Are we able to model the human auditory system in speech signal enhancement?

Daniel Zakševski, Gintautas Tamulevičius
Vilnius University, Institute of Data Science and Digital Technologies

Introduction

Human hearing possesses unique characteristics like frequency selectivity, temporal adaptation and auditory scene analysis which make us quite profficient at speech processing. Traditional methods, often relying on static filters, fail to capture this dynamic and selectivity. Deep learning algorithms, while promising, struggle in low SNR environments, unlike human hearing, and often rely on the same fixed structures, which limits their adaptability. This research explores the potential ways of integrating a holistic auditory model into deep learning for speech enhancement, aiming to improve performance in challenging acoustic conditions.

A view of the auditory system

The **outer ear** allows localization and is responsible for the collection of sound.

The **middle ear** amplifies and transmits sound vibrations to the inner ear.

The **inner ear**, primarily the cochlea, performs analysis of sound frequencies.

The **brainstem** and **brain** process auditory signals, allowing sound perception and comprehension.

Usual modeling of the auditory system

Often simplified to single-channel audio input, thus is usually skipped.

Typically modeled implicitly through basic preprocessing steps like pre-emphasis filtering or is entirely skipped.

It is commonly modeled using static filterbanks, such as MFCC or Gammatone filters.

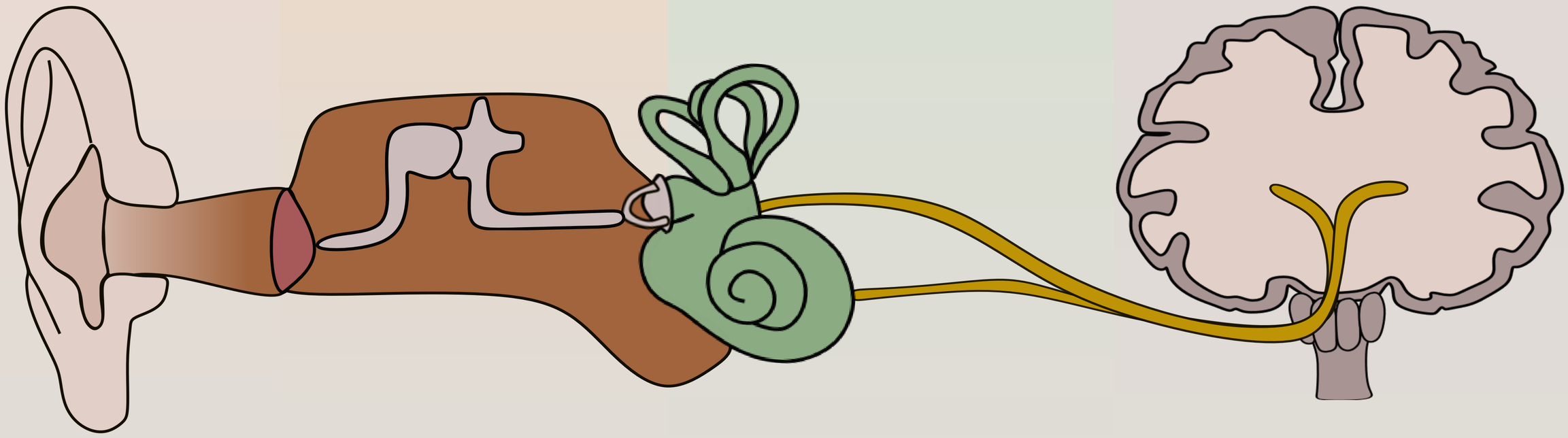
Focused on feature extraction through deep learning models, with limited incorporation of higher-level processes like auditory scene analysis and attention.

Common paradigms

End-to-End A single neural network implicitly models the entire auditory pathway, learning to enhance speech through an internal representation of the ear’s processing mechanisms.

Two-stage Employs a fixed filterbank to decompose the incoming sound into a representation analogous to the frequency analysis performed by the cochlea.

A neural network is used to process the frequency-domain representation of the sound, learning higher-level features.



Modeling suggestions

- **Binaural hearing:** utilize two-channel audio input and incorporate binaural processing mechanisms to leverage spatial cues for improved source separation and noise reduction.
- **Dynamic cochlear modeling:** instead of using static filters, develop deep learning models that can produce dynamic and adaptive filterbanks, mimicking the cochlea’s response to varying sound levels and frequencies.
- **Tonototopic organization:** explore architectures that explicitly incorporate the tonotopic organization of the cochlea and the brain, potentially through frequency-specific processing routes or attention mechanisms.
- **Loss functions:** design loss functions that reflect human auditory perception by incorporating psychoacoustic principles and measures of speech intelligibility.
- **Brainstem modulation:** investigate incorporating feedback mechanisms inspired by efferent pathways to modulate the sensitivity and selectivity of the model's processing units.
- **Spiking Neural Networks (SNNs):** employ a biologically-inspired architecture that mimics the dynamics and information processing mechanisms of biological neurons.

Contact: daniel.zaksevski@mif.vu.lt

15th Conference Data Analysis Methods for Software Systems
November 28 – 30, 2024, Druskininkai, Lithuania