بیست و هشتمین کنفرانس بین المللی کامپیوتر
انجمن کامپیوتر ایران
۵ و ۶ بهمن ۱۴۰۱ | دانشگاه صنعتی شریف

Analyzing the Use of Auditory Filter Models for Making Interpretable Convolutional Neural Networks for Speaker Identification

Hossein Fayyazi and Yasser Shekofteh

Faculty of Computer Science and Engineering
Shahid Beheshti University, Tehran, Iran

Presenter

Hossein Fayyazi

Outline

- Introduction
- Auditory Filter Models
- Model Architecture
- Results and Discussion
- Conclusion & Future Works

Introduction

understanding the function of different parts of the living organism's body

advances in artificial intelligence field of research

High complexity of some body parts, e.g., the brain

using some abstractions for making intelligent models, which can make the models uninterpretable

Deep Neural Networks (DNNs)

- high performance
- less reliability because the decisions made are ambiguous (black-box nature)

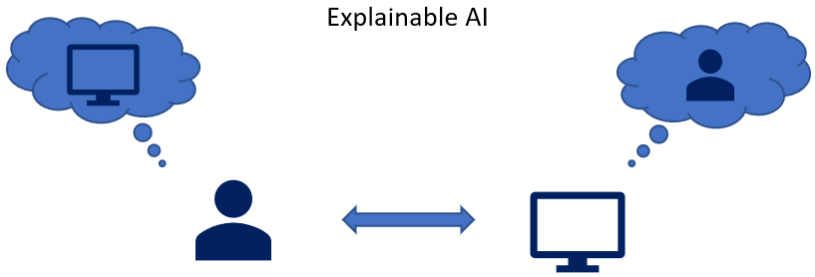
Solution

- making DNN models explainable
- E.g. use of meaningful filters in the first layer of Convolutional Neural Networks (CNNs)

Introduction

eXplainable AI (XAI)

- reveal the weaknesses of different solutions
- show future research directions
- help researchers learn more about the problem
- explainable systems can be derived from human's biological mechanisms



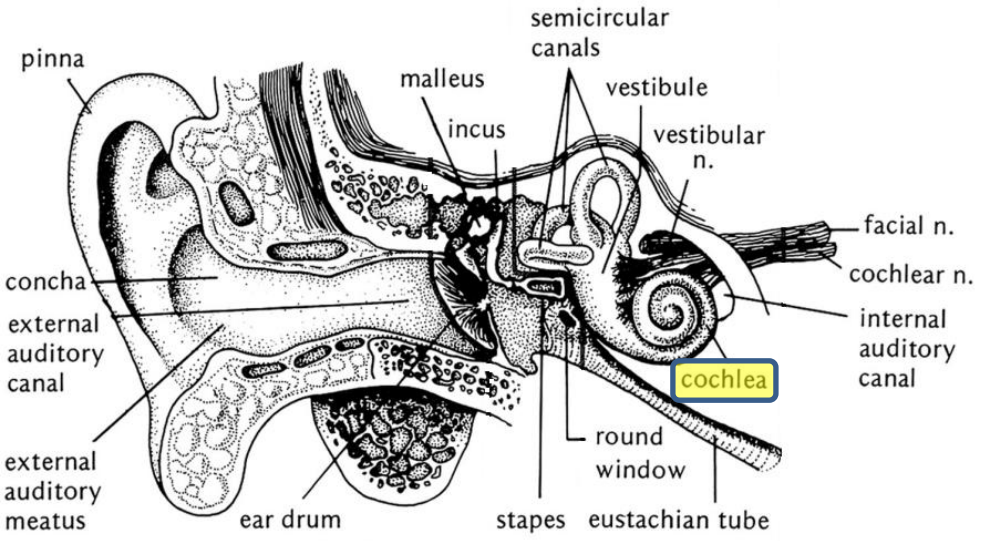
Goal

- Understanding hearing models and adapting learning models closely to human hearing
- examine the use of some auditory filter models as CNN front-ends
- evaluate the resulting filter banks in the Speaker Identification (SID) task

Auditory Filter Models

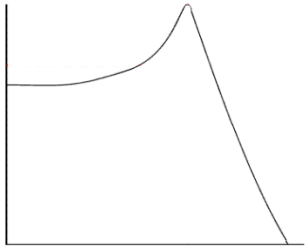
Modeling and explaining the human hearing mechanism

- auditory filtering in the cochlea corresponds to bandpass filtering
- each filter represents a location along the cochlear partition
- By analyzing different physiological and psychological experiments, different auditory filter models are proposed.

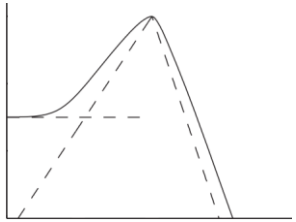


Auditory filter models

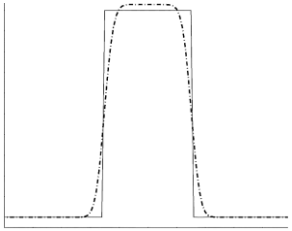
- **Simple resonance**
- can be considered as an order-1 gammatone
- has been frequently tried and usually rejected as a model of auditory filtering



- **rounded exponential (roex) family**
- are not corresponding to a real filter
- has no time-domain equivalent



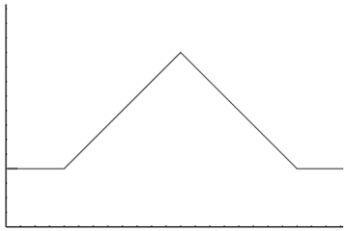
- **Rectangular**
- can be considered as the difference between two low-pass filters
- equivalent to a sinc function in the time domain



$$h_{rect}[n] = 2ABsinc(Bn)\,cos(2\pi f_c n)$$

Auditory filter models

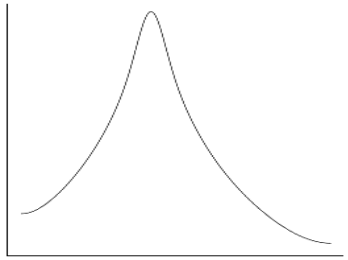
➢ **Triangular**



- Auditory filter shapes can be approximated by triangular filters
- Mel-filterbanks is commonly computed based on this filter type
- The impulse response of the triangular filter is represented in the time domain by the sinc2 function

$$h_{tri}[n] = Asinc^2(Bn)\,cos(2\pi f_c n)$$

➢ **Gammatone**



- popular in auditory modeling
  - a good match of impulse response to the physiologically derived revcor functions measured in the cochlear nucleus of cats
  - Few parameters of Gammatone-family filters match most of results from psychophysical and physiological experiments
- simple time domain response
- Complicated frequency domain formula

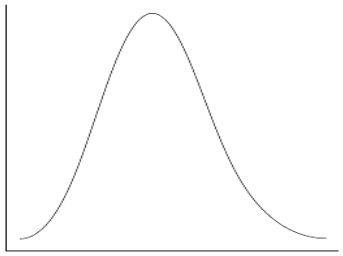
$$h_{gamma}[n] = An^{N-1}exp(-2\pi Bn)\,cos(2\pi f_c n)$$

Auditory filter models

➢ **Gaussian**

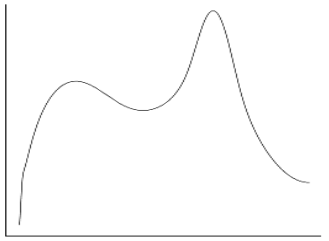
- a high-order gammatone

$$h_{gauss}[n] = Aexp(-n^2/2\sigma^2)\ cos(2\pi f_c n)$$

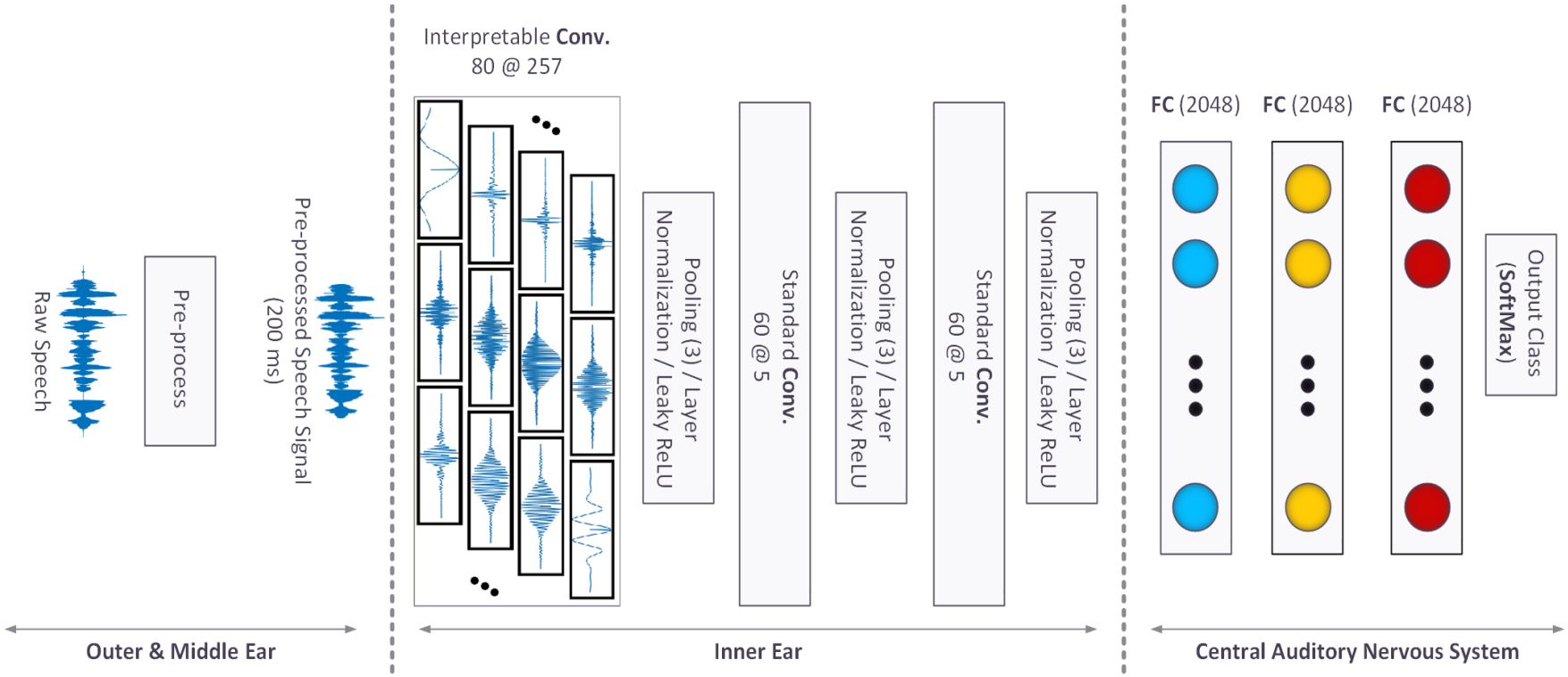


➢ **Cascaded filter**

- the auditory system does not respond linearly
- Cascading linear filters is a way to make more complicated filters that introduce some non-linearity to the model
- Here, we cascade two Gaussian filters to see the impact of introducing filter non-linearity to the final results



Model Architecture



Results and Discussion

Dataset

- TIMIT: a well-known dataset for speech processing tasks
- contains 16 kHz recordings of 630 speakers of eight major dialects of American English

- Five of the eight sentences of each speaker are used for training, two for validation, and one for testing
- a 630-class classification problem

Error rate

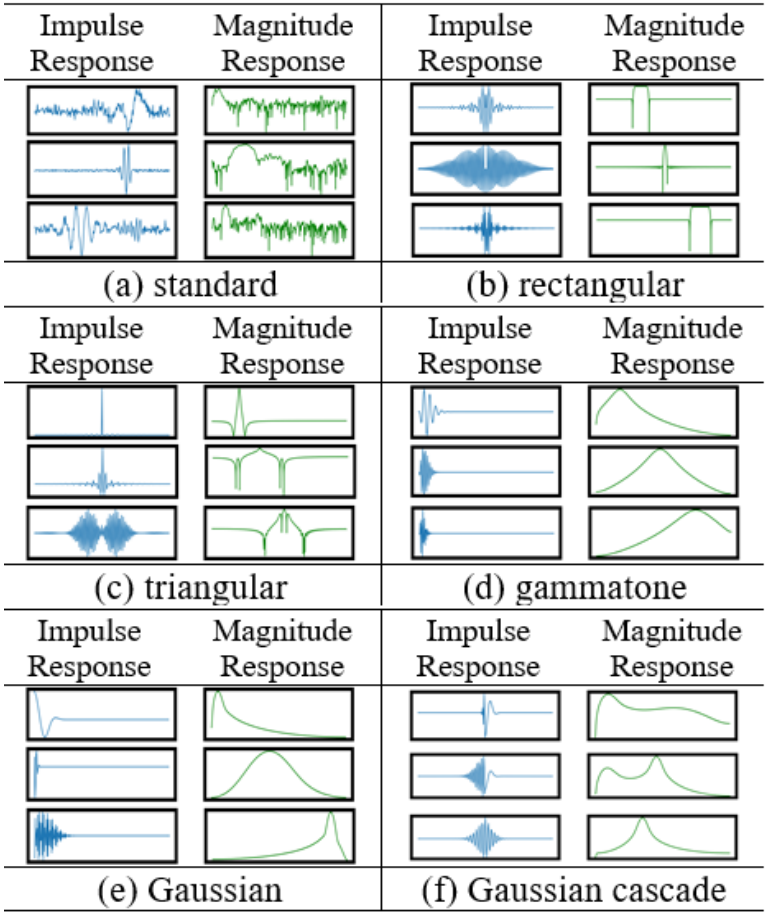
	Validation		Test	
	FER %	CER %	FER %	CER %
Standard (CNN)	51.22 ± 0.98	1.90 ± 0.08	50.72 ± 0.62	1.58 ± 0.27
Rectangular	48.36 ± 1.09	1.37 ± 0.25	48.25 ± 1.28	1.64 ± 0.36
Triangular	48.46 ± 1.57	0.98 ± 0.23	48.03 ± 1.14	1.16 ± 0.09
Gammatone	45.55 ± 1.79	1.32 ± 0.20	44.85 ± 1.32	1.27 ± 0.16
Gaussian	46.25 ± 1.55	1.16 ± 0.20	45.96 ± 1.42	1.53 ± 0.56
Gaussian Cascade	44.76 ± 0.31	1.11 ± 0.27	44.49 ± 0.27	0.95 ± 0.16

Results and Discussion

Time and frequency domains shapes

- audio filter models have a meaningful time domain shape and their magnitude response can be determined explicitly by a center frequency and bandwidth
- the standard filters have unfamiliar, noisy shapes with no meaning

➢ Specific filter types are a strong replacement for standard ones to have a better understanding of the decision made by a CNN model

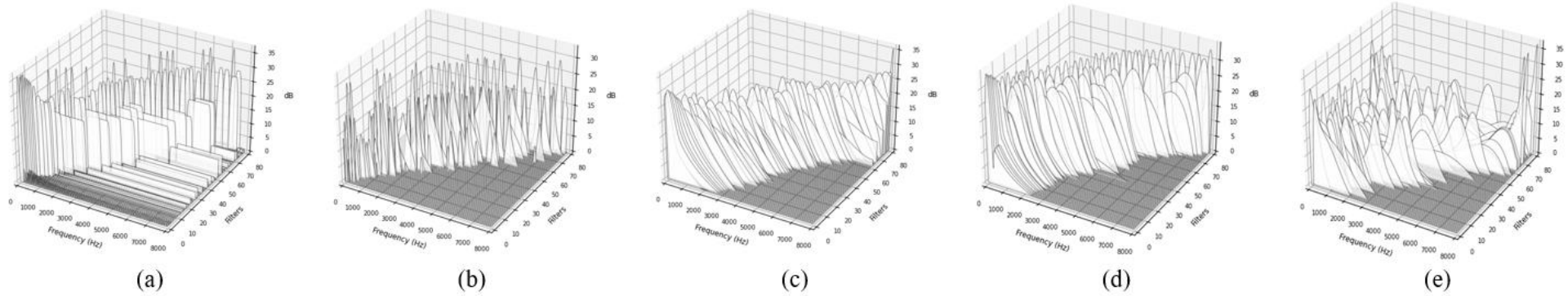


Some examples of learned filters

Results and Discussion

Learned filter bank

- The filters operate in very low frequencies are more than in high frequencies
- While filters with sharper peaks are placed in lower frequencies, the peaks become shallower at high frequencies
- These observations are consistent with the experiments that have been conducted on the filtering function of the human auditory system



(a) rectangular, (b) triangular, (c) gammatone, (d) Gaussian and (e) Gaussian cascade filter banks learned by different models

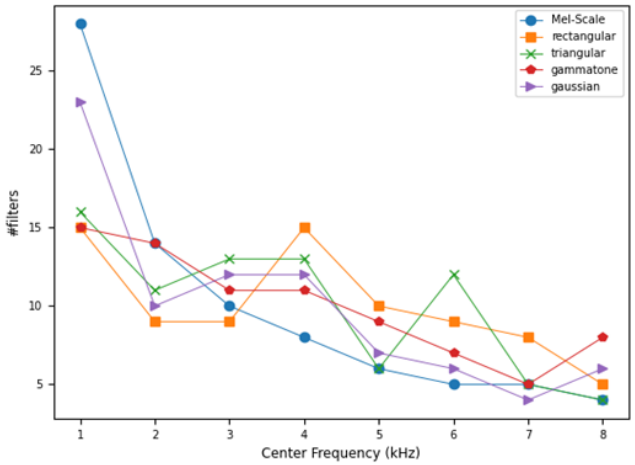
Results and Discussion

Center frequencies

- the overall trend of learned filter banks is as the Mel-scale one
- the importance of frequencies close to 2 kHz is considered less in all models
- the number of filters sensitive to high frequencies is not as low as the Mel-filterbanks

➢ In a specific application like SID, the fundamental frequency, below 1 kHz frequency, has more impact in distinguishing speakers than the two or three first formants of a speech signal

➢ Some information related to speaker recognition is spread in higher frequencies



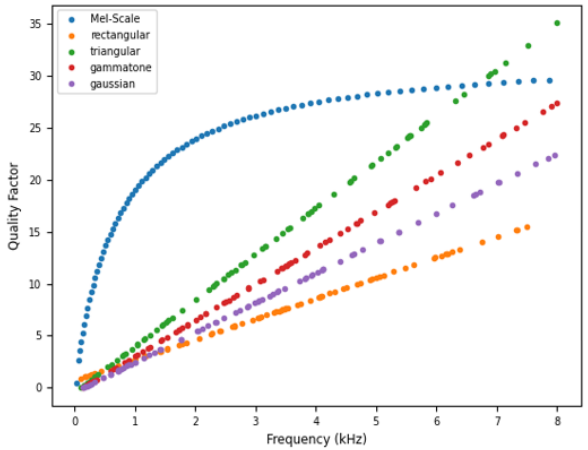
Distribution of center frequencies of learned filter banks

Results and Discussion

Quality Factor (QF)

- the fraction of the center frequency to the bandwidth of a filter
- used to examine both parameters at the same time

- the overall trend of QF for all filter types is incremental
- filters at high frequencies are further apart and have higher bandwidth
- The slope of the fitted lines of interpretable filters reveals the importance of higher frequencies in this specific task
- The number of filters in 0 ~ 1.5 kHz and 2.5 ~ 4 kHz frequency bands is more than the others, which demonstrates that these frequency bands create more distinction between different speakers



Fitted lines to Q-factor of learned filter banks

Results and Discussion

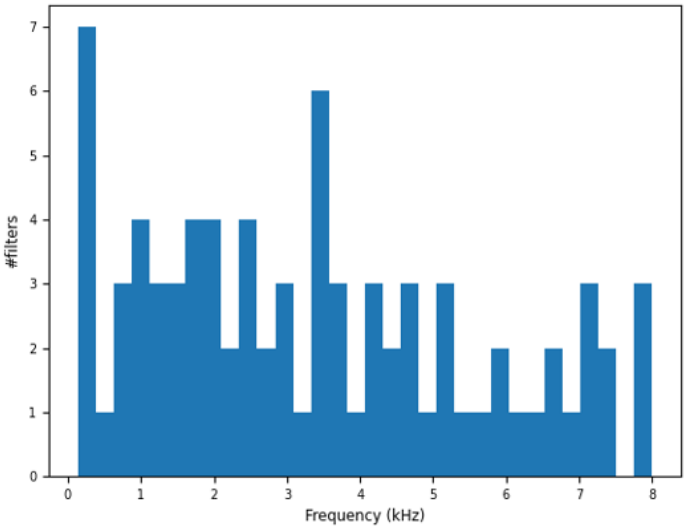
Frequency analysis from speech production view

- The unique characteristics of a speaker are encoded during speech production
- These features should be involved in the invariant factors in the physiology of the vocal tract

Some interesting findings presented in some researches:

- The function of different articulatory speech organs that make speaker-dependent features lead to non-uniformly distribution of these features in high frequency bands.
- glottis information is encoded between 100 Hz and 400 Hz
- The information of the piriform fossa is encoded between 4 kHz and 5 kHz.

➢ The proper filter bank is completely task-dependent



distribution of learned filters in different frequencies for gammatone model

- most filters operate in 0 ~ 250 Hz where glottis information is encoded.
- Other filters are operated in high frequencies with a non-uniform distribution and frequencies related to speech formants are not emphasized as much as Mel-filterbanks

Conclusions

The key operation of the cochlea in the inner ear

- The separation of the speech signal into different spectral bands
- can be simulated using linear or non-linear filter banks

This paper

- Analyzing the use of auditory filter models in the first layer of a CNN architecture for SID task

Some results

- The increase of the filter bandwidth with the center frequency of the filter is observed in learned filter banks
- Data-independent filter banks, such as Mel-scale, are general filters that must be customized for the task at hand
- Although the overall distribution of learned filters in different frequency regions is almost similar to Mel-scale, but there are also differences that, if considered, more optimal models can be achieved
- The use of these types of filters increases the efficiency of the model

Future Works

- Examining other DNN architectures, especially the models equipped with attention mechanisms, and analyzing the interpretability of the model

- Investigation of the functionality of auditory filter models in other speech processing tasks
  - speech recognition
  - speech enhancement

Thank You