Whisper

Redefining Hearing Aids

by Dwight Crow, Jan Linden, and Andreas Thelander Bertelsen

Abstract

The Whisper Hearing System is a learning hearing aid, a new kind of hearing aid that gets better over time. This continuous improvement is driven by regular software upgrades that keep patients and practices up to date with the latest technological advances. Continuous advancement of Whisper's performance is made possible by two major innovations that other hearing aids have not yet widely adopted: 1) the artificial intelligence technique called deep learning, and 2) mobile processor chips hundreds of times more powerful than existing hearing aids.

A Fundamentally New Hardware Approach

The Whisper Hearing System includes two BTE RIC style earpieces and the Whisper Brain, which is a pocket-sized device connected via an ultra-low latency proprietary wireless protocol to the earpieces. Patients can use the earpieces by themselves or in tandem with the Brain when it is nearby.

Whisper uses similar chips as those found in mobile phones and virtual reality devices for high performance audio defined entirely in software. These chips are hundreds of times more powerful than those in existing hearing aid devices, which use the legacy paradigm of building small, specialized but inflexible chips to perform specific audio functions.

Building a system for low latency integration of these powerful mobile processor chips required Whisper to innovate in hardware design and write new wireless protocols with much lower latency than Bluetooth. This technological foundation allows the Whisper Hearing System to run deep learning algorithms hundreds of times more complex than legacy hearing aids and to be easily upgraded. This would be similar to downloading a new app for your mobile phone.

Figure 1: The Whisper Hearing System, which includes the Whisper Brain (left), and BTE RIC-style earpieces (right).

Artificial Intelligence and Deep Learning

Deep learning is an artificial intelligence technique that became popular starting in 2012 (Krizhevsky et al., 2017, p. 87). It has produced record-breaking performance in transcribing speech (Wang et al., 2019, p. 1018), removing background noise (Kumar & Florencio, 2016), classifying sounds and scenes (Valenti et al., 2017, p. 1553), and even separating multiple simultaneous voices into distinct audio streams (Maciejewski et al., 2019). In short, deep learning has transformed what is possible in audio processing.

One example of deep learning delivering new possibilities is in the vexing area of improving speech clarity in noisy environments. Traditional signal processing techniques have been applied to speech clarity in academia for decades. Yet the Signal to Noise Ratio (SNR) has increased by only 4-5 dB in environments such as restaurants and cafes with many background voices (Paliwal et al., 2012, p. 287). In contrast, deep learning algorithms have been able to separate multiple voices with over 10 dB of SNR improvement (Isik et al., 2016). This is possible because deep learning provides a more powerful and precise approach to processing sound. With its ability to learn and recognize patterns, deep learning can target the patterns within the sound rather than solely using frequencies like traditional signal processing. As a result, deep learning can enhance voices in much more challenging listening environments, even when the background noise is overlapping significantly with the voices of interest.

Delivering clarity in the most challenging environments is one of the most exciting aspects of deep learning's promise. While legacy hearing aids have largely struggled to improve results in restaurants, cafes, and other crowded listening environments that involve background voices, deep learning has shown its strongest performance in exactly these situations. Recently, Whisper jointly published research with Mitsubishi replicating deep learning's success with improving speech clarity across 3000 audio speech samples recorded in real-world restaurants and cafes (Wichern et al., 2019). Across 3000 audio samples, the SNR improved significantly, and went from the SNR of a loud restaurant to an SNR equivalent of a typical household setting.

This means deep learning can make a huge difference for hearing health. And while this research with Mitsubishi focused specifically on using artificial intelligence for speech clarity, in real-world applications such as within a hearing aid, it is possible to layer on additional processing techniques. For example, microphone beamforming (i.e., directionality) provides even greater final speech clarity. As such, these additional processing algorithms deliver even larger benefits to users of a Whisper Hearing System in challenging situations.

But how is it possible for deep learning to yield exciting results like this? The key is the number of parameters a deep learning algorithm has and the number of operations it uses every time it runs. Deep learning has vastly more parameters and operations than traditional signal processing algorithms. This is the core of its improved performance, and it is worth explaining what these two concepts mean.

Parameters are numerical values within every sound processing technique that affect how the algorithm will behave on a given sound signal. These parameters are individual switches tuned to deliver the best performance given a variety of acoustic scenes. In general, the more parameters available to an audio system, the better that system will be in handling complex acoustic scenes like restaurants, cafes, and other noisy places.

Operations are the computational actions, such as addition or multiplication, that are used to calculate the outcome of any sound processing algorithm. The more operations an algorithm has, the more sophisticated use of its parameters it can make. While the parameters of an algorithm store what it has learned, the operations are needed to make it run.

As context, prior versions of signal processing utilized only a small handful of parameters and operations to achieve their results. This was necessary because they had to run on the small processors available in legacy hearing devices. For example, the Speech Presence Probability algorithm (Gerkmann & Hendricks, 2011) is a highly popular technique used to detect speech. It produced record-breaking performance as recently as 2011. But this algorithm has just 10 parameters, requires only hundreds of operations each time it is run, and is characteristic of many traditional techniques that can only hold a small amount of information when tuned over new audio.

In contrast, at the core of any deep learning model are the millions to billions of parameters that store the information each algorithm has learned (Reddi et al., 2019 and Heaven, 2020). This learning occurs during a process called training, where diverse audio data is given as input so the model can detect patterns and commonalities between different situations. For example, someone might train a deep learning algorithm to recognize the world "Hello." The first step would be to create a new deep

learning algorithm with a few million parameters. Next, the engineer would train the algorithm to predict whether the word “Hello” is present in many, many audio samples. On each sample, every parameter would be slightly adjusted to improve the prediction. After millions of predictions and small improvements, the algorithm would contain all the information needed to determine whether audio contained the word “Hello.”

Deep learning algorithms store information like the above in their parameters, and it is the number of parameters that determines how much an algorithm can learn. An algorithm with a few hundred parameters would stop learning after minutes of audio and would have very basic predictions. In contrast, an algorithm with millions of parameters could continue to learn while being trained on years of audio and completing very complex tasks.

This helps explain why deep learning models greatly outperform their traditional counterparts: the huge increase in parameters stores much more information about handling the challenging situations that traditional signal processing falls short on. Deep learning also helps explain why Whisper can continue to learn and improve compared to legacy hearing aids: it is the first hearing device with chips that can support deep learning algorithms with orders of magnitude more parameters, and therefore audio information storage, than legacy devices. This allows the Whisper Hearing System to learn information from orders of magnitude more distinct audio situations than other hearing aids. Whisper also enables upgrades employing these academic advances that require cutting-edge capabilities.

It would be natural to ask why legacy hearing devices do not also use powerful deep learning algorithms with a high number of parameters. The answer is simple: all legacy hearing aids, even modern ones, do not have the capability to run powerful deep learning algorithms of this size because they cannot run enough operations to support them with the chips available in an earpiece. Powerful algorithms like those in the Whisper Hearing System require more than 300 billion operations each second. In contrast, the two top manufacturers have chips that can support between 0.4 billion to 1.2 billion operations per second in 2020 (Welle & Bach, 2016). Even much vaunted AI-driven earpiece launches from existing manufacturers top out at about twice these numbers. This hardware is simply insufficient to run the full-scale deep learning algorithms discussed above.

"The increase in parameters allows the deep learning model to handle challenging situations.”

This gap leads to a large discrepancy in the quality of deep learning algorithms being used in learning hearing systems like Whisper versus legacy devices, even ones that state they are running deep learning on the earpiece. While legacy devices may describe their sound processing as deep learning, they must run algorithms with orders of magnitude less parameters and operations than a learning hearing system. This severely limits the performance of earpiece-based algorithms and creates a false equivalence in deep learning. As an analogy, mobile devices from 2005 and 2020 are both smartphones but have orders of magnitude different capabilities. This discrepancy is proportional to the difference between the algorithms on a Whisper Hearing System and a legacy hearing

device.

The next question follows naturally. How does Whisper enable deep learning algorithms with orders of magnitude greater capacity than legacy hearing devices?

Mobile Processors

Mobile processors are the computational chip at the core of smart devices from phones to VR headsets and smartwatches (Robertson, 2020). These processors can run hundreds of billions of operations per second and enable smart devices to run powerful deep learning algorithms. The Whisper Brain is a pocket-sized accessory that combines a powerful mobile processor with a proprietary, ultra-low latency wireless connection. Instead of only using a small, legacy chip in the earpiece, the Whisper Brain is the first hearing system to use a mobile processor to directly enhance audio.

Smartphones have driven the need for powerful mobile chips with impressive results. As can be seen in Figure 2, the computation available in mobile processors has increased by over 650X since 2008. This has enabled explosive growth in the capabilities of modern smart devices while leaving devices without mobile processors lacking new capabilities.

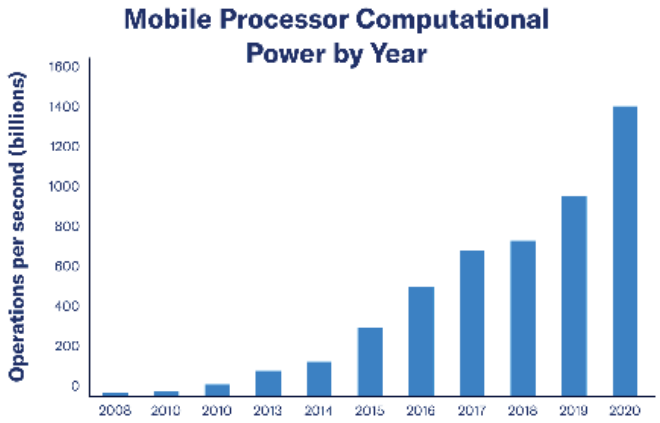

Figure 2: The computational power available in mobile processors from 2008 to 2020.

The more operations a mobile processor performs each second, the more powerful capabilities it provides. Examining smartphone processor speeds over the last 12 years illustrates this (Qualcomm, 2008-2020). For example, a smartphone processor from 2008 had around 2 billion operations per second and could load basic websites with HTML. A smartphone processor from 2014 had around 100 billion operations per second and could play graphical mobile games. And a smartphone processor from 2020 often had over 1 trillion operations per second and could do everything from recognize your voice to transcribe speech and put special effects into video calls. For any smart device, the more computational power it has, the more it can do.

In comparison, legacy hearing aids using earpiece chips only have the computational power of a phone from 2008 and are therefore limited to 2008 capabilities. Figure 3 illustrates the growing gap between legacy hearing devices and modern mobile processors. While the Whisper Brain is capable

of over 300 billion operations per second to run deep learning, the processor in a competing 2020 flagship legacy hearing device has only 375 million (Davis, 2017) operations per second available. This puts the competing product at less than 20% of the power of a phone from 2008.

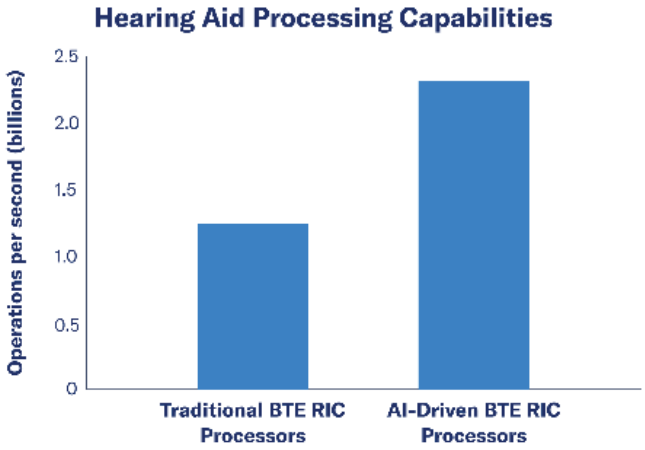

Figure 3: Processing capability of traditional BTE RIC hearing aids and AI-driven BTE RIC devices.

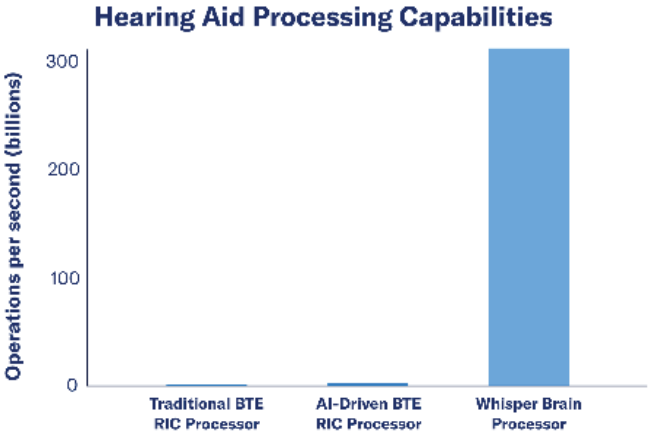

Figure 4: Processing capabilities of the Whisper Brain compared to traditional BTE RICs and AI-driven BTE RIC devices.

While legacy hearing devices lack powerful new capabilities, the Whisper Hearing System delivers significant hearing benefits to patients for the first time (Miller, 2020). The dedicated and ultra-low latency integrated mobile processor in the Whisper Hearing System receives new software functionality for regular upgrades improvements. It is also the first hearing system hardware that is able to run powerful deep learning algorithms because of this processor. One such algorithm the Whisper Hearing System runs is the proprietary deep learning powered Sound Separation Engine, which helps remove background noise. Figure 5 shows actual results of this deep learning algorithm run on a real-world restaurant audio with a primary speaker and loud background conversations. The upper spectrogram shows the original audio, and the lower spectrogram shows the enhanced primary voice elevated above the background noise. To provide this benefit, The Whisper Hearing System passed the full audio of the conversation in 4 ms segments to the Sound Separation Engine on the

Whisper Brain, where each 4 ms segment was directly enhanced with deep learning. This enhancement took over 300 billion operations per second and would have been impossible to perform using a legacy device.

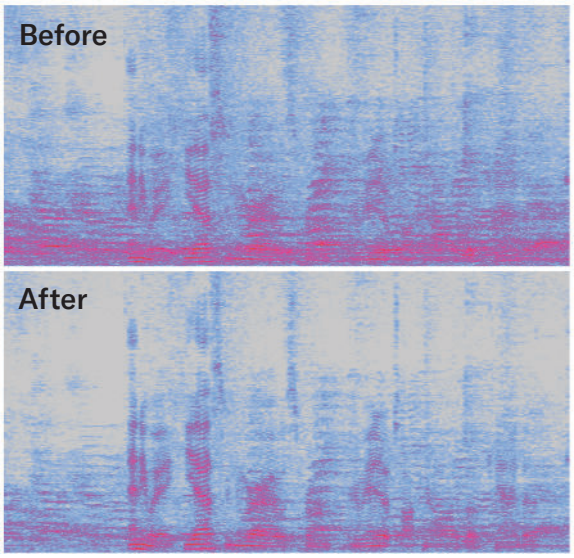

Figure 5: Actual Whisper noise reduction results. The spectrogram shows a visual representation of the signal with frequency on the y-axis and time on the x-axis. Whisper reduces background noise – even in vocal frequencies – by using deep learning to recognize the voice patterns in the noise.

The Whisper Hearing System can run algorithm like these for the benefit of the patient, because it has more integrated computational power than any other hearing aid. While Whisper can directly enhance a 4 ms segment of audio in 3.7 ms using billions of operations, the processor in a leading competitor would take 2,800 ms to perform this same task. Because running an algorithm hundreds of times per second is a requirement to enhance audio in the tight latency requirement of a premium hearing device, it is impossible for a legacy device to run a deep learning algorithm of this size and capability. As such, it is exciting for Whisper to bring algorithms and audio benefits like the Sound Separation Engine to audiology via the first device with an ultra-low latency mobile processor fully integrated into the hearing system.

Conclusion

Deep learning algorithms have transformed audio processing to deliver audio benefits such as record-breaking speech enhancement in restaurants and cafes. The power of a deep learning algorithm is proportional to the parameters and operations it contains. Parameters store what a deep learning algorithm has learned, and operations are the mathematical actions required to run the algorithm on new data. Because deep learning algorithms are only as powerful as the number of parameters and operations they employ, it is impossible to run powerful deep learning algorithms on earpiece-only hearing devices due to the small amount of parameters and rate of operations per second that earpieces can support.

The Whisper Hearing System is the first hearing device to include a dedicated, ultra-low latency, fully integrated mobile processor, contained in the Whisper Brain, in addition to earpieces. This makes it the first device that can support the large number of parameters and operations required for powerful deep learning algorithms. Its unique properties enable the hearing system to continuously improve by storing new information in its large parameter set. Whisper also possesses a processor sufficient to support other new upgrades. These properties and proprietary technology are the core of why Whisper is a new category of learning hearing aid that keeps getting better over time.

Dwight Crow is the co-founder and CEO of Whisper. Prior to Whisper, Dwight built the ecommerce segment at Facebook and helped drive over $1B per quarter in revenue. He was the founder of Carsabi, a machine learning based car sales aggregator which was acquired by Facebook in 2012. Dwight has a BS in Chemical Biology and Computer Science from the University of California, Berkeley.

Jan Linden is the head of engineering at Whisper. Jan is an expert in engineering management with a focus on audio and video technology. He has over 25 years of R&D experience in speech processing and speech coding at Chalmers University of Technology (Sweden), University of California, Santa Barbara, SignalCom Inc., and Global IP Sound, Inc. Prior to Whisper, Jan worked at Nest and Google, among other companies. He holds a Ph.D. and an M.Sc. in Electrical Engineering from Chalmers University of Technology. He has published more than 25 papers and filed several patents.

Andreas Thelander Bertelsen is an industry-leading architect within signal processing and machine learning for hearing devices. He is the lead audio architect at Whisper who has more than a decade of experience in researching, developing and designing noise suppression systems for hearing devices. Andreas led the design of the noise suppression system in the successful Oticon Opn S and Oticon More series of hearing devices. His work has led to the authorship of numerous key patents in the field.

References

Davis, N. (2017, December 11). A New Wireless-Enabled Audio Processor for Hearing Aids and Cochlear Implants. All About Circuits. https://www.allaboutcircuits.com/news/new-wireless-enabled-audio-processor-dsp-hearing-aid-cochlear-implant/

Gerkmann, T., & Hendricks, R. C. (2011, October 1). Noise power estimation based on the probability of speech presence. IEEE Conference Publication. https://ieeexplore.ieee.org/document/6082266

Heaven, W. D. (2020, December 10). OpenAI's new language generator GPT-3 is shockingly good—and completely mindless. MIT Technology Review. https://www.technologyreview.com/2020/07/20/1005454/openai-machine-learning-language-generator-gpt-3-nlp/

Ho, J. (2015, August 24). Qualcomm Details Hexagon 680 DSP in Snapdragon 820: Accelerated Imaging. AnandTech. https://www.anandtech.com/show/9552/qualcomm-details-hexagon-680-dsp-in-snapdragon-820-accelerated-imaging

Isik, Y., Le Roux, J., Chen, Z., Watanabe, S., & Hershey, J. R. (2016, July 7). Single-Channel Multi-Speaker Separation using Deep Clustering. ArXiv.Org. https://arxiv.org/abs/1607.02173

Killion, M. C. (1997). SNR Loss: I Can Hear What People Say, But I Can't Understand Them. The Hearing Review, 4(12), 8–14. https://www.etymotic.com/media/publications/erl-0037-1997.pdf

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84–90. https://doi.org/10.1145/3065386

Kumar, A., & Florencio, D. (2016, May 9). Speech Enhancement In Multiple-Noise Conditions using Deep Neural Networks. ArXiv.Org. https://arxiv.org/abs/1605.02427

Maciejewski, M., Wichern, G., McQuinn, E., & Le Roux, J. (2019, October 22). WHAMR!: Noisy and Reverberant Single-Channel Speech Separation. ArXiv.Org. https://arxiv.org/abs/1910.10279

Miller, R. (2020, October 15). Whisper announces $35M Series B to change hearing aids with AI and subscription model. TechCrunch. https://techcrunch.com/2020/10/15/whisper-announces-35m-series-b-to-change-hearing-aids-with-ai-and-subscription-model/

Oticon More Technology. (2021). Oticon. https://www.oticon.com/professionals/brainhearing-technology/more-technology

Paliwal, K., Schwerin, B., & Wójcicki, K. (2012). Speech enhancement using a minimum mean-square error short-time spectral modulation magnitude estimator. Speech Communication, 54(2), 282–305. https://doi.org/10.1016/j.specom.2011.09.003

Ramsdell DA. The psychology of the hard-of-hearing and the deafened adult. In: Davis H, editor. Hearing and deafness. New York: Holt, Rinehart, & Winston; 1947. pp. 459–473. [Google Scholar]

Reddi, V. J., Cheng, C., Kanter, D., Mattson, P., Schmuelling, G., Wu, C., Anderson, B., Breughe, M., Charlebois, M., Chou, W., Chukka, R., Coleman, C., Davis, S., Deng, P., Diamos, G., Duke, J., Fick, D., Gardner, J. S., Hubara, I., … Zhou, Y. (2019, November 6). MLPerf Inference Benchmark. ArXiv.Org. https://arxiv.org/abs/1911.02549

Robertson, A. (2020, February 25). Qualcomm reveals a headset design for its latest VR chips. The Verge. https://www.theverge.com/2020/2/25/21147912/qualcomm-snapdragon-xr2-vr-ar-xr-platform-referenc e-design-announcement

Roux, J. L., Wisdom, S., Erdogan, H., & Hershey, J. R. (2018, November 6). SDR - half-baked or well done? ArXiv.Org. https://arxiv.org/abs/1811.02508

Valenti, M., Squartini, S., Diment, A., Parascandolo, G., & Virtanen, T. (2017). A convolutional neural network approach for acoustic scene classification. 2017 International Joint Conference on Neural Networks (IJCNN), 1547–1554. https://doi.org/10.1109/ijcnn.2017.7966035

Wang, D., Wang, X., & Lv, S. (2019). An Overview of End-to-End Automatic Speech Recognition. Symmetry, 11(8), 1018. https://doi.org/10.3390/sym11081018

Welle, J. N., & Bach, R. (2016). The Velox Platform. Oticon. https://www.oticon.com/-/media/oticon-us/main/download-center/white-papers/15555-9940-velox-whitepaper.pdf

Wichern, G., Antognini, J., Flynn, M., Zhu, L. R., McQuinn, E., Crow, D., Manilow, E., & Le Roux, J. (2019, July 2). WHAM!: Extending Speech Separation to Noisy Environments. ArXiv.Org. https://arxiv.org/abs/1907.01160

Wikipedia contributors. (2021, February 18). Adreno. Wikipedia. https://en.wikipedia.org/wiki/Adreno

Wu, Y.-H., Stangl, E., Chipara, O., Hasan, S. S., Welhaven, A., & Oleson, J. (2018). Characteristics of Real-World Signal to Noise Ratios and Speech Listening Situations of Older Adults With Mild to Moderate Hearing Loss. Ear & Hearing, 39(2), 293–304. https://doi.org/10.1097/aud.0000000000000486