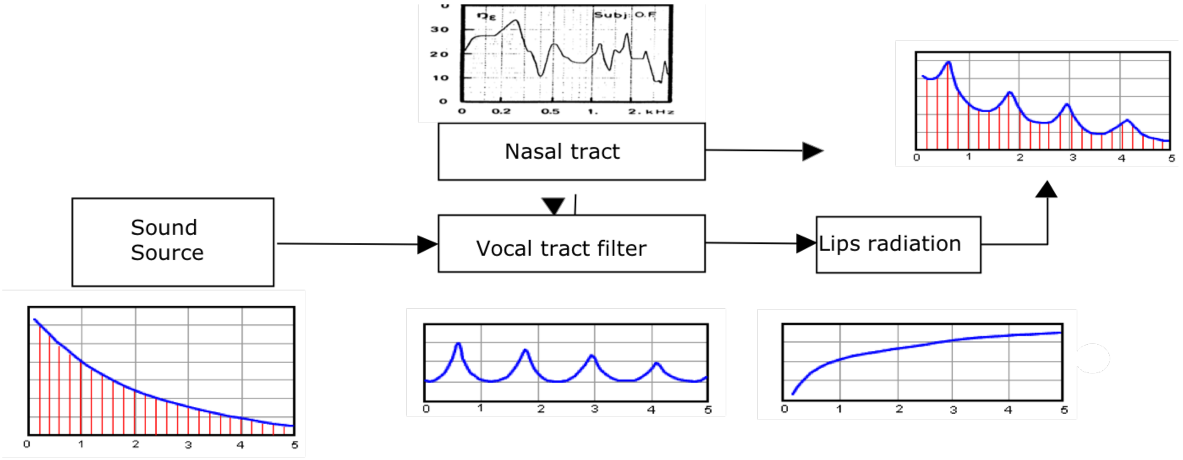
Feature/Representation Learning Overview

Conventional feature extraction

1. Quasi-stationarity (windowing, time-frequency resolution)
   - window size, typically 20-30 ms
   - window shift, typically 5-10 ms
2. Speech production knowledge

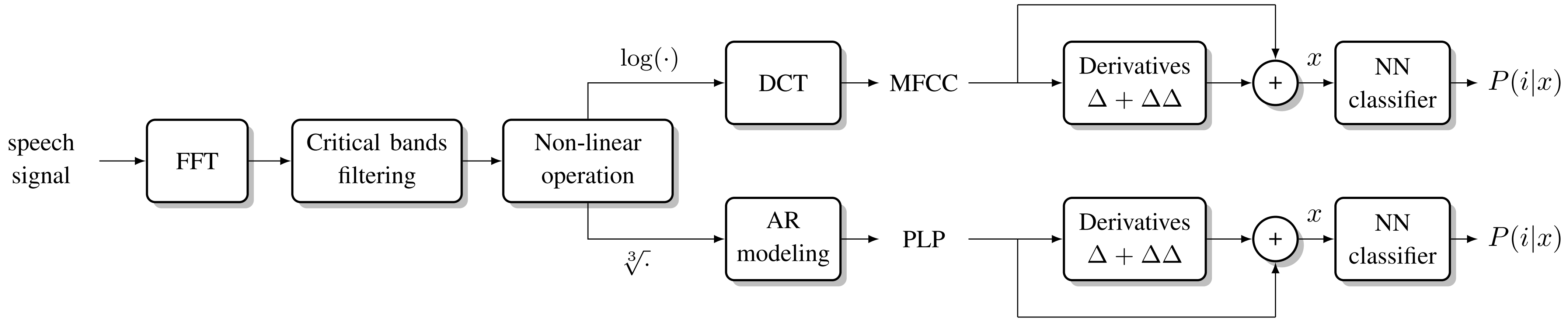


Credits: Lindqvist-Gauffin, Sundberg, Stevens, Mannel

3. Sound perception knowledge
   - Auditory modeling, such as, critical bands (filterbanks applied on spectrum), non-linear compression, equal loudness curve weighting

Conventional Acoustic Modeling

- Conventional cepstral features extraction process and modeling:



Note: Neural networks is an example classifier here. There are several types of classifiers. Similarly, classification is a pattern recognition problem. There are other types of pattern recognition problems

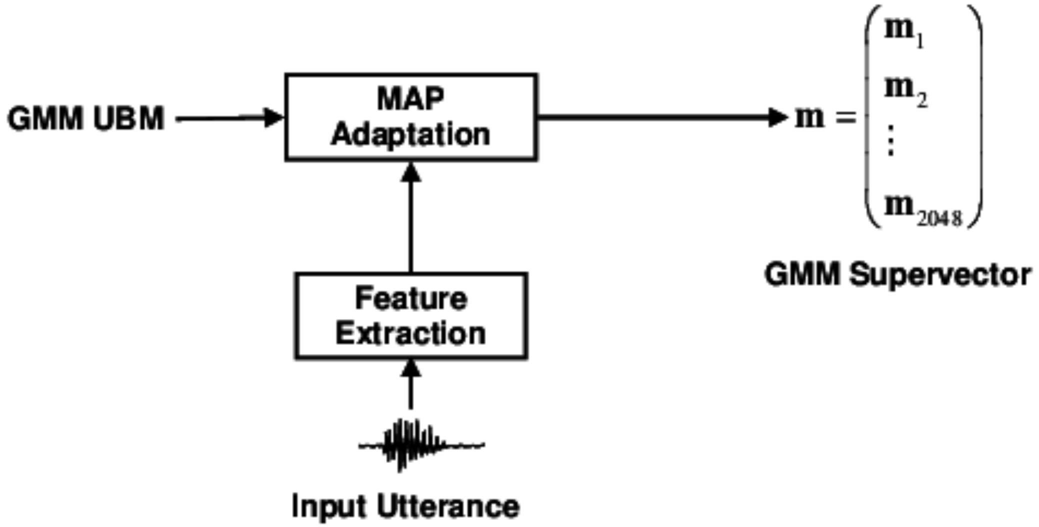
Clustering-based feature representation (1)

- Cluster the short-term hand-crafted feature vectors such as, cepstral feature, linear prediction coefficients using k-Means, Gaussian mixture modeling (GMM)
- Typically use the parameters as representations
- Examples:
  - Isolated word recognition by clustering isolated word patterns
  - Speaker verification using GMM supervectors

GMM UBM (universal background model) is trained on lots of "unseen" speakers data

MAP denotes Maximum Aposteriori

m_j denotes mean vector of Gaussian mixture j

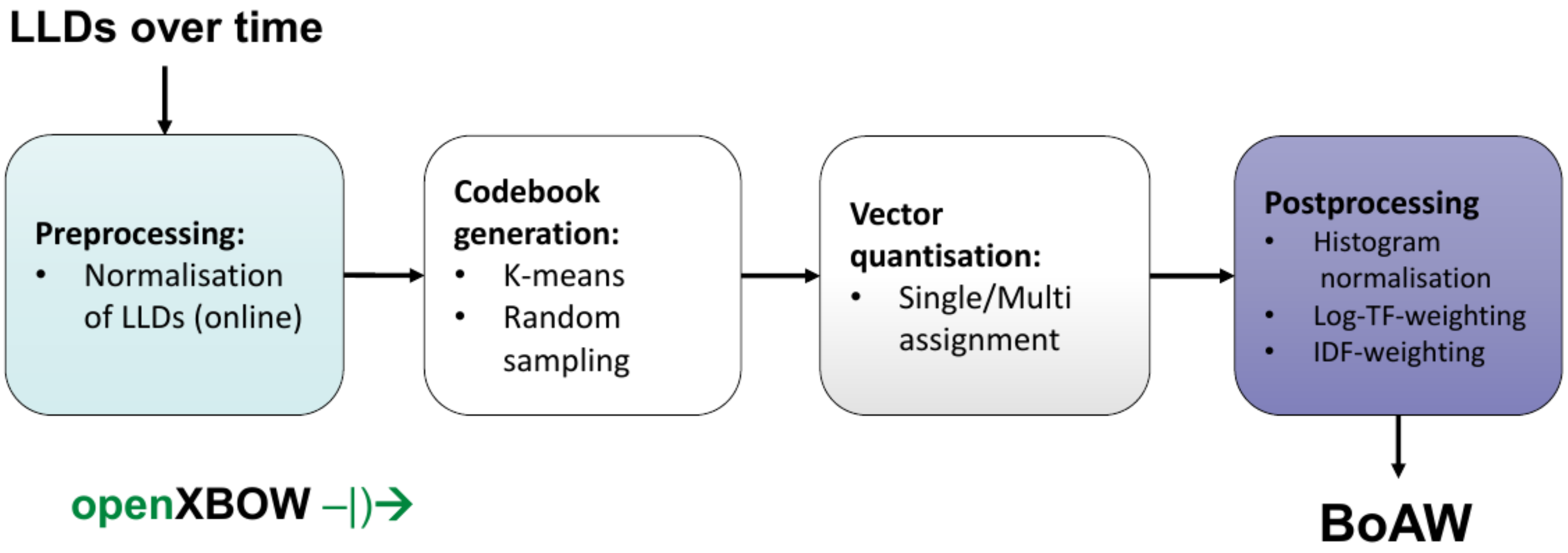


Low dimensional "i-vector" representation builds on top of GMM supervector by applying factor analysis

Source: Joe Campbell

Clustering-based feature representation (2)

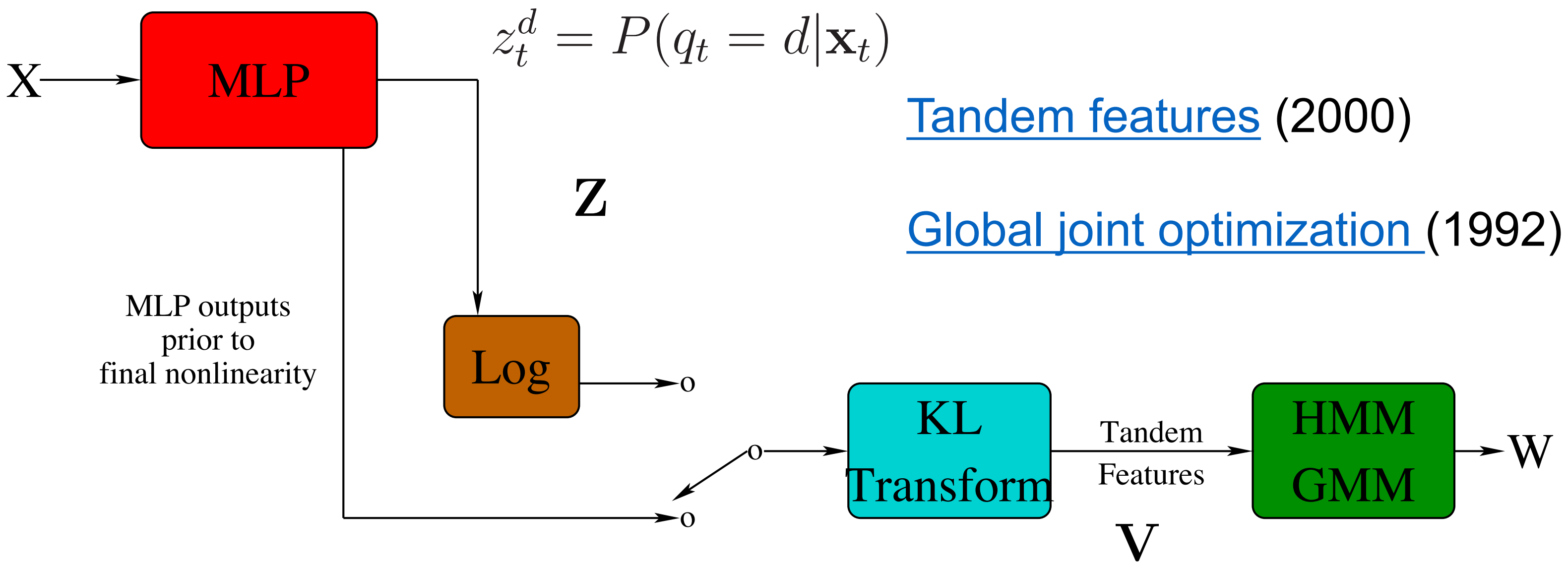
Bag-of-Audio-Words (BoAW) representation for paralinguistic speech processing, such as, emotion classification, affective rating, atypical speech detection/classification.



LLDs denote short-term frame level features, e.g., MFCCs, F0, Formants

Supervised learning-based representations (1)

Neural network-based features for speech recognition



Tandem features (2000)

Global joint optimization (1992)

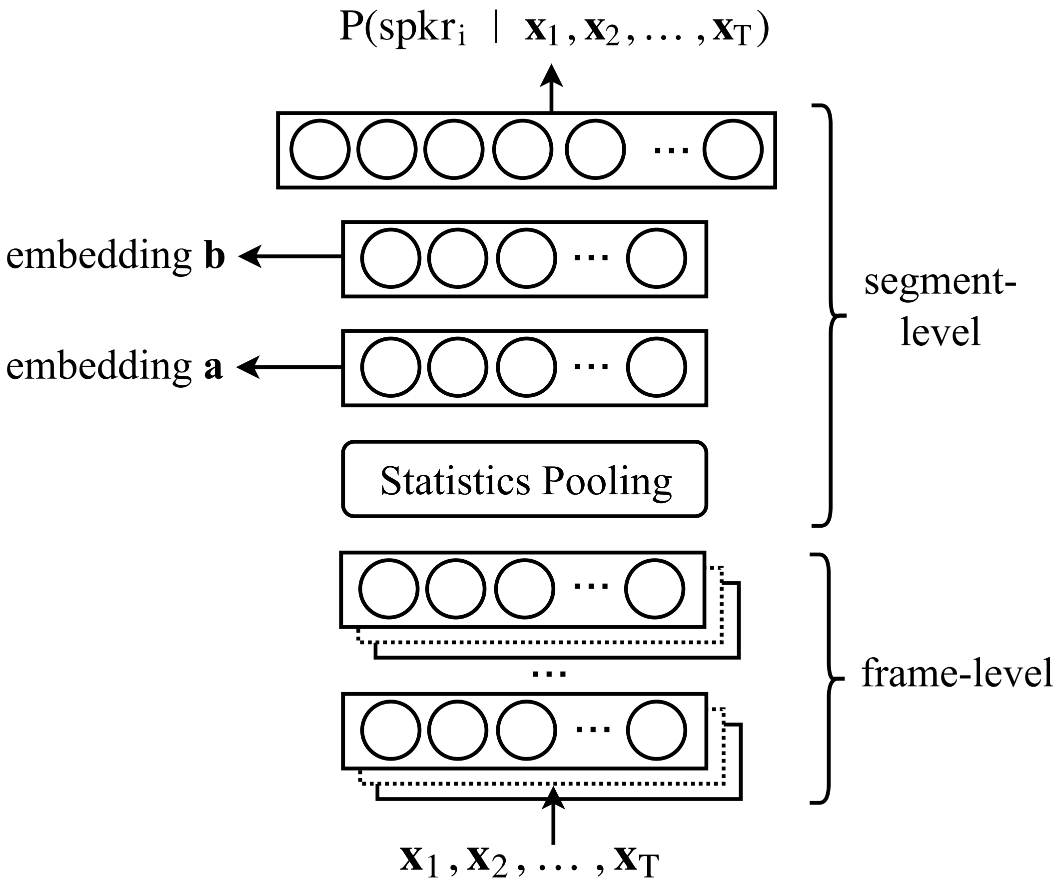
X denotes sequence of cepstral/spectral feature vectors

- Output layer representation
- Hidden layer representation
- Bottleneck representation

- GMMs
  - Generative approach
  - Easy parameter adaptation methods
  - Effectively model context-dependent phone units (late 1990s-early 2000s)
  - Sophisticated tools
- ANNs
  - Discriminative approach
  - No prior assumption about data
  - Modeling long temporal context relatively easy, enabling integrating auditory processing knowledge

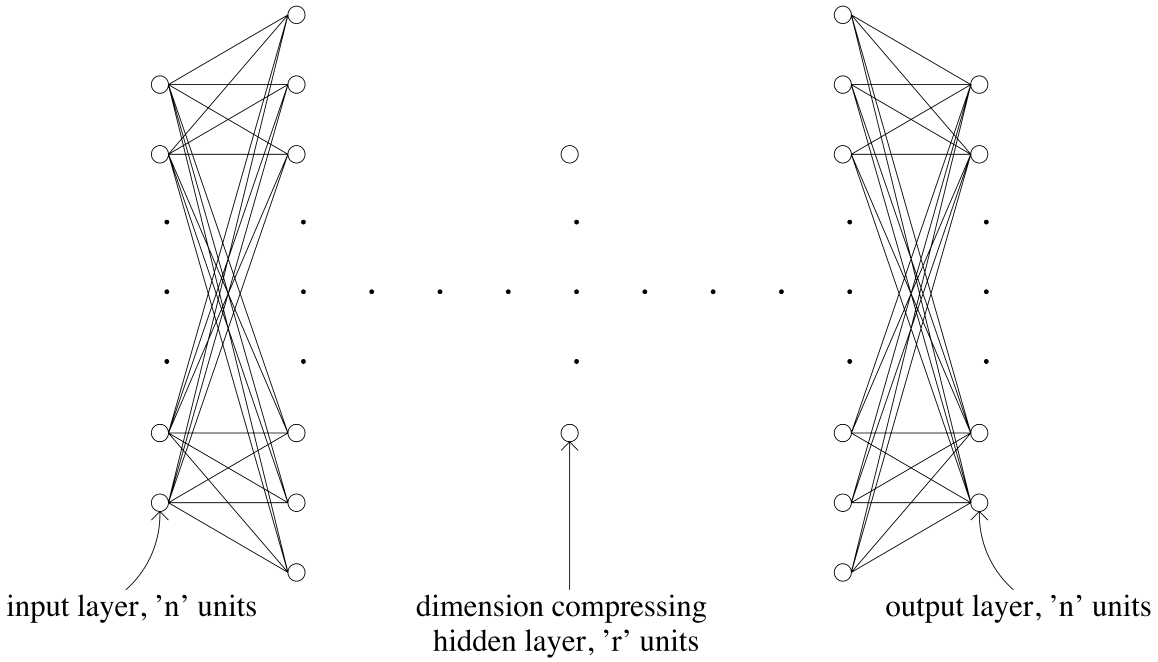
Supervised learning-based representations (2)

Neural features for speaker recognition task.



- Snyder et al. Deep Neural Network Embeddings for Text-Independent Speaker Verification, in Proc. of Interspeech 2017
- Snyder et al. X-vectors: Robust DNN Embeddings For Speaker Recognition Speaker Recognition, in Proc. of Interspeech 2018.

Auto-encoding/Auto-association



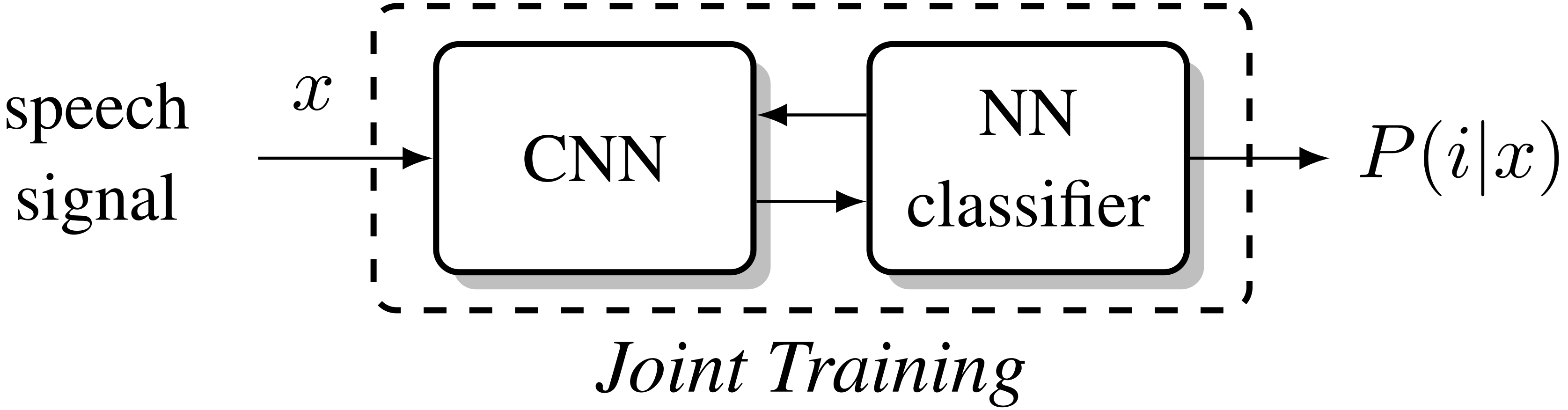
- Can be interpretted as principal component analysis
- With only one hidden layer, equivalent to singular value decomposition (SVD)
- Can do better than standard SVD with different topologies (e.g., more hidden layers)
- Different types auto-encoders, e.g., variational auto-encoders (VAE), vector quantization VAE (VQ-VAE), denoising autoencoders

Bourlard and Kemp, Auto-association by multilayer perceptrons and singular value decomposition, Biological Cybernetics, 1988

Ikbal, Misra and Yegnanarayana, Analysis of autoassociative neural networks, in Proc. of IJCANN, 1999

Bourlard and Kabil, Autoencoders reloaded, Biological Cybernetics, 2022.

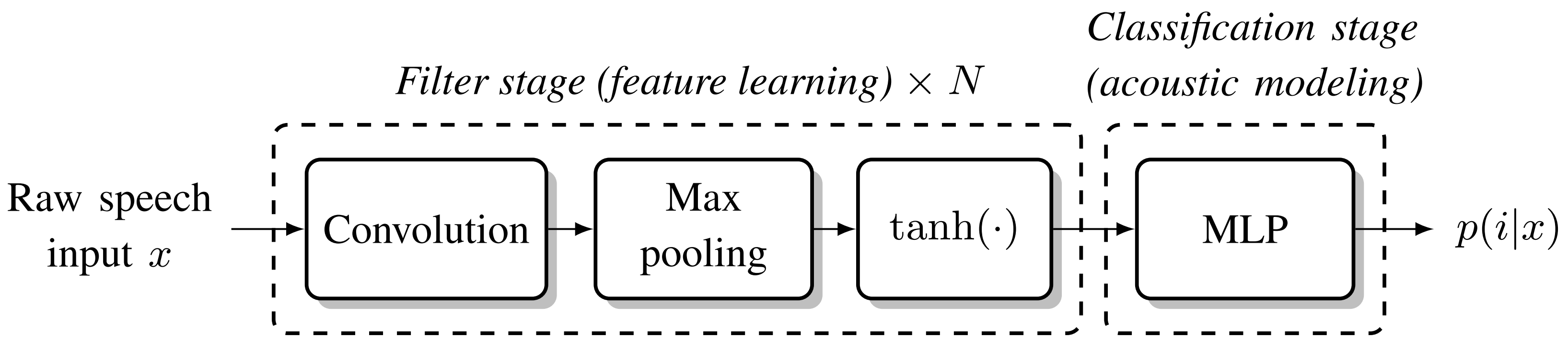
End-to-end acoustic modeling using CNNs (1)



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

Palaz, Magimai-Doss and Collobert,"End-to-end acoustic modeling using convolutional neural networks for HMM-based automatic speech recognition", Speech Communication, 2019
D. Palaz,"Towards End-to-End Speech Recognition", EPFL PhD Thesis, 2016

End-to-end acoustic modeling using CNNs (2)

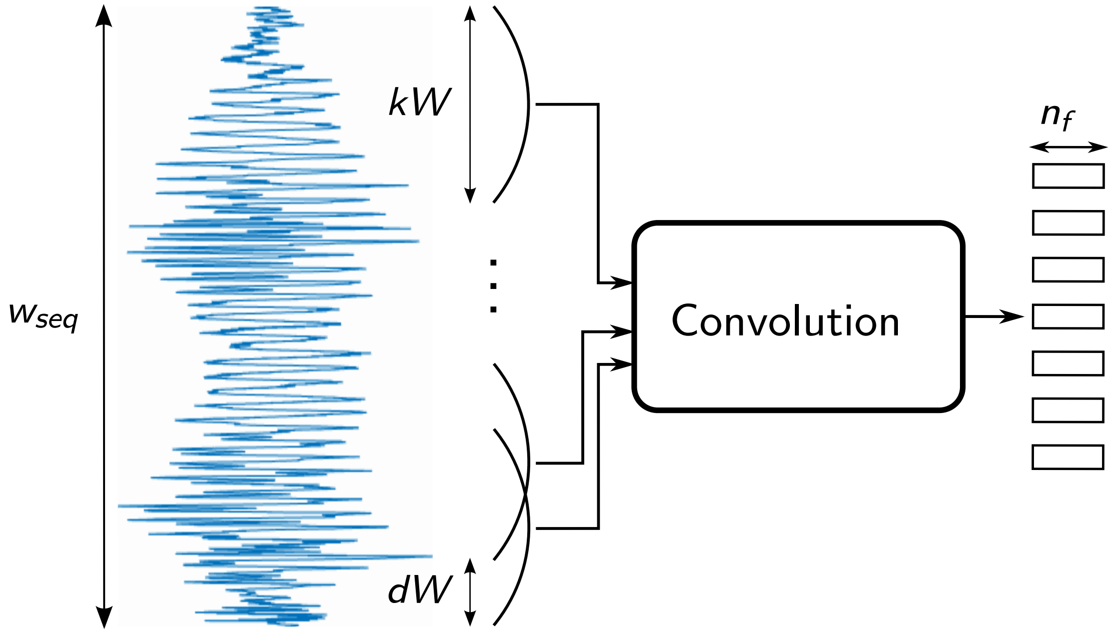


Minimal prior knowledge

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

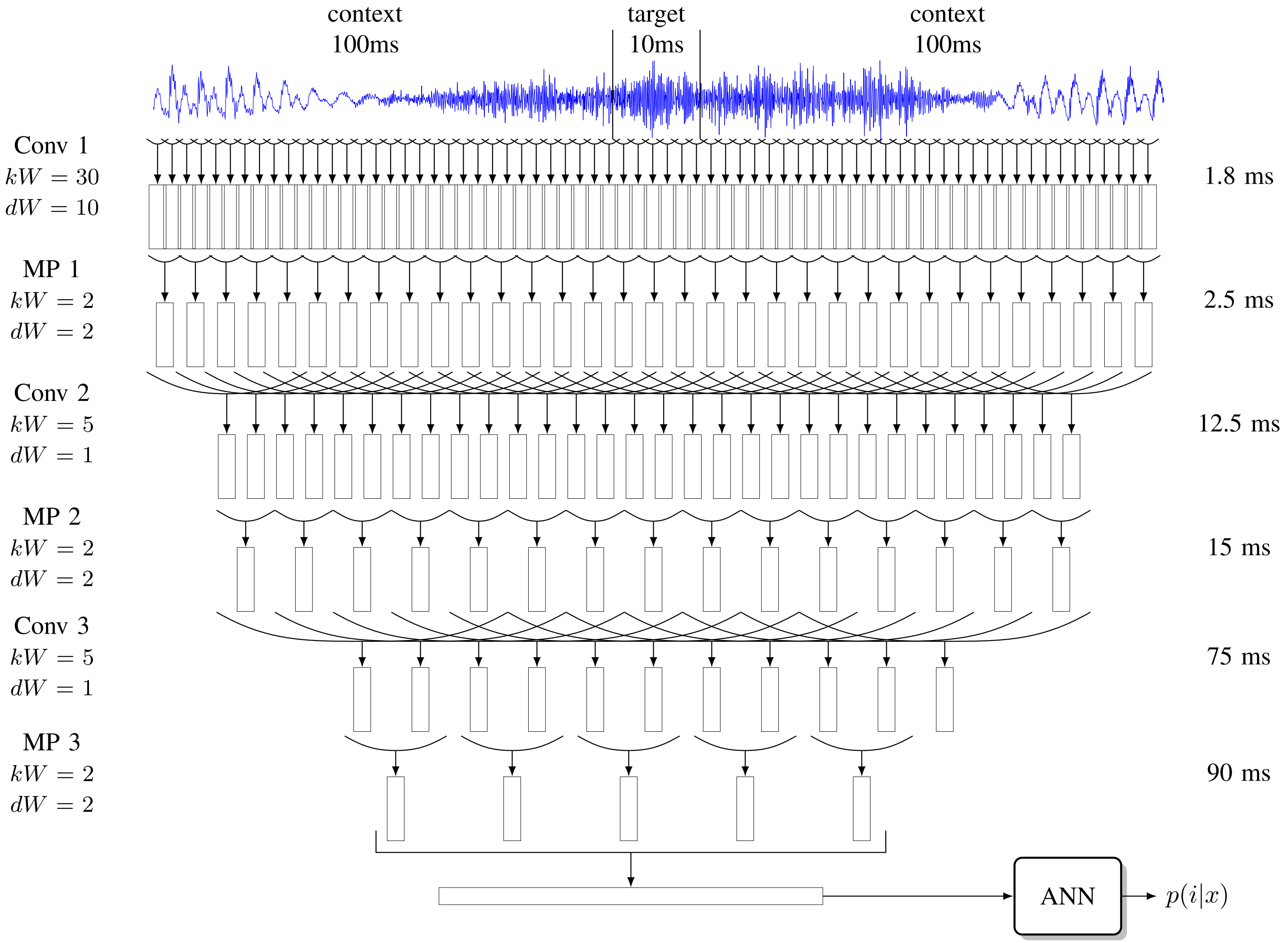
Determined in a data-driven manner.
All the stages are trained jointly using back-propagation with a cost function based on cross entropy.

First convulation layer illustration



- 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

Illustraton of CNN trained for speech recognition



Application	w_{seq}	kW	# of conv. layers	# of hidden layers
Speech reco.[7]	250-310 ms	sub-seg	3-5	1-3
Speaker reco.[8]	$\approx$ 500 ms - 2.5 s	seg, sub-seg	2-6	1
Presentation attack detection[9]	$\approx$ 300 ms	seg	2	1 or none
Gender reco.[10]	250-310 ms	seg, sub-seg	1-3	1
Paralinguistic[11,12]	250-500 ms	seg, sub-seg	3-4	1
Breathing Patt. Est.[13]	3-4 s	sub-seg	4	1

[7] Palaz, Magimai.-Doss, and Collobert, "End-to-end acoustic modeling using convolutional neural networks for HMM-based automatic speech recognition," Speech Communication, Vol. 108, April 2019, Pages 15–32.

[8] H. Muckenhirn,"Trustworthy speaker recognition with minimal prior knowledge using neural networks", PhD Thesis No. 7285, Ecole polytechnique fédérale de Lausanne (EPFL), Switzerland, 2019.

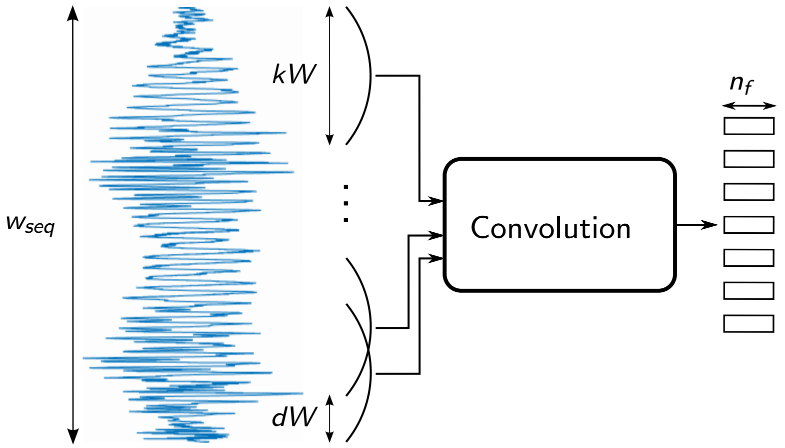
[9] Muckenhirn, Magimai-Doss, and Marcel, "End-to-End Convolutional Neural Network-based Voice Presentation Attack Detection," in Proceedings of the IEEE International Joint Conference on Biometrics (IJCB), 2017.

[10] Kabil, Muckenhirn, and Magimai.-Doss, "On learning to identify genders from raw speech signal using CNNs," in Proceedings of Interspeech, 2018.

[11] Purohit et al., "Towards learning emotion information from short segments of speech," in Proc. of ICASSP, 2023.

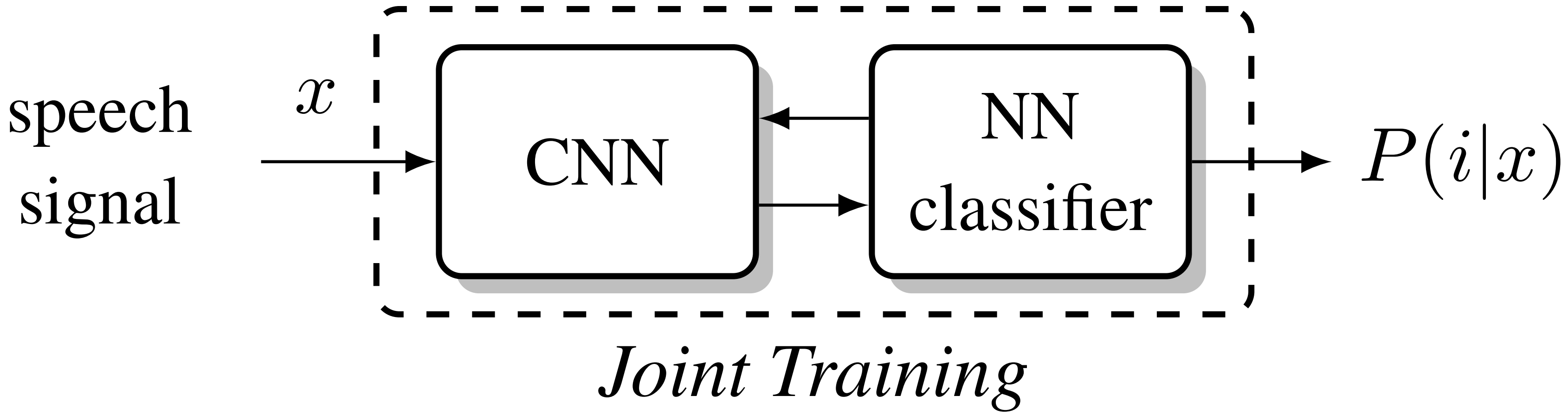
[12] Dubaganta, Vlasenko and Magimai.-Doss,"Learning voice source related information for depression detection", in Proc. of ICASSP, 2019.

[13] Nallanthigal et al., "Deep learning architectures for estimating breathing signal and respiratory parameters from speech recordings," Neural Networks, 2021.



- 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

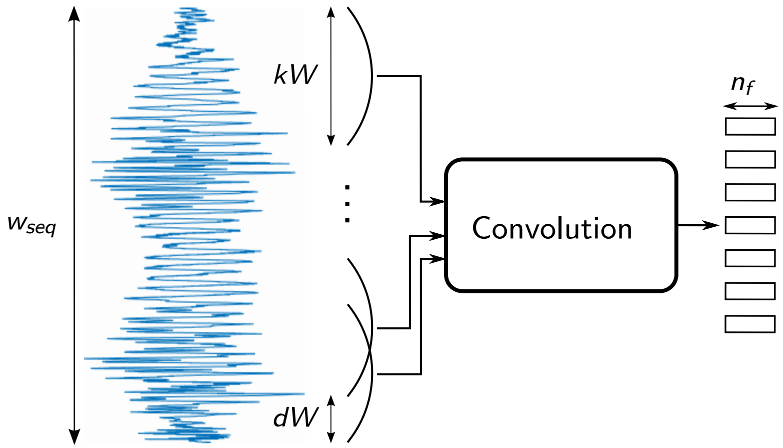
Central question



<u>What information does such systems learn?</u>

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

Filter analysis: first convolution layer

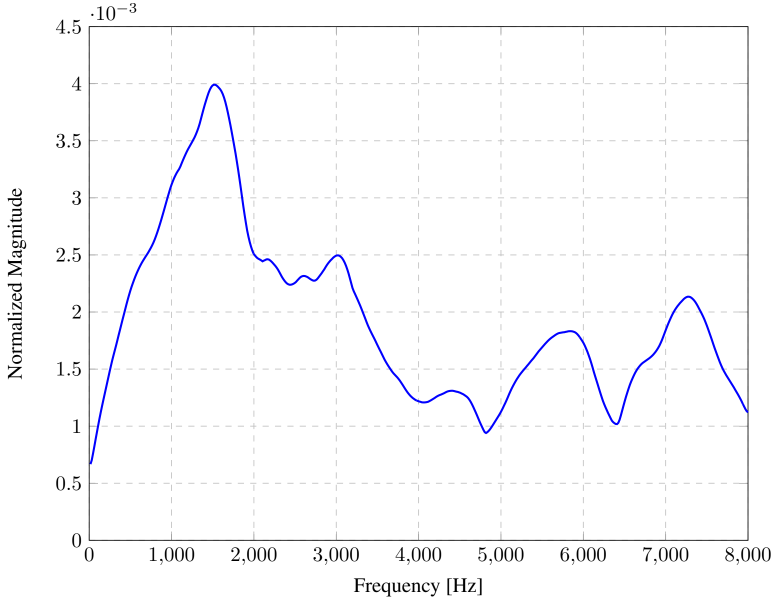


- w_{seq}: Input speech signal with temporal context
- kW: Window size
  - Sub-segmental (< 1 pitch period)
  - Segmental ($1 - 3$ pitch periods)
- dW: Window shift (< 1 pitch period)
- n_f: number of filters

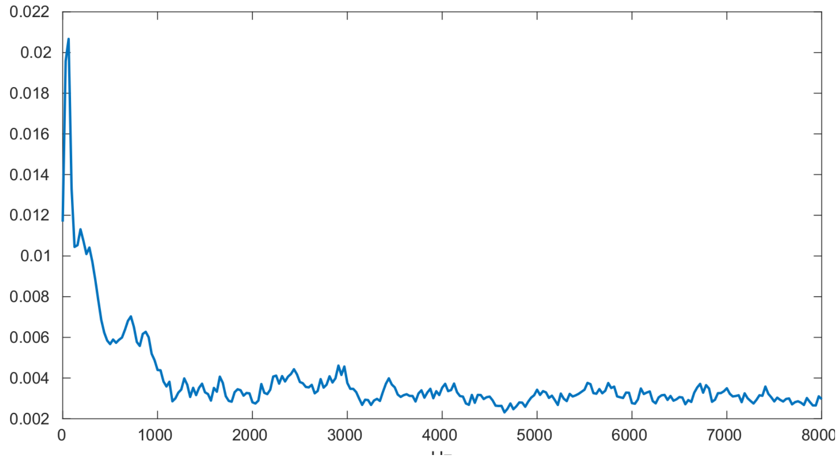
- Cumulative frequency response of filters

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

Speech recognition
kW = 1.8 ms



Speaker recognition
kW = 18 ms

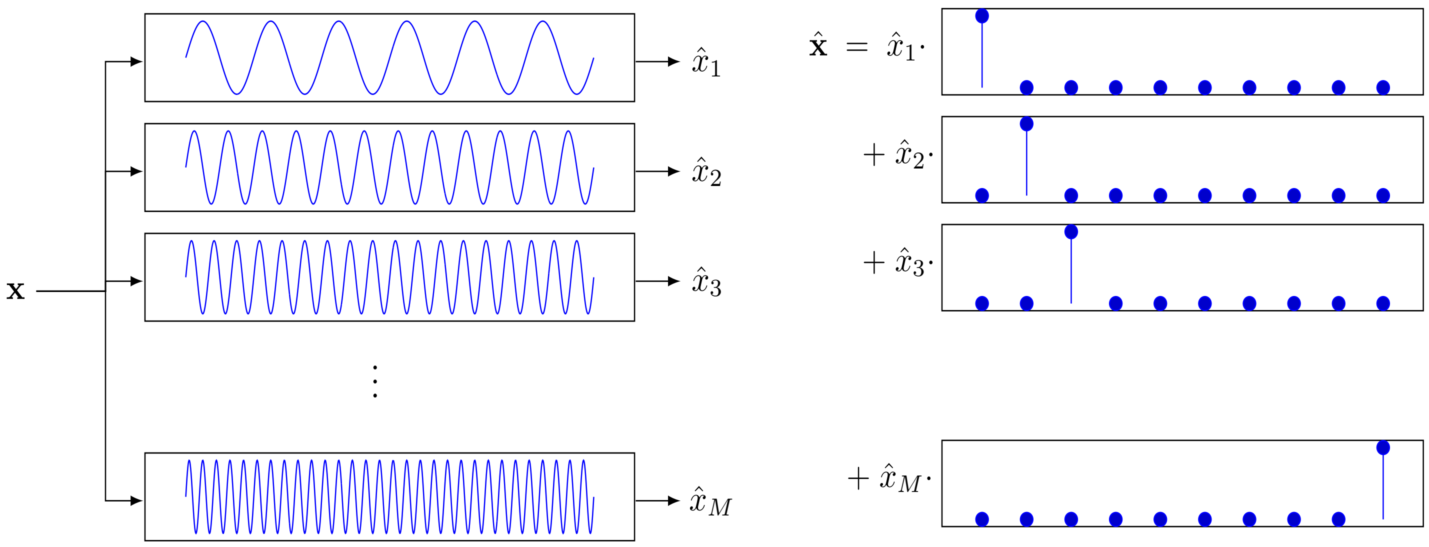


Between 1st and 2nd conv. layers: DFT analogy

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

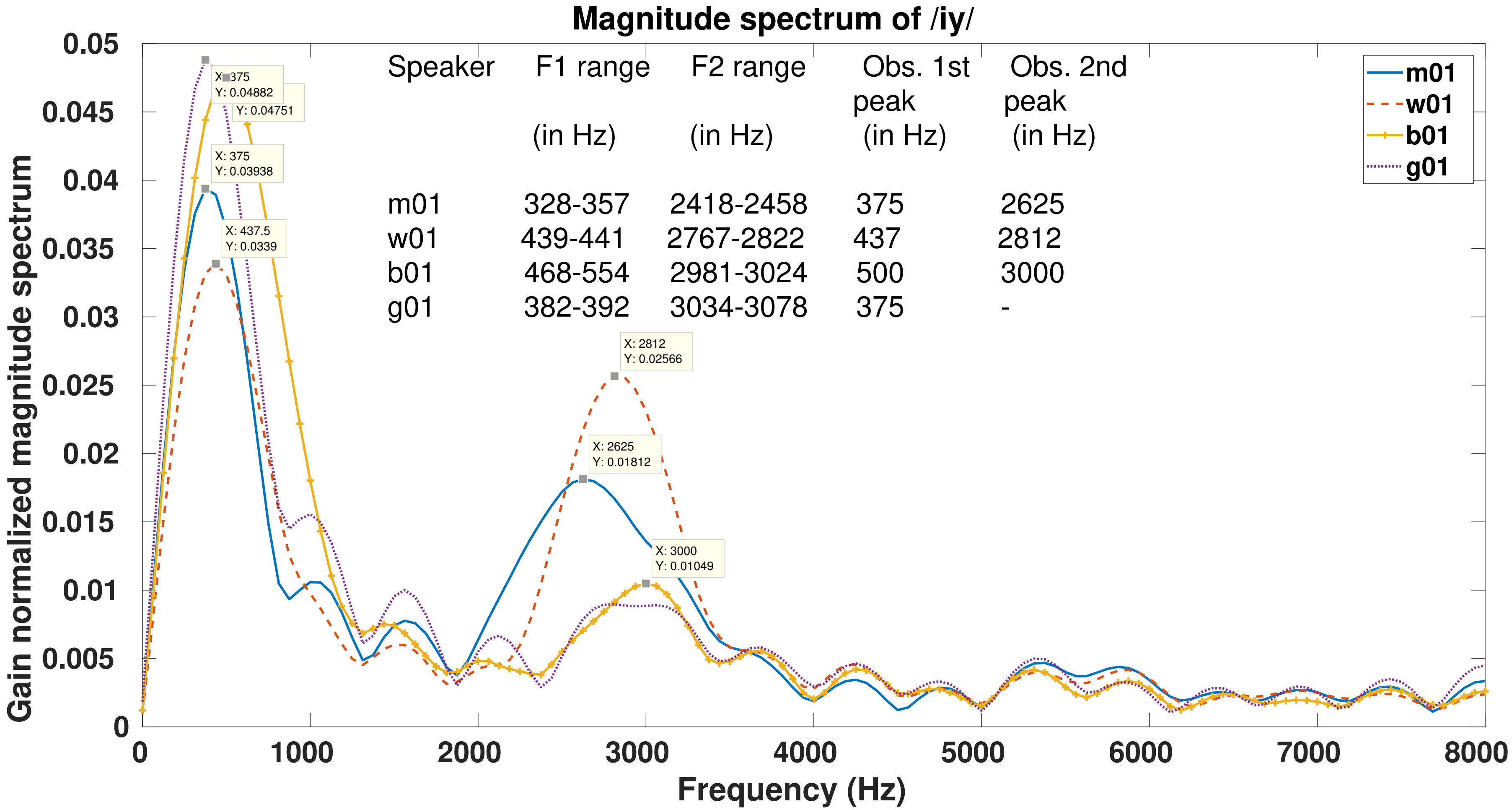
$$\mathcal{X} = \sum_{m=1}^{M} \langle \mathbf{x}, f_m \rangle \ \mathrm{DFT}[f_m],$$

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 and $kW = M$ then $\mathcal{X}$ is DFT of $\mathbf{x}$.



$$\mathcal{X} = \sum_{m=1}^{M} \hat{x}_m \cdot \mathcal{F}_m$$

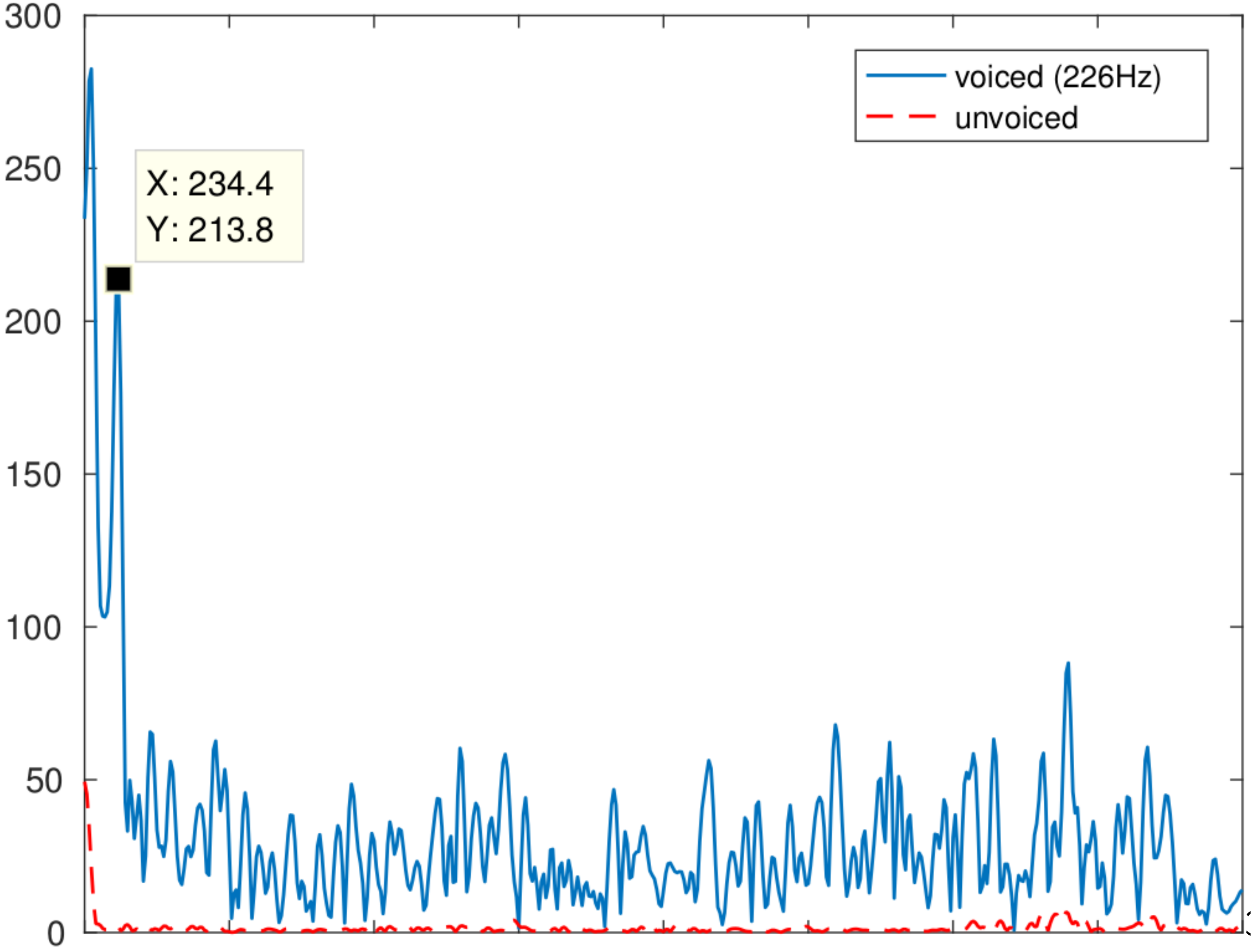
Between 1st and 2nd conv. layers: Speech Reco.



CNN trained on Wall Street Journal corpus

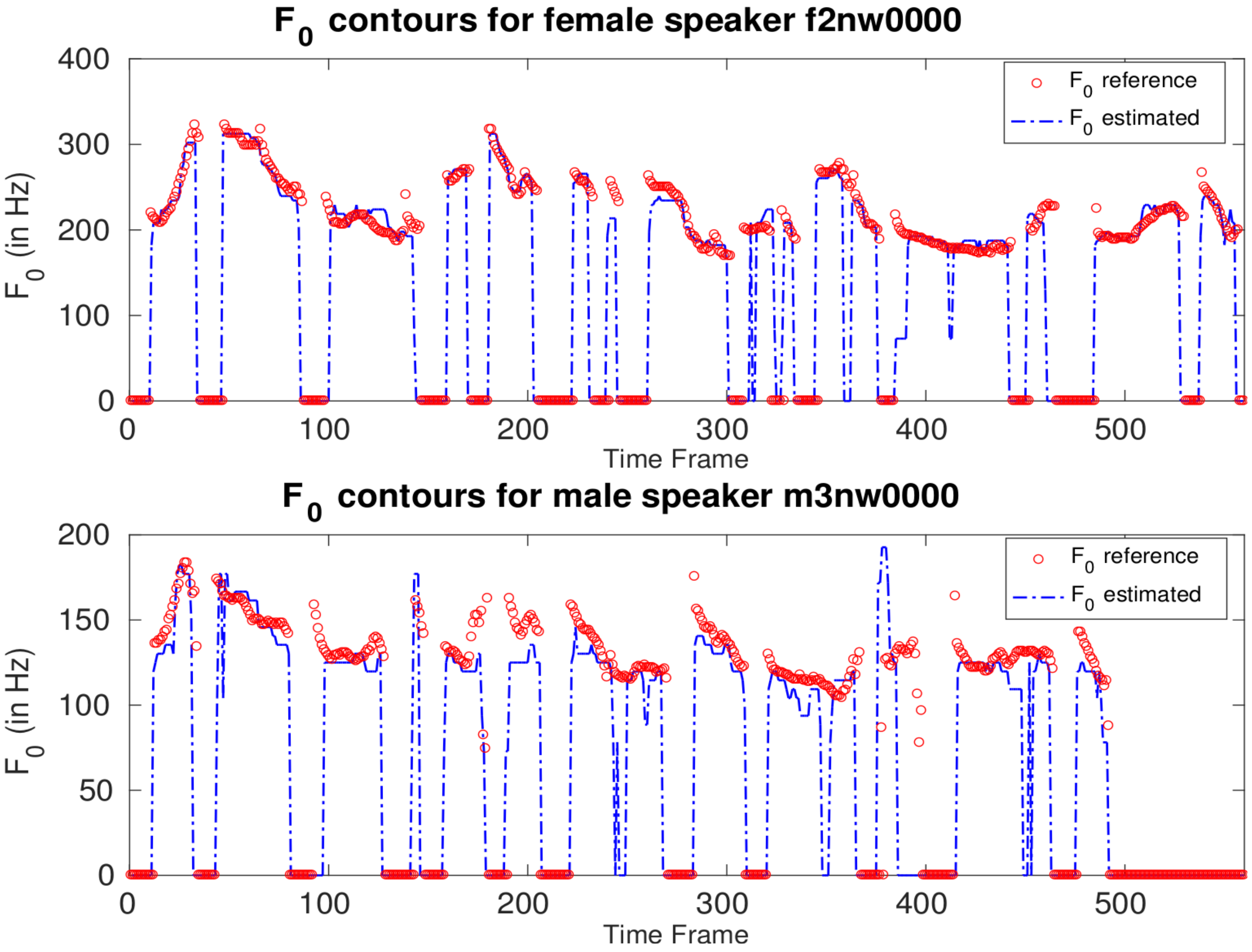
Spectral response of /iy/ from American English Vowel dataset.

Between 1st and 2nd conv. layers: Speaker Reco. (1)



kW =18 ms

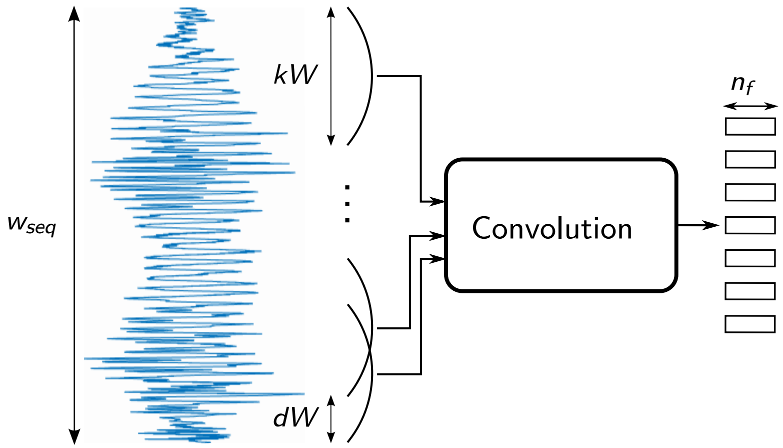
Between 1st and 2nd conv. layers: Speaker Reco. (2)



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

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

Filter analysis: first convolution layer

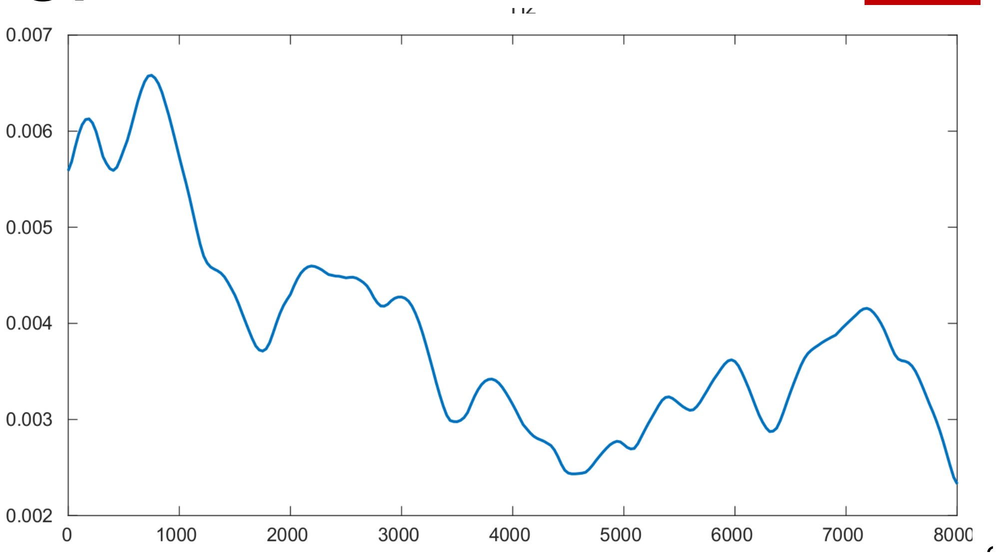


- w_{seq}: Input speech signal with temporal context
- kW: Window size
  - Sub-segmental ($<$ 1 pitch period)
  - Segmental ($1-3$ pitch periods)
- dW: Window shift ($<$ 1 pitch period)
- n_f: number of filters

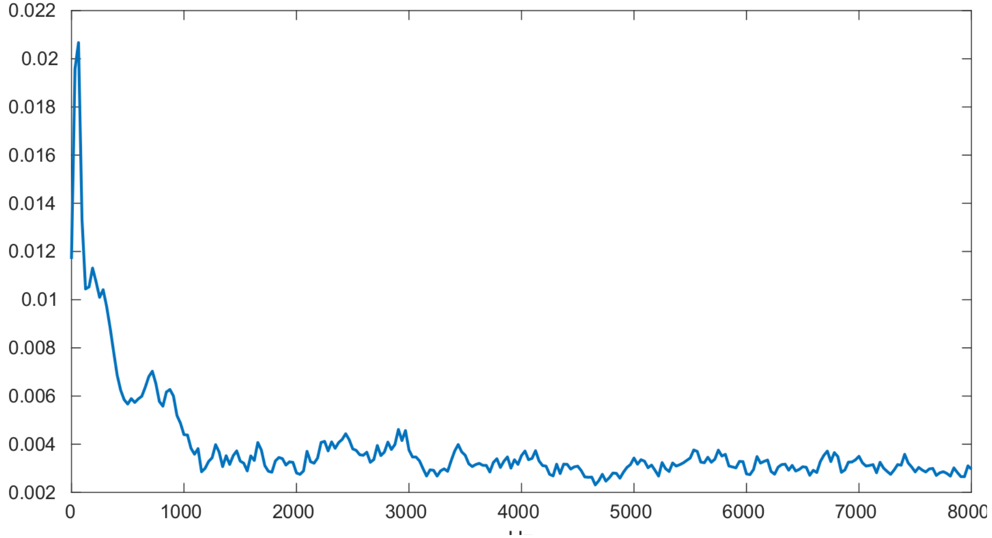
- Cumulative frequency response of filters

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

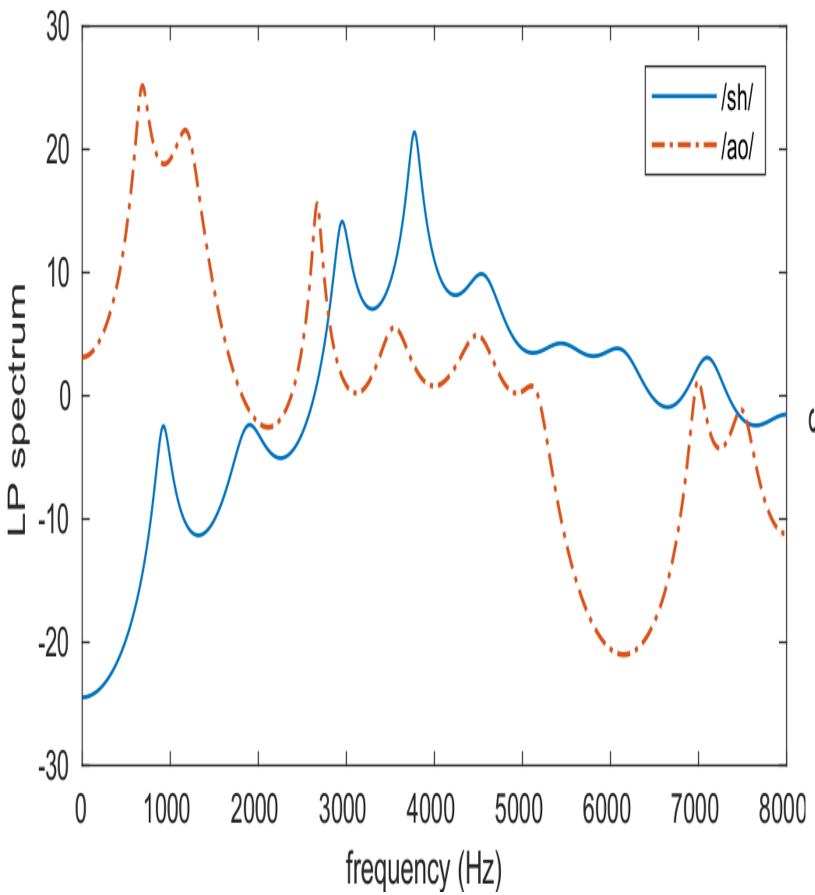
Speaker recognition
kW = 1.8 ms



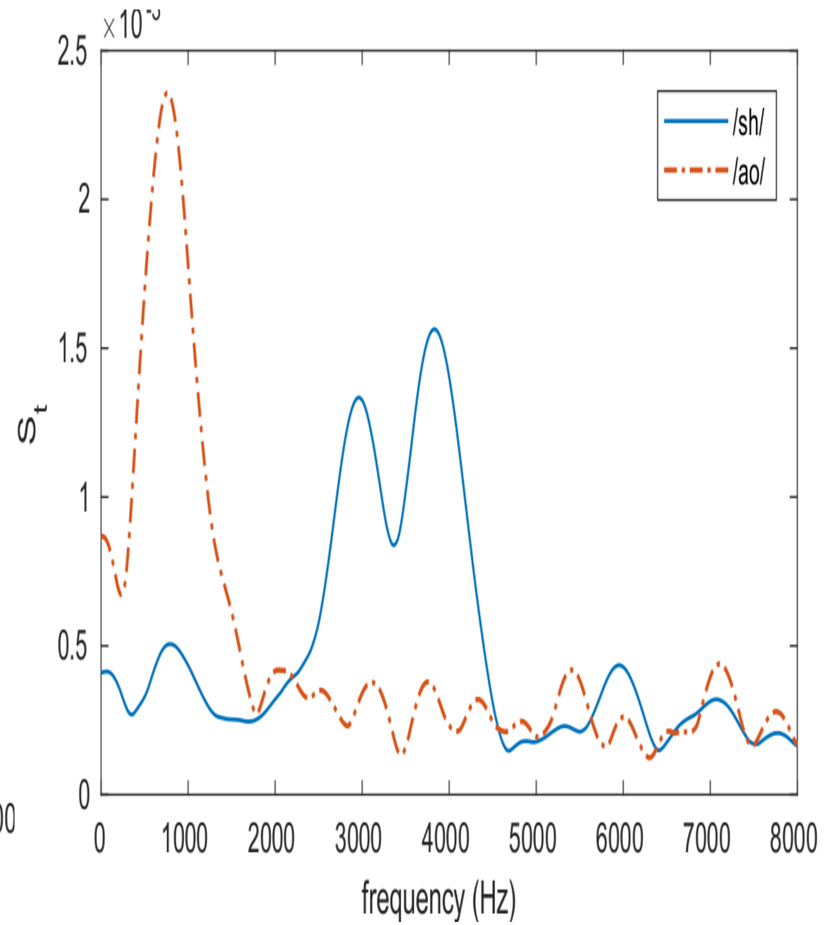
Speaker recognition
kW = 18 ms



Between 1st and 2nd conv. layers: Speaker Reco. (3)



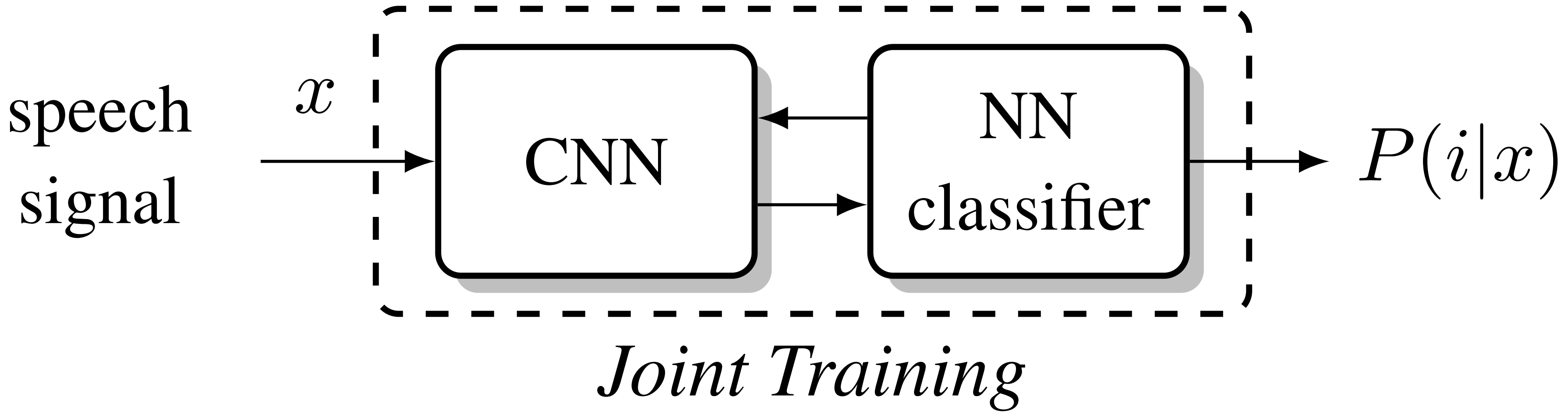
LP Spectrum



$|\mathcal{X}|$

Muckenhirn, Magimai-Doss and Marcel, [On learning vocal tract system related speaker discrimination information from raw speech signal using CNNs], in Proc. of Interspeech 2018

Central question



What information does such systems learn?

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

Visualization in Computer Vision

Visualization of what is captured by neural networks is a very active field of research for image recognition tasks.

Three approaches:

- input perturbation-based methods
- reconstruction-based methods
- gradient-based methods

Gradient-based visualization

Input (image, waveform...): $\mathbf{x} = [x_0 \ldots x_{N-1}]$.
Output unit corresponding to class c (before softmax layer): y^c.

Gradient:

$$\frac{\partial y^c}{\partial x_n},$$

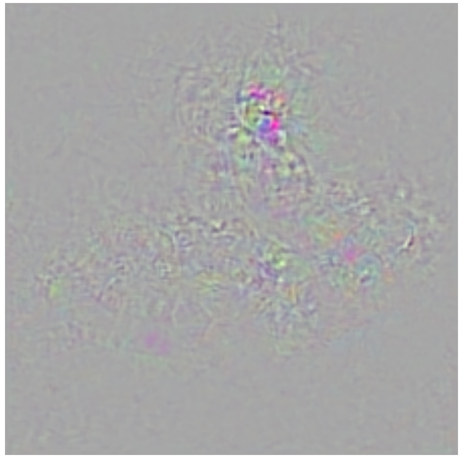
$n = 0, \ldots, N-1$

- It measures how much a small variation of each pixel value will impact the prediction score.
- It yields a "relevance" map of the same size as the input.

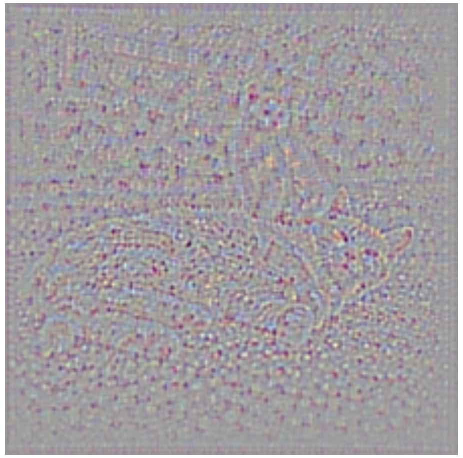
Example visualizations

Different gradient-based methods: differ on how the gradient is computed at a ReLU layer.

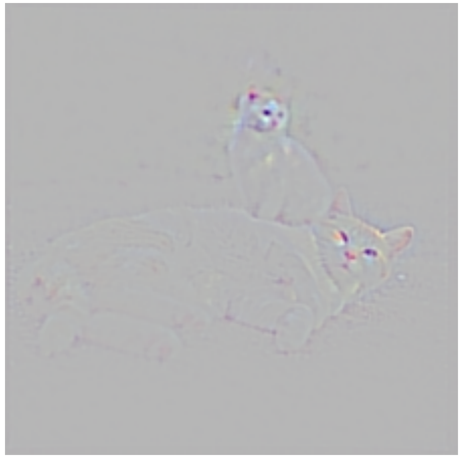
original image



saliency map



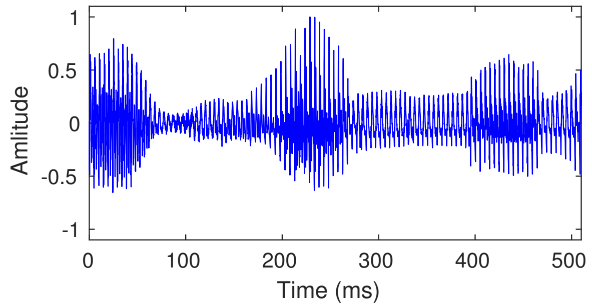
deconvnet



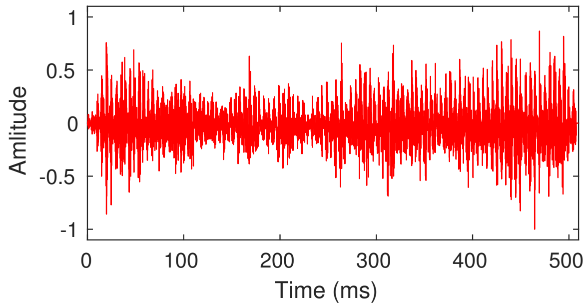
guided backpropagation

Gradient-based visualization for speech

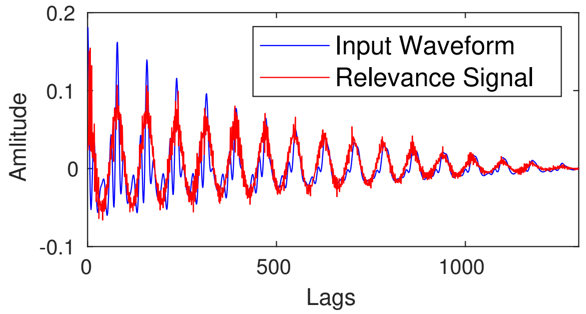
- Given an input speech-output class pair and the trained system, what is the contribution of each sample on the output score?[18]



Original Signal



Relevance signal

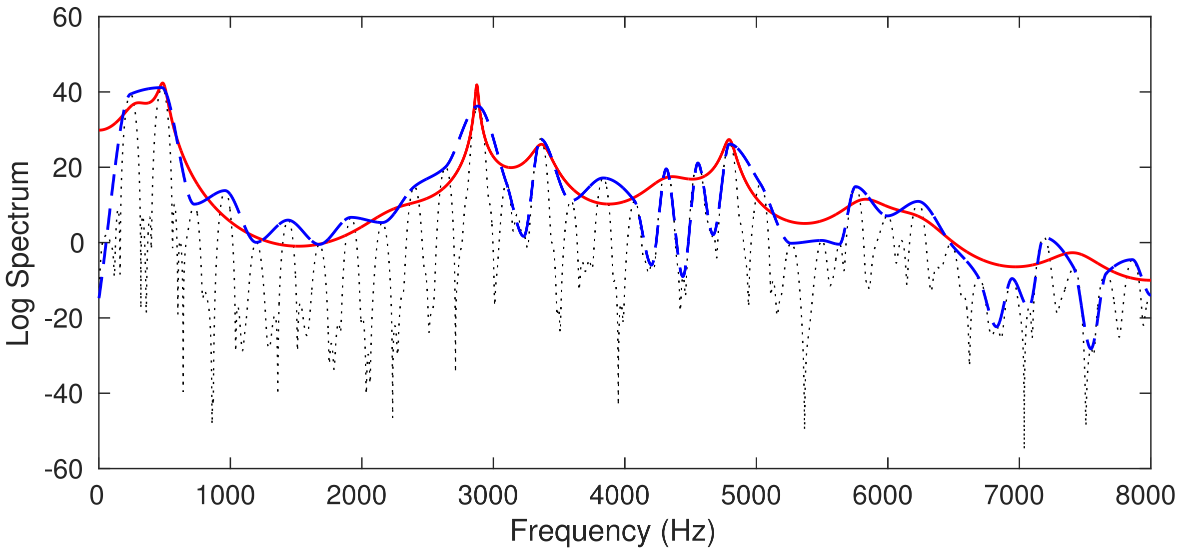


Autocorrelation

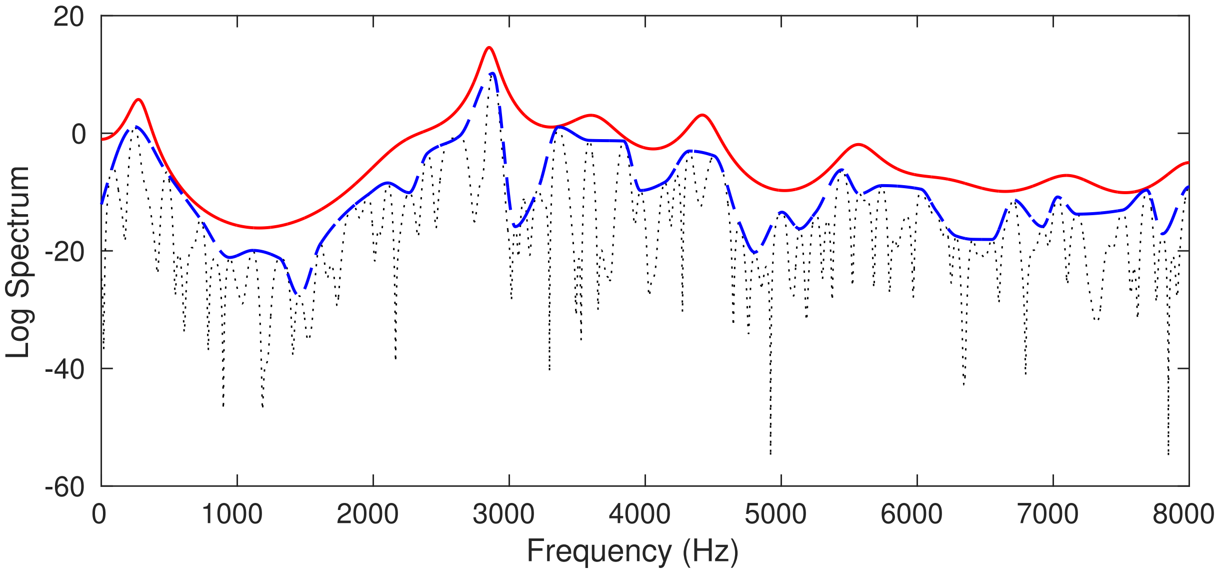
- Unlike images relevance signal for speech signals are not visually interprettable.
- Interpretation can be done in spectral domain (can be shown theoretically)

[18] H. Muckenhirn et al.,"Understanding and Visualizing Raw Waveform-based CNNs ," Proc. of Interspeech, 2019.

Whole network analysis: speech recognition (1)



Original Spectrum of /iy/



Relevance signal spectrum of /iy/

Whole network analysis: speech recognition (2)

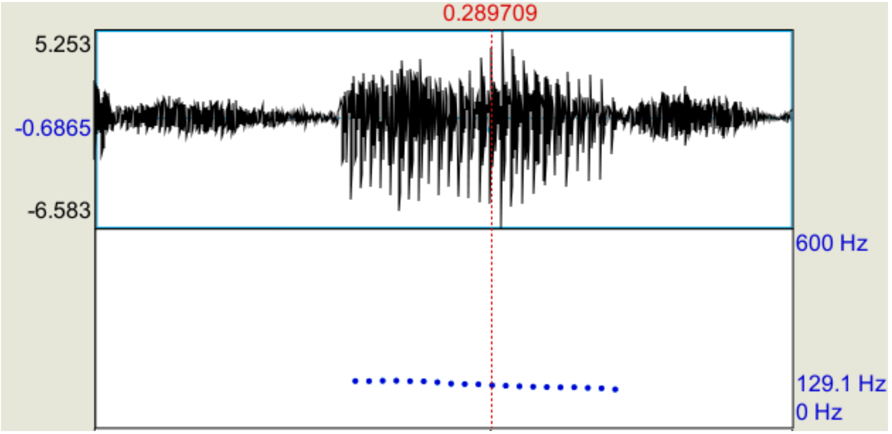
- Analysis of CNN trained on TIMIT phone recognition task on American English Vowel (AEV) dataset
- F0, F1 and F2 estimated automatically for the relevance signal for the steady state regions and compared to the values specified on the original study.

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.

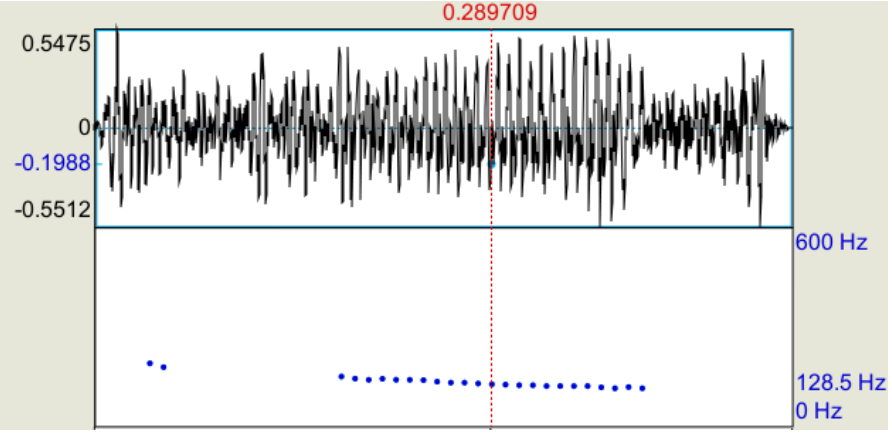
		/ah/	/eh/	/iy/	/oa/	/uw/
F0	F	93	91	91	94	92
	M	92	90	89	93	90
F1	F	90	92	93	91	93
	M	88	92	92	89	93
F2	F	94	94	94	95	94
	M	94	93	94	94	93

3

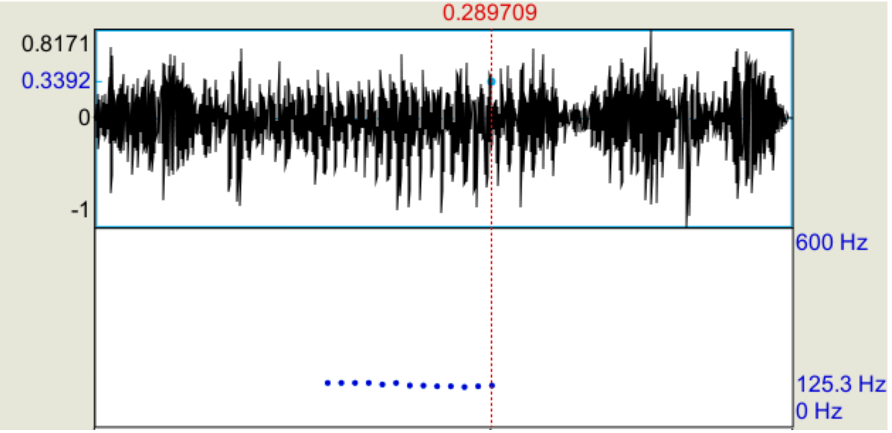
Whole network analysis: speaker recognition (1)



Original signal

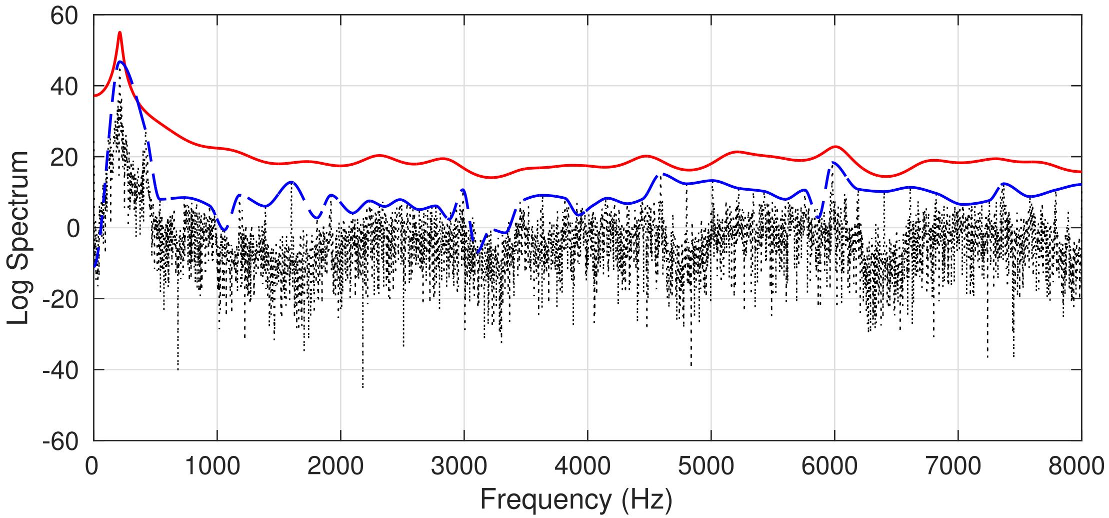


Segmental modeling

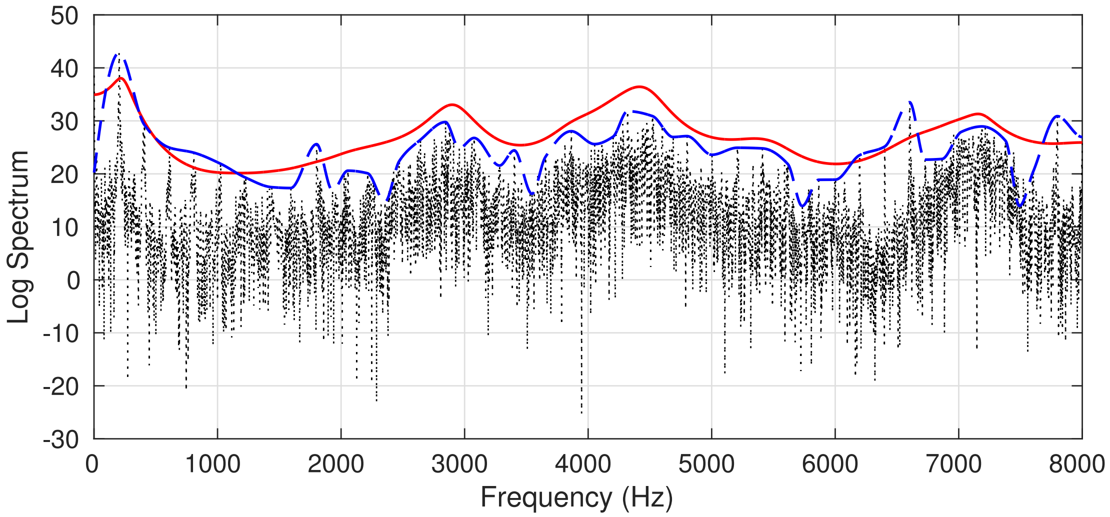


Sub-segmental modeling

Whole network analysis: speaker recognition (2)



Segmental modeling

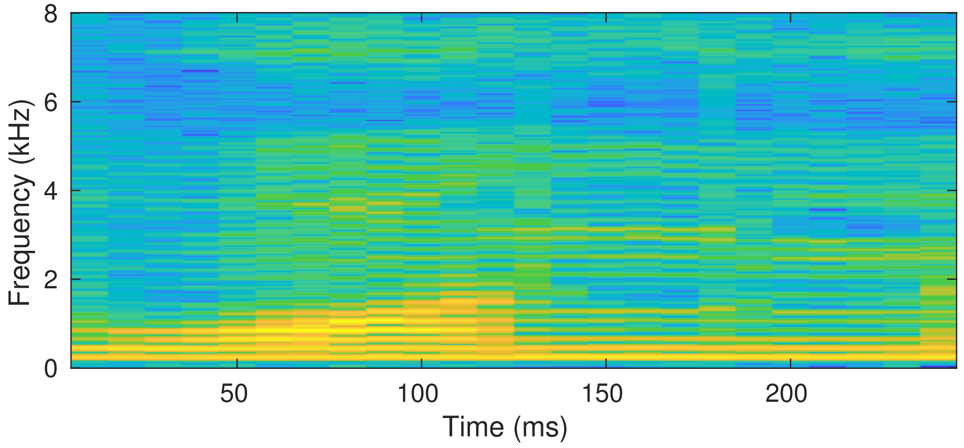


Sub-segmental modeling

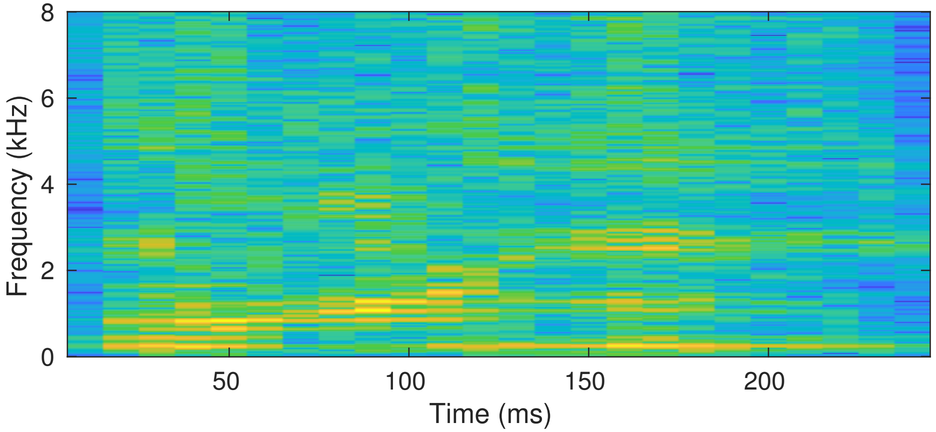
Utterance level average spectrum (long term average spectrum)

Whole network analysis: speech and speaker

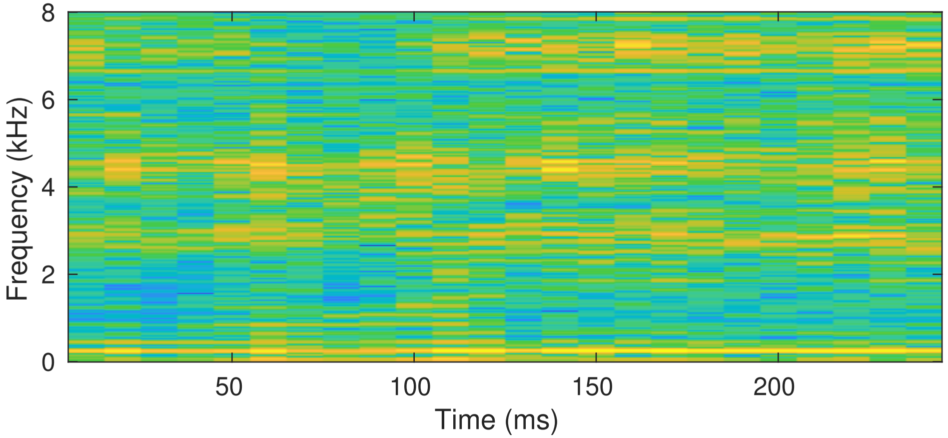
Original signal spectrogram



Phone CNN relevance signal spectrogram



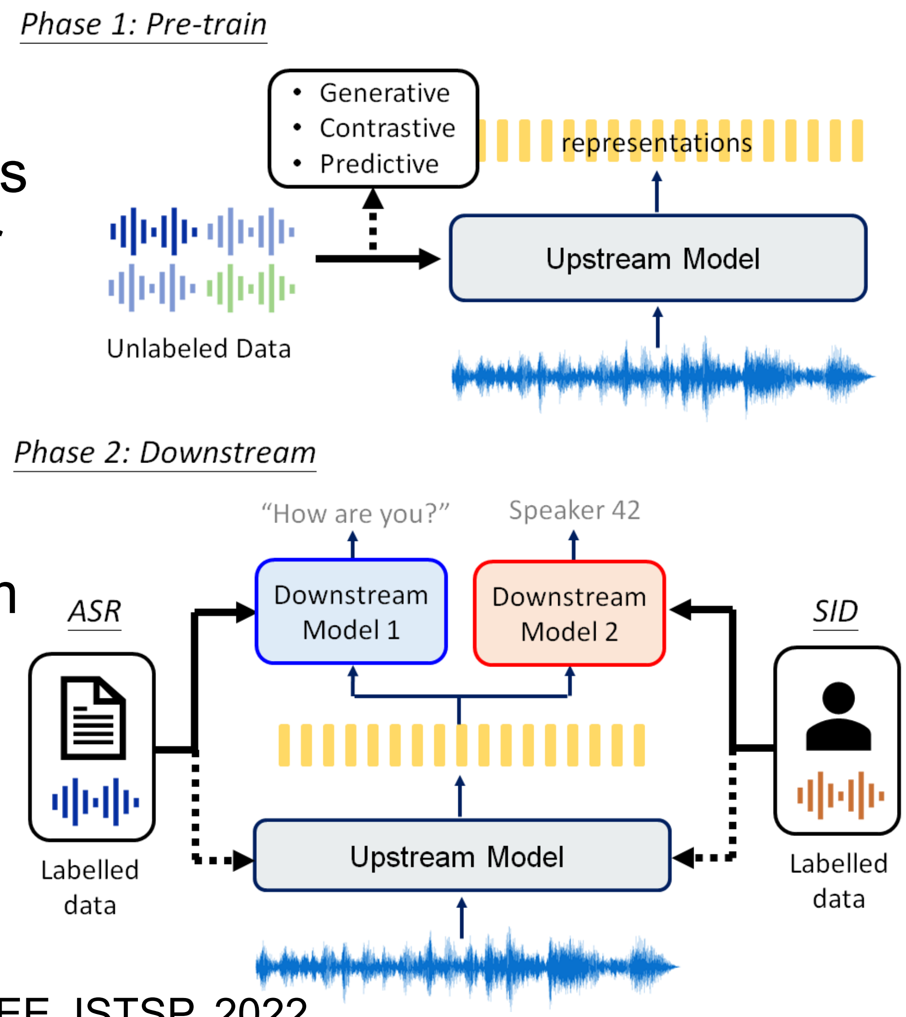
Speaker CNN relevance signal spectrogram



Self-supervised speech representation (1)

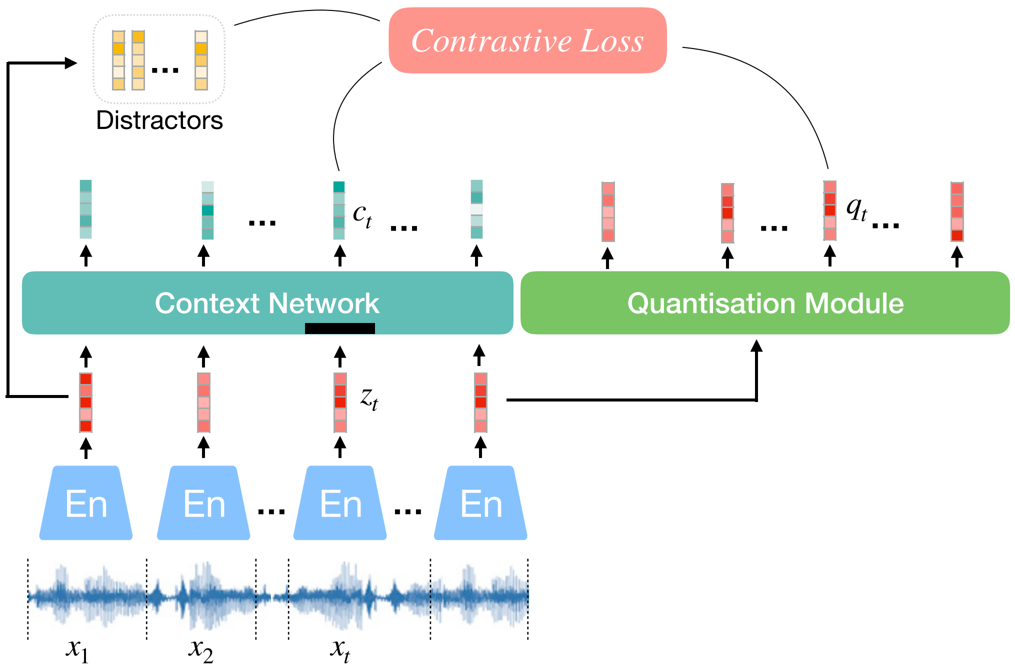
Combines several understandings from speech processing and takes inspiration from text and computer vision domains

- Replacing hand-crafted features by CNN-based encoder
- Temporal context modeling using transformers and attention mechanism
- Hierarchical information processing
- Reconstructing information by clustering latent representations or acoustic features, and predicting them
- Different types of modeling: generative, contrastive, predictive, Siamese
- Multiple views of the data

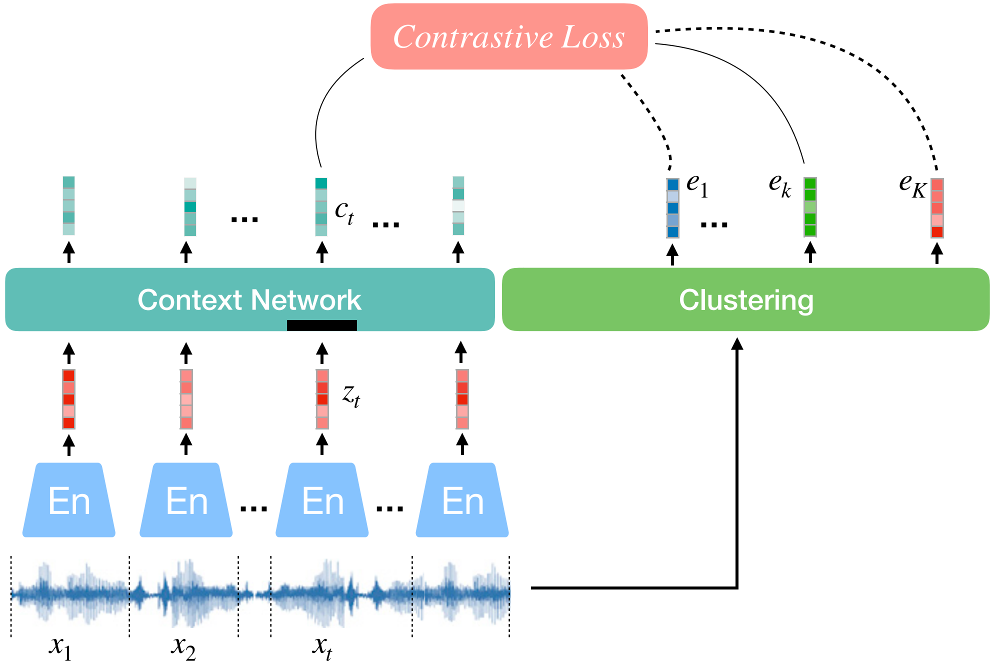


Liu et al. "[Audio self-supervised learning: A survey]", Patterns, Vol. 3, 2022.
Mohammed et al. "[Self-Supervised Speech Representation Learning: A Review]", IEEE JSTSP, 2022

Self-supervised speech representation (2)



Wav2vec 2.0



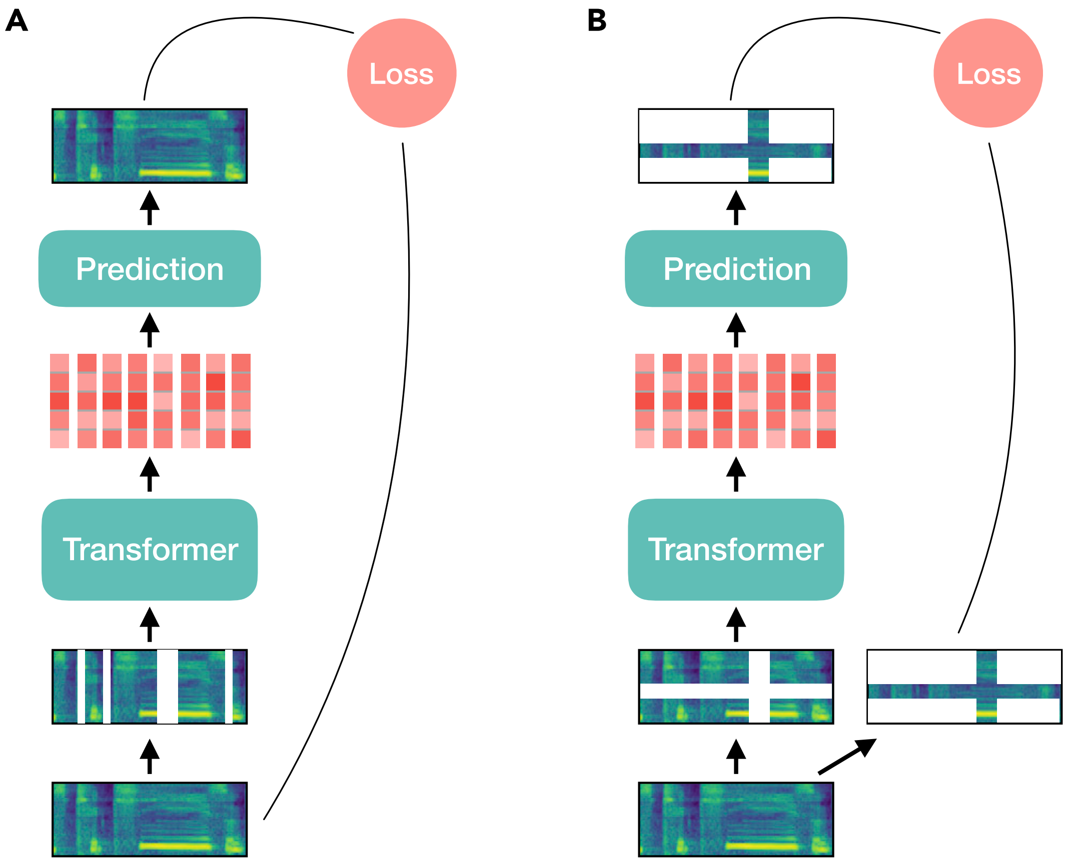
HuBERT

En: CNN-based encoders

Liu et al. "Audio self-supervised learning: A survey", Patterns, Vol. 3, 2022.

Self-supervised speech representation (3)

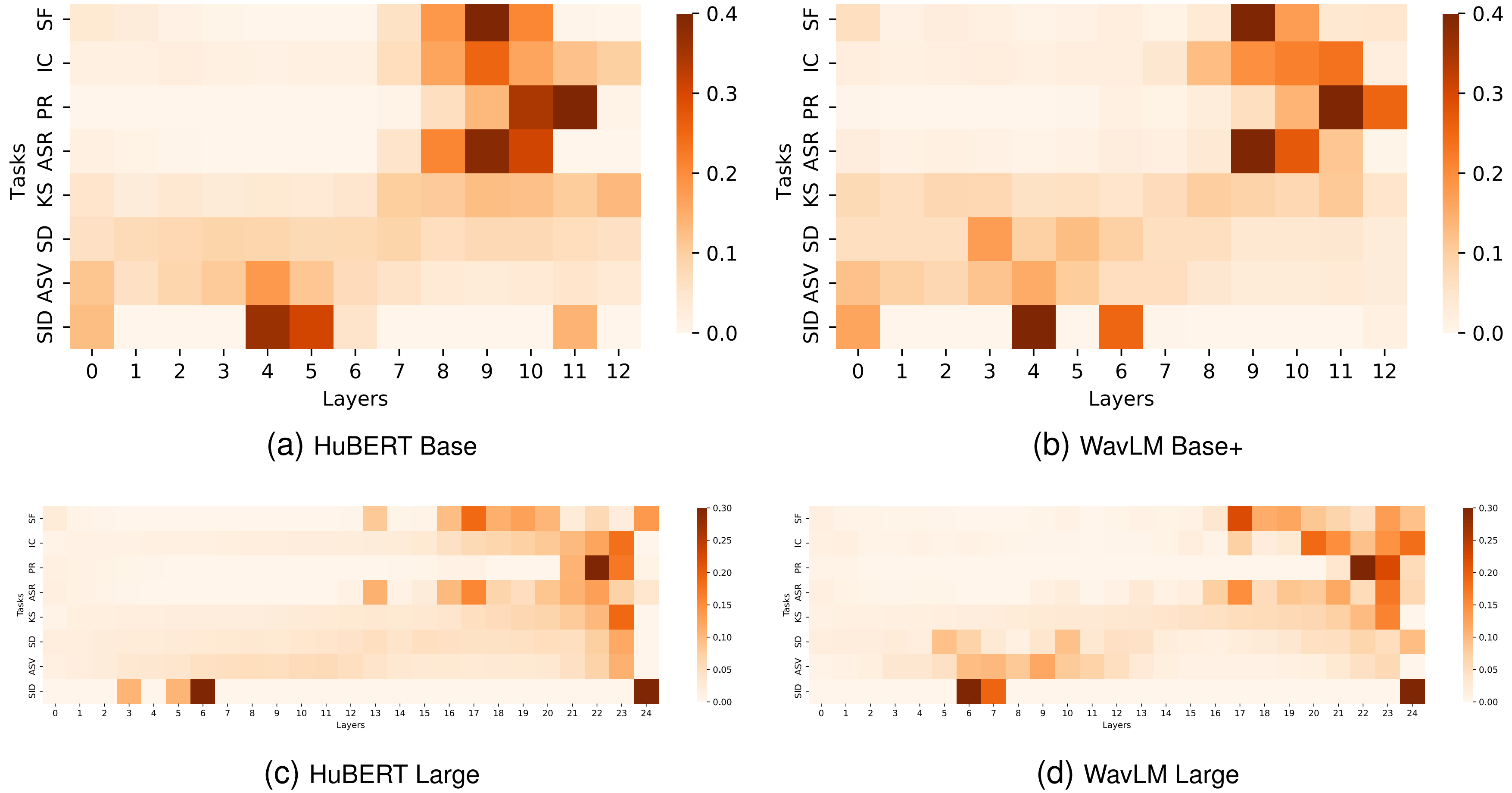
Masked prediction-based approach



Mockingjay DAPC

Liu et al. "Audio self-supervised learning: A survey", Patterns, Vol. 3, 2022.

Self-supervised speech representation (4)



Layerwise-Taskwise analysis on SUPERB Benchmark

SID: Speaker identification
ASV: Speaker verification
SD: Speaker diarization
KS: Keyword spotting
ASR: Speech recognition
PR: Phoneme recognition
IC: Intent classification
SF: Slot filling

Chen et al. [“WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing”](), IEEE JSTSP, 2022.

Self-supervised speech representation (5)

Open questions

- Are all layers meaningful for all tasks?
  - How to find that out and/or select?
- Efficient adaptation
- Model compression
- Interpretability/explainability
  - What kind of information are the layers modeling and how?
  - Why and where the model fails?
- Trustworthy?
- Universal speech processing truly feasible with one single model?

Thank you for your attention!