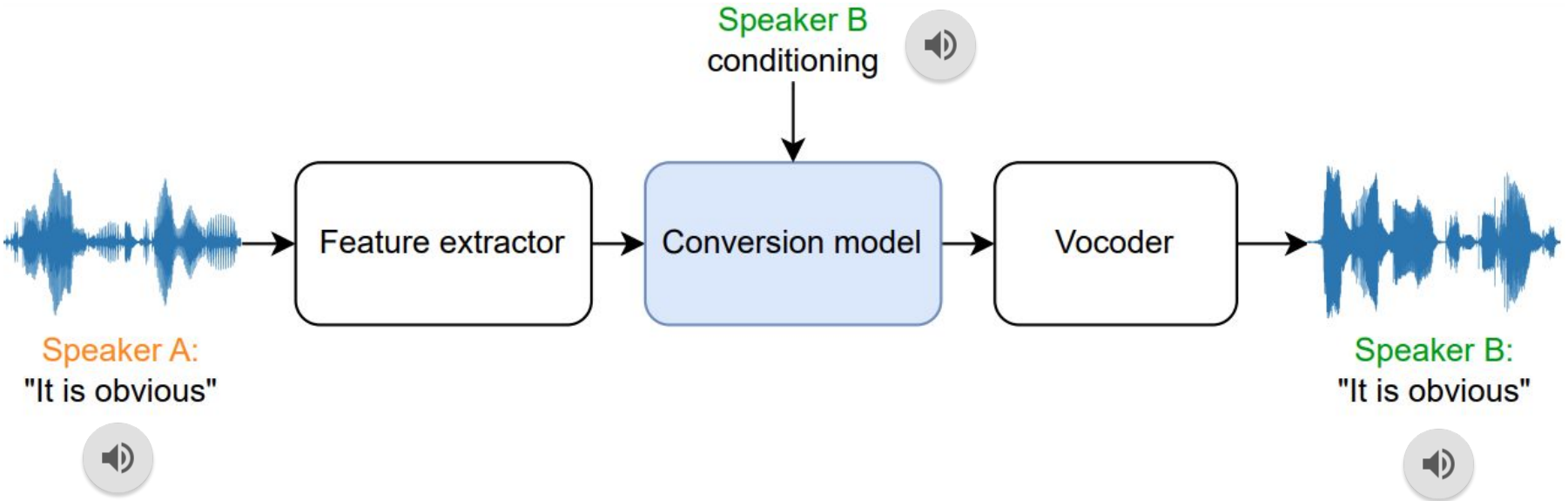
Voice conversion with just* nearest neighbors

Matthew Baas, Benjamin van Niekerk, Herman Kamper

July 2023

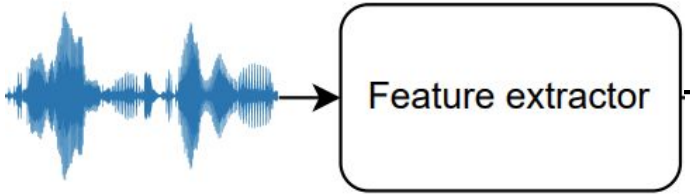
What is voice conversion?

Voice conversion: transforming source speech into a target voice, while keeping the words unchanged.

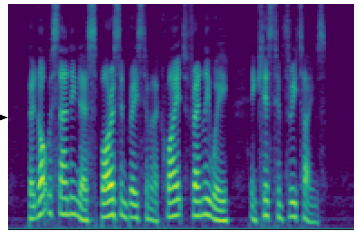


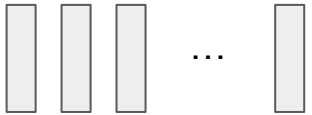
Feature extraction

Convert speech into a more disentangled form



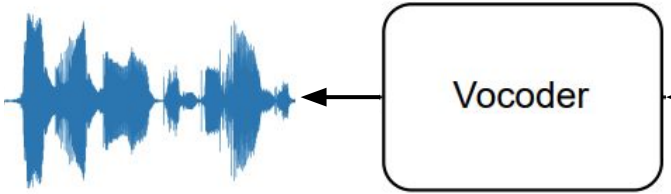
Mel-spectrogram





SSL model speech features

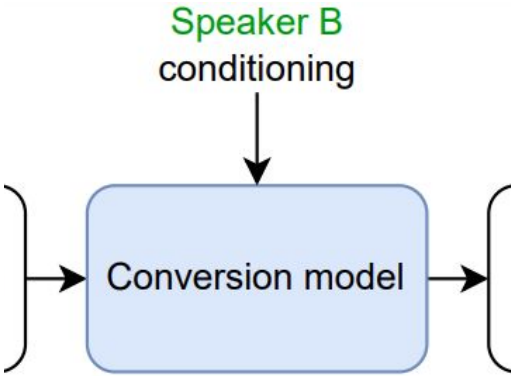
Convert the disentangled representation back to waveform

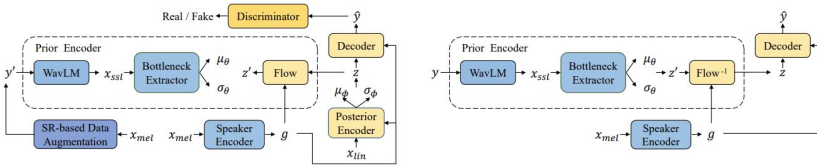


Key insight: linear distances here correspond to differences in speech sounds (phones)

Conversion model

- Recent systems can work in any-to-any settings
- But, they are increasingly complex and hard to build upon
- The conversion model often has special techniques to disentangle speaker from content





FreeVC [2022]

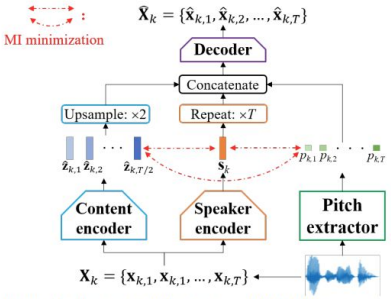
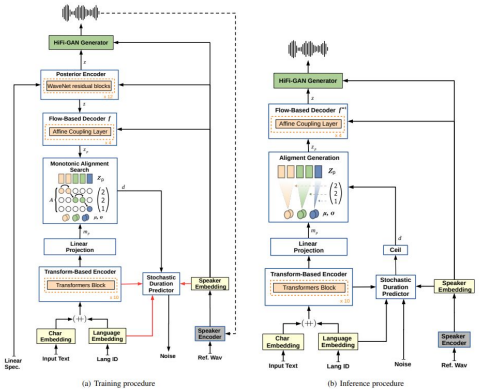


Figure 1: *Diagram of the proposed VQMIVC system.*

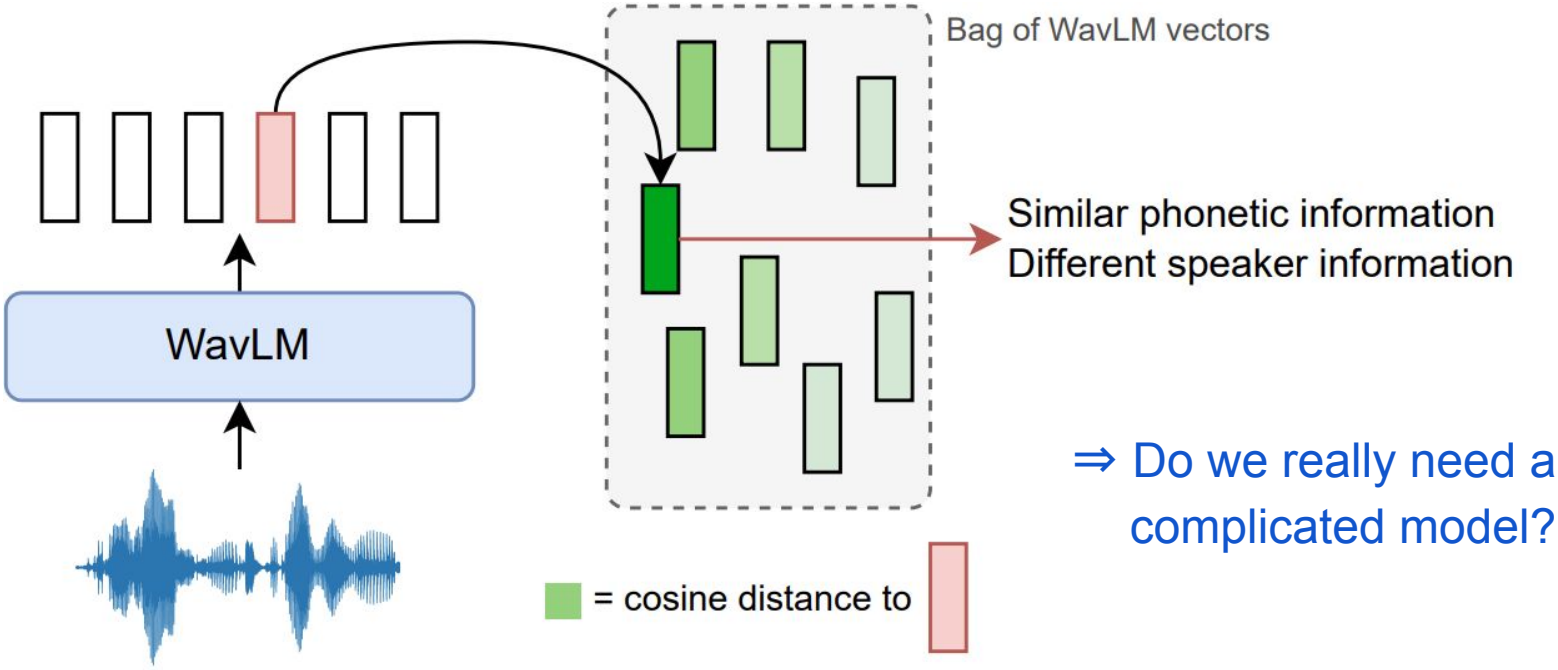
VQMIVC [2021]



YourTTS [2023]

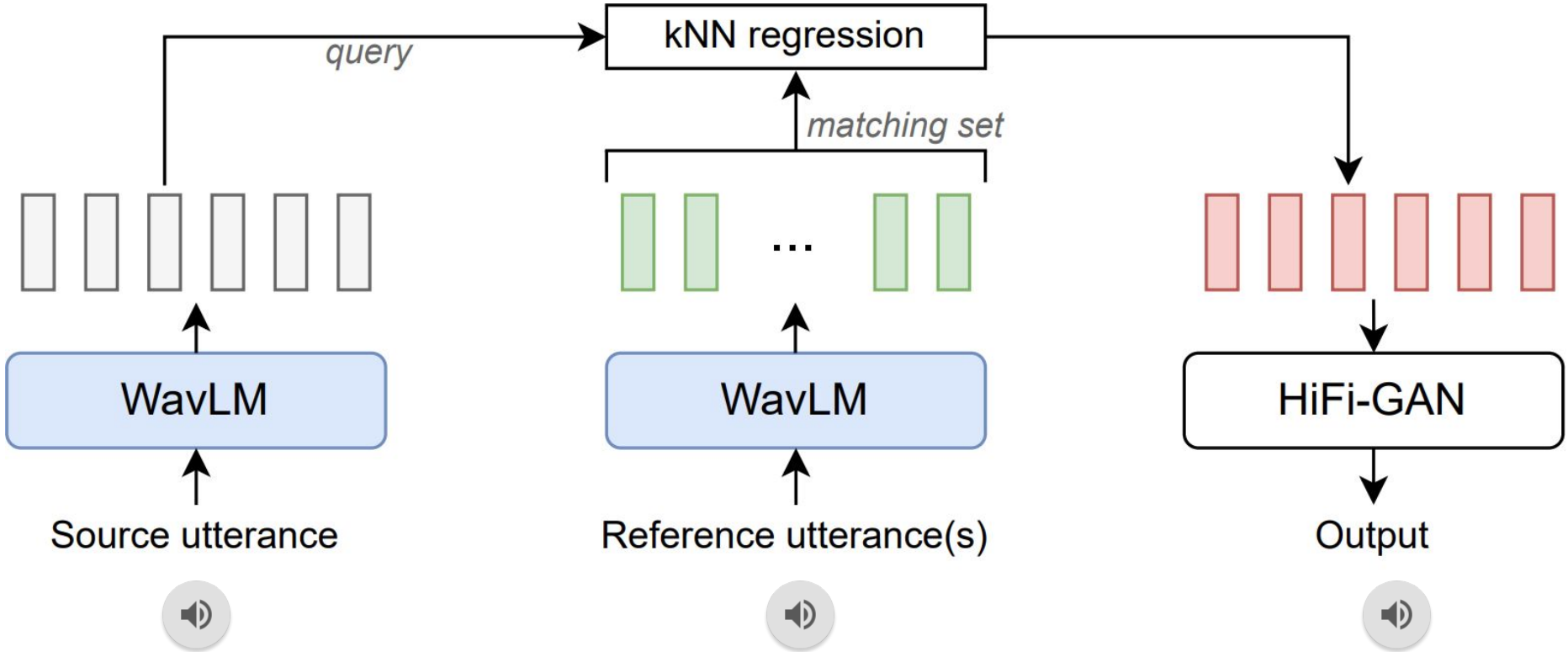
Key idea

WavLM features *linearly encode* what sound is being said!



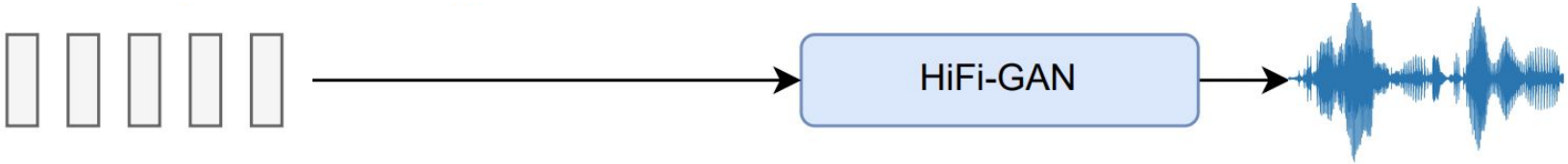
⇒ Do we really need a complicated model?

Keeping it simple: kNN-VC

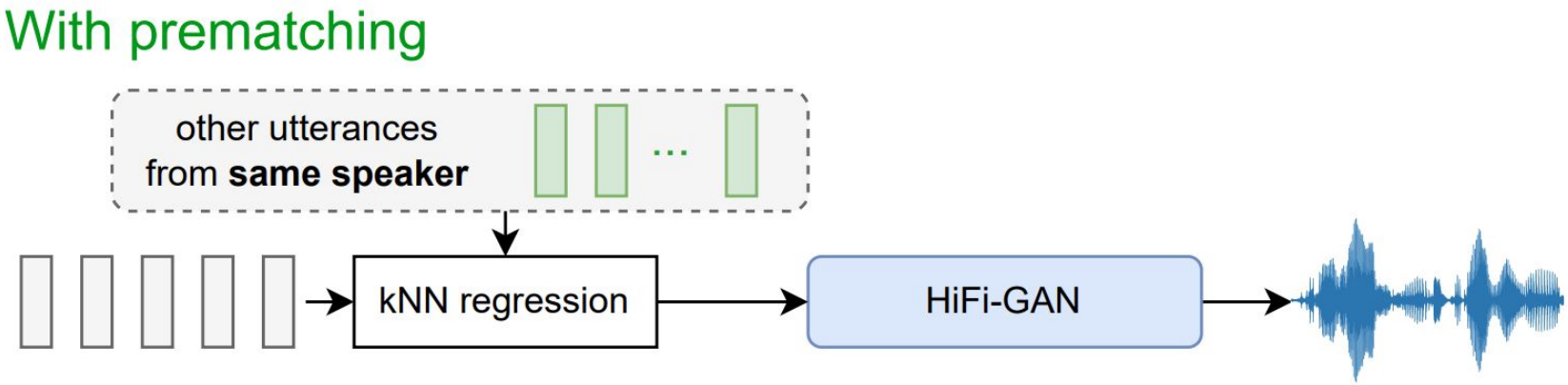


Prematching

Problem: vocoding kNN features computed leads to distorted temporal information.



Solution: train HiFi-GAN on *prematched features* to better follow inference conditions.



Evaluations

Table: Results with $k = 4$ measuring the intelligibility (WER), naturalness (MOS), and speaker similarity (EER, SIM) of the converted speech.

Model	WER ↓	EER ↑	MOS ↑	SIM ↑
Testset Topline	5.96	—	4.24	3.19
VQMIVC	59.46	2.22	2.70	2.09
YourTTS	11.93	25.32	3.53	2.57
FreeVC	7.61	8.97	**4.07**	2.38
kNN-VC	**7.36**	**37.15**	**4.03**	**2.91**

$\Longrightarrow$ similar naturalness/intelligibility, but superior target speaker similarity when compared to existing methods, while being much simpler.

Fun results

kNN is non-parametric ⇒ source and reference can be any audio clip!

Cross-lingual conversion

Source  Reference  Output 

Whispered music conversion

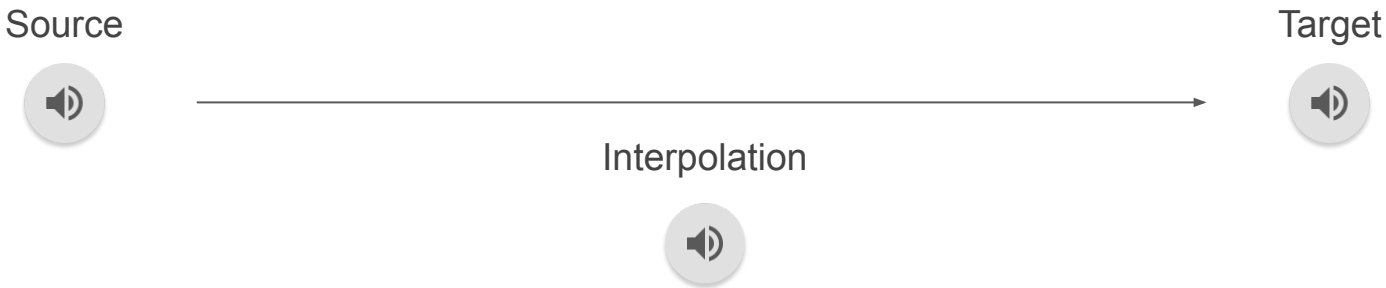
Source  Reference  Output 

Human-to-animal conversion

Source  Reference  Output 

Fun results

We can even interpolate between voices
(generating new unseen voices in between)



Interpolation demo by Everett Cheng
(https://eccheng.github.io/ml/audio/vc/2023/07/04/knn-vc-morph.html),
check out the link for more samples!

Conclusion

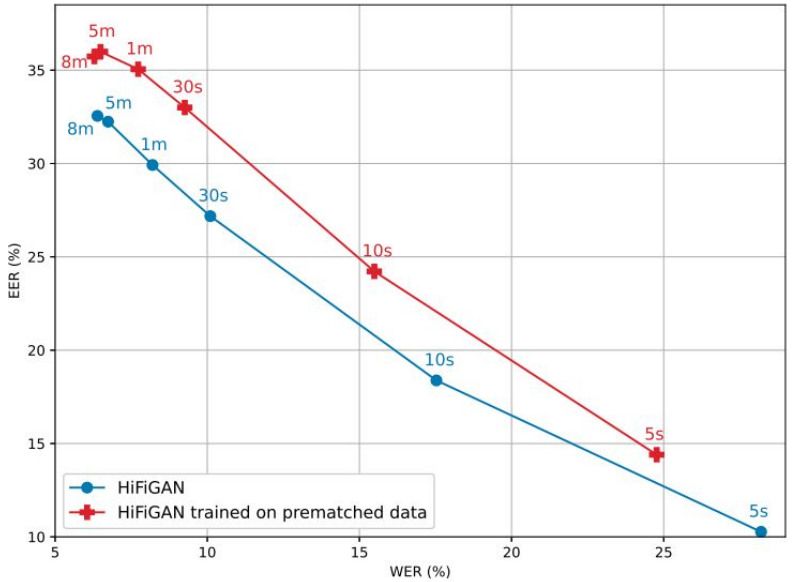
- We do not need complex methods for convincing voice conversion 😊
- Just kNN on WavLM features achieves compelling results
- kNN-VC is easy to use and evaluate – fostering better comparisons
- **Future**: investigation kNN-VC in more diverse domains

Full paper, more samples, code:

Thank You!

Results (detailed)

Figure: Performance tradeoff with varying amounts of reference data.



$\Longrightarrow$ Prematching helps: regardless of amount of reference data, prematched vocoder sounds more natural and closer to target speaker.

$\Longrightarrow$ Less reference data hurts but is still intelligible with as little as $\sim$10 sec of audio; more reference data helps, plateauing at $\sim$5 min.