Explicitly disentangling speaker from speech

Implicit disentanglement all aspects of speech

This requires that we explicitly design how to discard speaker information…

What if we can *learn* separation between speaker information and the rest of speech ⇒ implicit disentanglement

Improving implicit speech disentanglement with GANs

Matthew Baas
Supervisor: Herman Kamper
August 2022

MediaLab, Electrical & Electronic Engineering Department, Stellenbosch University

Disentanglement

Explicit disentanglement

Speech synthesis model designer after much studying…

Speaker identity is different from the rest of speech by **X**

Design cool new model using idea **X** to disentangle and modify **speaker identity**

Limitation: human **expert designed demarcations** of different characteristics of speech

Can we try learn them instead?

Implicit disentanglement

Good disentanglement that is still easy to control?

Design cool new model that *should discover a controllable representation that disentangles* **speaker identity.**

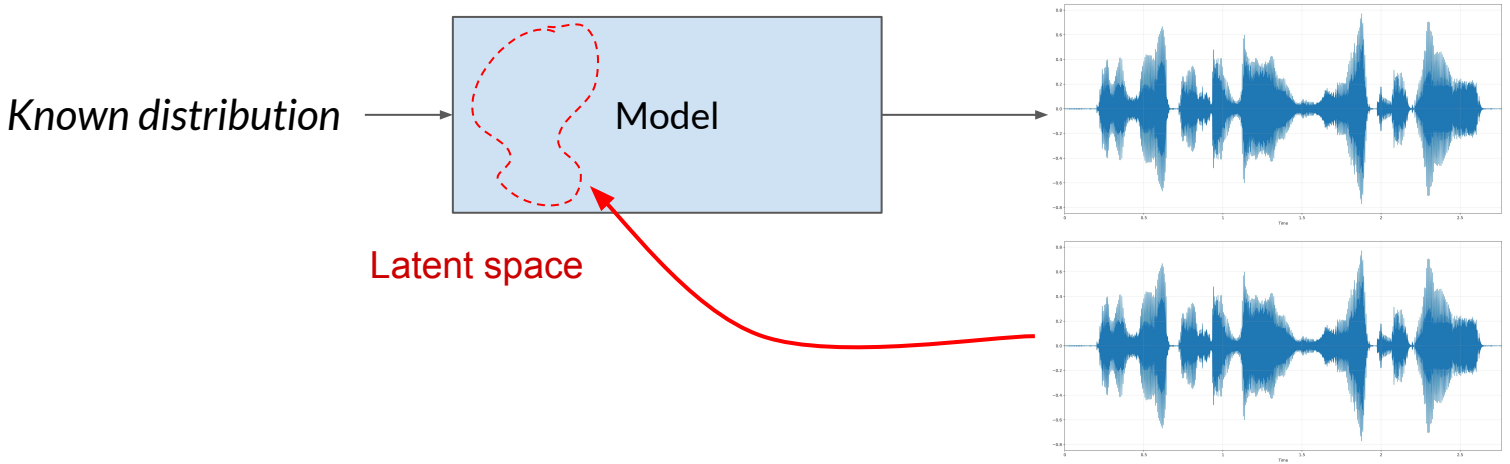
Task

1. Minimal assumptions about speech
2. Tractable training
3. Practical for inference

Unconditional speech synthesis task

Common setup

We also want

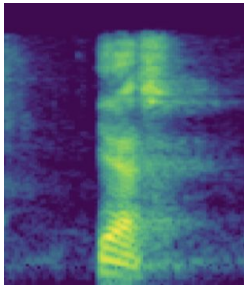


Related efforts

Similar effort in **unconditional image synthesis**:

- Modelling a latent space from noise
  - StyleGAN
  - BigGAN
  - PPOGAN
- Encourage disentanglement
  - StyleGAN2
  - StyleGAN3
- Latent space metrics
  - FID
  - Path length
  - Linear separability
- Inverse mapping methods
  - Noisy projection
  - Pivot tuning

Overview of our approach

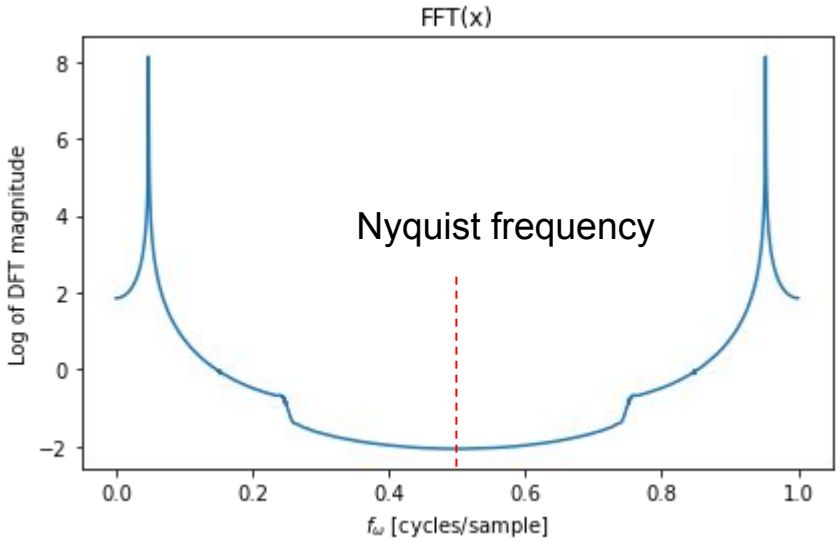


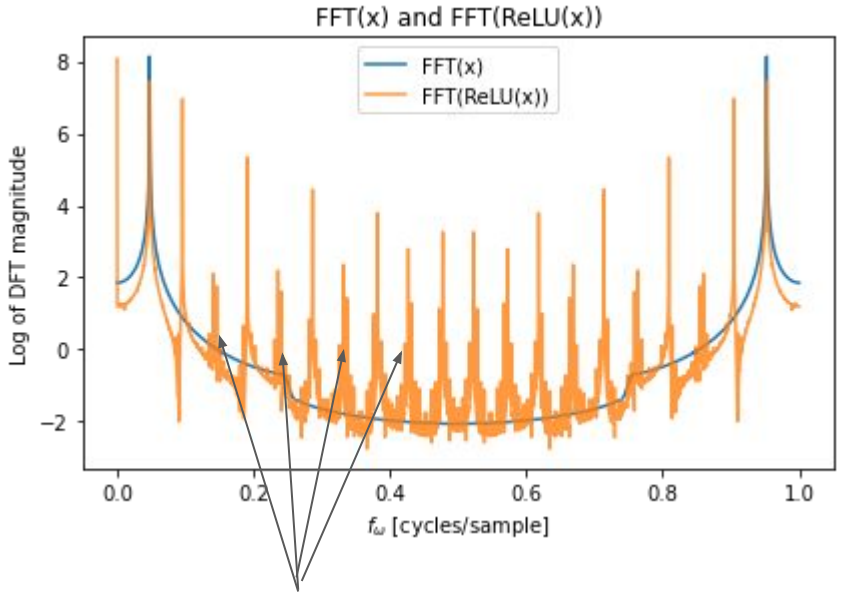
- Google Speech Commands dataset
- Great synthesis quality and diversity
- **Many tricks!**
- Apply to unseen tasks zero-shot

- Voice conversion
- Speech editing
- Speech enhancement
- Speaker verification

Anti-aliasing in GANs

x is single sinusoid activation.





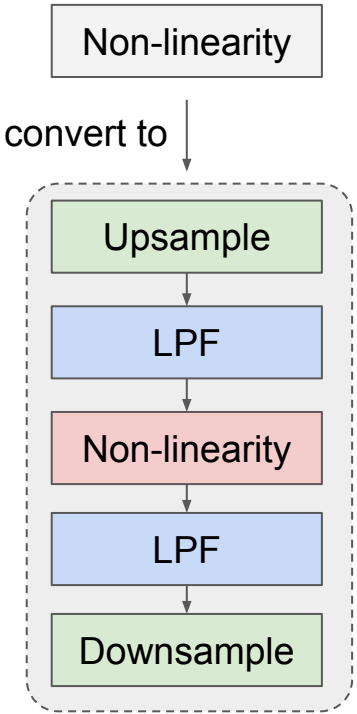
Aliasing! ⇒ need anti-aliasing filters!

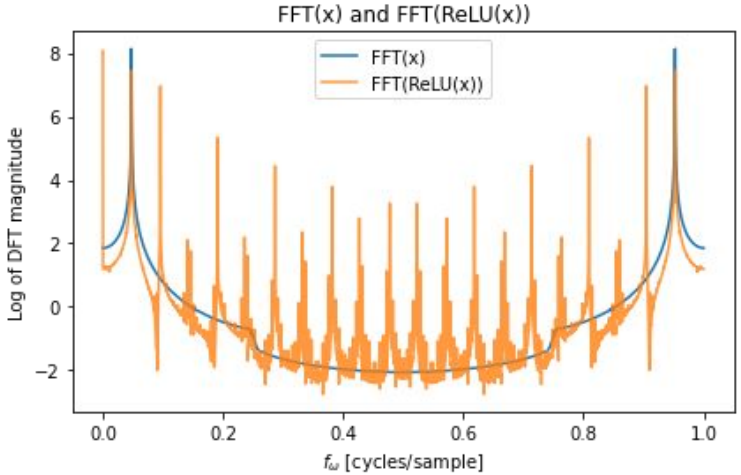
Aliasing causes:

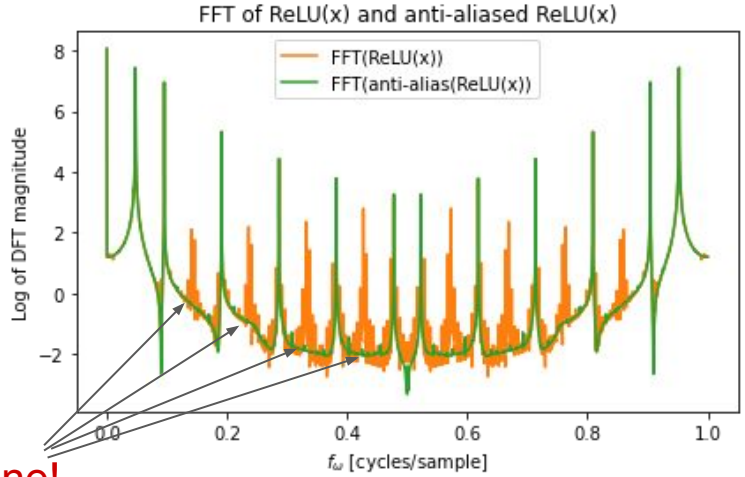
- GAN to hide artifacts in aliasing
- Poorer reconstruction quality when finally vocoding to audio

Anti-aliasing in GANs

StyleGAN3 proposed:

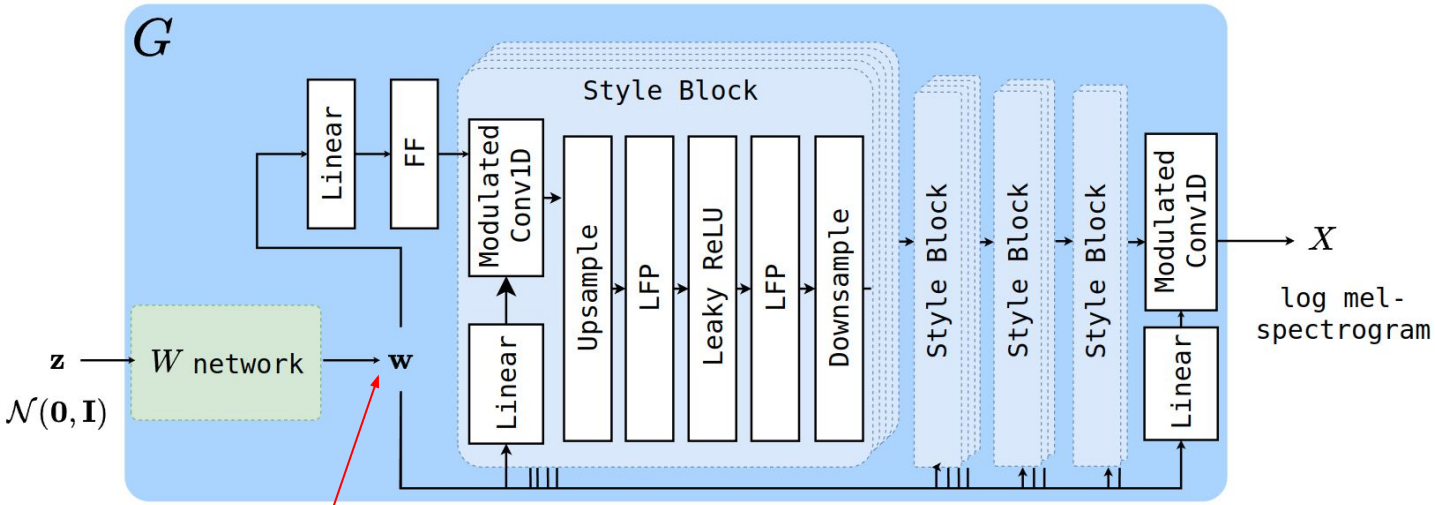






Aliasing gone!

Model & GAN tricks

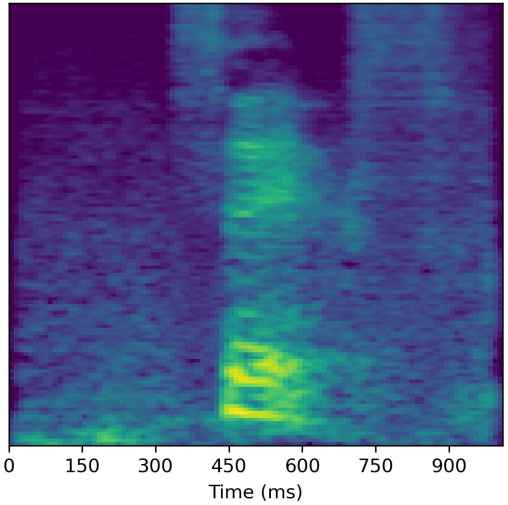


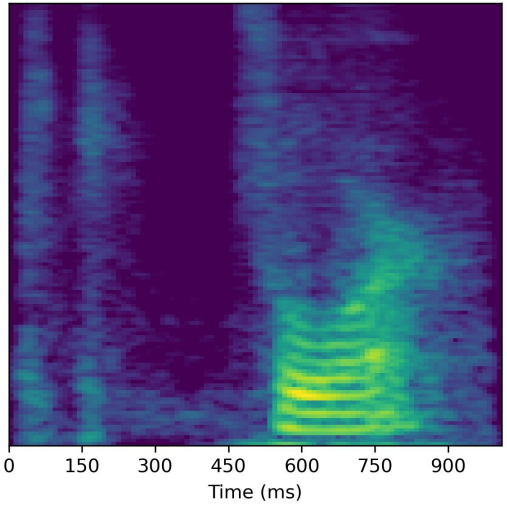
Disentangled latent space

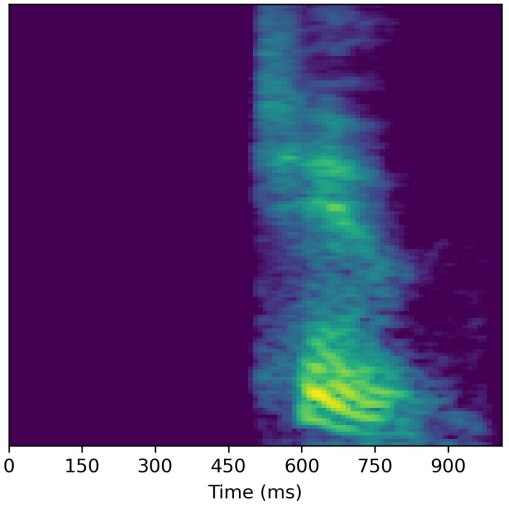
Tricks:

- R1 regularization
- EMA generator weights
- Adaptive discriminator augmentation
- Equalized learning rates
- Lower W network learning rate
- Adaptive discriminator updates

Unconditional samples

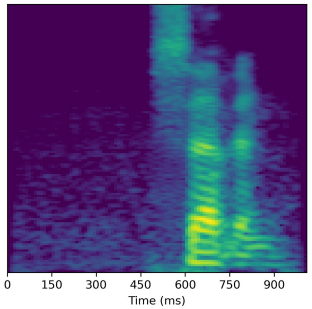






Voice conversion

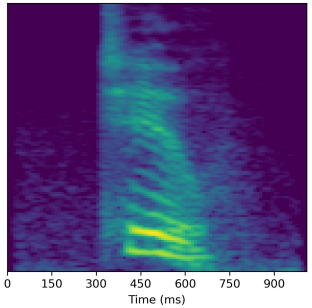
Source utterance



projection

$\mathbf{w}_1$

Utterance with target speaker

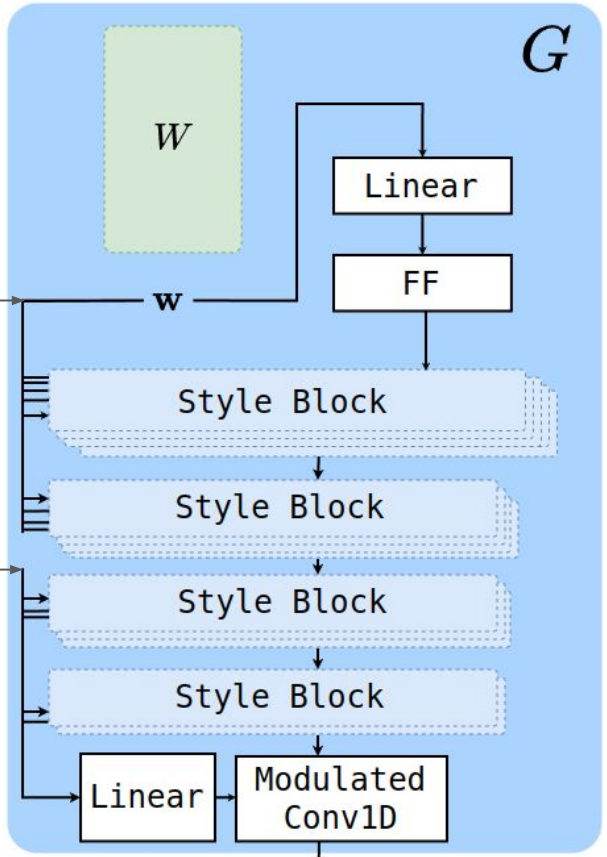


projection

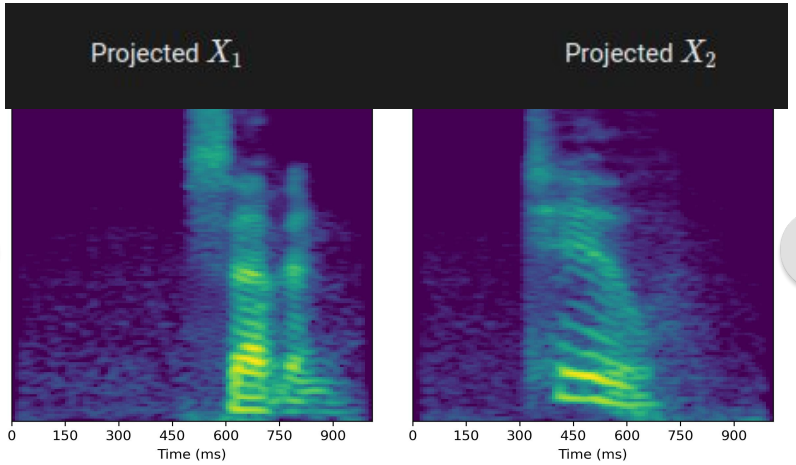
$\mathbf{w}_2$

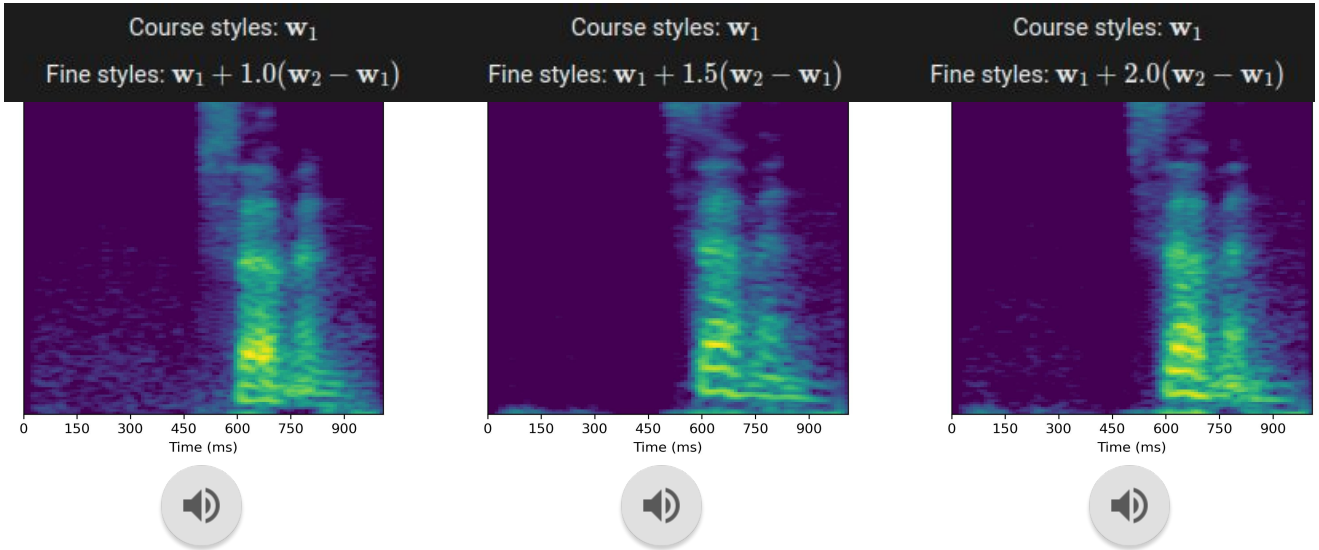
Course styles

Fine styles



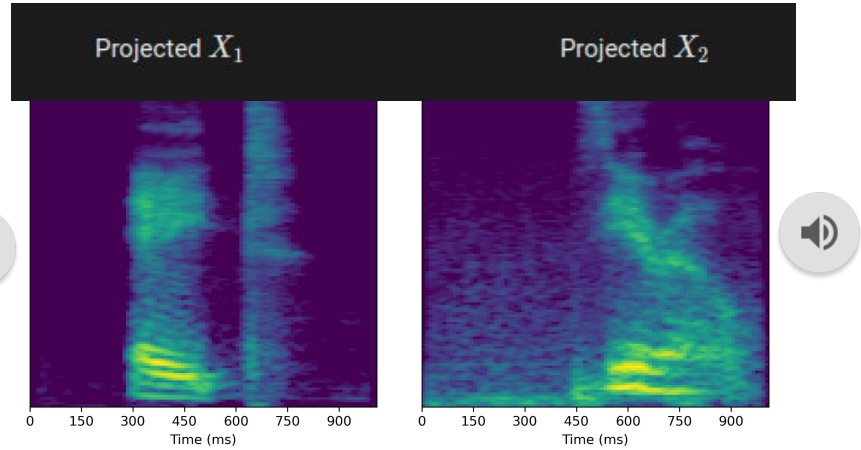
Voice conversion

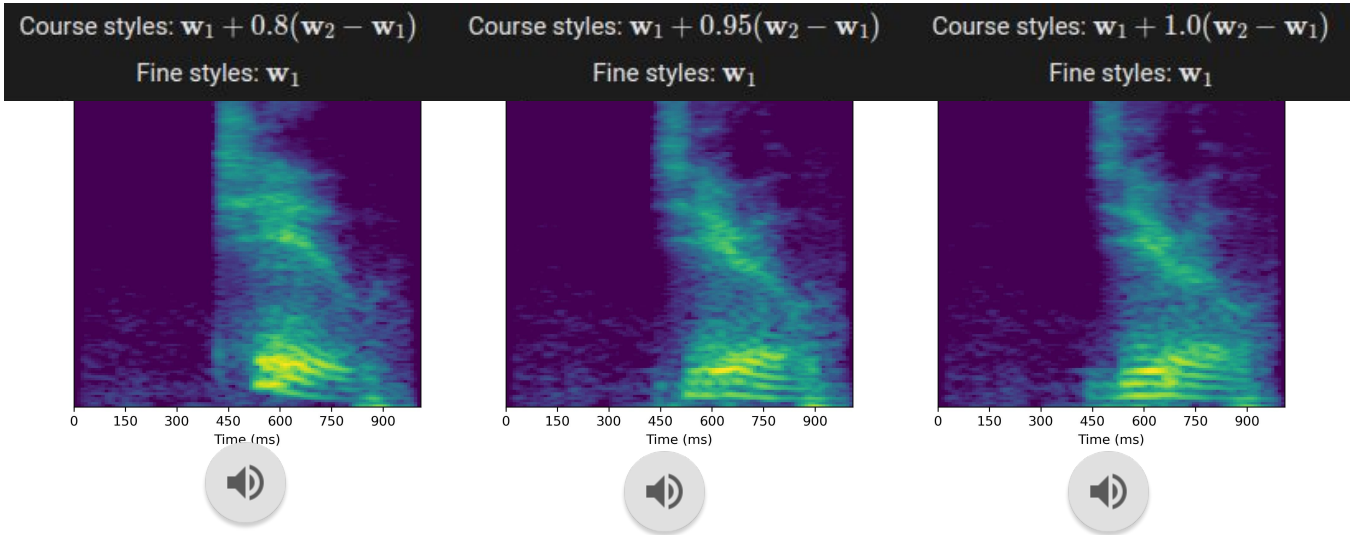




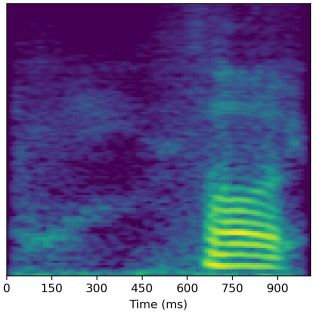
Digit conversion

Just do the opposite!

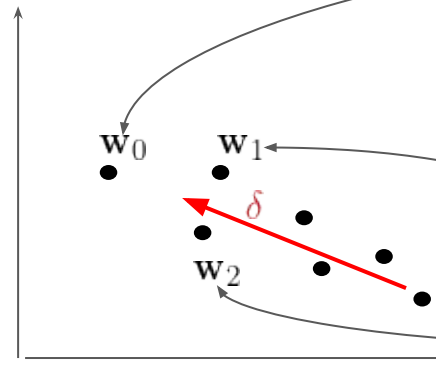




Speech enhancement



projection



\+ noise

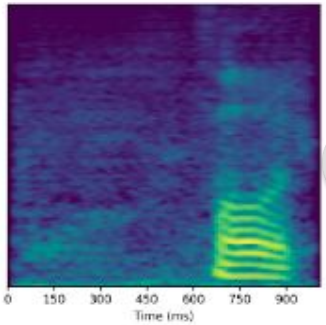
\+ noise

…

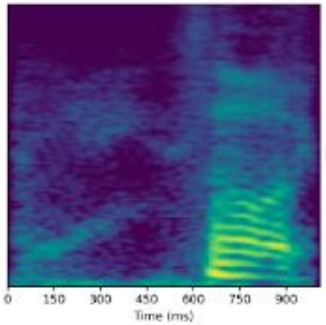
We now know the direction of decreasing noise in W-space

δ

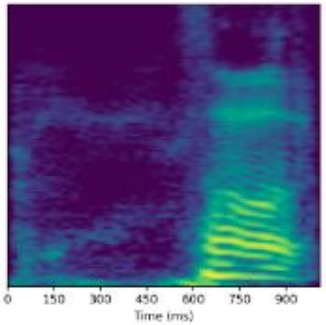
All styles: $\mathbf{w}_0 - 3\delta$



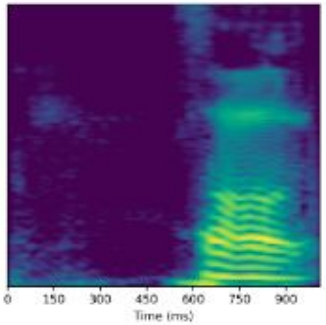
All styles: $\mathbf{w}_0 + 3\delta$



All styles: $\mathbf{w}_0 + 6\delta$



All styles: $\mathbf{w}_0 + 9.0\delta$



Limitations & future directions

1. Other synthesis tasks phrased as latent manipulations / perceptual tests?
2. Does not work well on out-of-domain data
3. Struggles to scale
4. Projection method not sophisticated

Thank you