An introduction to
MoreSound Intelligence™

MoreSound Intelligence (MSI) is the new "brain-inspired sound scene clarifier" feature in Oticon More™ hearing aids. MSI consists of several sub-features, and this tech paper aims to provide you with a deeper insight into all of these as they specifically pertain to Oticon More 1.

Firstly, you will be presented with an overview of the complete MSI feature, and then you will dig into the individual sub-features one by one in the order they appear in the processing flow.

Some important highlights:

• Virtual Outer Ear – the new true-to-life pinna model effective in easy environments, with three settings in Oticon Genie 2 for user preference

• Neural Clarity Processing – the Deep Neural Network is trained in real-life sound scenes in the development phase to optimally support the brain, and is embedded in the new Polaris™ platform

• Sound Enhancer – dynamic gain primarily for speech, given in complex environments, with three settings in Oticon Genie 2 for user preferences

oticon
life-changing technology

Sound scenes are dynamic, complex, and unpredictable, and it is the brain's role to handle this compexity; to hear, and to create meaning from it all. The processing in the hearing aid needs to provide a good signal to help the brain interpret it. However this does not happen by limiting the sound scene by applying noise reduction and directionality. The brain needs access to more information from its immediate surroundings to aid the brain's natural way of working. More of the full perspective of sounds, in order to get more out of life. (O'Sullivan et al., 2019; Hausfeld et al., 2018; Puvvda & Simon, 2017)

Oticon's MoreSound Intelligence introduces a quantum leap in sound scene processing giving access to the full sound scene with clear contrast and balance.

MoreSound Intelligence

MoreSound Intelligence (MSI) is an advanced feature that include several sub-features. In the following paragraphs, this tech paper will describe all the different sub-features.

Figure 1 shows the different steps and sub-features in MSI. Firstly, the sound scene is scanned and analysed. Based on that analysis, coupled with the settings in the fitting software (Oticon Genie 2), the sound scene is then passed on to Spatial Clarity Processing and Neural Clarity Processing. It follows one of the two paths: the path for easy environments or the path for difficult environments. What we get at the output end of the MSI block is a cleaned-up signal that is ready for further adjustment by the hearing aid (e.g. amplification).

For more information please see Brændgaard, M. 2020. The Polaris Platform. Oticon tech paper - on the complete processing flow for Oticon More, and Løve, S. 2020. Optimal Fitting of Oticon More. Oticon Whitepaper - on fitting Oticon More in Genie 2.

Throughout MSI, the input sound processing is done in 24 channels. Compared to previous premium Oticon hearing instruments, the extra number of channels provide double the precision in a frequency range that includes the 1.5-5 kHz frequency channels (Brændgaard, 2020), which are the most important for speech sounds.

In addition to having the extra channels for more precision, the 24 channels are also linked. This means that in the Deep Neural Network, all channels can see the processing taking place in the other channels. This minimizes the risk of artefacts being created by a type of sound being categorized wrongly in just one channel. By minimizing artefacts the sound quality is improved.

Scan and Analyse

In order to make the correct processing of the different sound sources, MSI needs to know the exact details of the sound scene. Moreover, sound scenes are dynamic and sound sources are constantly moving and changing. To ensure that all the details are captured, the sound scene is scanned 500 times per second, to correctly map the different sound sources. Based on this scanning and mapping, MSI calculates the optimal signal to noise ratio (SNR) as well as noise levels.
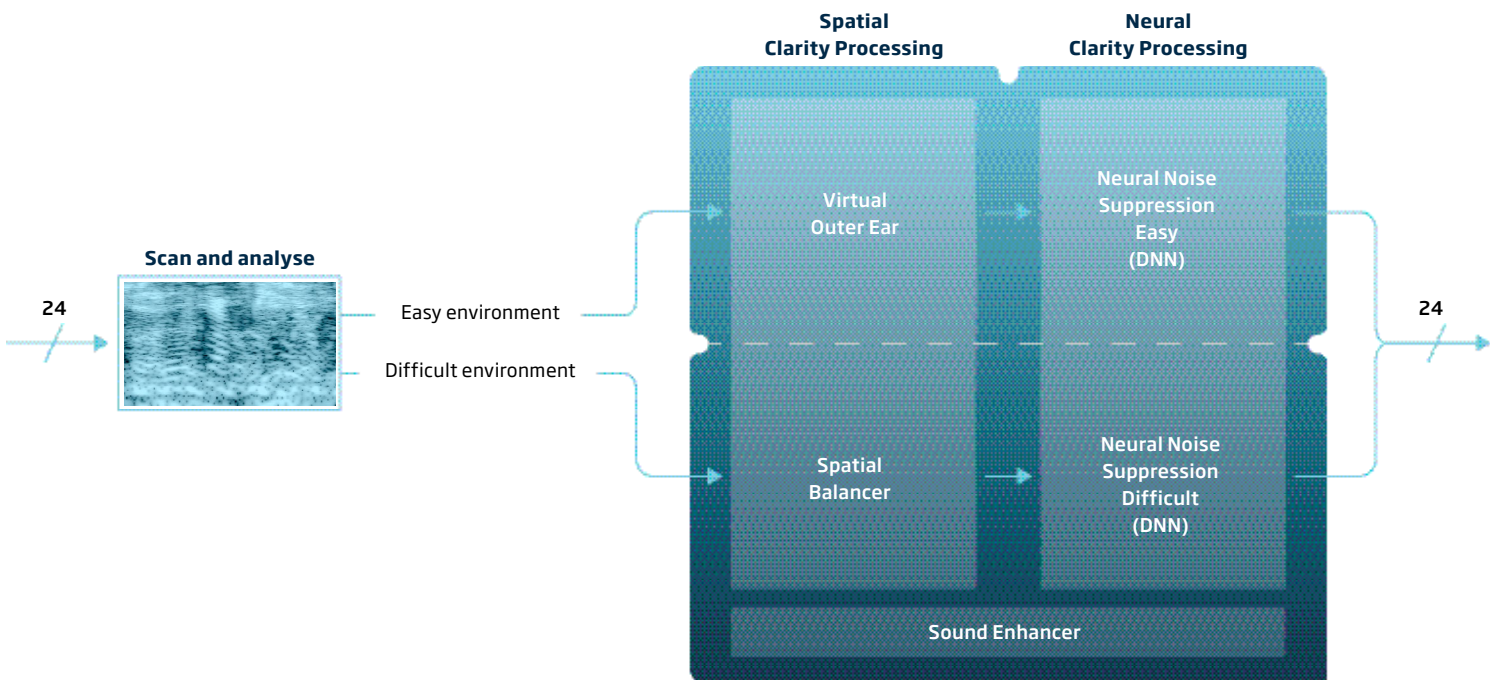

Figure 1: The Processing flow of MoreSound Intelligence

The SNR and noise level estimators have been updated such that they now run in 24 channels, and the SNR estimator now works over the broader range of -10 to +15 dB SNR. The SNR is the main driver used for distinguishing between easy and difficult environments, and the level of help provided by the system will be determined by both the SNR and the noise level estimates. The principle of the relationship between SNR, noise level, and help provided can be seen in figure 2.

Any changes in the sound scene will be detected when scanning, but only persistent changes (more than 2 secs) will make the hearing aid adapt the level of help.

MoreSound Intelligence uses the SNR to distinguish between easy and difficult environments, but the specific details differ from person to person. Where it draws the line between easy and difficult environments depends on the individual user's settings in Oticon Genie 2. That is, the settings made in Genie 2, which are based on user input during the fitting session, will determine which SNR must be present for a sound scene to be processed as an easy or difficult environment.

On the one hand, when the sound scene is determined to be an easy environment, the processing follows the flow (see figure 1) starting with the Virtual Outer Ear and then Neural Noise Suppression – Easy.
On the other hand, when the sound scene is determined to be a difficult environment, the processing follows the flow starting with the Spatial Balancer and then Neural Noise Suppression – Difficult, with the underlying help of the Sound Enhancer.

The full analysis of the sound scene is also used by Neural Clarity Processing for processing in the Deep Neural Network (see later in this paper).

Spatial Clarity Processing

Being able to place sound sources in the spatial environment is an important ability which becomes more difficult when hearing loss is present (Akeroyd, 2014).

The pinna helps us locate sounds in three dimensions – in distance, vertical (up/down), and horizontal (front/back). Depending on whether it is the interaural time difference, interaural level difference, or the head-related transfer function that is at play, some frequencies are more relevant for localisation than others. (Akeroyd, 2014).

We all have different ear sizes and pinna shapes and sound will therefore be modified in different ways when it enters the ear canal, depending on the anatomy of the ear. For example, due to the shape of the outer ear, some people will have more or less frontal focus than others.

When we place the hearing aid microphones behind the ear, the ability to utilise the natural spatial cues provided by the pinna is eliminated. This ability needs to be recreated by signal processing in the hearing aid.

Spatial Clarity Processing consists of two different features, Virtual Outer Ear and Spatial Balancer, which help recreate this spatial sensation in easy and difficult environments, respectively.
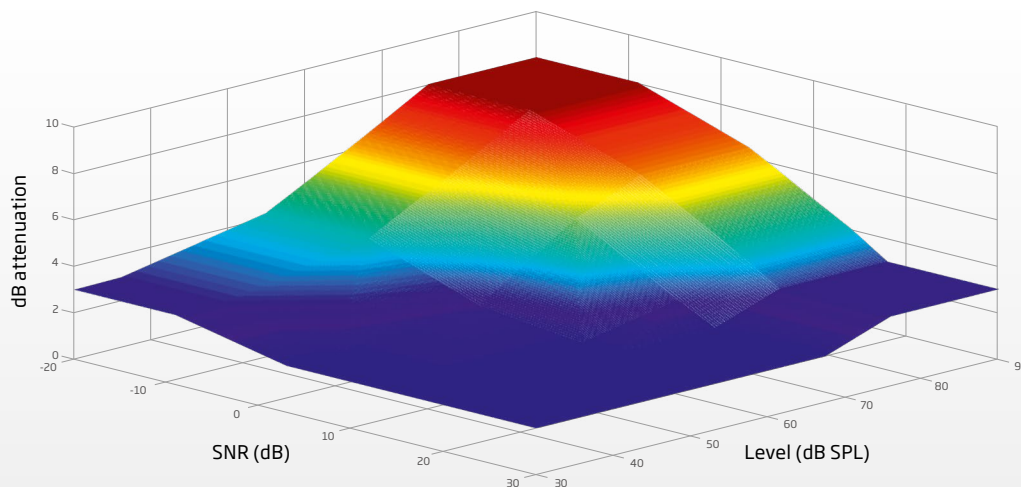

Figure 2. Principle of the relationship between SNR, noise level, and noise suppression. The further to the left on the SNR axis, the lower the SNR. The further to the right on the Level axis, the higher the level. The lower the SNR and the higher the level, the more help is provided by the system.

Virtual Outer Ear

The Virtual Outer Ear (VOE) are three different true-to-life pinna models that can be set by the hearing care professional in the fitting software based on the hearing aid user's spatial sound needs.

VOE helps the user recreate spatial awareness in easy environments.

To make the best signal processing compensation for the natural pinna, the head-related transfer functions (HRTF) were measured from different angles in the horizontal plane on 130 human ears. The Directivity Index (DI), which is a measure of the amount of directionality of the individual ears in each frequency band, can be calculated from the HRTFs. A higher DI equals a more directional ear.

The DI results can be seen in figure 3. The dark magenta line is the average, while the blue line is the ear with the smallest calculated DI (and by consequence the most omni directional). The grey line is the ear with the largest calculated DI (ear with the most front focus). The difference in DI between ears can be seen mainly between 2 and 5 kHz, and can be quite a large difference.

The information from the HRTFs can also be shown as a polar plot, which shows the response for a specific angle/frequency attained in VOE (figure 4). Again, the dark magenta line is the average, the blue is the smallest DI, and the grey is the largest DI.

The difference between ears ranges around 4 dB DI, from -2 dB to +2 dB, where most measurements are in the 0.5 dB DI area. This means that most people get a natural amplification from the outer ear of around 0.5-1 dB in the 2-5 kHz area.

Consequently, VOE is based on these measurements to create a pinna model as natural and accurate as possible.

The measurements have shown that the effect the pinna has on sound varies between ears. This means that the way a person is used to hearing sound will differ, depending on their outer ear anatomy. In addition, it is not possible to take a simple measurement of the outer ear and then infer how the user 'hears'. However, since we do know it will differ across individuals, the VOE has three different settings with slightly more or less frontal focus (see figure 5) and this can be set in the Oticon Genie 2 fitting software according to user preference. The slightly more frontal focus is created by letting less sound in from behind.

The three different settings are based on the following DI measurements:

- The average measurements are used for the Balanced setting (default) which is designed to create the best ratio between audibility of all sounds in the user's surroundings while being able to focus on speech from the front
- The highest measurements are used for the Focused setting (more focus on speech from the front)
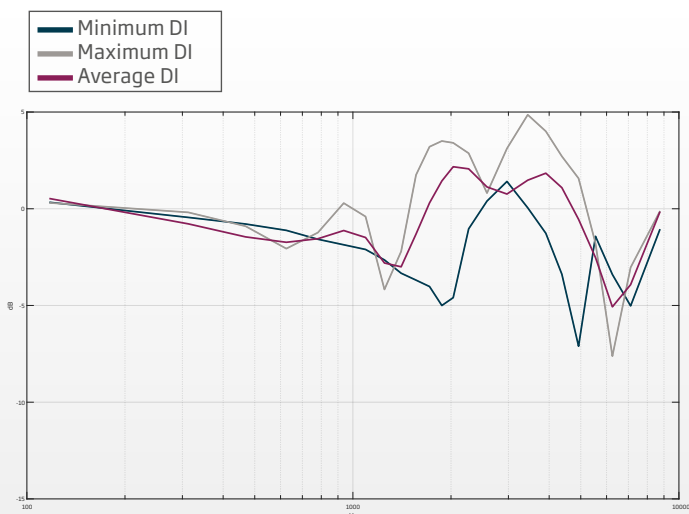- The lowest measurements are used for the Aware setting (access to all surroundings)

Figure 3. Directivity Index for human ears showing average (dark magenta), minimum (blue) and maximum (grey) DI.
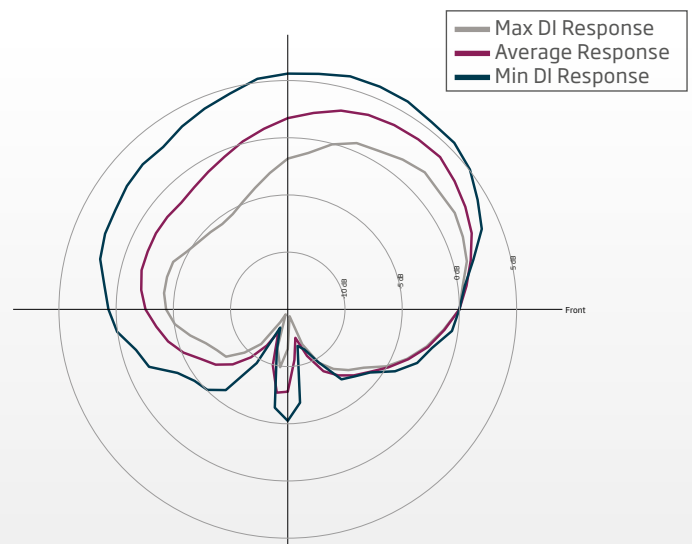

Figure 4. Polar plot for left ear showing the directionality of human ears averaged for 2-5 kHz. In this example, the person is facing the right-hand side.

The differences between the three settings are approximately 0.5 dB DI measured AI weighted*.

Spatial Balancer

Spatial Balancer is a more powerful feature than VOE when it comes to difficult environments. Spatial Balancer quickly balances distinct sound sources in the environment.

Spatial Balancer is provided with an omnidirectional signal and a back-cardioid signal from the two microphones. The omnidirectional signal provides all the sounds in the sound scene including the sounds from in front, which are often the most important signals to the user. The back-cardioid signal provides every sound from the sound scene except the sounds from in front. The two signals are constantly being compared to define the placement of the noise sources. Spatial Balancer uses a minimum-variance distortionless response (MVDR) beamformer to create polar plots making the most optimal balance for the given sound scene. The null direction in the polar plots is placed towards the most dominant noise sources to attenuate the noise and keep it in the background of the sound scene. The system can create individual null directions for each of the 24 frequency channels in the hearing aid which allows Spatial Balancer, in principle, to control 24 sound sources (48 in total around the head). Each null direction is updated 125 times per second.

Spatial Balancer increases the SNR by suppressing individual noise sources, placing them in the background and thereby creating a balanced sound scene.

Adding to the number of channels and thereby the number of possible nulls compared to prior Oticon products makes the system more precise in targeting the sound source. The null can be deeper which means the noise source can be pushed more into the background if needed.

AI weighted: Articulation Index weighted. This means that in the DI calculation, the frequencies are weighed based on the importance of speech understanding. The lowest and highest frequencies have less weight than mid frequencies.

Neural Clarity Processing – Deep Neural Network

Neural Clarity Processing is where the unique and dedicated development for Oticon More™ hearing aids resides – the Deep Neural Network (DNN). The DNN has, for a given sound scene, learned to recognize what should be put in the foreground (sounds of interest with a lot of information) and what should be put in the background (sounds of less interest with less information). In this way it creates better clarity and contrast between sound sources. This section of the paper will introduce you to neural networks and to the Oticon implementation of a Deep Neural Network.

You might not think about it, but neural networks are being used all over the world to tackle large complex questions. Based on your previous actions it could be suggesting a playlist or recommending things to buy. Such complex, multifactoral challenges can be addressed more easily by a neural network than by traditional approaches which rely of establishing set rules.

Before we dig into how the DNN is working in the hearing aid let's take a brief look at what it is in the brain the DNN mimics when learning. We will use a visual example of distinguishing between a cat and a dog as this is easier to relate to.

In our brains, we have areas within the cortex specialized in processing visual signals such as the primary visual cortex which contains over a hundred million neurons and many more connections between them. (Leuba G. & Kraftsik R, 1994.) We as humans, over our lifetimes, have become astoundingly efficient at making sense of what we see but the brain needs to learn before we can do this. The task of distinguishing between a cat and a dog is normally quite easy for an adult, but it is not necessarily easy to tell how we do it.
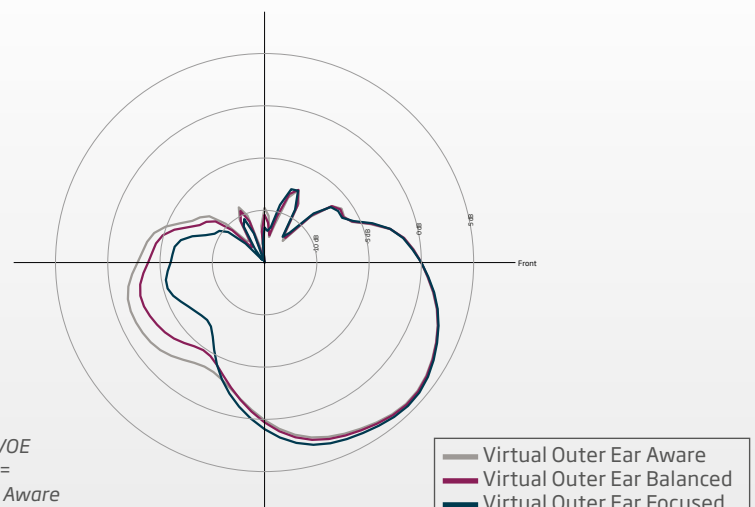

Figure 5. Three different settings of VOE shown at 3437.5 Hz. Dark mangenta = Balanced, grey = focused, and blue = Aware

Virtual Outer Ear Aware
Virtual Outer Ear Balanced
Virtual Outer Ear Focused

What are the rules we have set up in our brains that make us capable of handling this task? Perhaps it is something about the nose or the ears, or perhaps it is something completely different that we are not even conscious of. We quickly realize we are bound to get lost in a morass of exceptions and caveats trying to verbalise these rules. The same goes for when trying to define rules of sound recognition, but that is nonetheless what scientists have done for a long time when addressing noise reduction in hearing aids.

Handling of noise sources

Until now noise reduction in hearing aids has been handled by man-made algorithms that defined what was relevant speech and what was noise respectively by applying rules about, for instance, modulation of the sound and the direction from which the sound source originated. That could often mean that only speech from the front was considered relevant. Such an assumption may not be true in all cases, as it ignores the importance of the brain's access to the full sound scene and limits our ability to localize sounds outside of its effective range. (For more information on the importance of the brain's access to the full sound scene please see Man, B. and Ng, E. 2020. BrainHearing – The new perspective. Oticon Whitepaper.)

Distinguishing between objects is a multifactorial problem. Our brain leverage not just the features and patterns we can identify from the signal, but also long-term semantic memory (Rönnberg et al., 2013). After a lifetime of using our ears, our brain has stored a representation of many, many different sounds.

Our brain has effectively learnt, through mistakes and experience, a unique methodological approach to distinguish between relevant and non-relevant sounds. Neural networks have taken this approach. The structure of a DNN is partially inspired by how our brain is organized, namely the neurons and their corresponding synapses. The neural network uses the iterative learning from the huge quantity of real-world data (see further down for more info in the training process) to establish knowledge about sound and how to process it. The iterative learning of the DNN is applied instead of following a strict set of pre-established rules based on man-made algorithms and generalizing to the many different and complex sound scenes in the real world. This approach takes sound processing and noise handling out of the lab and into the real world.

What is a Neural Network?

Neural networks are a specific class of algorithms under the more general discipline of machine learning. The idea of machine learning is to take a large amount of data, known as training samples, and then develop a system which can learn from them. The uniqueness of neural networks stem from their architectural similarity to the brain. In the context of neural networks, there is a basic unit called the neuron. A neuron's purpose, much like a relay neuron in the brain, is to receive information, store it, and finally pass it on to the next neuron. A group of neurons form a layer, and multiple specialized, interlinked layers form the neural network consisting of an input layer at the start, hidden layers in the middle and output layer at the end. This forms the most basic class of neural networks (figure 6).
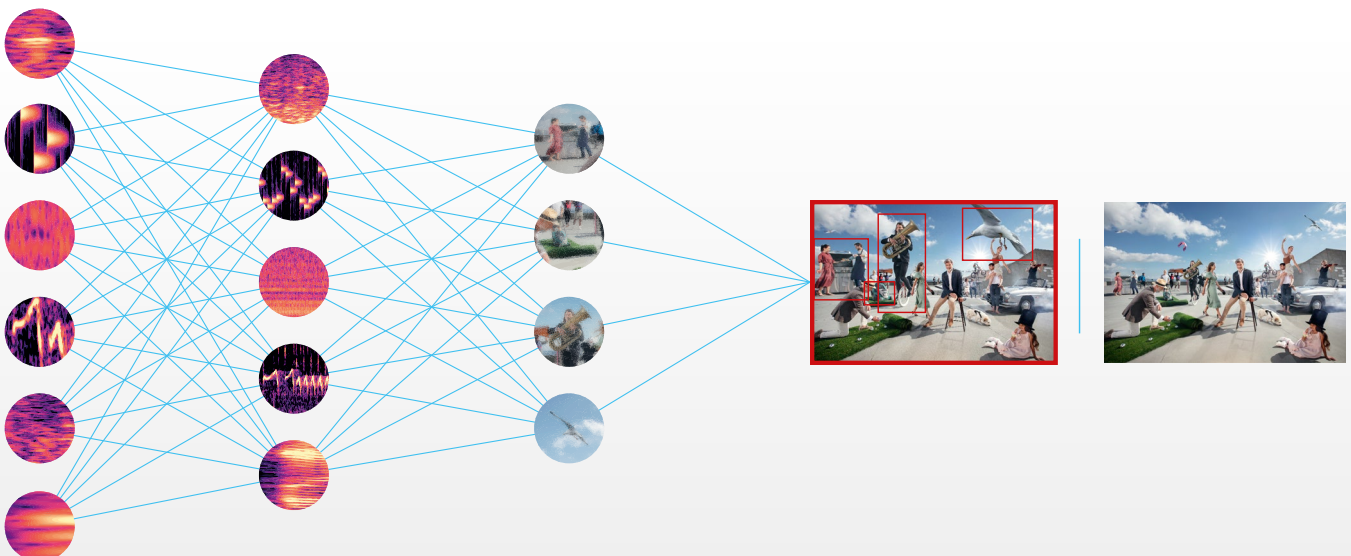

Figure 6. Conceptual illustration of a neural network.

Oticon's very own Deep Neural Network
Generally, the development of a Deep Neural Network (DNN) requires 3 fundamental stages: (1) Scope (2) Train and Learn, and (3) Test (see figure 7).

1. Scope
First and foremost, we scope the problem and define what we want to deliver to hearing aid users. In our case, we want to improve the hearing impaired listeners' access to the full sound scene, by creating a good neural code to better support the stages of Orient and Focus in the brain (Man & Ng, 2020).

To fulfil that purpose, we need to consider the nature of the data (the sound scenes). What are the characteristics of different sounds like speech and noise. Speech is dynamic and it changes all the time even when it is coming from one single person. The key feature of a speech signal is that there is a degree of continuity. For example, if you listened to your friend's voice, it is unlikely that his/her voice will suddenly change in tempo, or that the pitch will suddenly change. In contrast, the surrounding noise may be a mixture of clinking glasses, voices from different people in the background which all differ not just in pitch and tempo, but also how they vary over time. This is why we designed a neural network that is specialized in dealing with such dynamic signals – A Gated Recurrent Unit (GRU), which is a variant of Long-Short Term Memory (LSTM) neural networks.

Long-Short Term Memory (LSTM) neural networks, as its name suggests, has something to do with memory, both in the short and long term.

Memory is defined in psychology as the faculty of encoding, storing, and retrieving information (Squire, 2009). As we know, memory serves as a form of information storage system that can be preserved over time. Additionally, it can be accessed to allow facilitation of more efficient decision making. The LSTM and GRUs operate on this principle. The idea is to "link" the neural network over time having the network pass information to itself (figure 8).

The result is an algorithm that not only recognizes different features that sounds have in a single moment, but also how those sound features vary over time. The ability to incorporate information over time is precisely the unique attribute that we need to analyse a dynamic signal such as sounds.

The DNN consists of an input layer, hidden layers in which the processing is not visible, and an output layer with the result we can hear. The input and output layers have 24 neurons corresponding to the 24 processing channels.

2. Train
The goal of this stage is to train the DNN on sound scenes to a degree that it can solve the task it is designed for. A lot of data is required for the training. This data was recorded in different sound scenes across a wide array of listening environments that listeners would be exposed to in their everyday lives. We utilized a specialized spherical microphone, capable of capturing 360 degrees of sounds to provide the DNN with a spatially accurate and detailed sound scene and train it on the full sound scene.
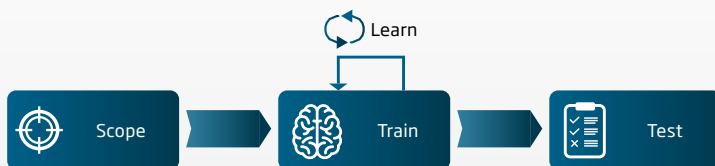

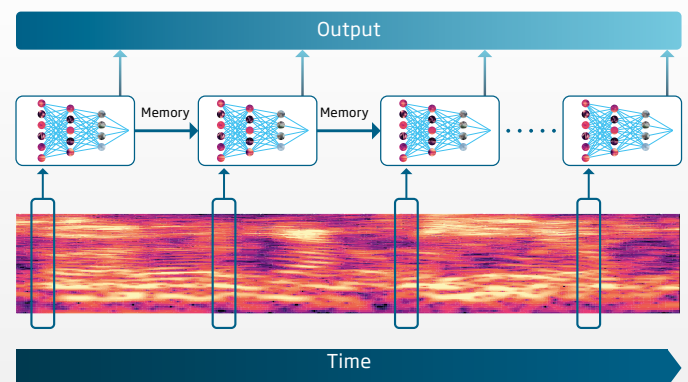Figure 7. Development stages of a Deep Neural Network – 1 Scope, 2 Train and Learn, and 3 Test.

Figure 8. The principle of a Long-Short Term Memory (LSTM) or Gated Recurrent Unit (GRU) Neural Network - linking the Neural Network over time having the network pass information to itself.

Some of the gathered data was allocated for the purpose of training the DNN, and some for testing. Data for testing will not be used until the test stage. The training data was provided to the DNN as input to learn about sound scenes.

The training process can be further subdivided into 4 looping steps: (A) Input (B) Forward Propagation (C) Output (D) Backwards Propagation (figure 9).

During the input step (A), the Deep Neural Network experiences exactly what was mentioned earlier. Neurons receive information of one sound scene and stores them. Next, forward propagation (B) utilizes the data from the input where each neuron passes on information to the next layer. Critically, the amount of information passed depends on the strength of the connections each of the neurons have with each other. This leads to the output prediction (C) of the objects that the DNN thinks it should enhance and suppress in the sound scene. However, just like anyone learning a new skill for the first time, the DNN makes mistakes. It may overemphasize the seagull for example (figure 9, first image under C).

Since this is a form of supervised learning, we instruct the DNN if it has made a mistake and should change. This action drives the process of backwards propagation (D), in which the DNN tweaks the individual connections between each neuron to better suppress the seagull next time. This process is then iterated for all the sound scenes and as a result, the Deep Neural Network starts to identify the features of each object to better distinguish between them. At the same time, scientists will also adjust specific features of the DNN, for example the rate at which the Deep Neural Network learns. This forms a natural symbiosis between the incredible adaptability of the DNN and the deep knowledge of scientists to best achieve our goal of delivering a good neural code. Over time, as the 4 stages iterate through all the 12 million sound scenes that we have fed into the DNN, its ability to emphasize and suppress the respective meaningful and non-meaningful objects improves until the scientists are satisfied. This is demonstrated by the increasingly accurate performance shown by the images at the output (figure 9 (C)).

3. Test

When the DNN has been fully developed, it is time to test how it performs with data it has not been exposed to before during the training stage. This step is critical since some DNNs might not perform well.

Though neural networks perform exceptionally well at completing the task at hand, the DNN may be too strict and specific to the data it has been trained on. For example, as shown above the DNN has learnt how much to suppress the seagull. What if there was a pigeon? They may still not be as meaningful to listeners at any given moment, but the Deep Neural Network may decide to enhance the pigeon, even though it suppresses the seagull. We therefore run into the problem of overtraining, where the DNN is trained to become very accurate, but fails to generalize to the real world. In other words, it only knows how to deal with data it has been trained on but falls short when exposed to new challenges. On the other hand, it may also just not perform well on the data it was trained on, making it too ambiguous (figure 10).
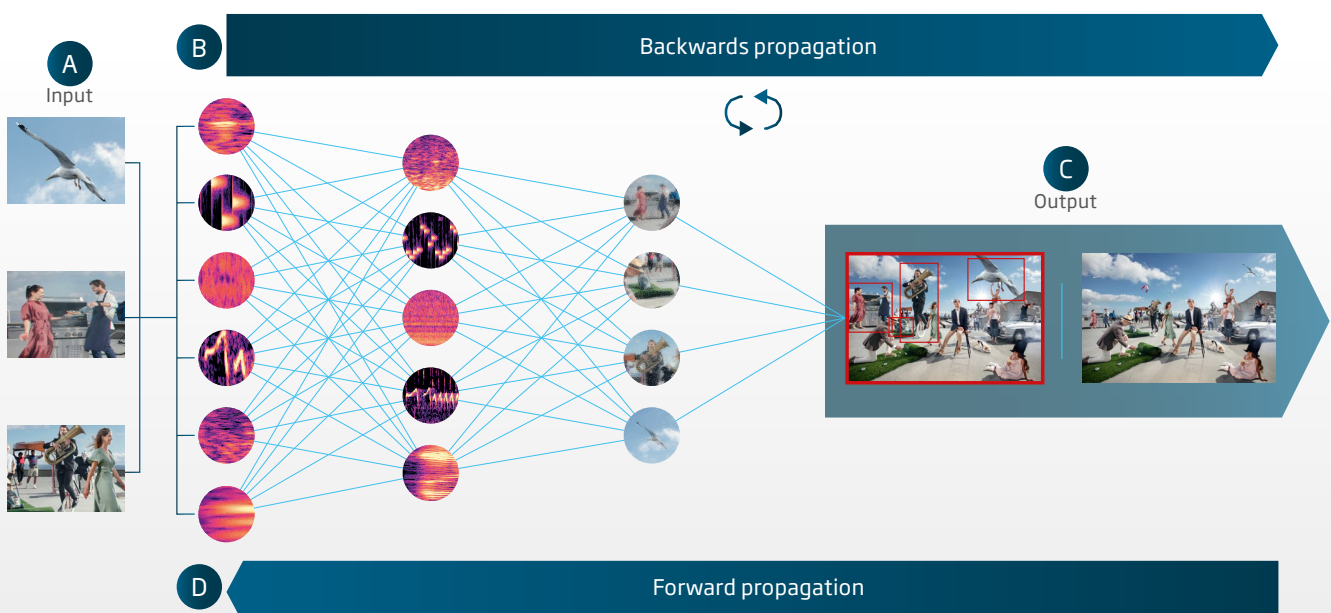

Figure 9. See text for explanation.

In order to avoid these problems, the test stage is crucial to provide us with a degree of confidence in how the DNN would perform in the real world, where sounds are highly variable and possibilities limitless. During the test stage, we scrutinize the DNN with the aforementioned testing data. Ideally, the DNN performs well even if tested with unknown data. If it does not end well, then the Deep Neural Network must be tweaked, or another DNN must be designed. In our case, we at Oticon have designed several versions of the DNN, each with its own unique attributes and selected the current one based on its performance in both the train and test stage.

The last test the DNN goes through is when it is implemented in the hearing aid as part of MSI being tested on the ears of hearing impaired listeners. The results of some of these tests will be described in a whitepaper expected to be available in December 2020.

A Deep Neural Network enables the sounds of the world to be handled precisely and automatically. This optimizes the way Oticon More makes sounds more distinct, and works seamlessly across varying listening environments. With this integrated intelligence, Oticon More has learned to recognize all types of sounds, their details, and how they should ideally sound – all in order to optimally support the brain.

Sound Enhancer

Normally the maximum effect of a noise suppression system has to be a compromise that works for all users, even though some users would have preferred more sound to be removed and some users find that too much sound has been removed. The sound processing in the hearing aid needs to make sure that the user is able to handle the environment and at the same time gets the right feeling of the sound scene to enjoy the atmosphere.

Sound Enhancer provides dynamic sound detail when noise suppression is active – mainly in difficult environments – which allows the output to be individualised.

The Comfort setting can be chosen for the full effect of the noise suppression system, while the Detail setting can be chosen for a strong level of connection to surroundings and attended talkers. This provides an option to customise for user preference.

Sound Enhancer is applied in the processing scheme after Spatial Clarity and Neural Clarity Processing. Sound Enhancer looks at the dynamic suppression done by Spatial Balancer and the DNN and then calculates how much sound to add to the signal based on the settings in Oticon Genie 2. Sound Enhancer follows the overall adaptive suppression done by Spatial Balancer and the DNN. It will not adapt to small changes all the time, but the bigger general changes in the sound environment.
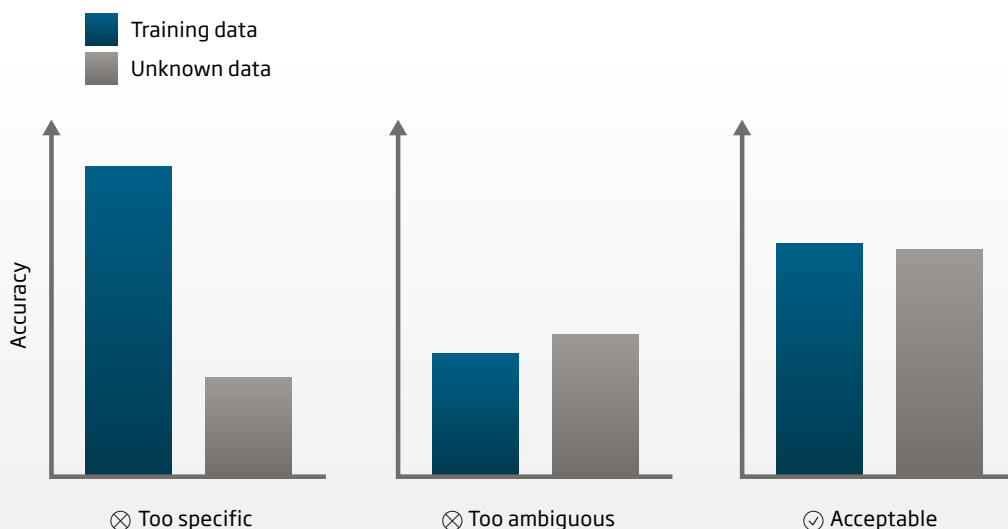

Figure 10. Accuracy of the DNN on training data and unknown data. High accuracy on training data but low on unknown data – DNN is too specific. Low accuracy on both training data and unknown data – DNN is too ambiguous. High accuracy on both training and unknown data - DNN is acceptable.

The added detail is mostly provided in the 1-4 kHz area. This means it will primarily enhance speech sounds. The compromise though is that it also enhances other types of sound. Figure 11 shows the greater emphasis on the speech region with a transition to the low and high frequency areas.

This frequency shaping means that the mid frequencies (1-4 kHz) with the most speech cues will be given more weight than the low and high frequencies. Due to this frequency shaping and the fact that noise is often low frequency and speech cues are mid frequency a slightly better contrast will be created between speech and noise when sound is added to the signal.

Balanced is the default setting and works for most people. Detail is the setting that provides the most sound with the greatest level of detail for users who prefer to have a strong level of connection to surroundings and attended talkers. Comfort is the setting with less sound and provides more comfort for the users who prefer to have overall listening effort reduced slightly by subtly subduing surroundings relative to attended talkers.

Figure 12 shows the output of Oticon More compared to Oticon Opn S. The purpose of the illustration is to show the relationship between the different settings, but the bars cannot be readily translated into familiar quantities since the feature is dynamic.

Perspective

MoreSound Intelligence scans the full sound scene, applies personal settings for the individual user, precisely organizes the sounds around the user, and uses the DNN to create contrast between the identified sounds. This is all done very fast and precisely.

MoreSound Intelligence is built on Oticon's BrainHearing™ philosophy. It gives access to the full sound scene, where individual sounds stand out in clear contrast to each other and provide the brain the good neural code it needs.
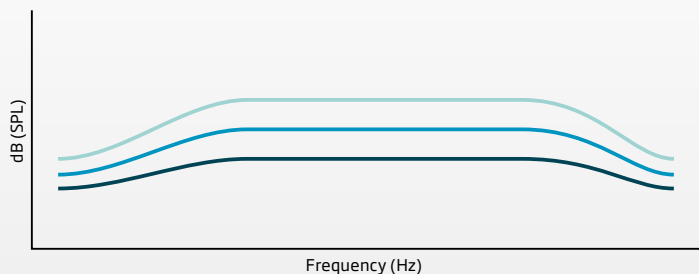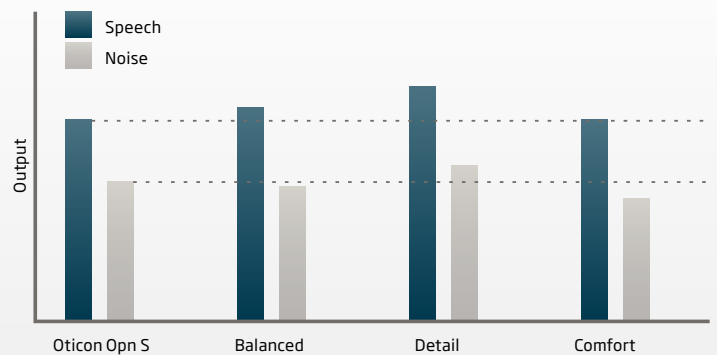

Figure 11. Frequency shaping for the sounds added by Sound Enhancer. Top line: Detail setting. Middle line: Balanced setting. Bottom line: Comfort setting

Figure 12. Relationship between the three different settings in Sound Enhancer compared to Oticon Opn S.

References

Akeroyd, M. A. (2014). An Overview of the Major Phenomena of the Localization of Sound Sources by Normal-Hearing, Hearing-Impaired, and Aided Listeners. *Trends in Hearing Vol. 18,* pp. 1-7.

Brændgaard, M. 2020. The Polaris Platform. Oticon tech paper

Hausfeld, L., Riecke, L., Valente, G., & Formisano, E. 2018. Cortical tracking of multiple streams outside the focus of attention in naturalistic auditory scenes. *NeuroImage, 181*, 617-626.

Leuba G; Kraftsik R (1994). "Changes in volume, surface estimate, three-dimensional shape and total number of neurons of the human primary visual cortex from midgestation until old age". *Anatomy and Embryology. 190* (4): 351–366. doi:10.1007/BF00187293. PMID 7840422. S2CID 28320951.

Løve, S. 2020. Optimal Fitting of Oticon More. Oticon Whitepaper

Man, B. and Ng, E. 2020. BrainHearing – The new perspective. Oticon Whitepaper

O'Sullivan, J., Herrero, J., Smith, E., Schevon, C., McKhann, G. M., Sheth, S. A., … & Mesgarani, N. 2019. Hierarchical Encoding of Attended Auditory Objects in Multi-talker Speech Perception. *Neuron, 104*(6), 1195-1209.

Puvvada, K. C., & Simon, J. Z. 2017. Cortical representations of speech in a multitalker auditory scene. *Journal of Neuroscience, 37*(38), 9189-9196.

Rönnberg, J., Lunner, T., Zekveld, A., Sörqvist, P., Danielsson, H., Lyxell, B., … & Rudner, M. (2013). The Ease of Language Understanding (ELU) model: theoretical, empirical, and clinical advances. *Frontiers in systems neuroscience, 7*, 31.

Squire, L. R. (2009). Memory and brain systems: 1969–2009. *Journal of Neuroscience, 29*(41), 12711-12716.

oticon
life-changing **technology**