Interpreting intermediate convolutional layers of generative CNNs trained on waveforms

Gašper Beguš & Alan Zhou

{begus,azhou314}@berkeley.edu

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

• Interpretability one of the main frontiers in AI research

• Most studies focus on vision

• 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

(Zeiler and Fergus, 2014)
Introduction

• Interpretability one of the main frontiers in AI research
• Most studies focus on vision
  ▶ 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

(Zeiler and Fergus, 2014)
Introduction

• Interpretability one of the main frontiers in AI research
• Most studies focus on vision
• 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

(Zeiler and Fergus, 2014)
Introduction

- Interpretability one of the main frontiers in AI research
- Most studies focus on vision
- 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

(Zeiler and Fergus, 2014)
Introduction

• Interpretability one of the main frontiers in AI research
• Most studies focus on vision
• 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.
- Our proposal requires no further processing of the outputs.

(Beguš and Zhou, 2022)
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)
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

- Conv1
- Conv2
- Conv3
- Conv4
- Conv4-zoomed
Averaging over feature maps

\[
\frac{1}{\|C\|} \sum_{i=1}^{\|C\|} C_i
\]  

(1)
Correlations

Normalized Time

F0 in Hz

Layer

Conv4
Out

1
2
3
4
5
Correlations

<table>
<thead>
<tr>
<th>1−5</th>
<th>6−10</th>
<th>11−15</th>
<th>16−20</th>
<th>21−25</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.5</td>
<td>5.0</td>
<td>7.5</td>
<td>10.0</td>
<td>2.5</td>
</tr>
<tr>
<td>5.0</td>
<td>7.5</td>
<td>10.0</td>
<td>2.5</td>
<td>5.0</td>
</tr>
<tr>
<td>7.5</td>
<td>10.0</td>
<td>2.5</td>
<td>5.0</td>
<td>7.5</td>
</tr>
<tr>
<td>10.0</td>
<td>2.5</td>
<td>5.0</td>
<td>7.5</td>
<td>10.0</td>
</tr>
</tbody>
</table>

Normalized Time

Intensity

Layer

Conv4
Out

1
2
3
4
5

Beguš & Zhou

ICASSP 2023

{begus,azhou314}@berkeley.edu
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 [dədæj] – 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

![Feature map 37: Similarity 0.470](image1)

![Feature map 58: Similarity 0.464](image2)

![Feature map 21: Similarity 0.462](image3)

![Feature map 61: Similarity 0.459](image4)
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


Thank you!

✉️: {begus, azhou314}@berkeley.edu
🐦: @BerkeleySCLab