



Attention Models and Auditory Transduction Features for Noise Robustness

Cathal Ó Faoláin¹, Andrew Hines¹

¹School of Computer Science, University College Dublin, Ireland

cathal.mellon-whelan@ucdconnect.ie, andrew.hines@ucd.ie

Abstract

Human abilities surpass current speech processing systems in complex, noisy environments. While popular inputs for Automatic Speech Recognition (ASR) systems, such as raw acoustic signals and Mel spectrograms, perform well in quiet conditions, their effectiveness declines in noise. A recently developed generative WaveNet-based model emulates human auditory transduction in real time, offering alternative input features through its “IHCogram” outputs. We investigate these IHCograms across various signal-to-noise ratios (SNRs) using state-of-the-art feature encoders. Our findings show that IHCograms significantly enhance phoneme recognition in noisy conditions with minimal computational overhead, regardless of the model encoder used. Additionally, we introduce our Attention Feature Encoder (AFE) models, which leverage the channel structure of IHCograms and demonstrate superior size and performance compared to existing feature encoders.

Index Terms: speech recognition, noise robust, auditory transduction

1. Introduction

Speech is a fundamental mode of human communication, and when reliability is ensured, it serves as a powerful tool for interacting with machines [1]. Despite advancements in Automatic Speech Recognition (ASR) systems, human performance in complex noisy environments remains superior for detection, analysis, and recognition. [2]. Humans have accuracy above chance level at Signal-to-Noise Ratios (SNRs) as low as -18 dB [3], while even moderate noise impacts ASR.

Modern ASR systems typically consist of two encoder stages: a feature encoder and a context encoder [4]. Together, these convert a speech signal into text [5]. The feature encoder extracts feature vectors that capture local, frame-level information, such as frequency and temporal details [1]. These feature vectors are then given to the context encoder, which models the relationships between them over time – incorporating essential temporal dependencies and contextual cues. Analysing how input choice affects a model’s potential requires examining the information accessibility within the feature vectors.

While feature encoders play a crucial role in extracting meaningful signal information, the choice of input representation significantly impacts their performance [1] [6]. Human auditory transduction evolved alongside speech, and is finely tuned to enhance speech encoding [7]. This has long been recognised, and has led to a popular preference for psycho-acoustic inspired features [1] such as Mel Spectrograms and MFCCs. More recently, Deep Neural Networks (DNNs) can identify and extract relevant features directly from the raw signal. While effective in quiet conditions, these features degrade

quickly in noise [2] [6], as they lack the signal-filtering phenomena that the human auditory system evolved to preserve speech information in complex listening environments.

The sound wave is transduced into neural signals by Inner Hair Cells (IHCs) in the cochlea: salient frequencies are amplified and their phase information is locked onto, alongside other phenomena [8] [9]. These phenomena make human speech recognition robust to noise before, and independent of, both context and language processing [3] [10]. Phenomenological models mimic these signal-filtering processes by reproducing the neural responses generated by auditory transduction [9]. Despite their high accuracy and state-of-the-art performance on various speech tasks [11] [12] [2], the computational costs of these models has made them unsuitable for real-time applications [13].

Recently, a WaveNet-based [14] deep-learning model [13] was trained to approximate the IHC outputs of a phenomenological model [15]. We will refer to this generative DNN as WavIHC. WavIHC operates in real-time, emulating key signal-filtering phenomena without the computational overhead of its predecessor. Previous studies have demonstrated the potential of phenomenological models in improving phoneme recognition in noise [2], but the computational complexity limited its application. WavIHC offers a computationally efficient alternative that retains the perceptually relevant information [13] [16].

In this paper, we use WavIHC’s output (IHCograms [13]) as a novel input feature for speech processing systems. We hypothesize that these features will provide significant noise robustness compared to traditional Mel Spectrograms or raw signal inputs, independent of the neural network architecture used for feature encoding. To evaluate this, we compare the performance of three state-of-the-art feature encoders along with our own novel AFE feature encoders on a phoneme recognition task using the TIMIT dataset [17] with added noise. Our results show that IHCogram features significantly reduce Phoneme Error Rate (PER) in noisy conditions, and the proposed AFE models outperform existing encoders in size and performance. These improvements add a modest overhead to inference time.

2. Modelling the Auditory Periphery

Inner Hair Cells (IHCs) are essential for converting sound waves into electrical signals for the brain. As sound waves pass through the outer, middle, and inner ear, they are filtered and processed. In the inner ear, sound travels along the basilar membrane, which vibrates at different frequencies from base to apex. This property makes each section sensitive to a specific Characteristic Frequency (CF). IHCs respond to these vibrations, each tuned to a particular CF. For more on human auditory transduction and the role of IHCs, see [18] [9] [8].

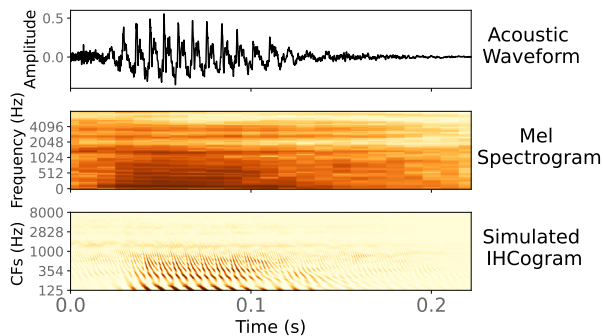


Figure 1: *Acoustic, Mel Spectrogram and Simulated IHC representations of the author saying the /p/ phoneme.*

IHC electrical potentials decode sound into CF channels [13], capturing variations in strength over time and encoding both temporal and spectral features of speech. In this way they are analogous to spectrograms. Unlike Mel spectrograms however, IHC potentials maintain precise timing, preserving critical temporal details for noise robustness. The resulting output (IHCogram) represents the acoustic signal as the electrical impulses that serve as input for the initial stages of human neural processing. An example is illustrated in Fig. 1.

Auditory transduction has evolved alongside speech, leading to efficient and noise-robust representations. A key feature of IHC responses is phase-locking, where IHCs synchronize with sound waves, even in noisy environments [8]. This response is particularly effective at lower frequencies, aligning with human speech fundamental frequencies [8].

WavIHC simulates the non-linear and adaptive processes that enhance IHCogram robustness, including phase-locking, non-linear tuning, two-tone suppression, compression and active amplification [18] [8] [9] [13]. This results in WavIHC providing a much more physiologically-inspired representation of the sound-wave compared to traditional signal-based inputs.

Phenomenological models aim to replicate experimentally observed auditory responses. However, this requires high sampling rates and complex computations – often exceeding real-time speeds. Despite this, their accuracy in simulating auditory transduction makes them invaluable when emulating human perception and performance is critical. For example, Bruce et al’s model [15] has been used in speech quality metrics [11], speech intelligibility [12], hearing aid prototyping [19] and robust automatic speech recognition ASR [2].

WavIHC, developed by Nagathil et al. [13], addresses this runtime limitation by emulating Bruce et al.’s model using a WaveNet-based architecture [14]. WavIHC was adapted for regression and taught to predict the IHC potentials of C CFs simultaneously. This model uses dilated causal convolutions to achieve a large receptive field and autoregression to incorporate past predictions into current outputs. With 7 layers, 4 stacks, and 127 residual and skip connections, WavIHC operates at a lower sampling rate of 16 kHz, leveraging the anti-aliasing properties of CNN architectures [20]. Its fully differentiable design ensures efficient performance on both GPUs and CPUs – enabling a real-time runtime [13] [16].

3. Method

Our aim is to assess the impact of using WavIHC’s IHCograms as an alternative to signal or Mel spectrogram inputs for deep learning speech recognition. By removing the context encoder

stages, we focus on the feature encoders, retraining them from scratch for phoneme recognition. We assess how feature input choice affects their ability to extract key features in both quiet and noisy environments.

To determine how input choice affects information accessibility within feature vectors, we investigate how the input affects a model’s ability to extract key speech features in noise. We use the feature encoder stages of five deep learning models – four designed for acoustic signals and one using Mel spectrograms for comparison. The four signal-based models are adapted to accept IHCogram inputs, and we analyse how this change impacts their performance on a phoneme recognition task at different signal-to-noise ratios (SNRs). By comparing the performance of the signal-based and IHCogram-based versions of the same models, we can attribute any differences in performance to the input type rather than the architecture.

Following the principle that linear separability is indicative of information accessibility for downstream tasks [5], a single fully connected layer is used to classify the feature encoder outputs into phonemes. In order to eliminate any potential biases from model architecture, pre-training or domain adaptation, we also standardize the training method and compare a comprehensive selection of models .

Our models are all trained from scratch using only clean speech via supervised learning. When a downstream task is known, this remains the most successful approach for learning task-specific representations [5], while including noise in training can degrade performance on both clean data and unseen noisy conditions due to domain mismatch [21].

We isolate and use the feature encoder stage of 3 state-of-the-art models: Contrastive Predictive Coding (CPC) [22], Wav2vec 2.0 (W2V) [4] and Autoregressive Predictive Coding (APC) [5]. CPC and W2V utilise CNN-based architectures but differ significantly in their sampling rates, context representations, and architectural designs (see Tables 1, 2 and Section 4.2 for details). Additionally, we developed two custom models, referred to as our Attention Feature Encoders (AFEs).

The Attention Feature Encoders (AFEs) models were developed to exploit IHCograms primary advantage: their meaningful channels. AFE and AFE2 are based on CPC and W2V, respectively, and incorporate Squeeze-and-Excitation (SE) modules [23] after CNN layers to explicitly model dependencies between frequency channels. The CNN layers extract temporal features while downsampling the signal, whereas the SE modules emphasize spectral features using attention mechanisms. This form of attention preserves CF channel separation – which is a characteristic of early neural processing [24]. Additionally, both AFE models replace dropout with batch normalization to improve stability and normalize feature distributions [25] [26].

4. Experimental Set-Up

4.1. Dataset

The speech dataset used is TIMIT [17], containing 16 kHz sampled audio recordings of 630 speakers reading 10 phonetically rich sentences. This dataset provides detailed time-aligned phonetic transcriptions, with phoneme labels collapsed into 39 classes following standard protocol [27]. To address TIMIT’s lack of a validation set, we re-split the dataset into training, validation, and test sets, with durations of 242, 32, and 48 minutes, respectively. The training set was further divided into 5 folds for cross-validation.

Noise Addition: Noise was added to our test set at 5 dB

Table 1: *Model parameters for training and testing.*

Model	Activation	Strides	Kernels
CPC	ReLU	(5, 4, 2, 2, 2)	(10, 8, 4, 4, 4)
W2V	GeLU	(5, 2, 2, 2, 2, 2, 2)	(10, 3, 3, 3, 3, 2, 2)
APC	ReLU	N/A	N/A
AFE	ReLU	(5, 4, 2, 2, 2)	(10, 8, 4, 4, 4)
AFE2	GeLU	(5, 2, 2, 2, 2, 2, 2)	(10, 3, 3, 3, 3, 2, 2)

increments, spanning an SNR range of 30 dB to -10 dB. Five noise types were used: *steady*, locally-generated white noise and *amplitude modulated* (children playing, air conditioner, dogs barking and street music) noise taken from the UrbanSound Dataset [28]. Non-overlapping segments were taken from the UrbanSound files after the noise event began and before it ended, ensuring they were longer than their randomly paired TIMIT audio files. These pairings remained consistent across different SNRs. All UrbanSound files were converted to standard WAV format before addition.

Our feature encoders sample sound at different rates (see Table 2). Audio files were framed at 10 ms and approximately 20 ms intervals, corresponding to each model’s sampling rate. Every frame was assigned a phoneme label based on the majority phoneme present, as indicated by the phonetic transcriptions. Batch sizes also varied by model: APC used a batch size of 25, while all other models used 4. Inputs and outputs were dynamically padded to improve GPU efficiency.

4.2. Models

Table 1 summarises the model parameters: activation functions, strides and kernels (where applicable). Table 2 describes the Feature Encoder models test configurations.

Nine feature encoders were tested including five variations of state-of-the-art models. Tests are labelled to denote input type: MEL_ for Mel spectrograms, SIG_ for acoustic signals, and IHC_ for WavIHC’s IHCograms. Each test was also trained separately. Most models have a layer size of 512, except AFE models, which use a size of 80.

Adapting encoders to accept IHCograms required minor changes. First, the input channel size was changed from 1 to 80, to match WavIHC’s tested output channels [13] [16]. Then, the encoders were placed atop a pre-trained, frozen WavIHC model.

All feature encoders were trained from scratch using supervised learning with cross entropy loss and the Adam optimizer for up to 100 epochs. Validation set performance and model configuration was saved every 4 epochs. Training stopped if validation loss increased 5 times consecutively, and the best performing model on the validation set was saved for testing.

4.3. Training and Testing

Models were trained using 5-fold cross-validation, with one fold serving as validation while the remaining four were used for training. This process was repeated five times, ensuring each fold was used as validation once. The folds were then combined for training, with validation performed on a held-out dataset. This approach provided a performance range for each model, as each was optimized on six different validation sets. Only clean speech was used for training. The model that performed best on their assigned validation set for each k-fold and full training set was saved. The saved best-performing models were tested on the clean test set, as well as the test sets corresponding to

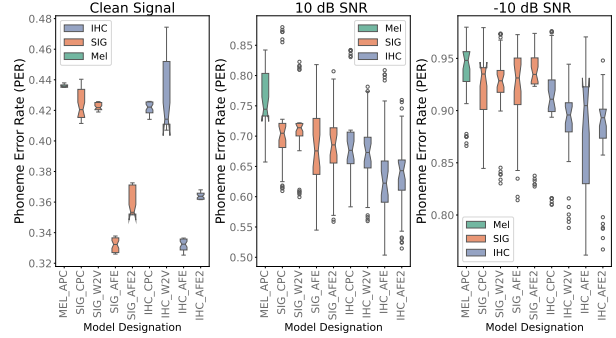


Figure 2: *Phoneme Recognition in Averaged Noise at 3 SNRs: Clean Signal, 10 dB and -10 dB. Shown is a box plot of the performance of each designation across all noise types at the given SNRs, which were selected purely to give an indication of the performance over the full SNR range tested (Clean to -10 dB). Note that the y-axis range is set based on the range of performance at that SNR, and neither the start and end points, nor the interval size, match the other SNRs.*

every SNR step for all noise types. Every model was therefore tested 6 times with each of the 10 test sets.

5. Results

5.1. Phoneme Classification in Noise

To determine whether using IHCograms improves phoneme recognition error in noisy environments, we analysed model performance in averaged noise, calculated as the average Phoneme Error Rate (PER) across five noise types at each SNR. Fig. 2 shows that models using IHCogram inputs consistently outperformed their signal-based counterparts as SNR decreased. At 10 dB SNR, IHCogram-based models had an average PER improvement of 3.9%, which remained at 3.3% at -10 dB SNR. MEL_APC performed comparably to CPC and W2V in quiet conditions but degraded rapidly in noise, achieving the worst performance of the tested models.

Our AFE models outperformed all tested state-of-the-art feature encoders, regardless of input type and despite their smaller model size. In quiet, SIG_AFE and SIG_AFE2 achieved average PERs of 33.2% and 35.9%, while IHC_AFE and IHC_AFE2 achieved 33.2% and 36.4%. This significantly outperformed the next best model, IHC_CPC which had an average PER of 42.1%. Across all SNRs, IHC_AFE2 performed best, with an average PER of 62.5% and more stable performance compared to IHC_AFE.

5.2. Noise Robustness Across Scenarios

The impact of using IHCograms varies depending on the noise type, as is shown in Fig. 3. The SNR threshold where IHCograms begin to significantly reduce PER differs by noise type, for example 25 dB for white noise and 10 dB for street music. However, two key trends emerge: 1) IHCograms significantly enhance phonemes recognition in noisy environments, with improvements continuing into even very low SNRs such as -10 dB, and 2) using IHCograms does not significantly decrease performance in quiet conditions – any difference lies well within the interquartile range of the signal-based models. Fig. 2 therefore highlights that IHCogram inputs provide robustness across diverse noise conditions, without compromising performance in quiet environmental conditions.

Table 2: ASR Feature Encoder Models. Average $t(s)$ is defined as the average test time required to process the 48 mins test set. Using IHCograms increases model size by WavIHC’s 3,529,156 parameters and increases the inference time.

Test	Feature Enc.	Input	Context (ms)	Blocks	Output (ms)	Architecture	Parameters	Avg $t(s)$
MEL_APC	APC [5]	Mel	25	3	10	MLP	586,791	15.50
SIG_CPC	CPC [22]	Signal	10	5	10	CNN	5,268,007	14.34
IHC_CPC		IHC	10	5	10		9,201,643	63.19
SIG_W2V	W2V [4]	Signal	25	7	~20	CNN	4,220,455	8.30
IHC_W2V		IHC	25	7	~20		8,154,091	56.19
SIG_AFE	AFE (ours)	Signal	10	5	10	CNN+SE	137,184	14.02
IHC_AFE		IHC	10	5	10		3,729,540	66.06
SIG_AFE2	AFE2 (ours)	Signal	25	7	~20	CNN+SE	113,834	9.70
IHC_AFE2		IHC	25	7	~20		3,706,190	56.40

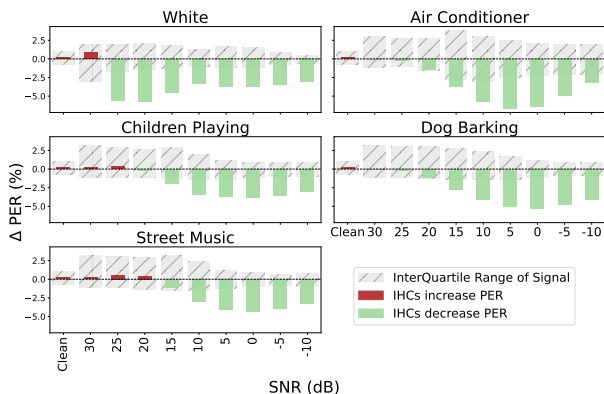


Figure 3: Average Effect of Using IHCograms instead of Signal as Input on PER. Average difference between IHCograms and signals for the same model type. The overall difference in PER is shown on the y-axis, and the SNRs ranging from high- to low- on the x-axis. The average interquartile range of the signal-based models shown for comparison.

AFE2 Spotlight: To avoid confounding variables from averaging multiple models, we focused on the performance differences between the signal- and IHCogram-based versions of AFE2. The results align with the trends in Fig. 3, highlighting the practical benefits of both the AFE2 architecture and IHCogram inputs. Notably, the IHC_AFE2 model starts with a high performance of 36.4% PER in quiet conditions and demonstrates a gradual decline as noise increases, illustrating its robustness in challenging acoustic environments.

5.3. WaveIHC Complexity Overhead

The computational complexity of using IHCograms stems from integrating the frozen WavIHC model and adapting the signal-based models to process the 80 channels of IHCogram inputs. This addition introduces 3,529,156 frozen parameters across all models, which, while not trainable, increases memory and inference costs. For trainable parameters, models like CPC and W2V see an increase of 404,480 parameters, due to having 512 hidden units in each layer. In contrast, our AFE models require only a 63,200 parameter increase, owing to their smaller size.

On average, processing a 48-minute test set required an additional 48.86 seconds when using IHCograms, as compared to signal-based models (see Table 2). This increase in inference time can be attributed solely to the integration of WavIHC, indicating a trade-off between improved auditory modelling and time complexity. However, this is well within real-time, with an

Table 3: PER in Noise for AFE2 Model. The best performing version at that SNR is bolded. PER is an average of the 6 models trained (see 4.4). ∞ is the clean signal (no noise).

dB	∞	30	25	20	15	10	5	0	-5	-10
White										
SIG	35.9	51.3	62.3	68.2	73.5	79.4	85.8	91.0	94.8	96.8
IHC	36.4	42.7	49.2	56.3	64.8	74.2	81.2	86.3	90.1	92.5
Air Conditioner										
SIG	35.9	37.3	39.6	44.8	54.8	67.6	78.6	86.0	90.5	93.2
IHC	36.4	38.1	40.5	45.1	53.1	62.6	71.6	78.3	83.8	87.9
Children Playing										
SIG	35.9	37.2	39.8	46.6	58.3	70.3	79.9	87.1	91.9	94.8
IHC	36.4	38.0	40.4	45.9	54.8	65.2	74.3	81.5	86.6	89.9
Dog Barking										
SIG	35.9	37.0	38.8	42.6	49.0	57.0	65.0	72.1	78.1	83.2
IHC	36.4	37.5	39.1	42.0	46.6	52.8	59.7	66.4	72.8	78.5
Street Music										
SIG	35.9	37.0	39.5	45.8	55.9	67.5	77.7	85.2	90.1	93.1
IHC	36.4	37.8	40.2	45.5	54.0	63.7	72.6	79.9	85.5	89.1

average real-time factor [1] (RTF) of 0.02 for the slowest model tested. Models were trained and tested on an NVIDIA GeForce RTX 4090 GPU.

6. Conclusions

We explored the use of WavIHC-simulated IHCograms [13] as alternatives to raw signals and Mel spectrograms across various SNRs. Our analysis showed that IHCogram inputs significantly improved phoneme recognition in noisy environments, reducing the phoneme error rate (PER) by an average of 3.9% at 10 dB SNR, with no negative effect in quiet conditions. These improvements were consistent across different architectures and were solely due to the input differences. While IHCogram features increased inference time and complexity, the overhead was minimal compared to downstream modeling [29] and remained within real-time limits. We also introduced our Attention Feature Encoder (AFE) models, which leverage the meaningful channel representations of IHCograms. Using Squeeze-and-Excitation (SE) modules, IHC_AFE2 achieved the best overall PER in both quiet and noisy conditions. Our findings indicate that IHCograms offer more robust features than traditional inputs. Future work will investigate whether these advantages persist with context modelling, focusing on using IHCogram features and AFE2 to help develop a fully end-to-end self-supervised ASR model optimized for noisy environments.

7. Acknowledgements

This publication has emanated from research conducted with the financial support of Taighde Éireann – Research Ireland under Grant number 18/CRT/6183. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission

8. References

- [1] M. Malik, M. K. Malik, K. Mehmood, and I. Makhdoom, “Automatic speech recognition: a survey,” *Multimedia Tools and Applications*, vol. 80, pp. 9411–9457, 2021, publisher: Springer.
- [2] M. S. Alam, M. S. A. Zilany, W. A. Jassim, and M. Y. Ahmad, “Phoneme Classification Using the Auditory Neurogram,” *IEEE Access*, vol. 5, pp. 633–642, 2017. [Online]. Available: <http://ieeexplore.ieee.org/document/7815325/>
- [3] G. A. Miller and P. E. Nicely, “An Analysis of Perceptual Confusions Among Some English Consonants,” *The Journal of the Acoustical Society of America*, vol. 27, no. 2, pp. 338–352, Mar. 1955. [Online]. Available: <https://pubs.aip.org/jasa/article/27/2/338/746012/An-Analysis-of-Perceptual-Confusions-Among-Some>
- [4] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations.”
- [5] Y.-A. Chung, W.-N. Hsu, H. Tang, and J. Glass, “An Unsupervised Autoregressive Model for Speech Representation Learning,” Jun. 2019, arXiv:1904.03240 [cs, eess]. [Online]. Available: <http://arxiv.org/abs/1904.03240>
- [6] W. Alkhalidi, W. Fakhr, and N. Hamdy, “Automatic speech/speaker recognition in noisy environments using wavelet transform,” in *The 2002 45th Midwest Symposium on Circuits and Systems, 2002. MWSCAS-2002.*, vol. 1, Aug. 2002, pp. 1–463. [Online]. Available: <https://ieeexplore.ieee.org/document/1187258/?arnumber=1187258>
- [7] Y. Shao, Z. Jin, D. Wang, and S. Srinivasan, “An auditory-based feature for robust speech recognition,” in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, Apr. 2009, pp. 4625–4628, iSSN: 2379-190X. [Online]. Available: <https://ieeexplore.ieee.org/document/4960661/?arnumber=4960661>
- [8] E. A. Lopez-Poveda, “Spectral processing by the peripheral auditory system: facts and models,” *International Review of Neurobiology*, vol. 70, pp. 7–48, 2005, publisher: Elsevier.
- [9] M. Rudnicki, O. Schoppe, M. Isik, F. Völk, and W. Hemmert, “Modeling auditory coding: from sound to spikes,” *Cell and Tissue Research*, vol. 361, no. 1, pp. 159–175, Jul. 2015. [Online]. Available: <https://doi.org/10.1007/s00441-015-2202-z>
- [10] G. A. Miller, G. A. Heise, and W. Lichten, “The intelligibility of speech as a function of the context of the test materials.” *Journal of experimental psychology*, vol. 41, no. 5, p. 329, 1951, publisher: American Psychological Association.
- [11] J. M. Kates and K. H. Arehart, “The hearing-aid speech perception index (HASPI),” *Speech Communication*, vol. 65, pp. 75–93, 2014, publisher: Elsevier.
- [12] A. Hines and N. Harte, “Improved speech intelligibility with a chimaera hearing aid algorithm,” in *Interspeech 2012*. ISCA, Sep. 2012, pp. 1468–1471. [Online]. Available: https://www.isca-archive.org/interspeech.2012/hines12_interspeech.html
- [13] A. Nagathil and I. C. Bruce, “WaveNet-based approximation of a cochlear filtering and hair cell transduction model,” *The Journal of the Acoustical Society of America*, vol. 154, no. 1, pp. 191–202, Jul. 2023. [Online]. Available: <https://pubs.aip.org/jasa/article/154/1/191/2902087/WaveNet-based-approximation-of-a-cochlear>
- [14] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, “WaveNet: A Generative Model for Raw Audio,” Sep. 2016, arXiv:1609.03499 [cs]. [Online]. Available: <http://arxiv.org/abs/1609.03499>
- [15] M. S. A. Zilany and I. C. Bruce, “Modeling auditory-nerve responses for high sound pressure levels in the normal and impaired auditory periphery,” *The Journal of the Acoustical Society of America*, vol. 120, no. 3, pp. 1446–1466, Sep. 2006. [Online]. Available: <https://pubs.aip.org/jasa/article/120/3/1446/899458/Modeling-auditory-nerve-responses-for-high-sound>
- [16] C. Faoláin and A. Hines, “Speech Feature Fidelity from a Generative Auditory Transduction Model,” in *2024 35th Irish Signals and Systems Conference (ISSC)*, Jun. 2024, pp. 1–6, iSSN: 2688-1454. [Online]. Available: <https://ieeexplore.ieee.org/document/10603120/?arnumber=10603120>
- [17] J. S. Garofolo, L. F. Lamel, W. M. Fisher, D. S. Pallett, N. L. Dahlgren, V. Zue, and J. G. Fiscus, “TIMIT acoustic-phonetic continuous speech corpus,” (*No Title*), 1993, publisher: Linguistic data consortium.
- [18] P. Avan, B. Büki, and C. Petit, “Auditory Distortions: Origins and Functions,” *Physiological Reviews*, vol. 93, no. 4, pp. 1563–1619, Oct. 2013. [Online]. Available: <https://www.physiology.org/doi/10.1152/physrev.00029.2012>
- [19] A. Hines and N. Harte, “Reproduction of the Performance/Intensity Function using image processing and an auditory nerve computational model,” 2010.
- [20] A. H. Ribeiro and T. B. Schön, “How Convolutional Neural Networks Deal with Aliasing,” in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2021, pp. 2755–2759, iSSN: 2379-190X. [Online]. Available: <https://ieeexplore.ieee.org/document/9414627/?arnumber=9414627>
- [21] Q.-S. Zhu, J. Zhang, Z.-Q. Zhang, M.-H. Wu, X. Fang, and L.-R. Dai, “A Noise-Robust Self-Supervised Pre-Training Model Based Speech Representation Learning for Automatic Speech Recognition,” in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022, pp. 3174–3178, iSSN: 2379-190X. [Online]. Available: <https://ieeexplore.ieee.org/document/9747379/>
- [22] A. v. d. Oord, Y. Li, and O. Vinyals, “Representation Learning with Contrastive Predictive Coding,” Jan. 2019, arXiv:1807.03748 [cs, stat]. [Online]. Available: <http://arxiv.org/abs/1807.03748>
- [23] J. Hu, L. Shen, and G. Sun, “Squeeze-and-Excitation Networks.”
- [24] N. Kraus and T. Nicol, “Brainstem origins for cortical ‘what’ and ‘where’ pathways in the auditory system,” *Trends in Neurosciences*, vol. 28, no. 4, pp. 176–181, Apr. 2005. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0166223605000470>
- [25] S. Ioffe and C. Szegedy, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift.”
- [26] S. Santurkar, D. Tsipras, A. Ilyas, and A. Ma, “How Does Batch Normalization Help Optimization?”
- [27] K.-F. Lee and H.-W. Hon, “Speaker-independent phone recognition using hidden Markov models,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, Nov. 1989, conference Name: IEEE Transactions on Acoustics, Speech, and Signal Processing. [Online]. Available: <https://ieeexplore.ieee.org/document/46546/?arnumber=46546>
- [28] J. Salamon, C. Jacoby, and J. P. Bello, “A Dataset and Taxonomy for Urban Sound Research,” in *Proceedings of the 22nd ACM international conference on Multimedia*. Orlando Florida USA: ACM, Nov. 2014, pp. 1041–1044. [Online]. Available: <https://dl.acm.org/doi/10.1145/2647868.2655045>
- [29] S. Gondi, “Wav2Vec2.0 on the Edge: Performance Evaluation,” Feb. 2022, arXiv:2202.05993 [cs]. [Online]. Available: <http://arxiv.org/abs/2202.05993>