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

Mathew Magimai Doss

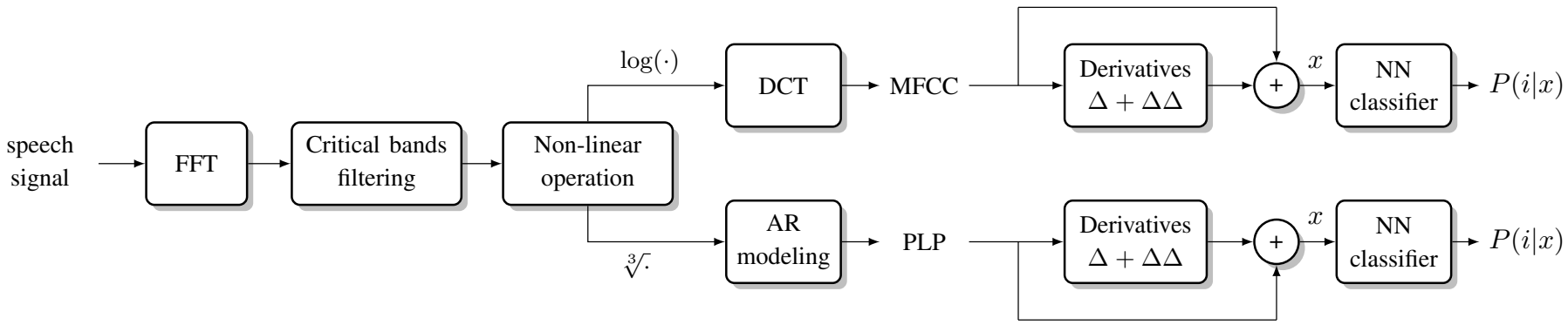
Collaborators: Vinayak Abrol, Selen Hande Kabil, Hannah Muckenhirn, Dimitri Palaz, D. S. Pavan Kumar

Idiap Research Institute, Martigny, Switzerland

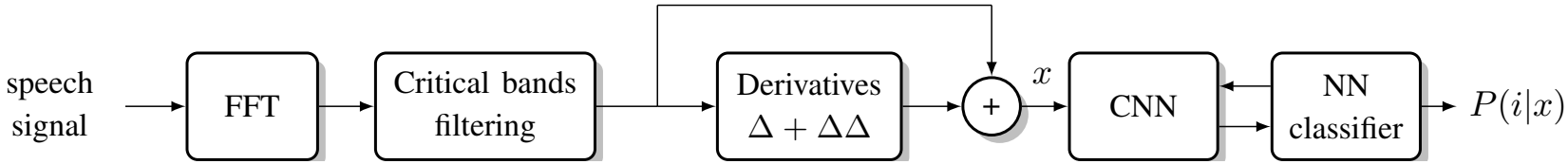
July 17, 2018

Conventional speech processing approach

- Conventional cepstral features extraction process:

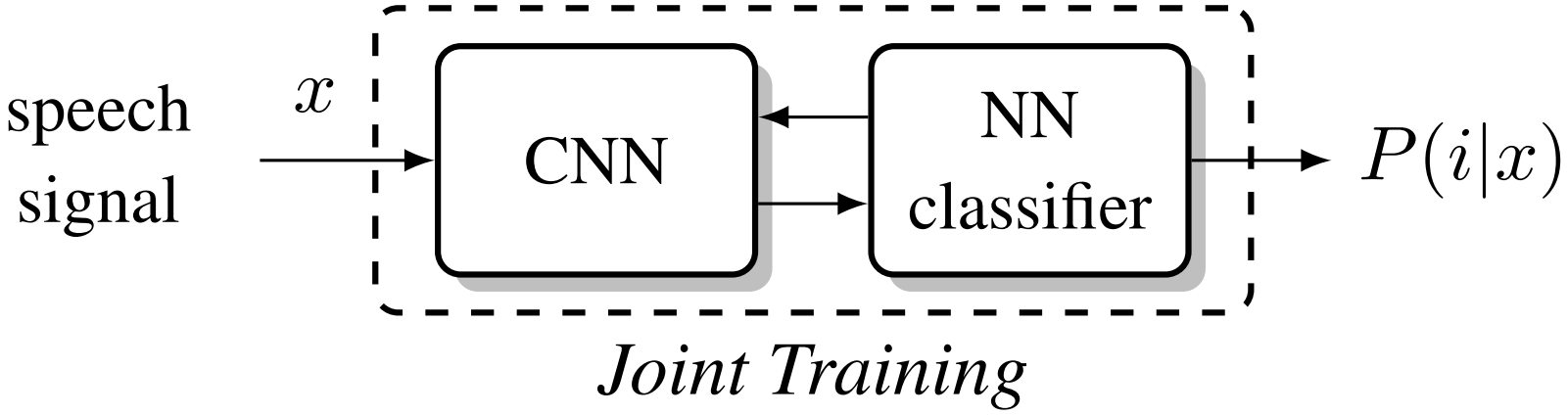


- Recent trend using Convolutional Neural Networks (CNN):



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

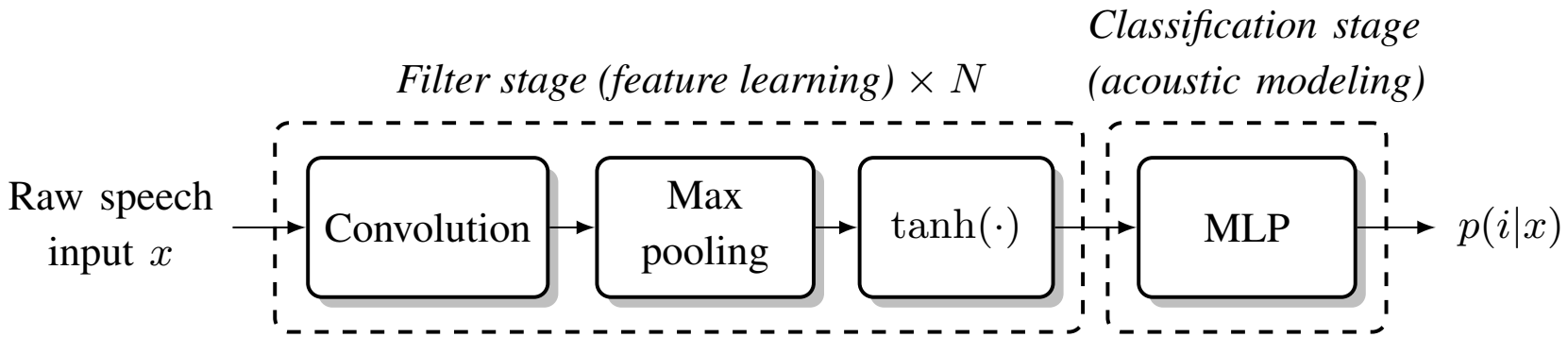
In this talk



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

CNN-based system using raw speech as input

Overview



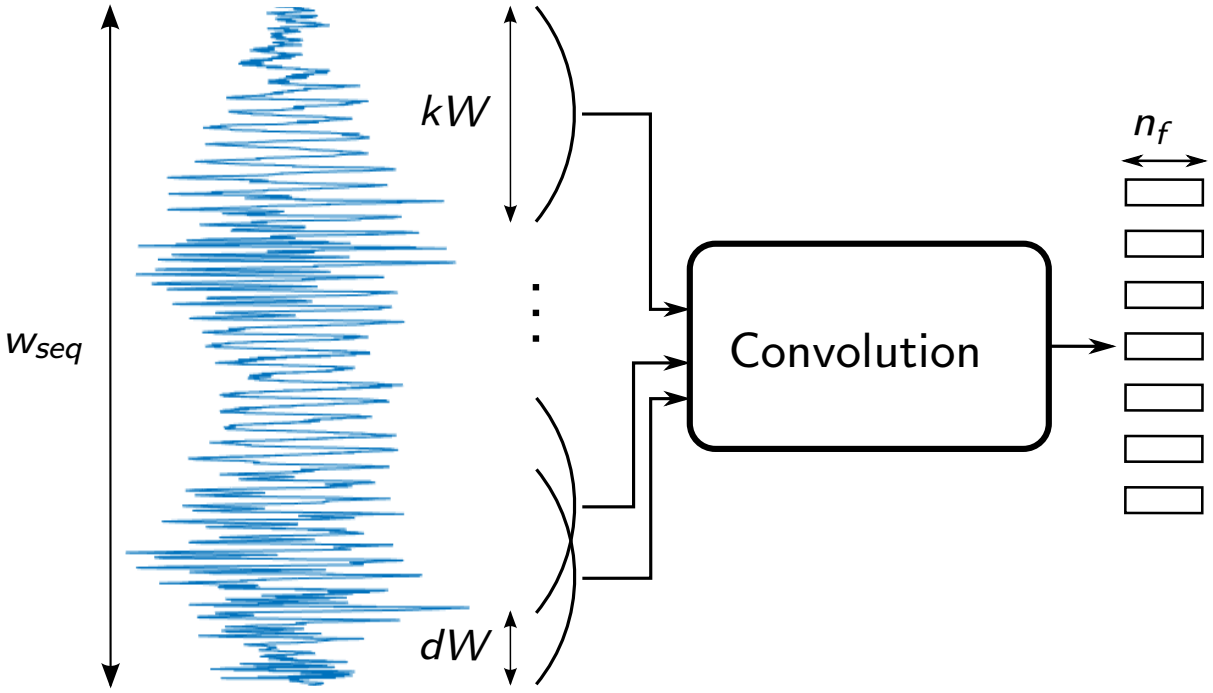
Minimal prior knowledge

- 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)
- n_f: number of filters

Speech processing applications

Application	w_{seq}	kW	# of conv. layers	# of hidden layers
Speech reco.[1,2]	250-310 ms	sub-seg	3-5	1-3
Speaker reco.[3,4]	$\approx$ 500 ms	seg, sub-seg	2-3	1
Presentation attack detection[5]	$\approx$ 300 ms	seg	2	1 or none
Gender reco.[6]	250-310 ms	seg, sub-seg	1-3	1
Paralinguistic[7]	250-500 ms	seg, sub-seg	3-4	1

[1] Dimitri Palaz, Ronan Collobert, and Mathew Magimai.-Doss, Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks, in Proc. of Interspeech, 2013.

[2] Dimitri Palaz, Mathew Magimai.-Doss, and Ronan Collobert, End-to-end acoustic modeling using convolutional neural net- works for automatic speech recognition, Idiap-RR Idiap-RR-18-2016, Idiap, 6 2016.

[3] Hannah Muckenhirn, Mathew Magimai.-Doss, and Sébastien Marcel, Towards directly modeling raw speech signal for speaker verification using CNNs, in Proc. of ICASSP, 2018.

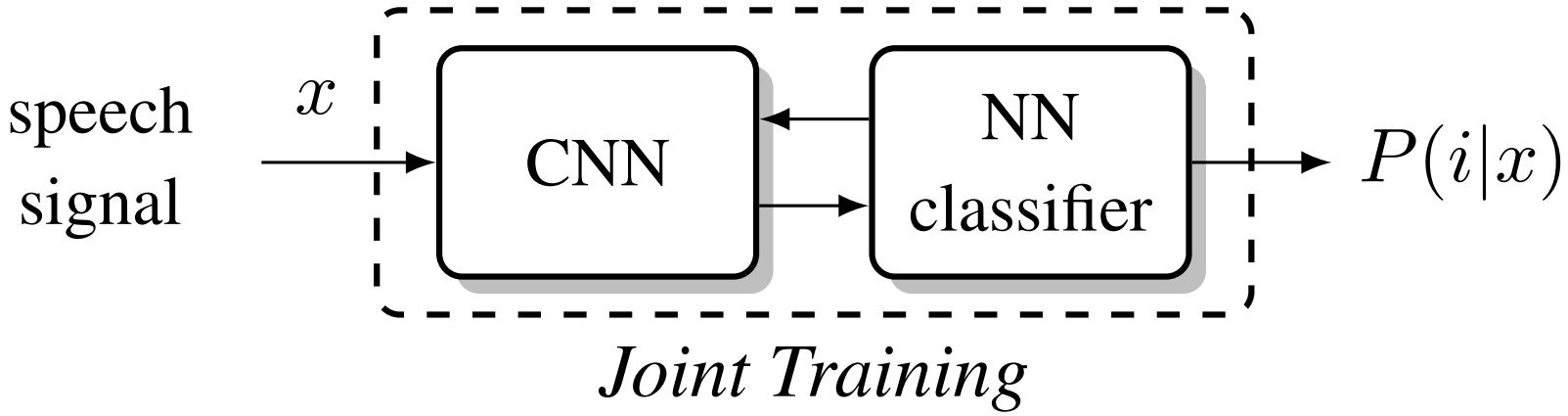
[4] Hannah Muckenhirn, Mathew Magimai.-Doss, and Sébastien Marcel, On learning vocal tract system related speaker dis- criminative information from raw signal using CNNs,, in Proc. of Interspeech, 2018.

[5] Hannah Muckenhirn, Mathew Magimai-Doss, and Sébastien Marcel, End-to-End Convolutional Neural Network-based Voice Presentation Attack Detection, in Proceedings of the IEEE International Joint Conference on Biometrics (IJCB), 2017.

[6] Selen Hande Kabil, Hannah Muckenhirn, and Mathew Magimai.-Doss, On learning to identify genders from raw speech signal using CNNs, in Proceedings of Interspeech, 2018.

[7] Bogdan Vlasenko, Jilt Sebastian, Pavan Kumar D. S., and Mathew Magimai.-Doss, Implementing fusion techniques for the classification of paralinguistic information, in Proc. of Interspeech, 2018.

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

$$F_{cum} = \sum_{m=1}^{M} \frac{F_m}{\|F_m\|_2}, \tag{1}$$

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

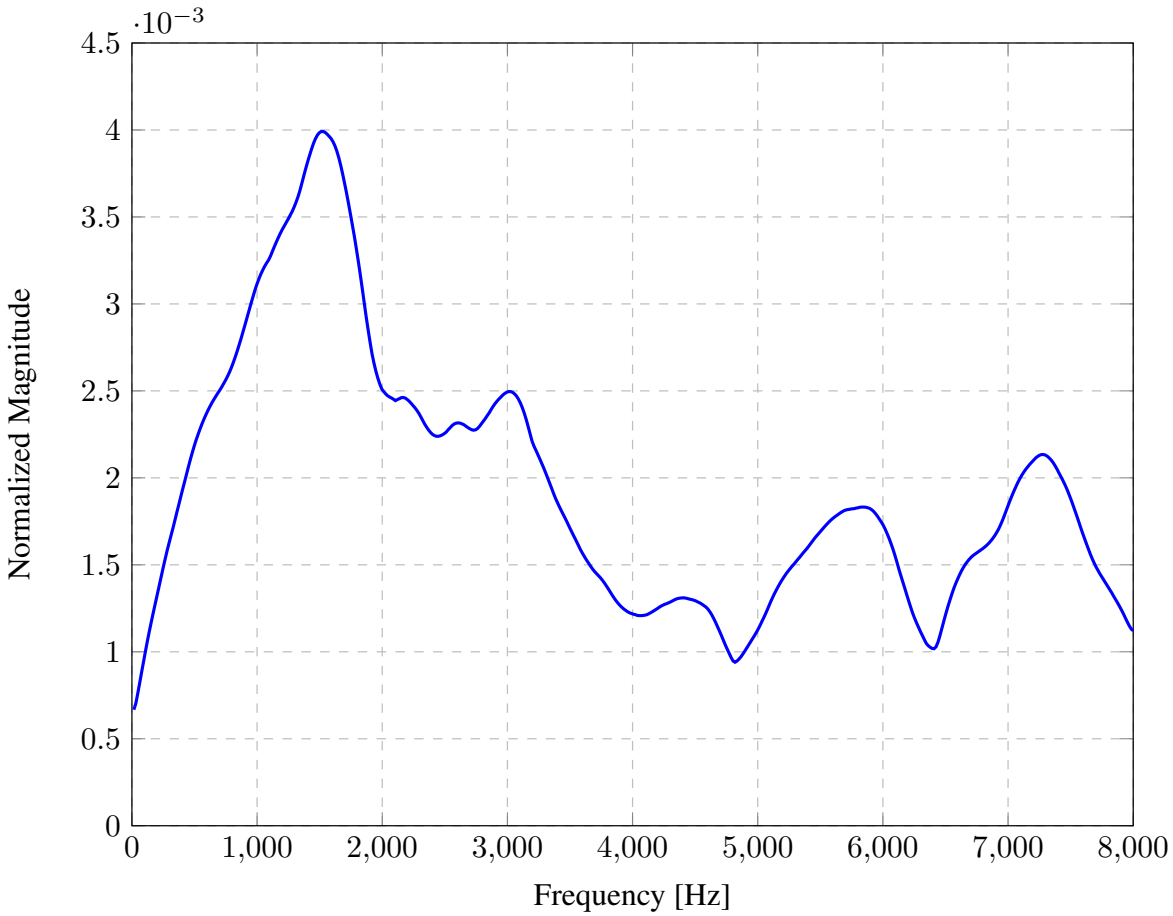
$$\mathcal{X} = \sum_{m=1}^{M} \langle \mathbf{x}, f_m \rangle \ \mathrm{DFT}[f_m], \tag{2}$$

where $\hat{x}_m = \langle \mathbf{x}, f_m \rangle$ is output of filter f_m and $\mathcal{X}$ is the spectral information modeled.
If $\{f_m\}$ were Fourier sine and cosine bases then $\mathcal{X}$ is DFT of $\mathbf{x}$.

Filter level analysis

Speech recognition: cumulative response

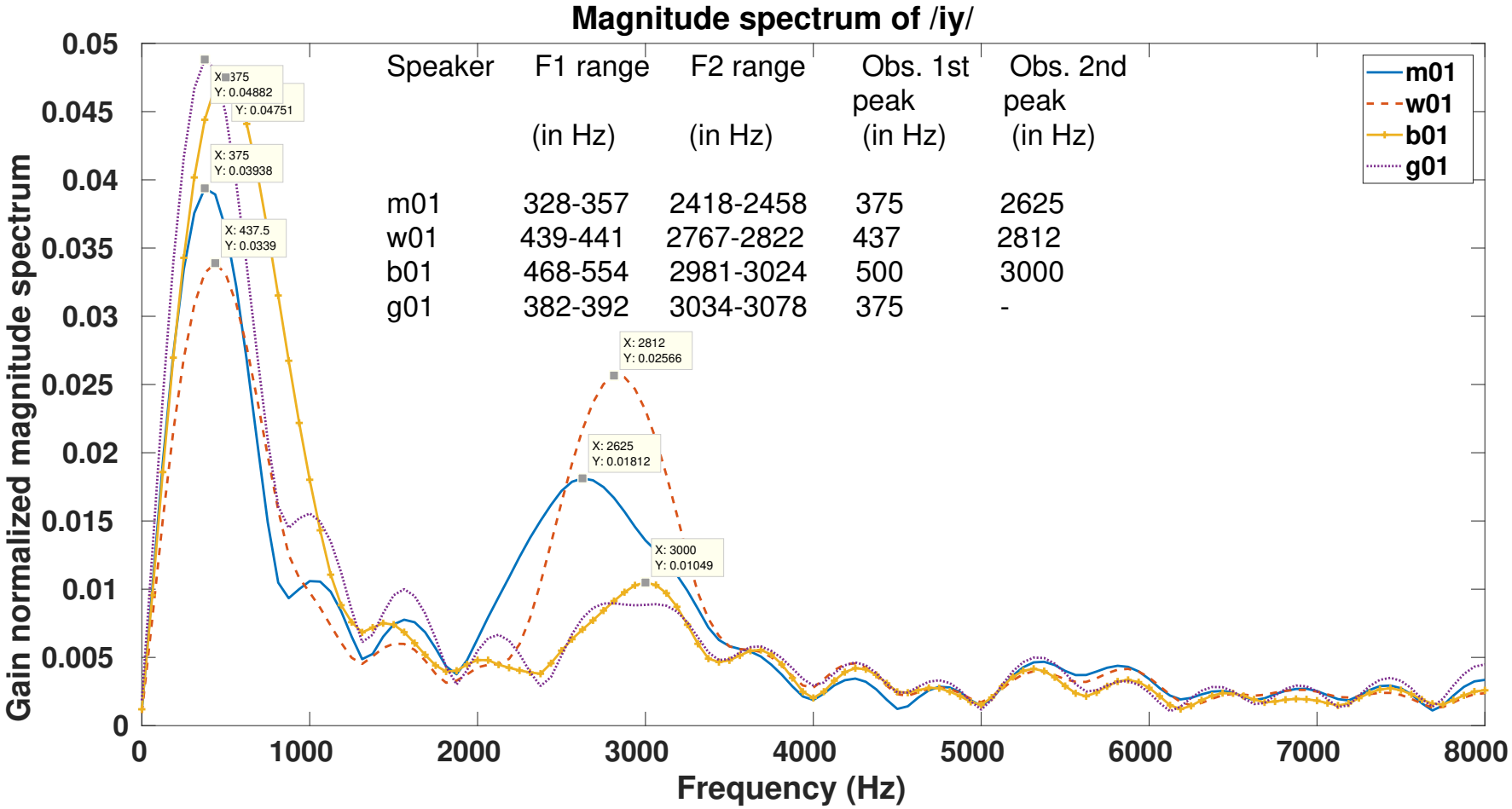
- Filters model sub-segmental speech
- Standard filterbank: constant-Q filters, i.e. flat response.



CNN trained on WSJ corpus

Filter level analysis

Speech recognition: spectral response $\mathcal{X}$ *foraframeofspeech*

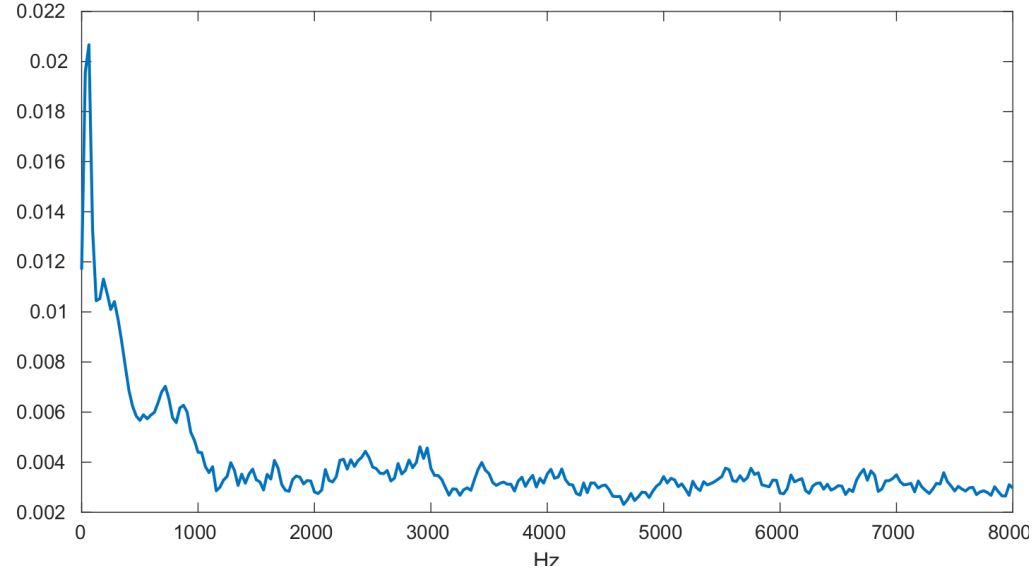


Speaker	F1 range (in Hz)	F2 range (in Hz)	Obs. 1st peak (in Hz)	Obs. 2nd peak (in Hz)
m01	328-357	2418-2458	375	2625
w01	439-441	2767-2822	437	2812
b01	468-554	2981-3024	500	3000
g01	382-392	3034-3078	375	-

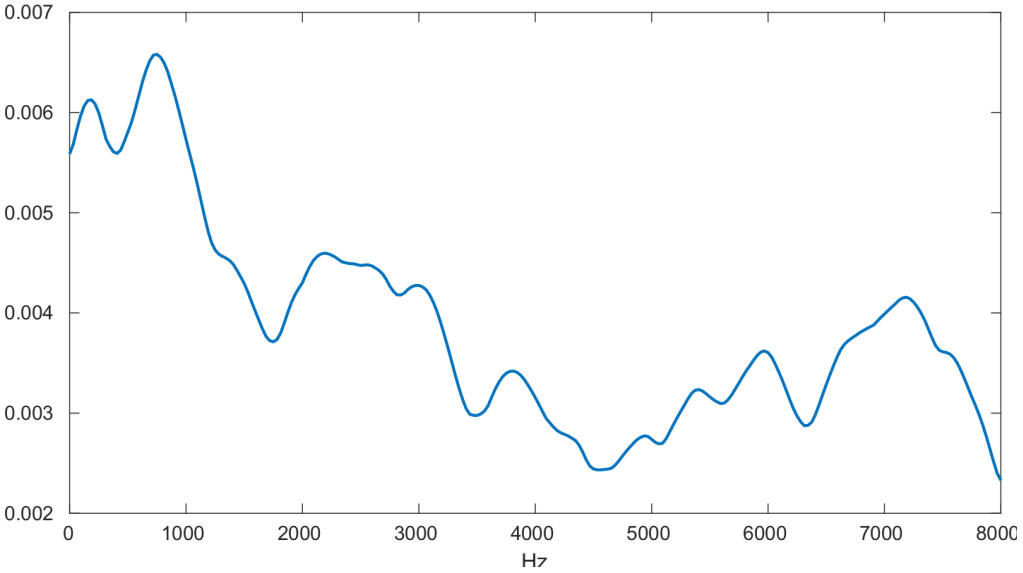
Spectral response of /iy/ from American English Vowel dataset.

Filter level analysis

Speaker recognition: cumulative response



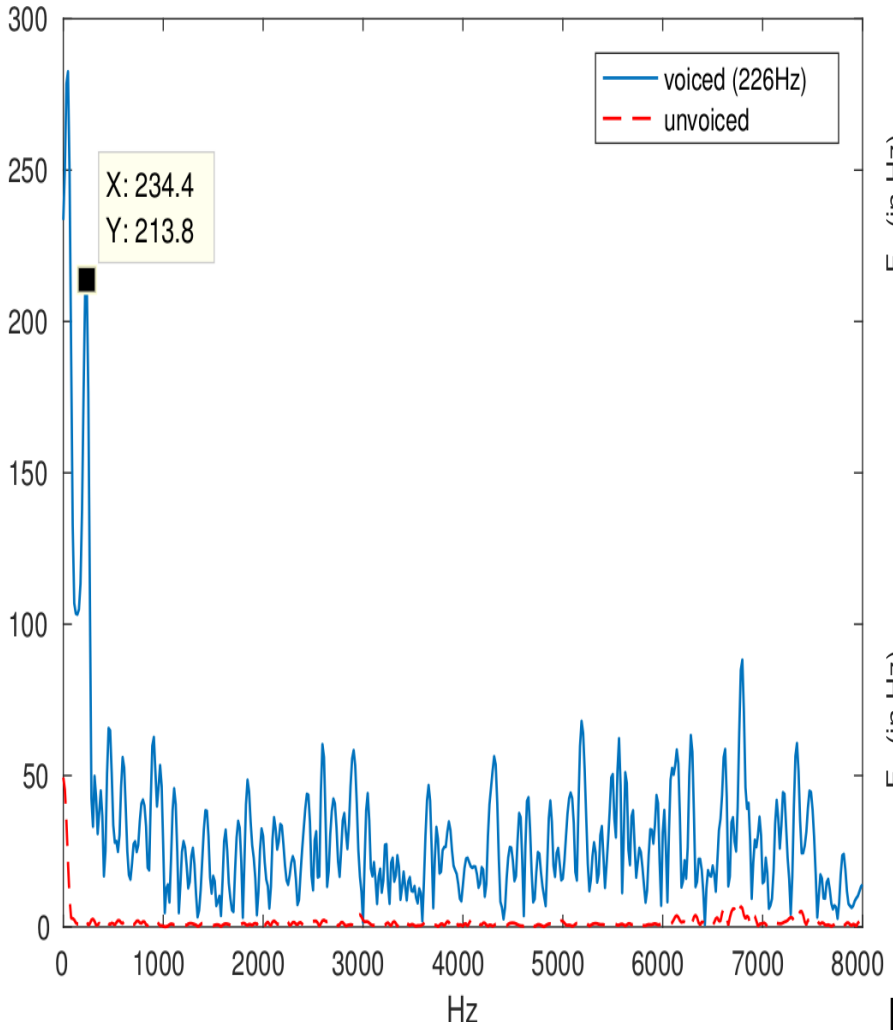
Segmental modeling



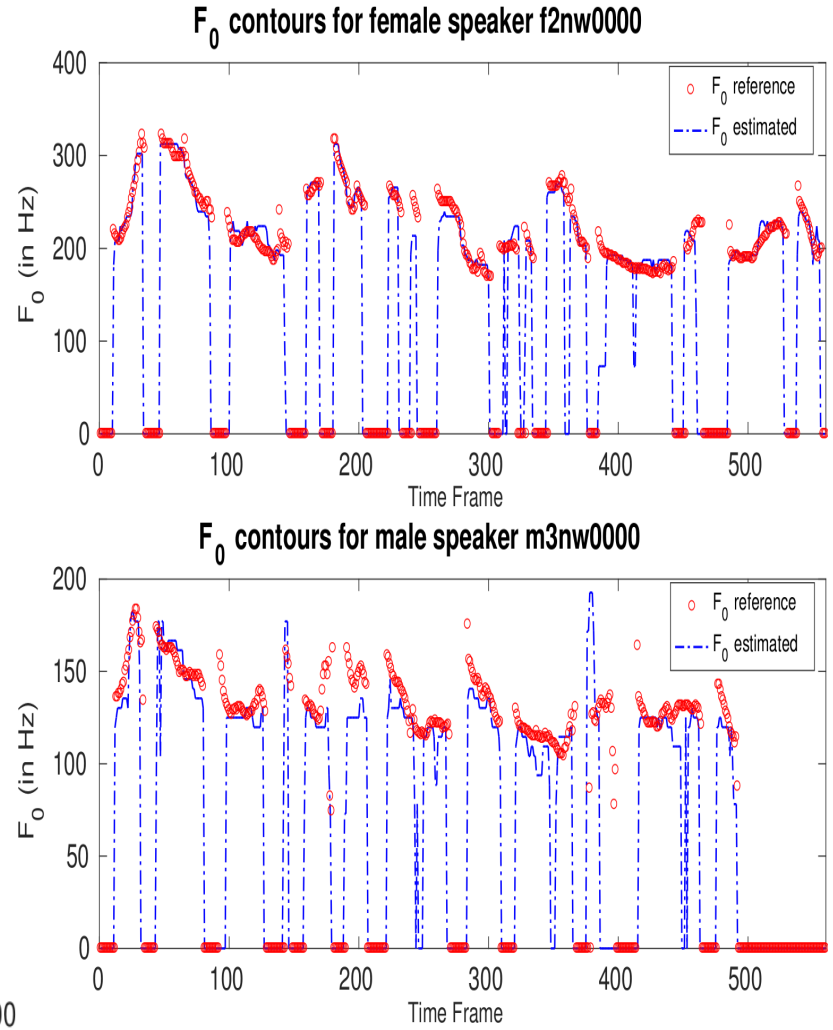
Sub-segmental modeling

Filter level analysis

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



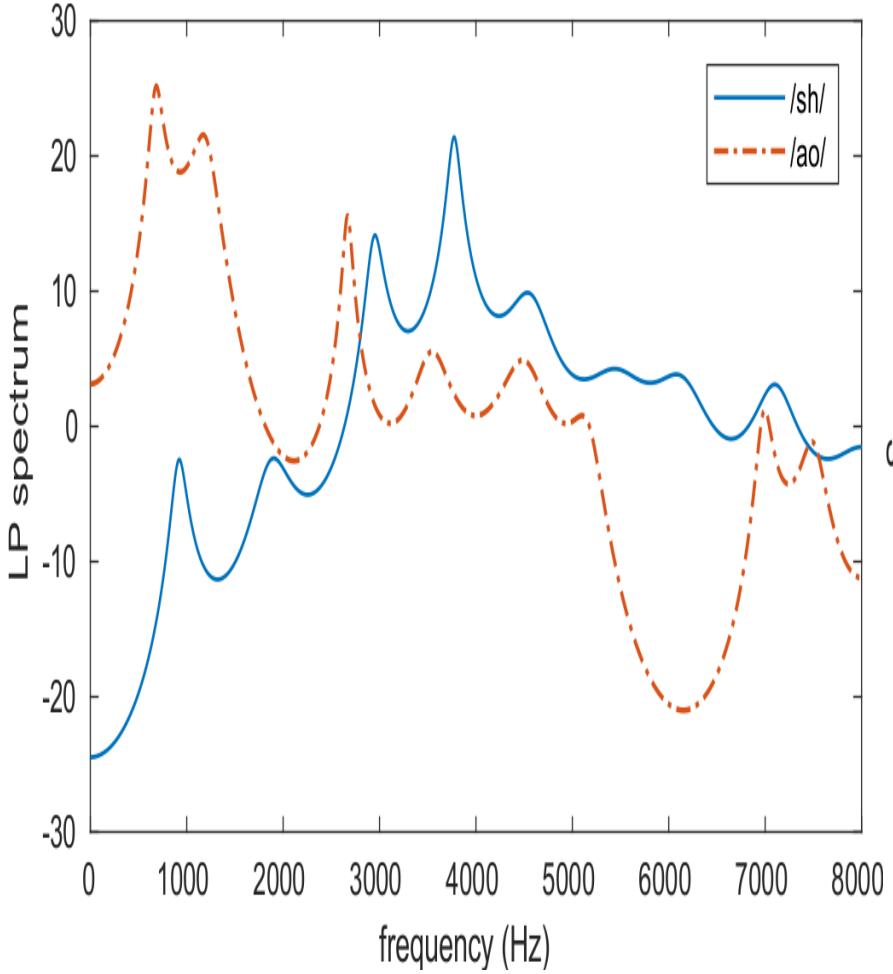
$|\mathcal{X}|$



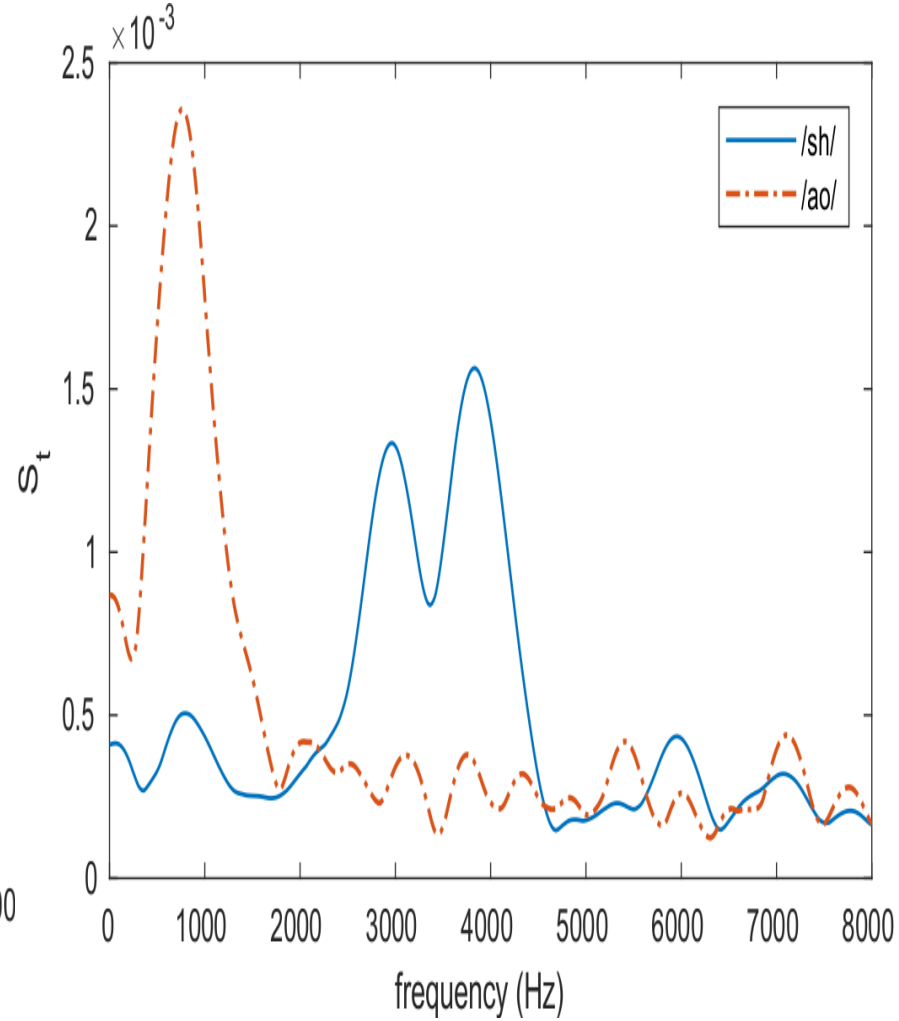
F0 contour estimated on Keele pitch database using the CNN-based speaker classifier trained on Voxforge.

Filter level analysis

Speaker recognition: spectral response $\mathcal{X}$ *of a frame of speech*. (Sub-segmental modeling)

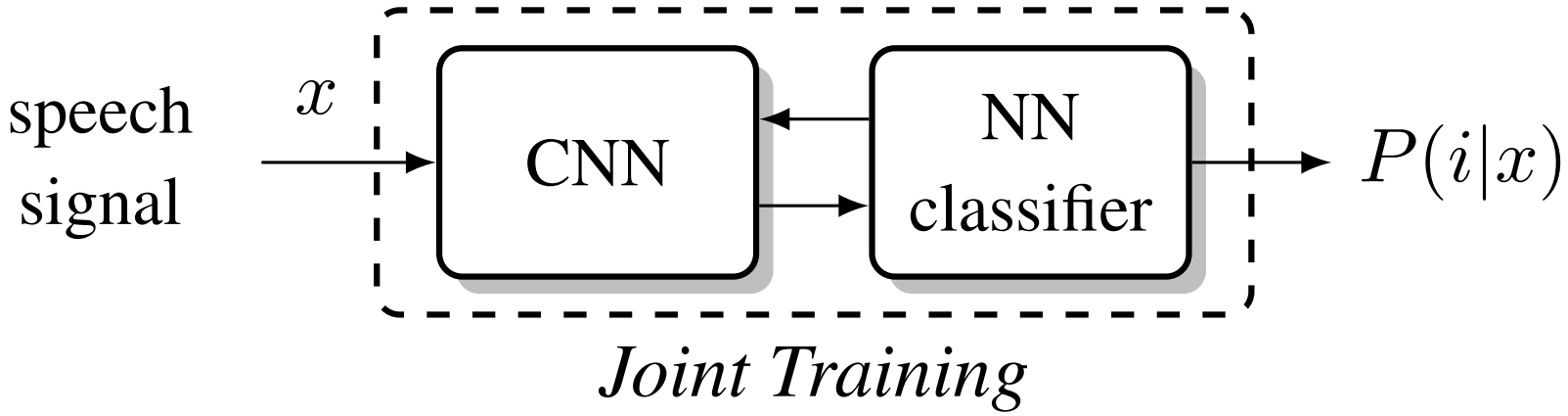


LP Spectrum



$|\mathcal{X}|$

In this talk



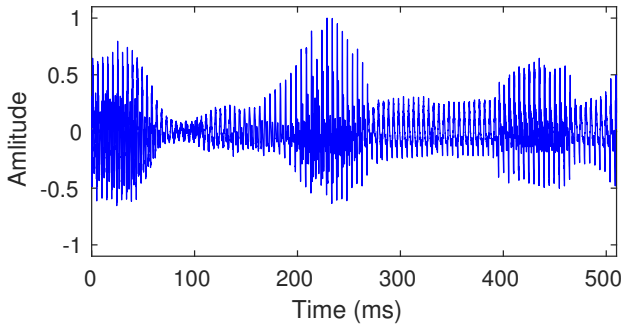
What information does such systems learn?

- Filter level analysis
- **Whole network level analysis**

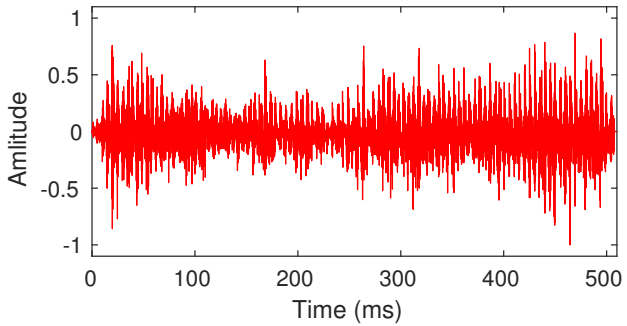
Whole network analysis

Gradient-based visualization

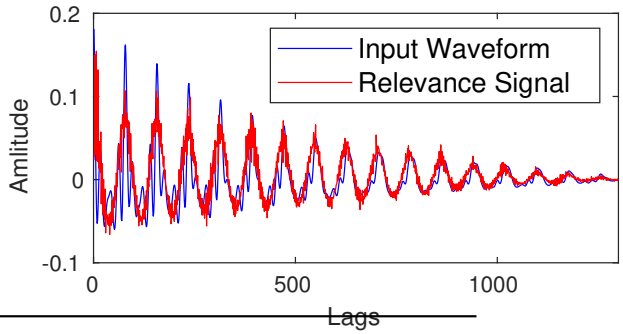
- In computer vision research, given an input image-output class pair and the trained system, finding contribution of each pixel in the image on the output score. (*guided backpropagation*)
- Given an input speech-output class pair and the trained system, what is the contribution of each sample on the output score?[8]



Original Signal



Relevance signal

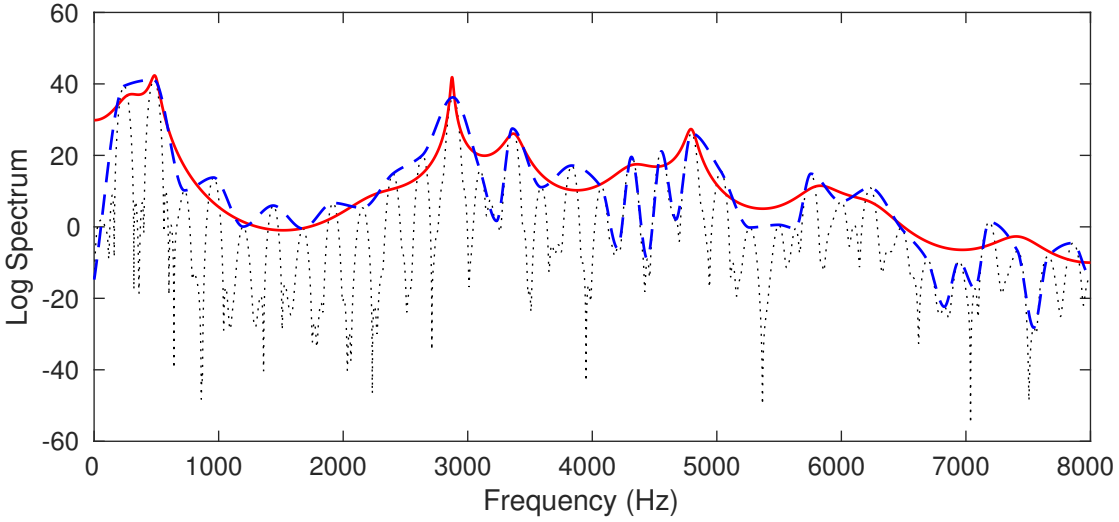


Autocorrelation

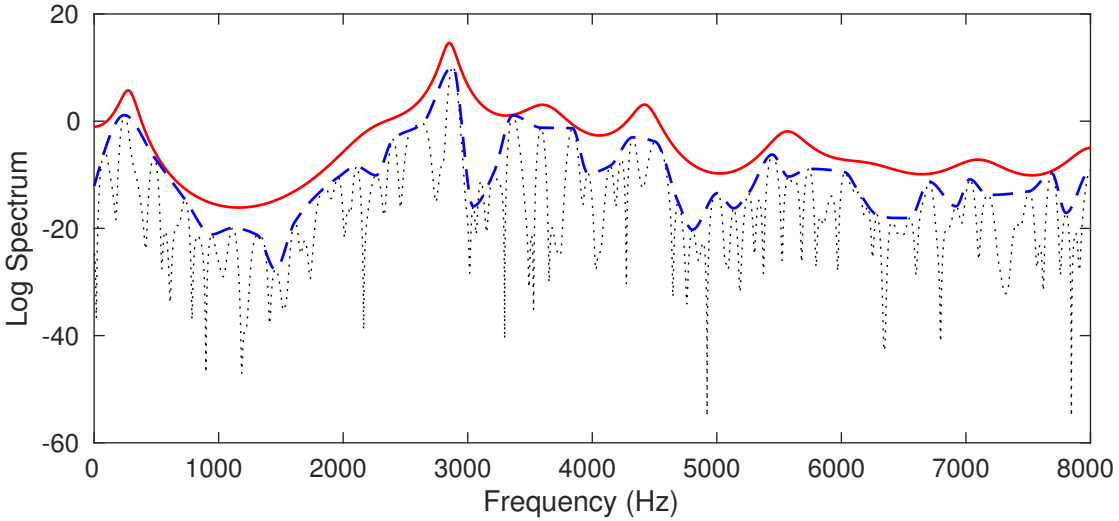
[8] H. Muckenhirn et al.,"Gradient-based spectral visualization of CNNs using raw waveforms ," Idiap Research Report Idiap-RR-11-2018, 2018. (submitted to SLT 2018)

Whole network analysis

Case study on speech recognition (1)



Original Spectrum of /iy/



Relevance signal spectrum of /iy/

Whole network analysis

Case study on speech recognition (2)

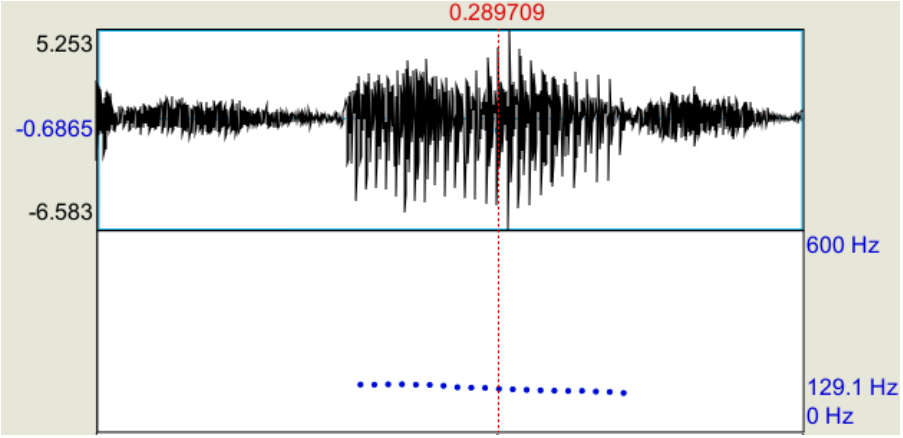
- Analysis of CNN trained on TIMIT phone recognition task on American English Vowel (AEV) dataset
- F0, F1 and F2 estimated automatically for the relevance signal for the steady state regions and compared to the values specified on the original study.

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.

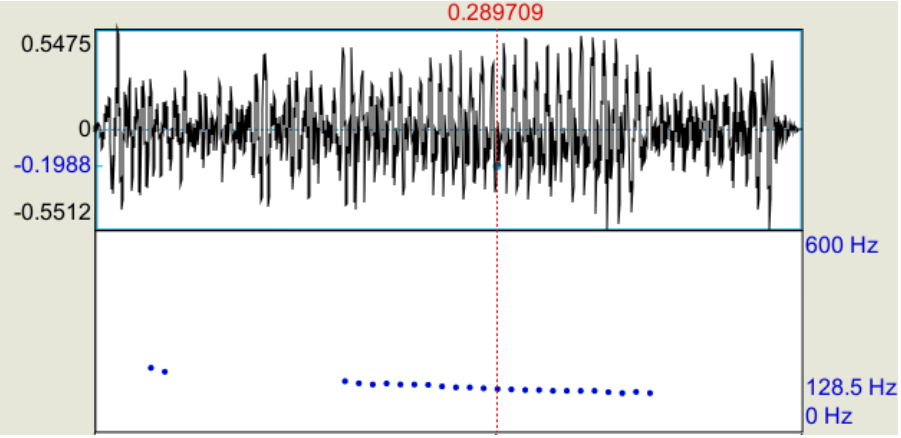
		/ah/	/eh/	/iy/	/oa/	/uw/
F0	F	93	91	91	94	92
	M	92	90	89	93	90
F1	F	90	92	93	91	93
	M	88	92	92	89	93
F2	F	94	94	94	95	94
	M	94	93	94	94	93

Whole network analysis

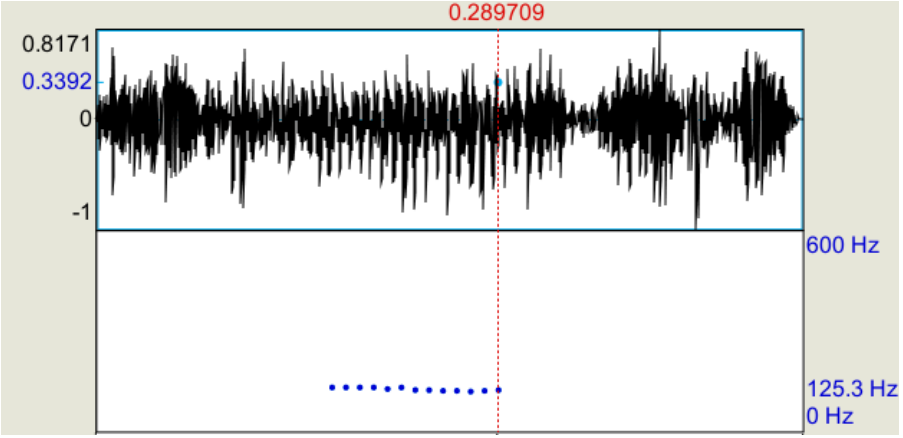
Case study on speaker recognition (1)



Original signal



Segmental modeling

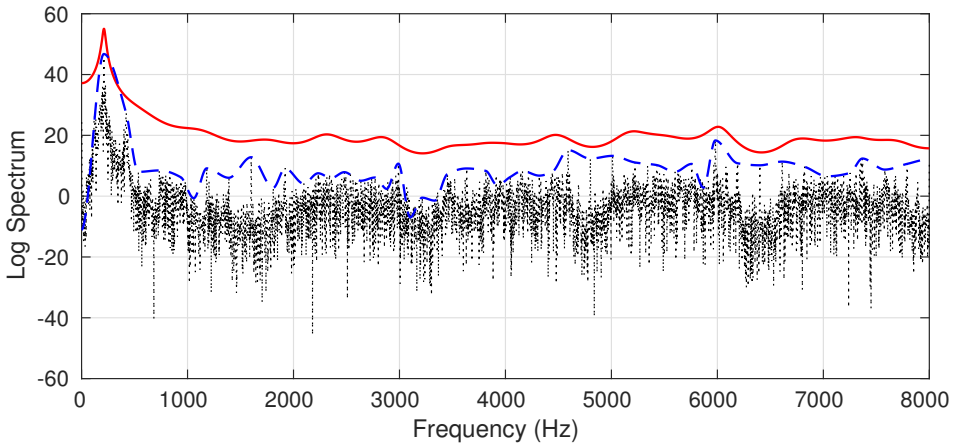


Sub-segmental modeling

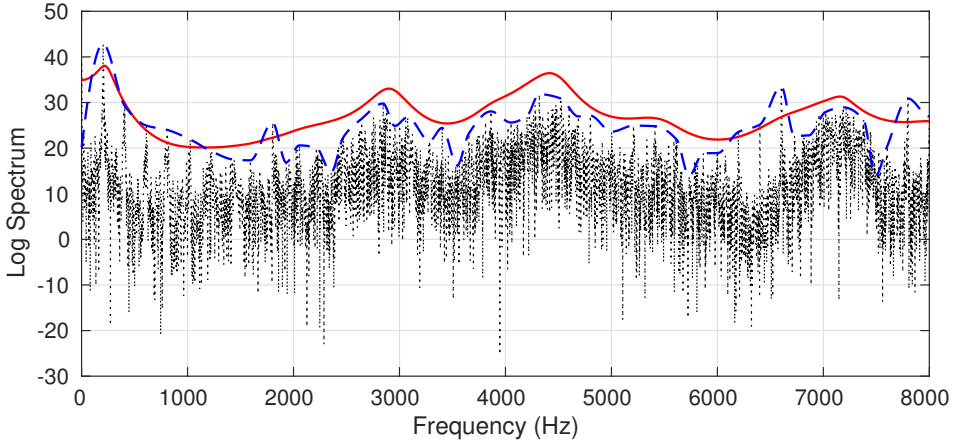
Whole network analysis

Case study on speaker recognition (2)

Utterance level average spectrum



Segmental modeling



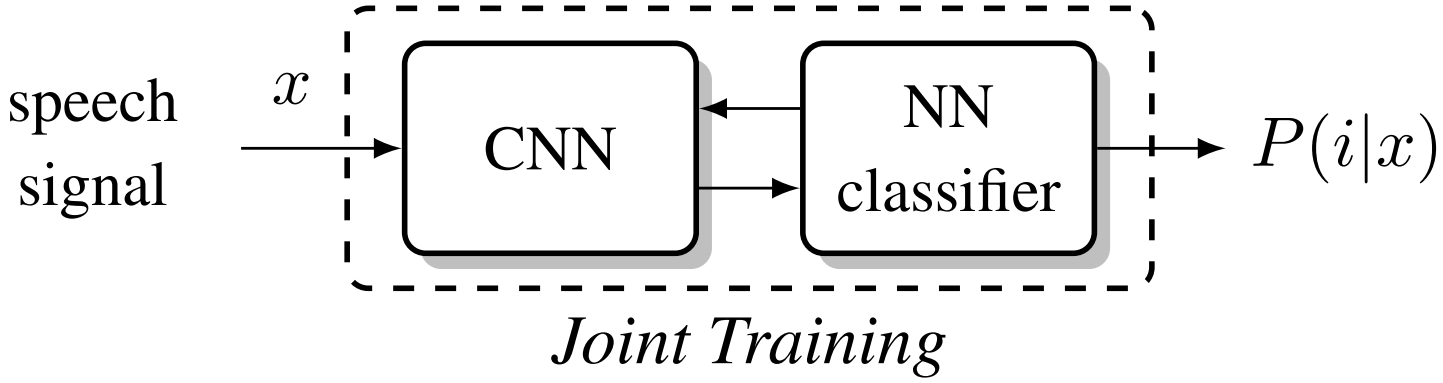
Sub-segmental modeling

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

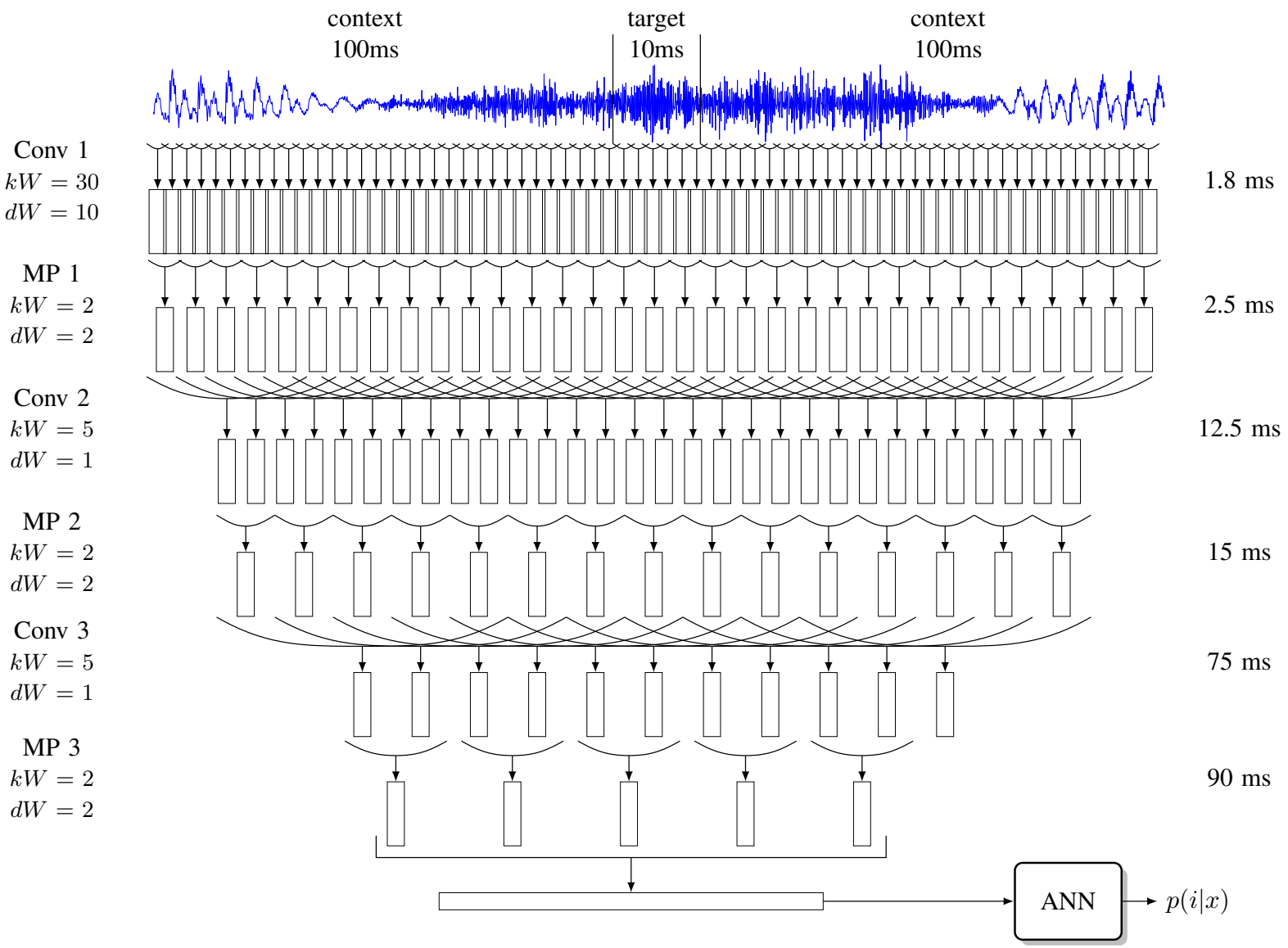
The End

Thank you for your attention!

Questions ?

CNN-based system using raw speech as input

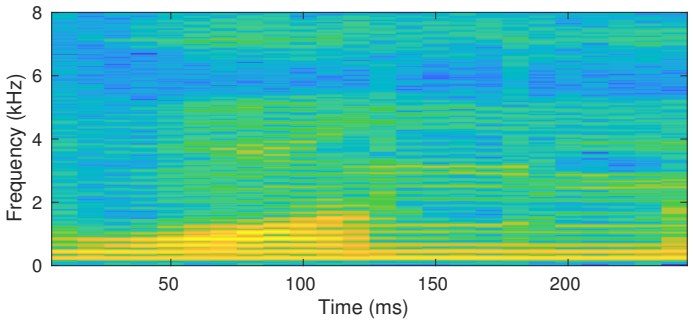
Detailed view for one example



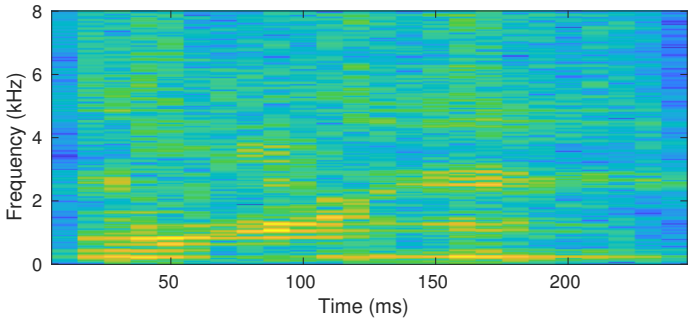
Whole network analysis

Speech recognition versus Speaker recognition

Original signal spectrogram



Phone CNN relevance signal spectrogram



Speaker CNN relevance signal spectrogram