Learning to Unlearn and Relearn Speech Signal Processing using Neural Networks: current and future perspectives

Mathew Magimai Doss

Collaborators: Vinayak Abrol, Selen Hande Kabil, Hannah Muckenhirn, Dimitri Palaz, D. S. Pavan Kumar

Idiap Research Institute, Martigny, Switzerland

July 17, 2018
Conventional speech processing approach

▶ Conventional cepstral features extraction process:

- Conventional cepstral features extraction process:
  - FFT
  - Critical bands filtering
  - Non-linear operation
  - DCT
  - log(·)
  - AR modeling
  - MFCC
  - Derivatives \( \Delta + \Delta \Delta \)
  - Derivatives \( \Delta + \Delta \Delta \)
  - CNN
  - NN classifier
  - \( P(i|x) \)

▶ Recent trend using Convolutional Neural Networks (CNN):

- Recent trend using Convolutional Neural Networks (CNN):
  - FFT
  - Critical bands filtering
  - Derivatives \( \Delta + \Delta \Delta \)
  - CNN
  - NN classifier
  - \( P(i|x) \)

1. Quasi-stationarity (windowing, time-frequency resolution)
   ▶ Motivated from speech coding analysis-synthesis studies
2. Speech production knowledge
3. Speech perception knowledge
In this talk

- Can help in overcoming limitations of conventional short-term speech processing
- Can help in better understanding speech signal characteristics in a task specific manner
Minimal prior knowledge

- Short-term processing
- Feature extraction can be seen as a filtering operation
- Relevant Information can be spread across time

Determined in a data-driven manner. 
All stages are trained jointly using back-propagation with a cost function based on cross entropy.
CNN-based system using raw speech as input

Illustration of the first convolutional layer

- $w_{seq}$: Input speech signal with temporal context
- $kW$: Window size
  - Sub-segmental ($< 1$ pitch period)
  - Segmental ($1 - 3$ pitch periods)
- $dW$: Window shift ($< 1$ pitch period)
- $nf$: number of filters
### Speech processing applications

<table>
<thead>
<tr>
<th>Application</th>
<th>$w_{seq}$</th>
<th>$kW$</th>
<th># of conv. layers</th>
<th># of hidden layers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech reco. $^{1,2}$</td>
<td>250-310 ms</td>
<td>sub-seg</td>
<td>3-5</td>
<td>1-3</td>
</tr>
<tr>
<td>Speaker reco. $^{3,4}$</td>
<td>≈ 500 ms</td>
<td>seg, sub-seg</td>
<td>2-3</td>
<td>1</td>
</tr>
<tr>
<td>Presentation attack detection $^{5}$</td>
<td>≈ 300 ms</td>
<td>seg</td>
<td>2</td>
<td>1 or none</td>
</tr>
<tr>
<td>Gender reco. $^{6}$</td>
<td>250-310 ms</td>
<td>seg, sub-seg</td>
<td>1-3</td>
<td>1</td>
</tr>
<tr>
<td>Paralinguistic $^{7}$</td>
<td>250-500 ms</td>
<td>seg, sub-seg</td>
<td>3-4</td>
<td>1</td>
</tr>
</tbody>
</table>

In this talk

What information does such systems learn?

- Filter level analysis
- Whole network level analysis
Filter level analysis
First convolution layer

- Cumulative frequency response of filters

\[ F_{\text{cum}} = \sum_{m=1}^{M} \frac{F_m}{\|F_m\|_2}, \]  

(1)

where \( F_m \) is the DFT of filter \( f_m \) and \( M \) is number of filters.

- Response of filters to input speech by interpreting learned filters collectively as a spectral dictionary

\[ \mathbf{X} = \sum_{m=1}^{M} \langle \mathbf{x}, f_m \rangle \text{DFT}[f_m], \]  

(2)

where \( \hat{x}_m = \langle \mathbf{x}, f_m \rangle \) is output of filter \( f_m \) and \( \mathbf{X} \) is the spectral information modeled.

If \( \{f_m\} \) were Fourier sine and cosine bases then \( \mathbf{X} \) is DFT of \( \mathbf{x} \).
Filter level analysis
Speech recognition: cumulative response

- Filters model sub-segmental speech
- Standard filterbank: constant-Q filters, i.e. flat response.

![Graph showing normalized magnitude vs frequency for a CNN trained on WSJ corpus.](image)

CNN trained on WSJ corpus
### Filter level analysis

Speech recognition: spectral response $\chi_{foraframeofspeech}$

#### Spectral response of /iy/ from American English Vowel dataset.

**Magnitude spectrum of /iy/**

<table>
<thead>
<tr>
<th>Speaker</th>
<th>F1 range</th>
<th>F2 range</th>
<th>Obs. 1st peak (in Hz)</th>
<th>Obs. 2nd peak (in Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>m01</td>
<td>328-357</td>
<td>2418-2458</td>
<td>375</td>
<td>2625</td>
</tr>
<tr>
<td>w01</td>
<td>439-441</td>
<td>2767-2822</td>
<td>437</td>
<td>2812</td>
</tr>
<tr>
<td>b01</td>
<td>468-554</td>
<td>2981-3024</td>
<td>500</td>
<td>3000</td>
</tr>
<tr>
<td>g01</td>
<td>382-392</td>
<td>3034-3078</td>
<td>375</td>
<td>-</td>
</tr>
</tbody>
</table>

- **Gain normalized magnitude spectrum**
- **Frequency (Hz)**: 0 1000 2000 3000 4000 5000 6000 7000 8000
- **Spectral response of /iy/ from American English Vowel dataset.**
Filter level analysis
Speaker recognition: cumulative response

Segmental modeling
Sub-segmental modeling
Filter level analysis

Speaker recognition: spectral response $\mathcal{X}$ (Segmental modeling)

F0 contours estimated on Keele pitch database using the CNN-based speaker classifier trained on Voxforge.

| $\mathcal{X}$ |

F0 contours for female speaker f2nw0000

F0 contours for male speaker m3nw0000
Filter level analysis

Speaker recognition: spectral response $\chi$ of a frame of speech. (Sub-segmental modeling)
In this talk

What information does such systems learn?

- Filter level analysis
- Whole network level analysis
Whole network analysis
Gradient-based visualization

▶ In computer vision research, given an input image-output class pair and the trained system, finding contribution of each pixel in the image on the output score. (*guided backpropagation*)

▶ Given an input speech-output class pair and the trained system, what is the contribution of each sample on the output score?  

---

Whole network analysis
Case study on speech recognition (1)
Whole network analysis
Case study on speech recognition (2)

- Analysis of CNN trained on TIMIT phone recognition task on American English Vowel (AEV) dataset
- F0, F1 and F2 estimated automatically for the relevance signal for the steady state regions and compared to the values specified on the original study.

Table: Average accuracy in (%) of fundamental frequencies, and formant frequencies of vowels produced by 45 male and 48 female speakers, estimated from relevance signal of AEV dataset.

<table>
<thead>
<tr>
<th></th>
<th>/ah/</th>
<th>/eh/</th>
<th>/iy/</th>
<th>/oa/</th>
<th>/uw/</th>
</tr>
</thead>
<tbody>
<tr>
<td>F0</td>
<td>F</td>
<td>93</td>
<td>91</td>
<td>91</td>
<td>94</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>92</td>
<td>90</td>
<td>89</td>
<td>93</td>
</tr>
<tr>
<td>F1</td>
<td>F</td>
<td>90</td>
<td>92</td>
<td>93</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>88</td>
<td>92</td>
<td>92</td>
<td>89</td>
</tr>
<tr>
<td>F2</td>
<td>F</td>
<td>94</td>
<td>94</td>
<td>94</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>94</td>
<td>93</td>
<td>94</td>
<td>94</td>
</tr>
</tbody>
</table>
Whole network analysis
Case study on speaker recognition (1)

Original signal

Segmental modeling

Sub-segmental modeling
Whole network analysis
Case study on speaker recognition (2)

Utterance level average spectrum

Segmental modeling

Sub-segmental modeling
Whole network analysis
Listening to relevance signal

- Relevance signal obtained from speaker recognition CNN (segmental modeling)
- Relevance signal obtained from speech recognition CNN
- Original signal
Summary

- Can help in overcoming limitations of conventional short-term speech processing
  - Allows both segmental modeling and sub-segmental modeling
- Can help in better understanding speech signal characteristics in a task specific manner
  - Relevance signal can be analyzed using conventional signal processing techniques to gain insight
- Work under progress to understand how the neural network is modeling the relevant information
  - Potentially provide new algorithms for speech signal processing
Thank you for your attention!

Questions?
CNN-based system using raw speech as input

Detailed view for one example

Conv 1
$kW = 30$
d$W = 10$

MP 1
$kW = 2$
d$W = 2$

Conv 2
$kW = 5$
d$W = 1$

MP 2
$kW = 2$
d$W = 2$

Conv 3
$kW = 5$
d$W = 1$

MP 3
$kW = 2$
d$W = 2$

$\text{ANN } p(i|x)$
Whole network analysis
Speech recognition versus Speaker recognition

Original signal spectrogram

Phone CNN relevance signal spectrogram

Speaker CNN relevance signal spectrogram