# Interpreting intermediate convolutional layers of generative CNNs trained on waveforms


Gašper Beguš & Alan Zhou

{begus,azhou314}@berkeley.edu


Speech & Computation

Berkeley
UNIVERSITY OF CALIFORNIA

ICASSP 2023
4 - 10 JUNE, RHODES ISLAND, GREECE

TASLP 6701 (MLSP-L3.4), June 6, 2023

# Introduction

- Interpretability one of the main frontiers in AI research
- Most studies focus on vision (Zeiler and Fergus, 2014)
- Spoken language is an ideal testing ground
  - Speech is more interpretable than vision
  - Humans discretize continuous physical property (speech sounds) into representations with various degrees of complexity
  - Generation data in speech is easier to access

# Introduction

- Interpretability one of the main frontiers in AI research
- Most studies focus on vision (Zeiler and Fergus, 2014)
- Spoken language is an ideal testing ground
  - Speech is more interpretable than vision
  - Humans discretize continuous physical property (speech sounds) into representations with various degrees of complexity
  - Generation data in speech is easier to access

# Introduction

- Interpretability one of the main frontiers in AI research
- Most studies focus on vision (Zeiler and Fergus, 2014)
- Spoken language is an ideal testing ground
  - Speech is more interpretable than vision
  - Humans discretize continuous physical property (speech sounds) into representations with various degrees of complexity
  - Generation data in speech is easier to access

# Introduction

- Interpretability one of the main frontiers in AI research
- Most studies focus on vision (Zeiler and Fergus, 2014)
- Spoken language is an ideal testing ground
  - Speech is more interpretable than vision
  - Humans discretize continuous physical property (speech sounds) into representations with various degrees of complexity
  - Generation data in speech is easier to access

# Introduction

- Interpretability one of the main frontiers in AI research
- Most studies focus on vision (Zeiler and Fergus, 2014)
- Spoken language is an ideal testing ground
  - Speech is more interpretable than vision
  - Humans discretize continuous physical property (speech sounds) into representations with various degrees of complexity
  - Generation data in speech is easier to access

# Introduction

## Proposal

1. **A technique to interpret and visualize intermediate layers in generative CNNs** (trained on raw speech data in an unsupervised manner)
2. Any acoustic property can be tested (where it is encoded)
   - F0, intensity, duration, formants, and other acoustic properties
   - test where and how CNNs encode various types of information
3. Combine this technique with linear interpolation of a model's latent space to show a **causal relationship** between individual variables in the latent space and activations in a model's intermediate convolutional layers

# Introduction

## Proposal

1. **A technique to interpret and visualize intermediate layers in generative CNNs** (trained on raw speech data in an unsupervised manner)
2. Any acoustic property can be tested (where it is encoded)
   - F0, intensity, duration, formants, and other acoustic properties
   - test where and how CNNs encode various types of information
3. Combine this technique with linear interpolation of a model's latent space to show a **causal relationship** between individual variables in the latent space and activations in a model's intermediate convolutional layers

# Introduction

## Proposal

1. **A technique to interpret and visualize intermediate layers in generative CNNs** (trained on raw speech data in an unsupervised manner)
2. Any acoustic property can be tested (where it is encoded)
   - F0, intensity, duration, formants, and other acoustic properties
   - test where and how CNNs encode various types of information
3. Combine this technique with linear interpolation of a model's latent space to show a **causal relationship** between individual variables in the latent space and activations in a model's intermediate convolutional layers

# Advantages of the new approach

- Manipulating and interpolating individual latent variables well beyond training range while visualizing intermediate layers
- Observing the causal relationship between individual variables in the latent space and linguistically meaningful units in intermediate layers
- Testing which acoustic properties are encoded at which layer via correlations
- Testing not only encoding of acoustic properties or words, but also of phonological processes and higher-level morphophonological processes such as reduplication
- Unsupervised generative models trained on raw speech

(Beguš and Zhou, 2022)

- Our proposal requires no further processing of the outputs

# Advantages of the new approach

- Manipulating and interpolating **individual latent variables well beyond training range** while visualizing intermediate layers
- Observing the causal relationship between individual variables in the latent space and linguistically meaningful units in intermediate layers
- Testing which acoustic properties are encoded at which layer via correlations
- Testing not only encoding of acoustic properties or words, but also of phonological processes and higher-level morphophonological processes such as reduplication
- Unsupervised generative models trained on raw speech

(Beguš and Zhou, 2022)

- Our proposal requires no further processing of the outputs

# Advantages of the new approach

- Manipulating and interpolating **individual latent variables well beyond training range** while visualizing intermediate layers
- Observing the causal relationship between individual variables in the latent space and linguistically meaningful units in intermediate layers
- Testing which acoustic properties are encoded at which layer via correlations
- Testing not only encoding of acoustic properties or words, but also of phonological processes and higher-level morphophonological processes such as reduplication
- Unsupervised generative models trained on raw speech

(Beguš and Zhou, 2022)

- Our proposal requires no further processing of the outputs

# Advantages of the new approach

- Manipulating and interpolating **individual latent variables well beyond training range** while visualizing intermediate layers
- Observing the **causal relationship** between individual variables in the latent space and linguistically meaningful units in intermediate layers
- Testing which acoustic properties are encoded at which layer via correlations
- Testing not only encoding of acoustic properties or words, but also of phonological processes and higher-level morphophonological processes such as reduplication
- Unsupervised generative models trained on raw speech

(Beguš and Zhou, 2022)

- Our proposal requires no further processing of the outputs

# Advantages of the new approach

- Manipulating and interpolating **individual latent variables well beyond training range** while visualizing intermediate layers
- Observing the **causal relationship** between individual variables in the latent space and linguistically meaningful units in intermediate layers
- Testing which acoustic properties are encoded at which layer via correlations
- Testing not only encoding of acoustic properties or words, but also of phonological processes and higher-level morphophonological processes such as reduplication
- Unsupervised generative models trained on raw speech

(Beguš and Zhou, 2022)

- Our proposal requires no further processing of the outputs

# Advantages of the new approach

- Manipulating and interpolating **individual latent variables well beyond training range** while visualizing intermediate layers
- Observing the **causal relationship** between individual variables in the latent space and linguistically meaningful units in intermediate layers
- Testing which acoustic properties are encoded at which layer via correlations
- Testing not only encoding of acoustic properties or words, but also of phonological processes and higher-level morphophonological processes such as reduplication
- Unsupervised generative models trained on raw speech

(Beguš and Zhou, 2022)

- Our proposal requires no further processing of the outputs

# Advantages of the new approach

- Manipulating and interpolating **individual latent variables well beyond training range** while visualizing intermediate layers
- Observing the **causal relationship** between individual variables in the latent space and linguistically meaningful units in intermediate layers
- Testing which acoustic properties are encoded at which layer via correlations
- Testing not only encoding of acoustic properties or words, but also of **phonological processes** and higher-level morphophonological processes such as reduplication
- Unsupervised generative models trained on raw speech

(Beguš and Zhou, 2022)

- Our proposal requires no further processing of the outputs

# Advantages of the new approach

- Manipulating and interpolating **individual latent variables well beyond training range** while visualizing intermediate layers
- Observing the **causal relationship** between individual variables in the latent space and linguistically meaningful units in intermediate layers
- Testing which acoustic properties are encoded at which layer via correlations
- Testing not only encoding of acoustic properties or words, but also of **phonological processes** and higher-level morphophonological processes such as reduplication
- Unsupervised generative models trained on raw speech

(Beguš and Zhou, 2022)

- Our proposal requires no further processing of the outputs

# Advantages of the new approach

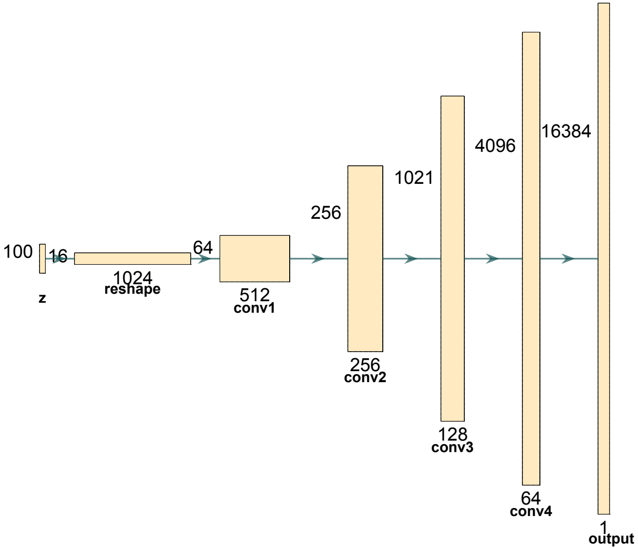
- Manipulating and interpolating **individual latent variables well beyond training range** while visualizing intermediate layers
- Observing the **causal relationship** between individual variables in the latent space and linguistically meaningful units in intermediate layers
- Testing which acoustic properties are encoded at which layer via correlations
- Testing not only encoding of acoustic properties or words, but also of **phonological processes** and higher-level morphophonological processes such as reduplication
- Unsupervised generative models trained on raw speech

(Beguš and Zhou, 2022)

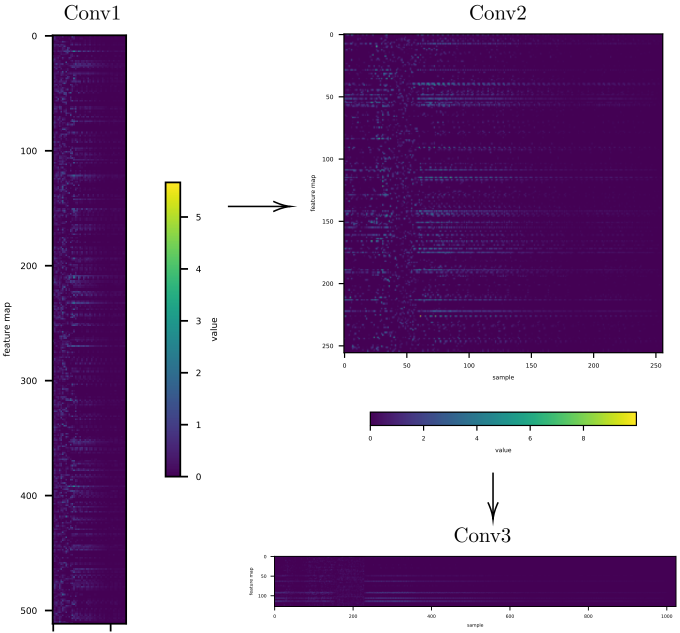
- Our proposal requires no further processing of the outputs

# The model

(Donahue et al., 2019; Beguš, 2021a)



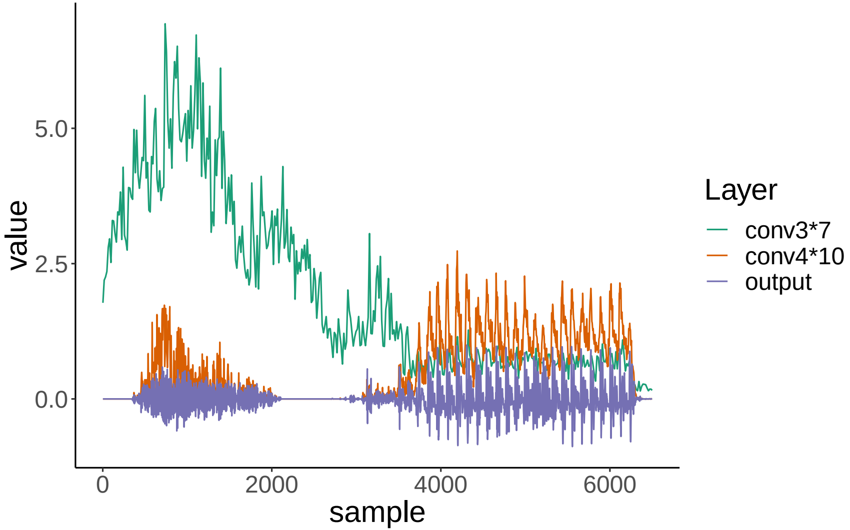
# Individual feature maps



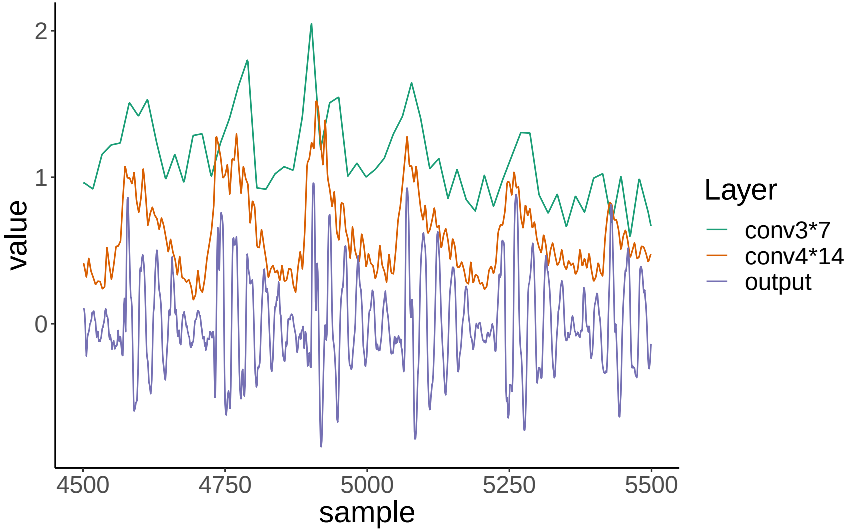
# Averaging over feature maps

$$\frac{1}{\|C\|}\sum_{i=1}^{\|C\|} C_i \tag{1}$$

# Layers

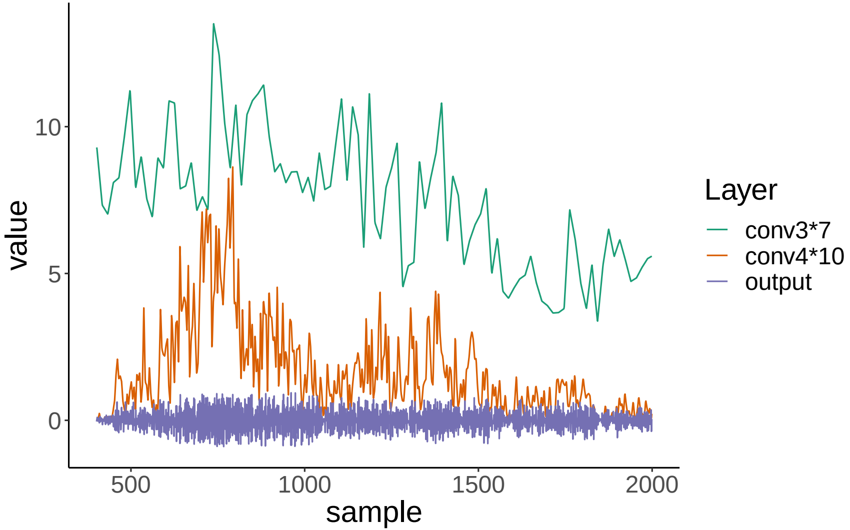

value
5.0
2.5
0.0
0
2000
4000
6000
sample
Layer
conv3*7
conv4*10
output

# Layers



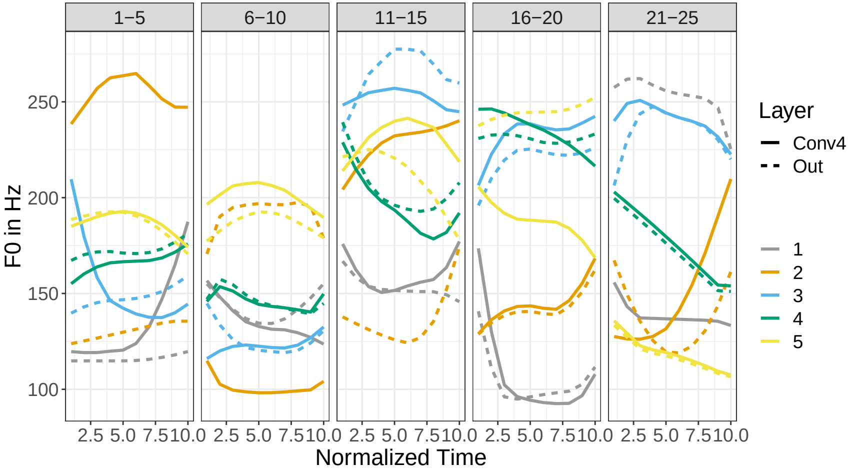
2
1
0
value
4500
4750
5000
5250
5500
sample
Layer
conv3*7
conv4*14
output

# Layers

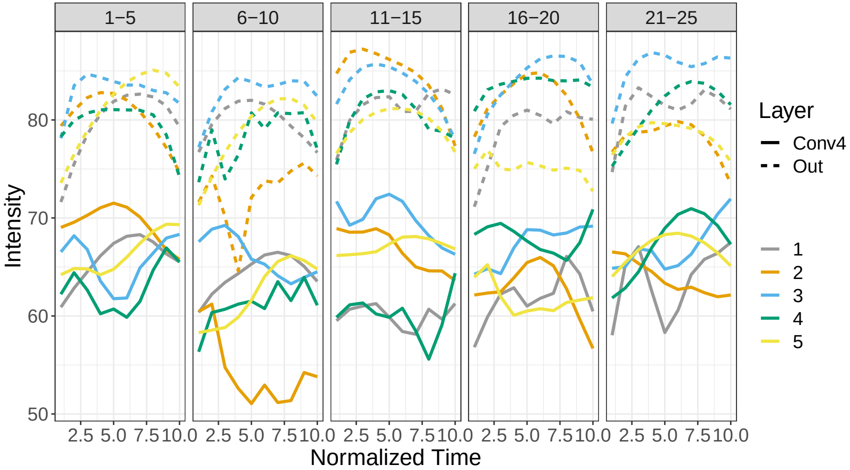


value
10
5
0
500
1000
1500
2000
sample
Layer
conv3*7
conv4*10
output

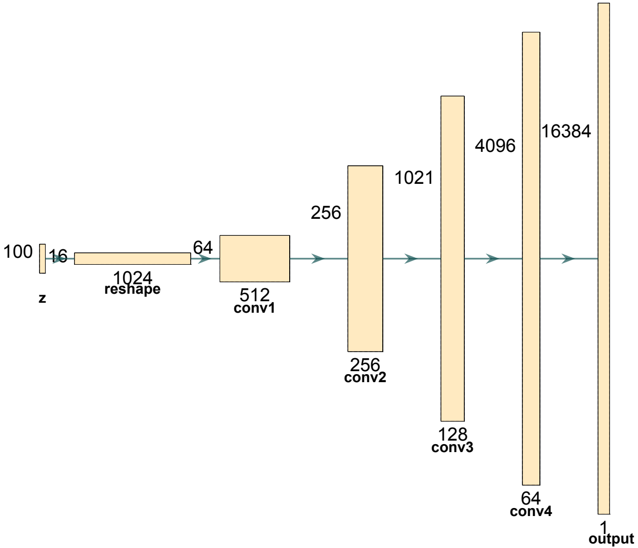
# Correlations



# Correlations

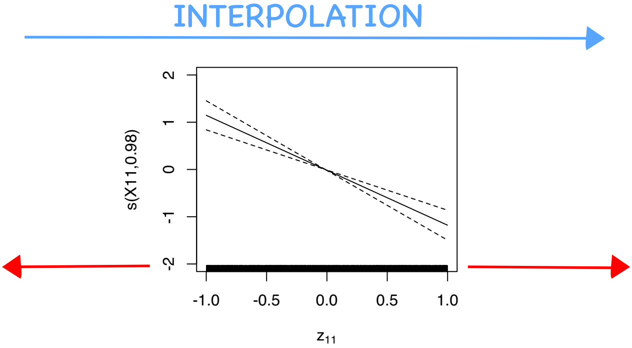


# The model (Beguš, 2020)

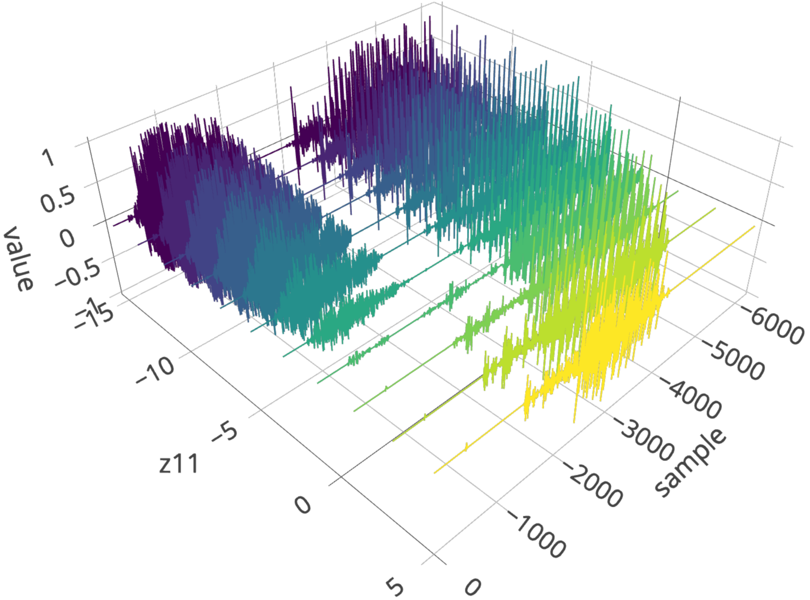


# The model (Beguš, 2021a)

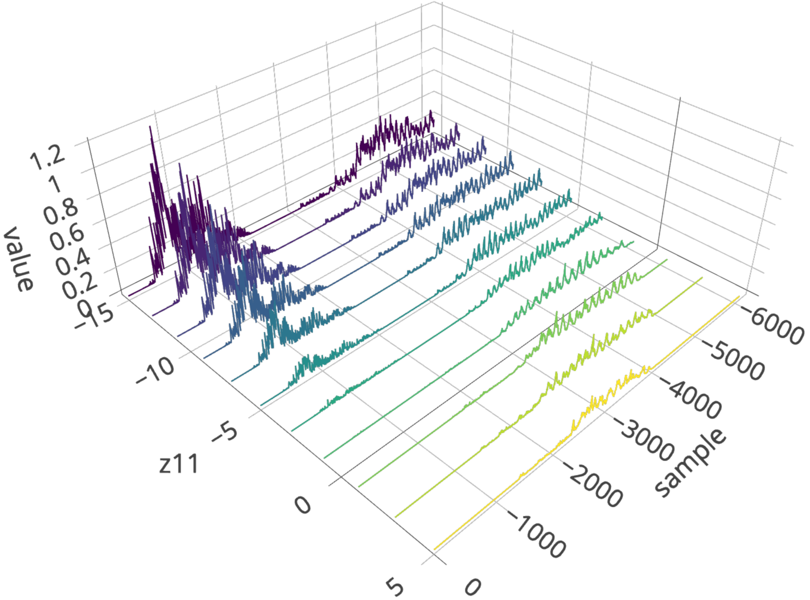
- Find a single variable in the latent space $z$ that correspond to [s]



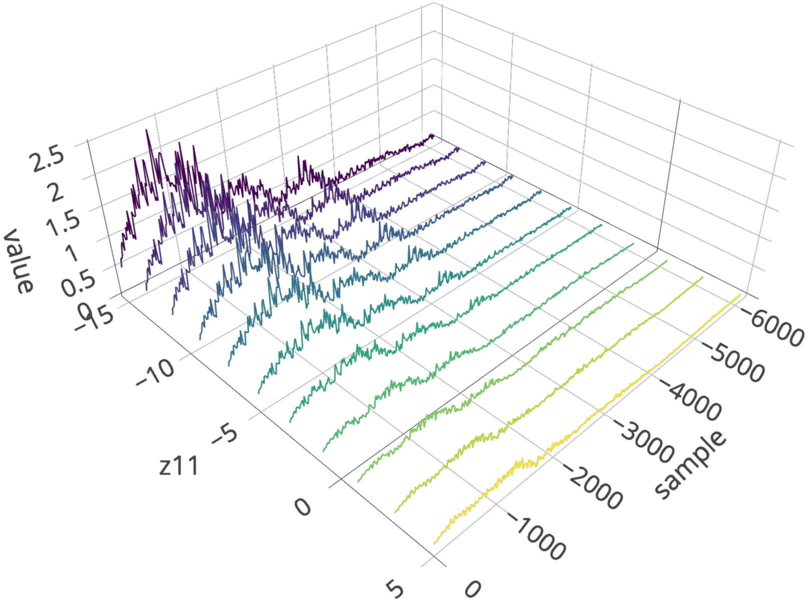
# Interpolation and a causal relationship #STV – Out



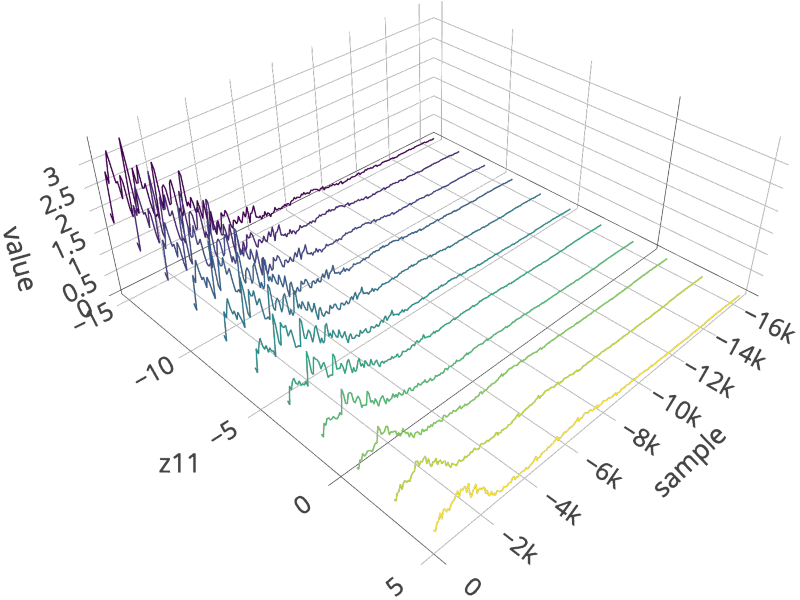
# Interpolation and a causal relationship #STV – Conv4



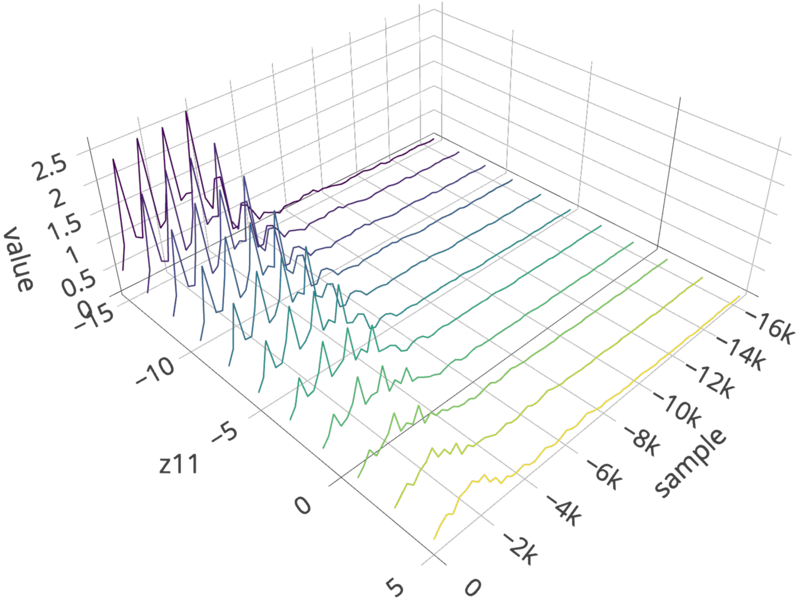
# Interpolation and a causal relationship #STV – Conv3



# Interpolation and a causal relationship #STV – Conv2



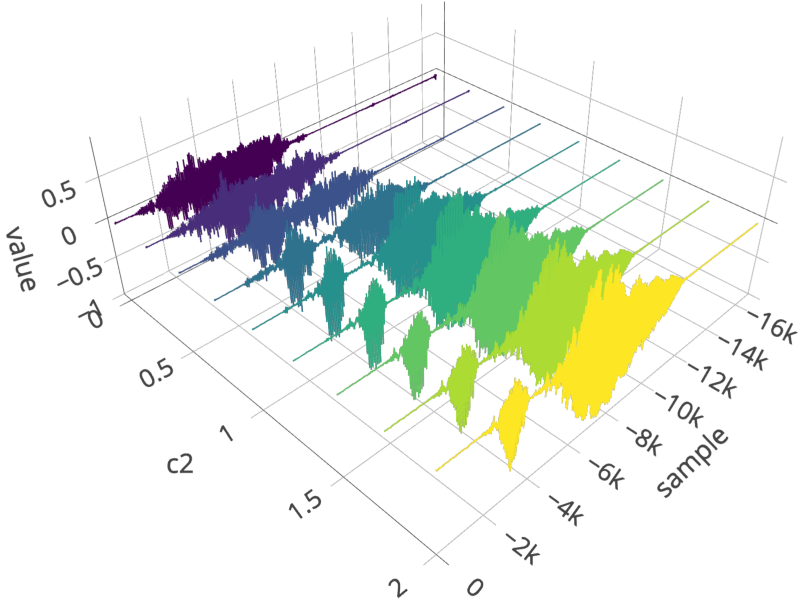
# Interpolation and a causal relationship #STV – Conv1



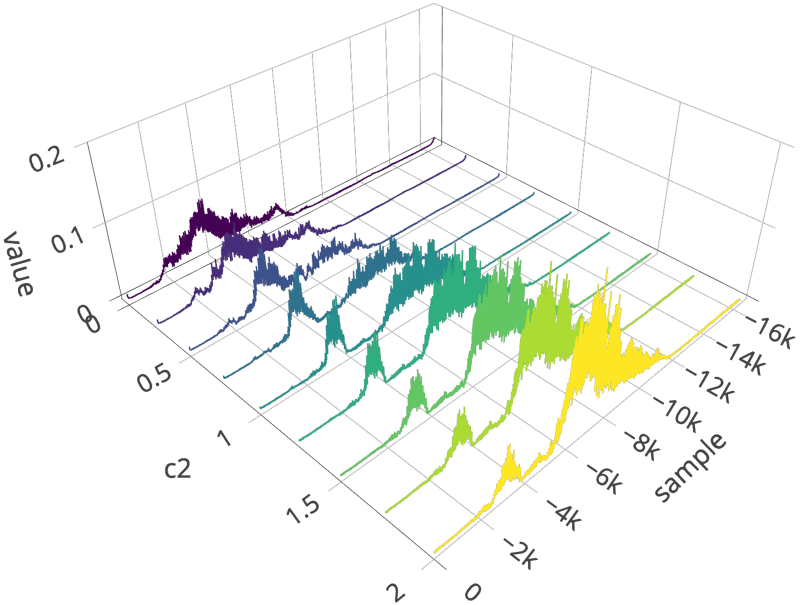
# More complex processes

- Reduplication (copying) one of the most complex processes in speech
- Learned from speech (Beguš, 2021b)

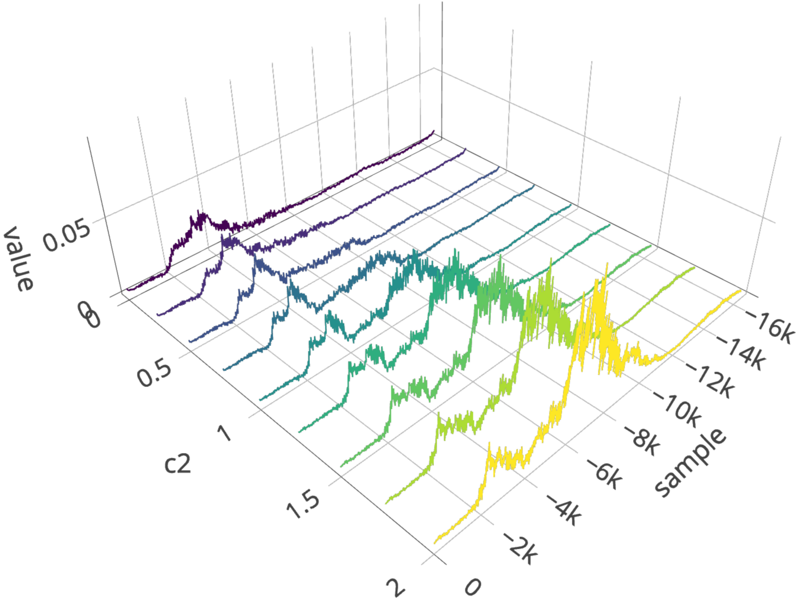
# Interpolation and a causal relationship [dədaj] – Out



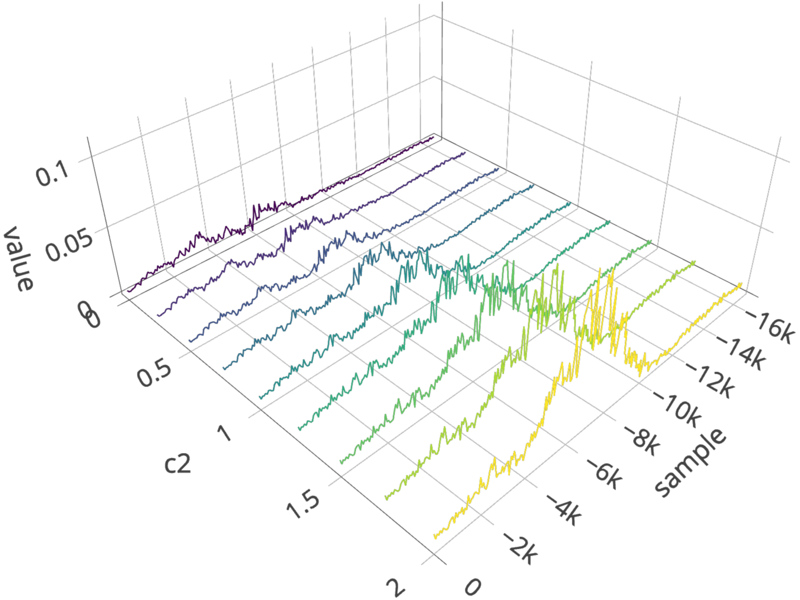
# Interpolation and a causal relationship [dədaj] – Conv4



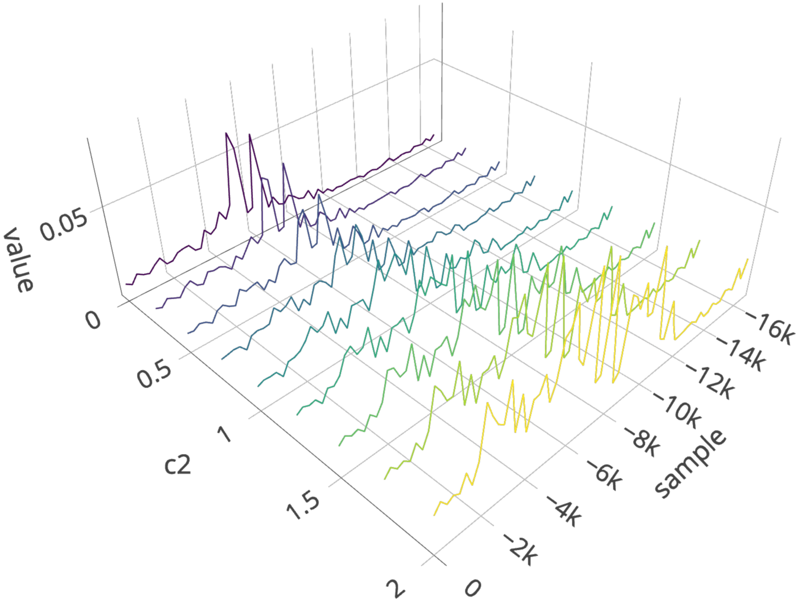
# Interpolation and a causal relationship [dədaj] – Conv3



# Interpolation and a causal relationship [dədaj] – Conv2



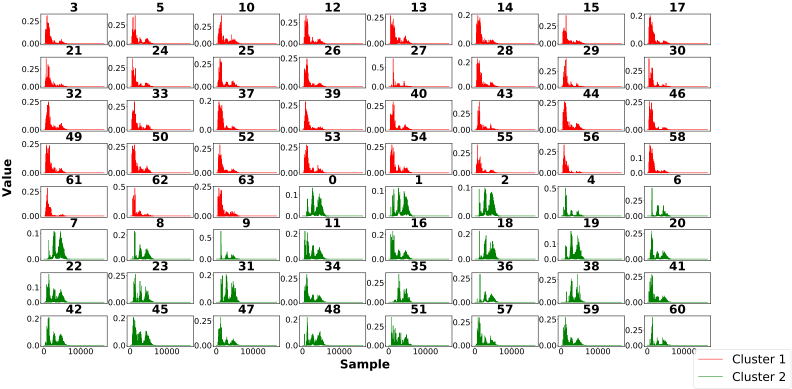
# Interpolation and a causal relationship [dədaj] – Conv1



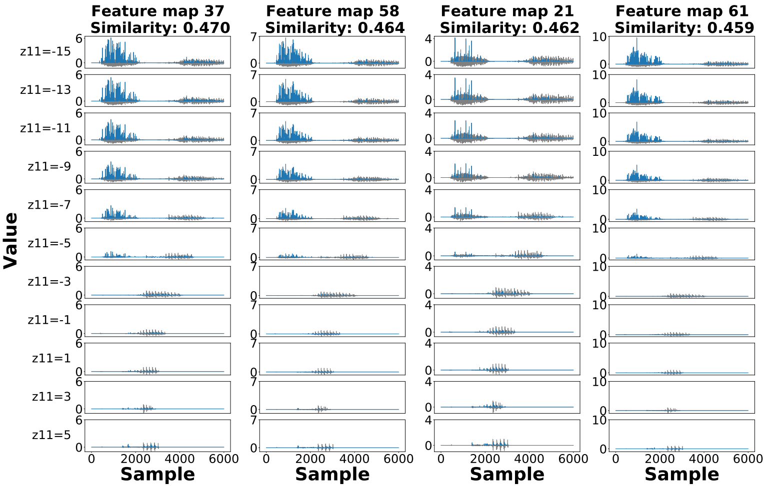
# Individual feature maps

- Lower-frequency properties such as acoustic envelope are encoded in earlier convolutional layers and that properties with frequencies higher than acoustic envelope (such as F0 or formant structure) get added on top of the envelope outline in the later layers.

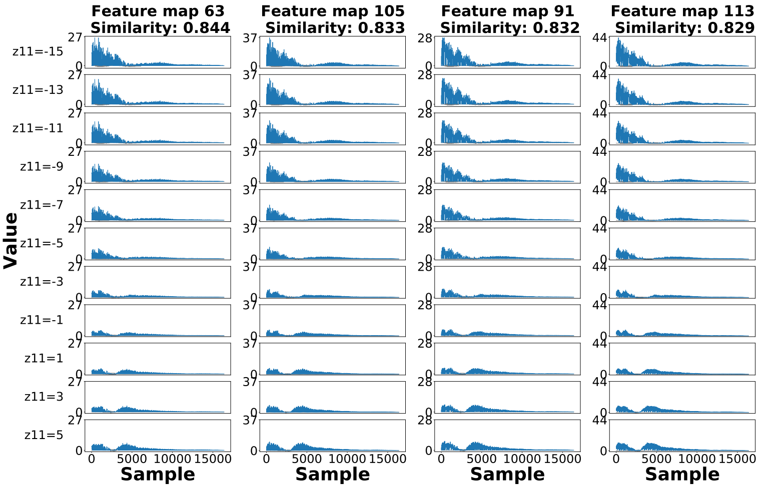
# Individual feature maps



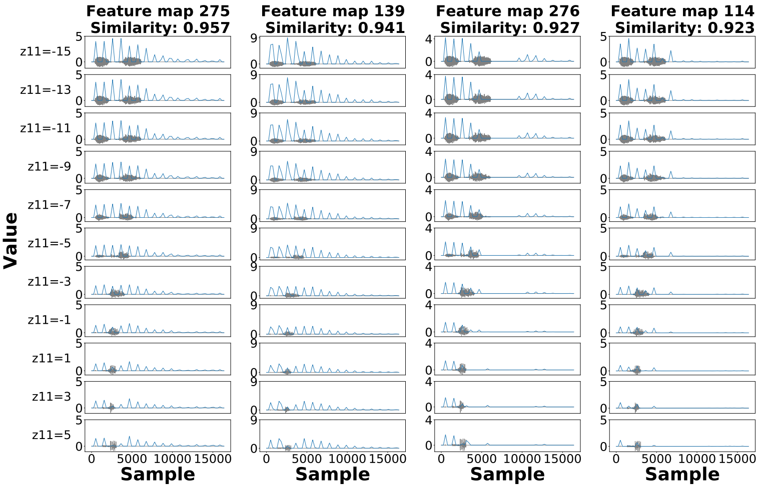
# Individual feature maps—causal interpolation



# Individual feature maps—causal interpolation



# Individual feature maps—causal interpolation



# Conclusions

- We can analyze which acoustic properties are encoded in which intermediate convolutional layers
- Understanding how phonological processes are encoded will be increasingly important as unsupervised speech technology systems become available in languages other than English
- Exploration of the causal relationship between individual latent variables and intermediate convolutional layers by manipulating and linearly interpolating latent variables to values outside of the training

# Future directions

- Other properties such as acoustic correlates of gender, dialects, race, or socioeconomic background can be probed with the same techniques as well.
- A diagnostic for improving the performance of CNNs trained on speech
- Brain–artificial neural network comparison (Beguš et al., 2023)

# References I

Beguš, G., Zhou, A., Zhao, T. C., 2023. Encoding of speech in convolutional layers and the brain stem based on language experience. Scientific Reports 13 (1), 6480. URL https://doi.org/10.1038/s41598-023-33384-9

Beguš, G., 2020. Generative adversarial phonology: Modeling unsupervised phonetic and phonological learning with neural networks. Frontiers in Artificial Intelligence 3, 44. URL https://www.frontiersin.org/articles/10.3389/frai.2020.00044/abstract

Beguš, G., 2021a. CiwGAN and fiwGAN: Encoding information in acoustic data to model lexical learning with Generative Adversarial Networks. Neural Networks 139, 305–325. URL https://www.sciencedirect.com/science/article/pii/S0893608021001052

Beguš, G., 10 2021b. Identity-Based Patterns in Deep Convolutional Networks: Generative Adversarial Phonology and Reduplication. Transactions of the Association for Computational Linguistics 9, 1180–1196. URL https://doi.org/10.1162/tacl_a_00421

Beguš, G., Zhou, A., 2022. Interpreting intermediate convolutional layers in unsupervised acoustic word classification. In: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 8207–8211.

# References II

Donahue, C., McAuley, J. J., Puckette, M. S., 2019. Adversarial audio synthesis. In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net, pp. 1–16.
URL https://openreview.net/forum?id=ByMVTsR5KQ

Zeiler, M. D., Fergus, R., 2014. Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (Eds.), Computer Vision – ECCV 2014. Springer International Publishing, Cham, pp. 818–833.

# Thank you!


: {begus,azhou314}@berkeley.edu

: @BerkeleySClab

Speech & Computation

Berkeley
UNIVERSITY OF CALIFORNIA