Mazen Elias [147]
Faculty of Engineering and Information Technology, University of Pecs, Pecs, Hungary
Mhd Mouayad Ahmad Abdalla [148]
Faculty of Engineering and Information Technology, University of Pecs, Pecs, Hungary
Zoltán Sári [149]
Faculty of Engineering and Information Technology, University of Pecs, Pecs, Hungary

MODELING HUMAN PITCH PERCEPTION USING COCHLEAR MODELS AND DEEP LEARNING

DOI: 10.5937/PIM26508E

Abstract:
This study investigates the simulation of human perception of pitch using cochlear computational models and deep learning algorithms, to develop a model that can predict the fundamental frequency along with other harmonic components from preprocessed signals that carry the properties of cochlear processing. The gammatone filterbank cochlear model was implemented to process musical audio files and compute cochlear-inspired spectrograms and features, followed by a hybrid deep neural networks model that was trained to predict the fundamental frequency and four other harmonics. The implemented cochlear model achieved high accuracy in conducting biologically inspired preprocessing tasks, and the deep learning model succeeded in learning unique patterns from the cochlear-analyzed data, generating accurate predictions. Achieving these results can inspire and motivate development in the fields of cochlear implants, and sound and music engineering.

Keywords: *Pitch perception, cochlear models, deep learning, fundamental frequency, harmonics.*

Introduction

Humans can achieve a high level of sound processing through the auditory system, which performs tasks that lead to a complex perception. The peripheral auditory system processes sound through a complicated series of amplitude modulation and frequency filtering, followed by neuronal processing in the auditory pathways, resulting in the discrimination of the pitch, timbre, and loudness, which leads to the unique experience of sound.

Pitch is the measure of the frequencies that we perceive, and most of the frequency processing happens in the cochlea; with its tonotopic membrane and its various biological filters, it allows for a special kind of

[147] mazenelias20@gmail.com, ORCiD: 0009-0004-7527-5460
[148] abdalla.mouayad@mik.pte.hu, ORCiD: 0009-0004-7105-7685
[149] sari.zoltan@mik.pte.hu, ORCiD: 0000-0001-6961-6063

processing that shapes our hearing properties helping in polyphonic hearing and enabling us to focus on major voices and frequencies even in the presence of noise and other sounds.

Millions of people of all ages worldwide suffer from hearing loss due to a variety of reasons that can include genetic factors, ear infections, or age-related changes. The degree of hearing loss can vary ranging from difficulties of understanding speech to complete hearing loss [1], [2].

Cochlear implants (CI) have been a life-changing solution for those with severe to profound hearing loss. These artificial implants bypass the different parts in the ear conducting the sound information to the auditory nerve directly. Despite the astonishing advancements in the technology of these implants, they still struggle with many challenges mostly pitch-related. People with cochlear implants often experience hearing artificial sounds more like robotic or mechanical sounds that lack many perceptual properties, they also face difficulties with appreciating music because of the low frequency resolution offered by cochlear implants.

Further research and development is highly needed in the direction of testing and employing cochlear models in perceptual sound processing applications. Modeling the pitch perception is one of the most complicated directions in perceptual processing, as many applications fall with many limitations including the spectral and temporal resolution, the challenges of real-time processing, and computational complexity. Achieving an artificial intelligence-based model that can process pitch information from cochlear preprocessed audio signals can provide significant insight into developing cochlear implants processing, and sound and music engineering field.

1. Related Work

Previous studies showed that gammatone filterbank models can provide physiological accuracy and efficiency in the implementation in cochlear applications, while also giving advantages over FFT-based filter processing [3], [4]. Deep learning approaches also achieved strong performance in pitch estimation applications; Kim et al. (2018) [5], developed the CREPE algorithm that shifted the pitch estimation field to a whole new perspective, as they developed a deep CNN model that can learn directly from raw audio signals, while Chou et al. (2018) [6], introduced a hybrid neural network based on the duplex model of pitch perception.

These studies indicate the value of the use of both cochlear-inspired models and deep learning models, but they also reveal the need for a framework that brings these directions together to apply in pitch perception. This gap led to the presented work in combining cochlear computational modeling with deep learning for pitch perceptual prediction.

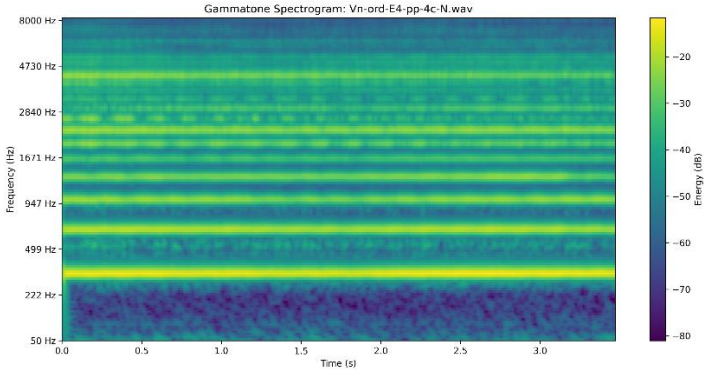
2. Methodology

2.1. Overview

The main objective was to model human perception of pitch using a combination of computational cochlear modeling and deep learning. To achieve this, a systematic methodology was developed, as it involved conducting many stages of testing, with the aim of answering the research questions and developing a robust model that is built upon real testing data.

2.2. Preprocessing and Data Preparation

The dataset used in this study is the TinySOL dataset, which contains 2913 isolated instrumental audio files covering a broad pitch range [7].



Picture 1. *Gammatone spectrogram.*

The preprocessing model that was used is Gammatone filter-bank, which is a linear model that is based on designing filter banks with shaper cutoff for higher frequency than the cutoff for lower frequencies. The bandwidths of the filters in this model are proportional to each center frequency and based on the ERB scale [4]. All files were processed with the gammatone filter-bank to extract cochlear-inspired spectrograms together with spectral and global features (Picture 1). After feature extraction, global normalization was applied to all model inputs, including the spectrograms and global descriptors, by computing the mean and standard deviation of the training data and using Z-score normalization.

2.3. Deep Learning Model

A hybrid multi modal architecture was developed combining convolutional neural networks (CNN) to process spectral features, dense neural networks (DNN) to process global features, and long short-term memory (LSTM) networks to integrate temporal information. The model allows for parallel processing of the f0 and each of the harmonic components ensuring that the model learns genuine patterns and relationships between the features and the labels. The first component of the deep learning architecture uses layers of convolutional neural networks to extract spectral and temporal patterns from the gammatone spectrograms. Two convolutional blocks were used with 32 and 64 filters respectively. Each block contains two convolutional layers, batch normalization, Rectified Linear Unit (ReLU) activation function, max pooling, and dropout. The second component of the architecture is composed of fully connected layers (dense layers) where every neuron in a layer is connected to every neuron from the previous layer. The third component employs a dual-LSTM layers to integrate information across time, combining CNN processed features with the dense layers output, while providing parallel paths to process each harmonic individually. This model is built as a regression problem rather than a classification problem, having significant implications in designing the architecture, selecting the loss function, and choosing the metrics. For this task, the mean squared error (MSE) is used as a loss function. For validation, the mean absolute error (MAE) and the mean absolute percentage error (MAPE) were used to assess the performance of each learning epoch.

3. Results

The results of the testing predictions were evaluated using statistical and psychoacoustic evaluation methods to provide a comprehensive view of the performance.

3.1. Model Performance

A comparative analysis was conducted to assess the performance of the model. The training was set to 150 epochs with enabling early stopping to avoid overfitting. The best model was saved at epoch 72 with a learning rate of 0.000063, achieving training loss of 0.1306, validation loss of 0.0121, training set mean absolute error (MAE) for f0 of 0.1601, and validation set mean absolute error (MAE) for f0 of 0.0307. This

indicates strong generalization, robust data normalization, and suggests high precision in pitch prediction.

The tests indicate that the f0 prediction was more accurate in general than the prediction of other harmonics, and the error increased with higher harmonics, indicating that the model is giving less importance for higher harmonics with the loss weights and the properties of the gammatone spectrograms.

Table 1: *Global evaluation metrics*

Component	*MAE (Hz)*	*MAPE (%)*	*Median AE (Hz)*	*Median APE (%)*
F0	9.38	3.51	5.19	1.84
H2	22.21	5.68	15.16	2.27
H3	40.46	6.90	31.45	2.92
H4	59.25	4.53	30.61	2.88
H5	73.72	6.29	51.36	3.48

3.2. Psychoacoustic Evaluation

To evaluate the perceptual properties of the predicted harmonics, harmonics ratios were computed and compared with the ideal (integer multiples) relations between harmonics. The relations in the predictions seem slightly shifted from the ideal relations, which can be attributed to the slight deviations in real-life harmonics from the ideal relation.

Table 2: *Psychoacoustic evaluation of harmonics relations*

Harmonic	*Mean Predicted Ratio*	*Mean Ratio Error (%)*	*Median Ratio Error (%)*	*% Within 10% of True Ratio*
H2/F0	2.07	5.18	2.38	83.86
H3/F0	3.16	6.79	3.42	75.26
H4/F0	3.99	3.78	2.36	92.63
H5/F0	5.24	5.91	4.07	85.61

4. Discussion

4.1. Key Findings

During the practical tests, many observations have been made through a series of tests and evaluations. Gammatone filterbank cochlear model

was able to provide a well-constructed and accurate perceptual representation of spectral and temporal data with a spatial resolution that is in direct relationship with the number of channels. Auditory spectrograms (spectral-temporal representation) provide well-informed data organization that is suitable for machine learning applications. CNN worked well with the spectrogram-represented data and was able to capture harmonics patterns. DNN was suitable for learning patterns from global features which helped in forming the predictions. LSTM layers were an optimal choice for the temporal processing of the outputs of the previous levels. Using the MSE as a loss function enabled the model to prioritize larger errors over smaller errors which is perceptually plausible as small frequency deviations are less noticeable. The hybrid model was impressively able to capture frequency patterns from the cochlear-modeled data and make accurate predictions. Studying other harmonic components along with the fundamental frequency was an important contribution, while the harmonic information helped linking the frequency patterns.

Compared to traditional signal processing and probabilistic methods in estimating pitch like YIN (De Cheveigné & Kawahara 2002) [8] and pYIN (Mauch & Dixon 2014) [9], this project achieved a close accuracy with deep learning although the cochlear-inspired processing introduced higher computational complexity. Furthermore, compared to deep learning models like CREPE (Kim et al. 2018) [5], good accuracy was achieved by treating the frequency as a regression task which allowed for finer resolution, unlike the classification-shaped task in CREPE.

4.2. Limitations

Despite the good performance and the high accuracy of the model, some limitations still persist and can be addressed. Processing the audio with gammatone filterbank presents a moderate computational complexity that may limit real-time implementation. This can be overcome by trading off the high resolution. It is also important to adress that the model is trained on clean, noise free data, which might cause confusion and inaccurate predictions if employed for noisy samples.

4.3. Future Work

Following the proposed work, there are several directions for development and further exploration, like building a specialized dataset that is optimal for cochlear modeling objectives, while offering a variety

of sounds and pitch levels. A crucial next step would also be the expansion of the work to noisy environments and realworld applications.

5. Conclusion

This research aimed to develop a model that simulates the human's cochlea in pitch-related information analysis. Through the implementation of the Gammatone cochlear model and hybrid deep neural networks, the model succeeded with predicting the fundamental frequency and four other harmonics. The results demonstrated the feasibility of integrating computational cochlear models with machine learning algorithms for perceptual pitch information detection. These contributions can support future developments in cochlear implant processing and music engineering enriching the auditory research field.

REFERENCES

[1] Møller, A. R. (2006). Hearing: Anatomy, Physiology, and Disorders of the Auditory System. Amsterdam: Academic Press.

[2] Haile, L. M., Orji, A. U., Reavis, K. M., Briant, P. S., Lucas, K., Steinmetz, J. D., Ong, L. K., Hay, S. I., and Mokdad, A. H. (2023). Hearing Loss Prevalence, Years Lived With Disability, and Hearing Aid Use in the United States From 1990 to 2019: Findings From the Global Burden of Disease Study. Ear and Hearing, 45(1), 257–267.

[3] Tabibi, S., Kegel, A., Lai, W. K., and Dillier, N. (2017). Investigating the use of a Gammatone filterbank for a cochlear implant coding strategy. Journal of Neuroscience Methods, 277, 63–74.

[4] Islam, R., and Tarique, M. (2024). Investigating the Performance of Gammatone Filters and Their Applica-bility to Design Cochlear Implant Processing System. Designs, 8(1), 16.

[5] Kim, J. W., Salamon, J., Li, P., and Bello, J. P. (2018). CREPE: A Convolutional Representation for Pitch Estimation. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 161–165). Piscataway, NJ: IEEE.

[6] Chou, H., Chen, M.-T., and Chi, T.-S. (2018). A Hybrid Neural Network Based on the Duplex Model of Pitch Perception for Singing Melody Extraction. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 381–385). Piscataway, NJ: IEEE.

[7] Cella, C.-E., Ghisi, D., Lostanlen, V., Lévy, F., Fineberg, J., and Maresz, Y. (2020). TinySOL: An audio dataset of isolated musical notes (Version 6.0) [Data set]. Geneva: Zenodo.

[8] de Cheveigné, A., and Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. The Journal of the Acoustical Society of America, 111(4), 1917–1930.

[9] Mauch, M., and Dixon, S. (2014). PYIN: A fundamental frequency estimator using probabilistic threshold distributions. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 659–663). Piscataway, NJ: IEEE.