Phonak
Insight.

Revolutionary Speech Understanding with Spheric Speech Clarity

Hearing aid users often encounter challenges when faced with background noise. To date, artificial intelligence (AI) has not been directly employed to address this issue. Phonak Audéo Sphere Infinio featuring Spheric Speech Clarity is an AI-based solution designed to tackle denoising, effectively removing undesirable sounds while maintaining the speech signal in real-time, improving Signal-To-Noise ratio by up to 10dB. This success is powered by Phonak's Deep Spheric Optimized Neural Integrated Chip (DEEPSONIC™) and the Audio Quality Estimator DNN.

June 2024: Henning Hasemann, Alena Krylova

Key highlights

- Spheric Speech Clarity, introduced in Phonak Audéo Sphere Infinio, delivers unprecedented speech and noise separation by harnessing the power of deep neural networks.

- The DEEPSONIC™ neural network processing chip is unparalleled in the industry.

- The Audio Quality Estimator deep neural network (DNN) allows for the prediction of human ratings for audio at large scale and is a key enabler for the success of Spheric Speech Clarity.

A Sonova brand

PHONAK
life is on

Considerations for practice

- Spheric Speech Clarity differentiates between wanted and unwanted sound, removing the latter from the signal. This results in a 10dB enhancement in Signal-to-Noise Ratio (SNR) in complex group conversations (Raufer et al., 2024).
- Perceptual clinical study conducted at the Phonak Audiological Research Center (PARC) showed that the technical benefit of Spheric Speech Clarity translates into double the likelihood of understanding speech in a complex speech in noise scenario (compared to without) and speech understanding increased by up to 36.8% compared to two leading competitor devices (Wright, A., et al., 2004).
- Clinical study results confirm that Spheric Speech Clarity delivers perceptual benefit with double the likelihood of understanding speech in a complex noisy environment, from any direction (Wright et al., 2024).
- The Audio Quality Estimator allows for the prediction of human ratings for audio at large scale, supporting the accurate and reliable ratings of noise, sound quality and preference, this is instrumental for the development of deep learning approaches that aim to improve speech understanding, such as Spheric Speech Clarity.
- DEEPSONIC™ can accelerate a large variety of different neural network types and provides tremendous flexibility for future applications.

Introduction

Ambient noise has consistently posed a significant challenge and impediment to adequate speech understanding for hearing aid users. Despite significant improvements in hearing aid denoising capabilities, communication in noisy environments remains one of the scenarios where hearing aid users experience the lowest satisfaction with their devices (Appleton-Huber, 2022).

The hearing aids market has been witnessing a substantial influx of features based on artificial intelligence (AI) such as advancements in acoustic scene classification or user support for hearing aid configuration. However, hearing aid manufacturers have addressed denoising, the process of removing noise from a signal in real-time, primarily using conventional signal processing techniques, such as beamformers, rather than AI (Hasemann & Krylova, 2024). New findings indicate that deep learning, a specialized branch of AI, has the potential to significantly diminish background noise, leading to remarkable enhancements in speech intelligibility for individuals using hearing aids (Diehl et al., 2023).

Spheric Speech Clarity is a cutting-edge deep neural network (DNN) based solution developed by Phonak to further tackle the infamous cocktail party problem. To our knowledge, it is the best performing model of its kind on the market to date. Paired with the latest machine learning (ML) based AutoSense OS 6.0 (Appleton-Huber, 2015), which recognizes acoustic situations in real-time, Spheric Speech Clarity addresses the denoising task directly by differentiating between wanted and unwanted sounds and removing the latter from the signal. This results in 10dB enhancement in Signal-to-Noise Ratio (SNR) in complex group conversations (Raufer et al., 2024). Unlike rule-based systems that attempt to filter out undesirable sounds, Spheric Speech Clarity emulates human-like perception in sound recognition and processing. With this paper, we aim to provide an insightful look into Spheric Speech Clarity, its development, and underlying mechanisms.

Essential concepts in artificial intelligence

Artificial intelligence is an information technology concerned with simulating aspects of human intelligence, such as learning, reasoning, problem-solving, and perception, in machines. Some AI algorithms, like rule-based and expert systems, rely on logical deduction and/or use predefined rules and knowledge bases to solve problems. More modern AI approaches, such as machine learning and, in particular, deep learning, can learn complex behavior from examples and behave correctly in new situations (Chatterjee & Zielinski, 2022).

The "brain" of a machine learning application is a mathematical structure called a model, which contains adjustable variables called parameters. In the training process, the model is introduced to example data points referred to as training data. Throughout this process, the parameters undergo continuous adjustment, gradually leading to generalization – the model's ability to make accurate predictions on unseen data. Once the training is completed and the model has learned patterns and relationships from that data, it can be deployed for inference, or, in other words, put into practical use.

Neural networks are a specific kind of machine learning model that are known to be good at modeling perception-like tasks. Like most other machine learning approaches, they are trained with large numbers of examples of inputs and expected outputs. Their model structure is inspired by neurons in biological brains. Deep neural networks (DNNs) are particularly complex neural networks that have lately gained tremendous

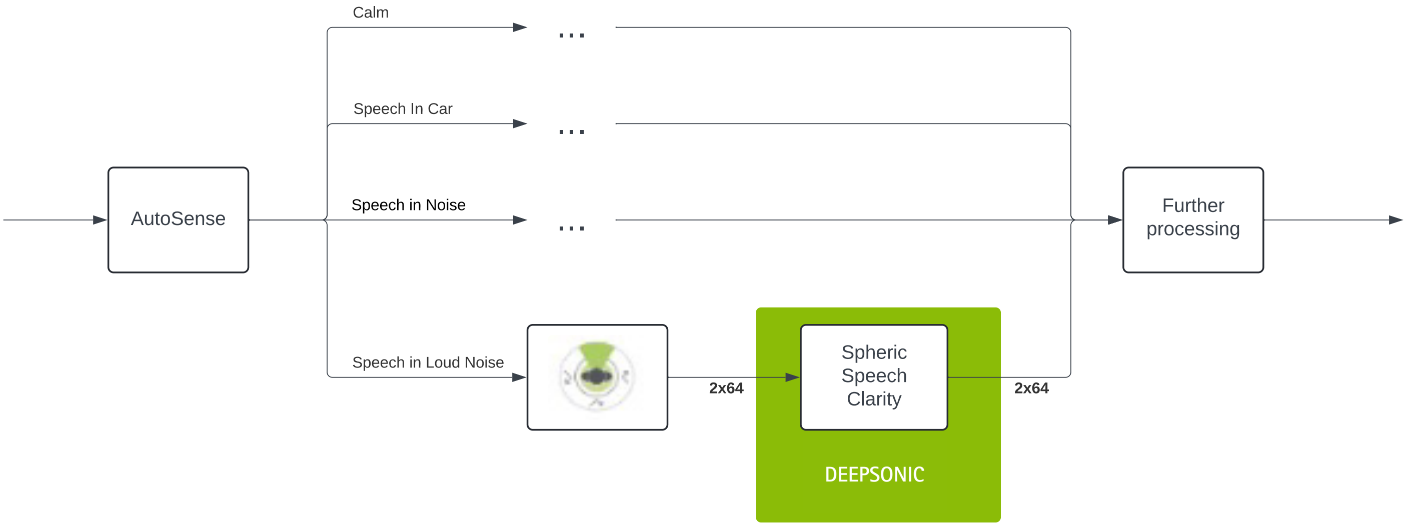

Figure 1. Signal flow. AutoSense OS chooses the correct path to process according to the situation and the individual fitting settings. In "Speech in Loud Noise", after spatial preprocessing, Spheric Speech Clarity is applied using the DEEPSONIC ™ chip to separate speech and noise.

popularity due to their ability to model very complex relationships of inputs and outputs and still generalize correctly from them.

Spheric Speech Clarity

Spheric Speech Clarity, introduced first with Phonak Audéo Infinio Sphere, is a new, proprietary DNN-based sound processing system that revolutionizes speech understanding in the most challenging acoustic environments. It is the first hearing aid technology that fully exploits the power of artificial intelligence to separate speech from noise.

This section takes a closer look at the systems involved in signal processing with Spheric Speech Clarity. We want to focus specifically on the most demanding auditory scenario of "Speech in Loud Noise" which implies situations with high background noise level, in which the goal is to recover as much speech understanding as possible. The following paragraphs walk through the signal processing, depicted in Figure 1 from left to right. In the final step, we further adjust the signal with the usual processing steps like gain application.

AutoSense OS

Signal processing starts with AutoSense OS by analyzing the sound scene to determine the type of processing required to maximize user benefit within the specific context.

Following the scene classification, the microphone is appropriately adjusted based on the user's fitting parameters to restore a degree of spatial awareness. The default setting is "fixed directional" as measures have shown it to yield the best processing results for the majority of users.

Spheric Speech Clarity

In the next step, Spheric Speech Clarity identifies and removes unwanted noise from the audio signal, retaining only the speech. This is the key step to drastically improve speech understanding. The major work horse of the speech/noise separation provided by Spheric Speech Clarity is a deep neural network with 4.5 million parameters which was specifically trained for this purpose. The Spheric Speech Clarity DNN was trained on 22 million sound samples to make it fit for all possible situations in which speech clarity is of interest. Spheric Speech Clarity receives as input a full spectrum of 64 frequencies, each of which has a real and imaginary component (you may also imagine it as frequency and phase). This spectrum contains the full audio signal information so the further processing can work with maximum precision. From this, the DNN computes a 64-frequency mask that separates speech from noise which is then applied to the audio signal.

DEEPSONIC™

As discussed in Hasemann & Krylova 2024, providing capable hardware that deep learning models require is not an easy task. The kind of computation required depends strongly on the neural network type such as CNNs (Venkatesan & Baoxin, 2017) or RNNs (Dupond, 2019) and can generally not be handled efficiently enough by standard digital signal processing (DSP) chips.

In practice, these computations are usually pretty resource-intense, as the "deep" part of the name refers to a long computational sequence. Thus, they need capable hardware to be computed efficiently. Both space and power are notoriously scarce in hearing aid devices, so designing hardware that can tackle this task poses a considerable challenge.

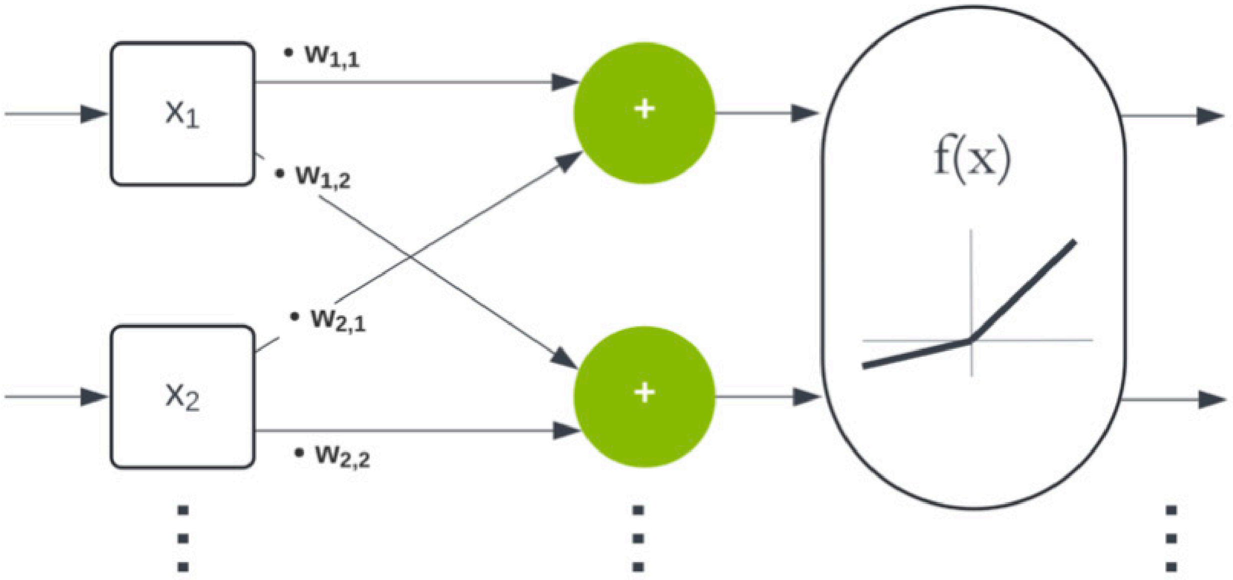

Figure 2. A simple layer of a neural network: The input values x1, x2, ... get multiplied with the parameters w1,1; w1,2; ... and summed up to produce intermediate outputs. In the next step, the intermediate outputs get passed through a nonlinear activation function f(x). A deep neural network chains multiple of these layers together.

To facilitate the 7,700 million operations per second necessary to compute the DNN in Spheric Speech Clarity, Phonak developed DEEPSONIC™, the most advanced deep neural network processing chip in a hearing aid yet. At time of launch, DEEPSONIC™ has a computational power that is a factor 53 higher than any chip used in the hearing aid industry. It can accelerate a large variety of different neural network types and thus provides tremendous flexibility for future applications. In the case of Spheric Speech Clarity, DEEPSONIC™ executes computations with a processing speed of 50MHz on 420 million transistors.

Deep Neural Networks

Neural networks are commonly introduced with analogy to neurons of biological brains. While this analogy is not wrong (after all it gives neural networks their name), it can sometimes feel a bit nebulous or downright mysterious. Also, the calculations in artificial neural networks, while related, are not the same as the ones in biological brains.

Beneath the surface, a neural network model consists of layers of simple computations like the one shown in Figure 2. That is, there are some inputs (here x1, x2, ...) and some parameters w1,1, w1,2, ...

A single layer is built from very simple operations (largely addition and multiplication), but the full neural network can become very complex due to chaining multiple layers together. When a neural network contains many layers, we call it a deep neural network.

The parameters (w) are numerical values we determine in training (see below). The input values (x) on the other hand are what goes into the network at run time. In the case of Spheric Speech Clarity, this is the full sound spectrogram containing 64 frequencies and phases, so we'll have x1 up to x128.

Training and evaluation workflow

Training a neural network model is an optimization process. Given a number of examples (input and expected output), we optimize the parameters such that the model "learns" to produce the expected outputs for all given inputs. For a denoising model the example inputs refer to snippets of noisy audio and the expected outputs to clean speech. If this is done correctly, the network will also work well on inputs that it has never seen before. In that case we say the model generalizes well.

This generalization is what makes this approach so useful: There are virtually infinite possible combinations of noises and human voices saying different things out there. Even with the largest possible dataset we could never hope to encounter all of them during training, so we need a model that can work well with situations that it has not seen before.

But how do we know how well a model generalizes? How do we determine that the model has been trained enough and that the output is "better" than the noisy speech in the input? How can we compare different models with each other?

One way to answer all these questions is of course given by the various kinds of user studies in which humans rate the audio in certain ways. For comparing many models in different training stages this approach does not scale well, so we complement it with an automated evaluation procedure.

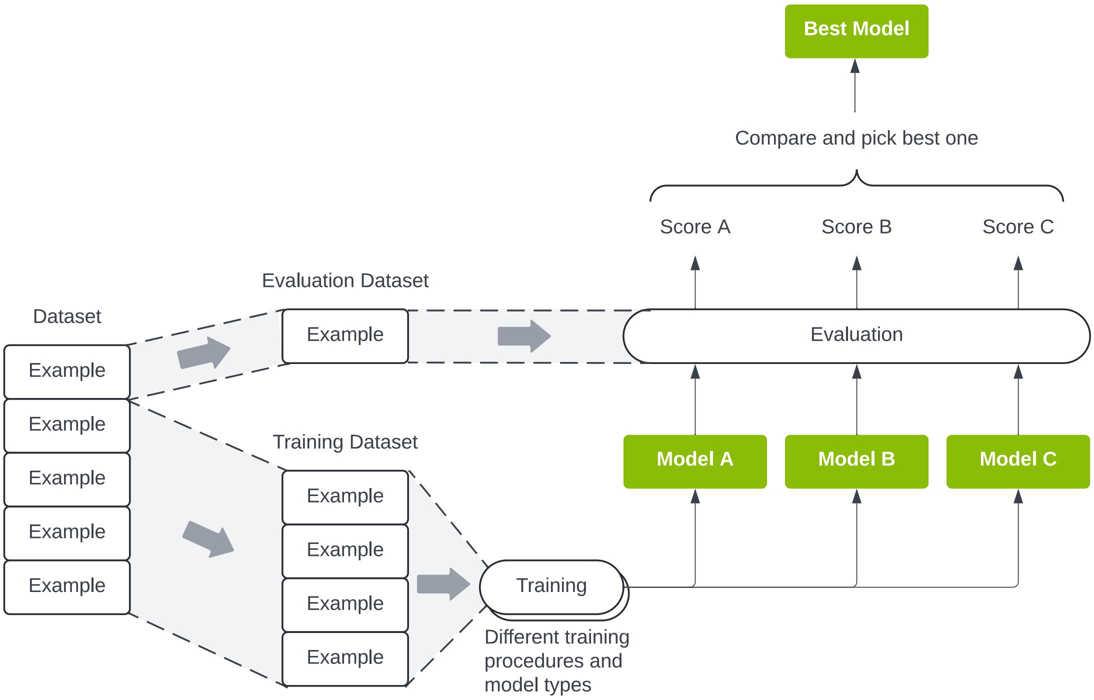

Figure 3. The initial dataset with examples is split into a small evaluation dataset and a larger training dataset. The training dataset is used to produce one or more trained models, whereas the evaluation dataset is used to test the models on data they have not encountered in the training process. This way, we can measure how well the models generalize to new situations.

Figure 3 illustrates the general (simplified) workflow of the evaluation procedure: After creating a dataset of examples of noisy and clean speech, a small portion is split off, which we call the evaluation dataset. The rest is the training dataset. Then, several models are trained using the training dataset. To understand how well one of these models performs (for example, to decide whether we can stop training it), is evaluated. More precisely, a selection of noisy speech examples are selected from the evaluation dataset and examined to see what output the model produces on them. Afterwards, a metric is computed (see below) on the output, resulting in a score that indicates how "good" each model sounds.

Based on this score the decision on how to proceed and whether a certain approach is worth pursuing is made. Without this, we would have very little chance of finding a good DNN for any use-case in which generalization is important.

Training

We have discussed above that a DNN is intrinsically a mathematical computation with some parameters that we optimize in a training process. In this section, we will look closer at how training works for DNNs in general and Spheric Speech Clarity in particular.

Slightly simplified, a training procedure for a single DNN works like this:

1. Start with random model parameters
2. Take an example of inputs and desired outputs
3. Compute for this example the discrepancy of the networks output to the desired output
4. Slightly adjust all parameters in a direction that improves this discrepancy
5. Continue with step two using a different example

This process is repeated a lot of times and is thus very computation intense. DNN training needs expensive hardware in compute centers and can, depending on the model, take months of just computation to complete.
In the case of Spheric Speech Clarity, we need to do some extra work to account for the effects of executing the DNN on a tiny chip in a hearing aid: The hardware that we use to train the DNN is very different to the chip in the hearing aid that will eventually execute it. To address this, we need to take measures to ensure the model is trained with these differences in mind. Finally, an important concern is energy consumption. During training, we ensure that the model we develop is as energy efficient as possible, so that the customer can have the maximum possible battery lifetime despite the complex computations that are happening.

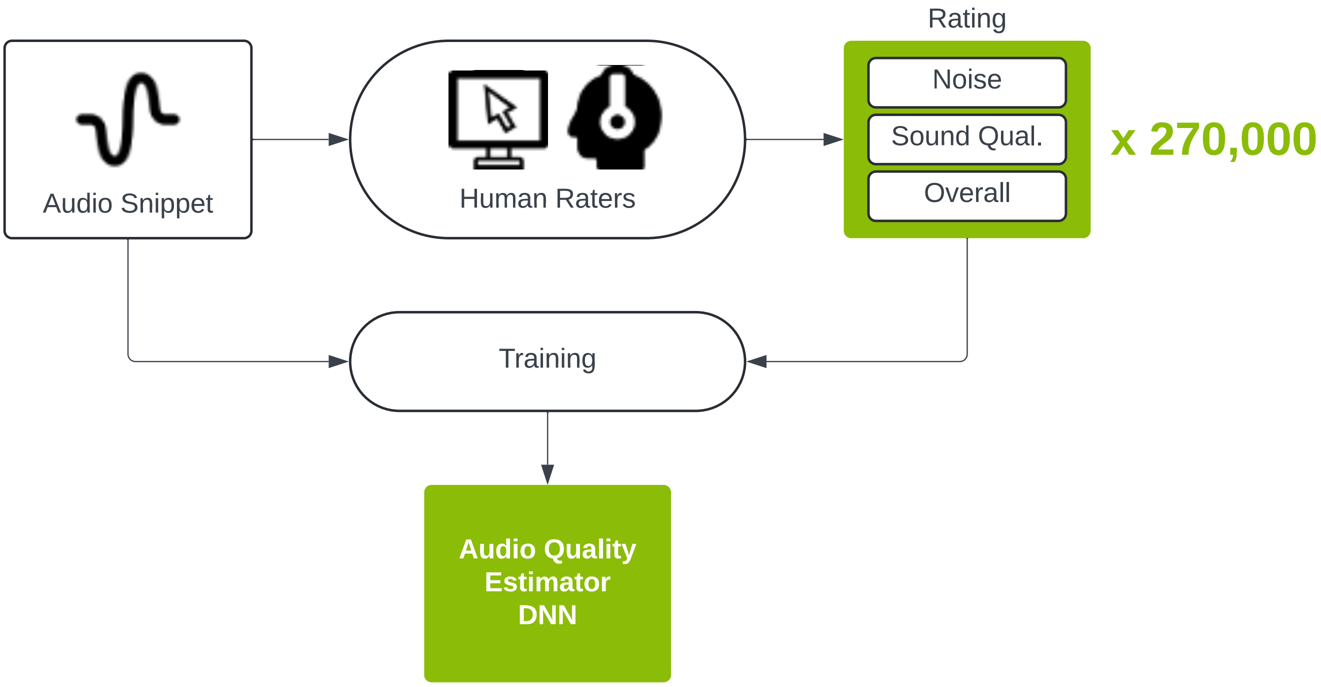

Figure 4. Training the Audio Quality Estimator DNN. First, we collected human ratings on a variety of audio scenes, then we trained the DNN to predict these ratings in the categories of Noise, Sound Quality and Overall score for arbitrary audio snippets.

Audio Quality Estimator

For evaluating a given snippet of audio data, for example, as part of the training for Spheric Speech Clarity, we are interested in measuring the following metrics:

- How much noise does it contain?
- How good is the sound quality? Are there any distortions?
- Overall, how much do people like how it sounds?

In a typical study on humans, we would collect these as Mean Opinion Scores (MOS) (ITU-T 2017). To obtain the best possible model for speech clarity it is crucial to accurately measure these scores for a large quantity of different models on many different sound scenes. Only this way can we ensure that the product reliably delivers clear speech in any situation. Measuring sound perception along these different axes allows us to choose the best trade-off between them.

Due to the large number of measurements, we need for monitoring and controlling our workflow of training DNNs for Spheric Speech Clarity, it would not be feasible to obtain all of them directly from human testers, a way to automate is required. This idea is not new: a variety of metrics exist today that try to measure the sound properties mentioned above or similar ones. For example, extended short-time objective intelligibility (ESTOI) algorithm, proposed by Jensen & Taal 2016, incorporates the ranking of spectrogram features based on their importance to speech intelligibility, which allows it to predict speech intelligibility without the need for expensive and time-consuming human listening tests.

Alas, the existing metrics do not cover all three aspects above that we are interested in, and we have found they do not provide helpful guidance for selecting good DNNs (but rather the selection process would uncover weaknesses in the metrics). In addition, these metrics are generally intrusive, that is, they require a clean speech reference to be available which makes them impossible to use for noisy real-world recordings.

To capture the human perception of audio in terms of noise, sound quality and overall preference, we trained a deep neural network model, the Audio Quality Estimator DNN. While the Spheric Speech Clarity DNN was trained to convert noisy audio into clean audio, this metric model was trained to mean opinion scores that human raters would give for a given audio snippet.

We had about 350 human raters rate the noise, sound quality and overall preference of 30,000 files resulting in a total of almost 1 million human ratings (each file was rated by 9 different raters, see Figure 4). The result is a DNN model that can predict human ratings of any piece of audio data without the need for a clean speech reference in the categories "noise", "sound quality" and "overall score" more accurately than any other available estimator. Figure 5 shows a comparison to various other metrics.

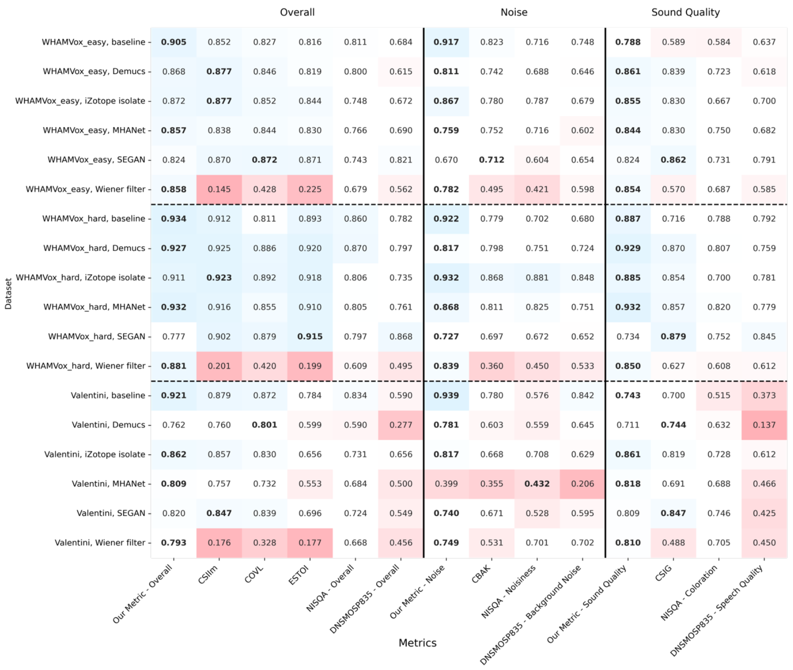

Figure 5. Pearson correlation of various metrics with human ratings on various datasets. Red shades indicate low correlations and blue shades high correlations. The numbers in bold are the highest correlations in each category. For more details refer to Diehl et al. 2022.

Verification

We believe that the Audio Quality Estimator DNN is one of Phonak's most outstanding and unique contributions to the industry and, together with the DEEPSONIC™ chip, one of the most important components that allowed us to bring Spheric Speech Clarity to the market.

In addition to using the Audio Quality Estimator during training, in the last phase of development, the Spheric Speech Clarity DNN underwent a long and rigorous series of tests to ensure we deliver a high-quality product to the customer. Table 1 provides an overview of the scope of the conducted biases testing.

Test	Description
Gender bias	Ensure male and female voices are treated equally.
Age bias	Ensure voices from of all age groups are treated equally.
Language bias	Ensure the various languages that are spoken by wearers are processed correctly and equally.
Speech impairment and emotional speech bias	Ensure voices with speech impairment or strong emotions are processed correctly and equally well to others.
Various listening test	Various listening tests and studies with people with and without hearing loss to assert the quality of the product

Table 1. Some of the additional tests conducted in the final phase of DNN training for Spheric Speech Clarity. Not included are the usual hardware-and medical testing procedures that happen in other stages of the quality assurance.

Conclusion

While numerous AI-based advancements have been introduced into the hearing aid market, to this date the technology has not been fully leveraged to address one of the primary challenges for users: comprehending speech in noisy environments.

Developed by world-class AI engineering experts, Phonak Audéo Sphere Infinio brings Spheric Speech Clarity, Phonak's pioneering deep learning solution to the market. Spheric Speech Clarity will allow millions of hearing aid users to enjoy social interactions in noisy environments like bars, restaurants, social gatherings, or public transport.

This software solution is enabled by two key advancements: DEEPSONIC™, the first DNN processing chip of its processing power to fit in a hearing aid and the Audio Quality Estimator DNN, which allows prediction of the perceived sound quality in audio snippets at large scale.

This Insight offered a glimpse behind the scenes of the Spheric Speech Clarity development. The hard work that the people at Phonak have put into Spheric Speech Clarity will change the lives of countless individuals for the better and set a new benchmark within the industry as the standard of excellence.

References

Åleskog, C., Grahn, H., & Borg, A. (2022). Recent Developments in Low-Power AI Accelerators: A Survey. Algorithms 2022, 15, 419. https://doi.org/10.3390/a15110419.

Appleton-Huber, J. (2022). What Is Important to Your Hearing Aid Clients… and Are They Satisfied? Retrieved March 18th, 2024, from https://hearingreview.com/hearing-loss/patient-care/counseling-education/what-important-to-your-hearing-aid-clients-are-they-satisfied.

Appleton-Huber, J. (2015). AutoSense OS – Benefit of the next generation of technology automation. Phonak Field Study News retrieved from https://www.phonak.com/evidence

Chatterjee, S., & Zielinski, P. (2022). On the Generalization Mystery in Deep Learning. arXiv preprint arXiv:2203.10036.

Diehl, P. U., Singer, Y., Zilly, H., Schönfeld, U., Meyer-Rachner, P., Berry, M., Sprekeler, H., Sprengel, E., Pudszuhn, A. & Hofmann, V. M. (2023). Restoring speech intelligibility for hearing aid users with deep learning. Sci Rep. 13(1), 2719. doi: 10.1038/s41598-023-29871-8.

Diehl, P. U., Thorbergsson, L., Singer, Y., Skripniuk, V., Pudszuhn, A., Hofmann, V. M., Sprengel, E. & Meyer-Rachner, P. (2022). Non-intrusive deep learning-based computational speech metrics with highaccuracy across a wide range of acoustic scenes. PLoS ONE 17(11): e0278170. https://doi.org/10.1371/journal.pone.0278170.

Dupond, S. (2019). A thorough review on the current advance of neural network structures. Annual Reviews in Control. 14, 200–230.

Hasemann, H., & Krylova, A. (2024). Artificial intelligence in hearing aid technology. Retrieved May 8th, 2024, from https://www.phonak.com/content/dam/phonak/en/evidence-library/white-paper/technical-paper/PH_Insight_ArtificialIntelligenceInHearingAidTechnology.pdf. ITU-T Rec. P.10/G.100 (2017) Vocabulary for performance, quality of service and quality of experience.

Jensen, J., Taal, C. H. (2016). An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers, in IEEE/ACM Transactions on Audio, Speech, and Language Processing. 24(11), 2009-2022. doi: 10.1109/TASLP.2016.2585878.

Raufer, S., Kohlhauer, P., Jehle, F., Kühnel, V., Preuss, M., & Hobi, S. (2024). Spheric Speech Clarity proven to outperform three key competitors for clear speech in noise. Phonak Field Study News retrieved from https://www.phonak.com/evidence

Reuther, A., Michaleas, P., Jones, M., Gadepally, V., Samsi, S., & Kepner, J. (2019). Survey and benchmarking of machine learning accelerators. In 2019 IEEE high performance extreme computing conference (HPEC),1-9. IEEE.

Venkatesan, R., & Li, B. (2017). Convolutional Neural Networks in Visual Computing: A Concise Guide. CRC Press. ISBN 978-1-351-65032-8.

Wright, A., et al "Spheric Speech Clarity applies DNN signal processing to significantly improve speech understanding from any direction and reduce the listening effort ." Phonak Field Study News in preparation expected August 2024.

Wright, A., Kuehnel, V., Keller, M., Seitz-Paquette, K., Latzel, M. (2024) "Spheric Speech Clarity applies DNN signal processing to significantly improve speech understanding from any direction and reduce the listening effort ." Phonak Field Study News retrieved from https://www.phonak.com/evidence