Breaking Through the 6ms Latency Barrier: A New Class of DNN for Zero-Latency On-Chip Conversation Enhancement

Introducing Chatable AI v3.0 Edge: *Latency-Free On-Chip Conversation Enhancement for TWS Earbuds and Hearing Aids*

Andrew J.R. Simpson, PhD
Chief Scientific Officer,
Chatable
London, UK.
andy@chatableapps.com

Giles H. Tongue
Chief Executive Officer,
Chatable
London, UK.
giles@chatableapps.com

Abstract—Super-human hearing has been described by prominent industry technologists as a 'moon-shot' because of major unsolved technical challenges associated with Artificial Intelligence (AI). The first unsolved challenge is AI processing latency - more than 6ms of latency is detectable and annoying for users. The second is AI processing requirements for on-chip deployment - the most effective AI architectures are too big to fit on-chip. Here, we introduce Chatable AI v3.0 Edge - a new class of Deep Neural Network (DNN) technology providing scientifically-validated Conversation Enhancement for True Wireless Stereo (TWS) Earbuds and Hearing Aids. The first On-Chip DNN to perform real-time (<6ms) direct processing of audio for Conversation Enhancement and the first with no audible latency.

Keywords—Super-Human Hearing, Conversation Enhancement, Deep Learning, Artificial Intelligence.

Conversation: More than Just Words

The sound of a voice is a rich source of information about the person who is talking to you. Encoded within the sound of a voice are subtle acoustic cues - sound qualities which capture physical characteristics of the person who is talking.

Thanks to your brain, these acoustic cues allow you to identify a person by their voice alone, locate the person in space, judge their proximity to you and even the direction they're facing. In fact, your auditory brain interprets a vast array of subtle acoustic cues encoded within the sound of a person's voice providing information about their size, shape, gender, age, state of health, arousal and emotional state. Humour, disappointment, excitement, happiness, sadness, irony, etc.

Therefore, it's clear that conversational speech is far more than just a question of intelligibility. It's more than just sending some words from A to B. For many, their voice is central to their personality, identity and self-expression.

Degraded: Conversation Experience

The amazing abilities of the brain to interpret these subtle acoustic cues comes from a combination of extraordinary sensory acuity in the ears and extraordinary neural resources dedicated to the task in the brain.

The acuity of human hearing is remarkable. In silent conditions, a normal hearing person is capable of hearing a mosquito flying at one meter. However, this tiny, delicate sound is almost impossible to detect if there are any other sounds happening at the same time. This is known as 'masking' and means that information has been lost by the brain.

With masking, more competing sound means more information is lost. At first, with gentle competing noise, the most subtle information starts to be lost. As the level of competing noise increases, more and more information is lost, until at last even words can become unintelligible.

We live in a bustling, noisy, densely populated and increasingly urban world. We spend much or even most of our time in enclosed and often brutally reverberant spaces. Even alone in a room, when you speak the sound you make radiates in all directions, reflects off walls and other surfaces like ripples in a pond and comes back to your ears disordered, overlapping and confused. Worse still, these urban spaces we live and converse in aren't empty and silent - they're full of people and other sources of competing noise all striving to be heard.

All of this combined causes extensive masking and irrevocable information loss. Intimacy between people is impaired. Where the voice is a fundamental means of self-expression, identity and individuality, this masking results in a fundamentally degraded experience of people in all their uniqueness and nuance. The subtle qualities that make each person and each moment of human connection unique are corrupted and the experience is degraded.

Conversation Enhancement for TWS Earbuds and Hearing Aids

With this in mind it's perhaps not surprising that according to recent industry research (FutureSource, 2021), Conversation Enhancement is the number one most requested feature for TWS Earbud users. For the same reasons, difficulty with speech in noise is also a long-standing problem for hearing aid users and a high priority for hearing aid manufacturers.

Conversation Enhancement means processing sound in real-time to undo the effects of environmental sound degradations, to prevent masking, to restore the information that was lost and thereby restoring the person-to-person experience to its fullest and most vivid.

Technically, Conversation Enhancement is an extremely difficult and as-yet-unsolved engineering challenge which requires selective acoustic enhancement of conversational speech signals which are mixed with other acoustic signals. This cannot be achieved with traditional signal processing methods (i.e., DSP) such as those methods employed in hearing aids or consumer hi-fi systems (i.e., graphic equalisers or multi-band compression). In fact, industry has been quick to realise that the natural solution to this problem is: Artificial Intelligence (AI). However, the problem remains unsolved and substantial barriers must be removed before this potential can be realised - barriers which are the subject of this whitepaper.

In order to solve this problem for a TWS Earbud or Hearing Aid, it is necessary to use AI to process incoming signals so as to emphasise conversational speech and attenuate interfering background signals in real-time, providing the user with an enhanced conversational experience. In both cases, the ideal solution would be on-chip in order that the user experience is not cumbersome. In both cases, minimal audio processing latency is preferable. Both AI latency and on-chip deployment are major technical challenges that must be overcome but have only partly been solved by industry AI methods to date.

AI: Current Industry DNN Methods

Deep Learning is an advanced form of AI and a subclass of machine learning. Deep Neural Networks (DNN) are artificial neural network architectures featuring deeply stacked layers of artificial neurons, trained using a variety of state of the art training methods. There are two prevailing industry approaches to the application of DNN to the problem of Conversation Enhancement: Inline DNN and Hybrid DNN, each with their pros and cons.

1. Inline DNN: Using DNN to Process Sound Directly

In this class of DNN application, a DNN takes in frames of audio samples, recorded from the microphone, and processes them in real-time to produce corresponding output frames of audio samples which are then routed to the receiver/loudspeaker so that the user can hear the processed sound.

The benefit of this 'inline' approach is that the considerable intelligence capability of Deep Learning neural networks can be directly applied to analysing and processing the sound itself in intelligent ways.

However, the downside of this approach is that it introduces substantial inherent latency - known as algorithmic latency - with up to 40ms algorithmic latency being common in leading industry products. Furthermore, current Inline DNNs have not yet become computationally efficient enough to fit on the chips used in TWS and Hearing Aids and on-chip application would not reduce algorithmic latency in any case.

2. Hybrid DNN: Using DNN to Control Traditional DSP

Alternate 'Hybrid' DNN industry approaches use DNN to passively conduct *sound analysis* - but, unlike Inline methods, the DNN do not directly process the sound heard by the user. The outputs of the Hybrid DNN are instead used to adjust control parameters of traditional hearing aid DSP (e.g., multi-band equalisation and compression).

This is broadly equivalent to having a tiny audiologist living inside your ear whose job is to constantly adjust the settings on your hearing aid in real-time to best suit the environment you are in. As long as the 'audiologist in your ear' makes relatively modest, steady and infrequent adjustments, the time it takes him/her to analyse the sound environment and make changes does not cause a latency that you will notice.

The advantage of this hybrid approach is that, unlike Inline DNN methods, because the DNN is not in the direct audio path the audio processing latency is the same as the traditional DSP latency. Furthermore, this approach is relatively computationally cheap due to greatly reduced complexity. As a result, this can be readily implemented on-chip.

However, the downside of this approach is that the potential sound processing itself is limited to what a traditional hearing aid DSP is capable of. Furthermore, any attempt to rapidly adjust the parameters of the hearing aid DSP inherently incorporates the algorithmic latency of the DNN into the audio path, resulting in prohibitive latency. Hence, real-time intelligent adjustments are limited to relatively slow acting instructions (such as changing the program selected on the DSP, or occasional adjustments to graphic equalisation gain settings). Thus, whilst this method can be applied on-chip, the overall efficacy of the approach is strictly limited.

Introducing a New Class of DNN: Chatable AI v3.0 Edge

The introduction of **Chatable AI v3.0 Edge** marks the emergence of a new class of On-Chip Inline Direct-Sound-Processing DNN for Conversation Enhancement because it is the first Inline Direct-Sound-Processing DNN to work without discernible latency, and the first to be on-chip.

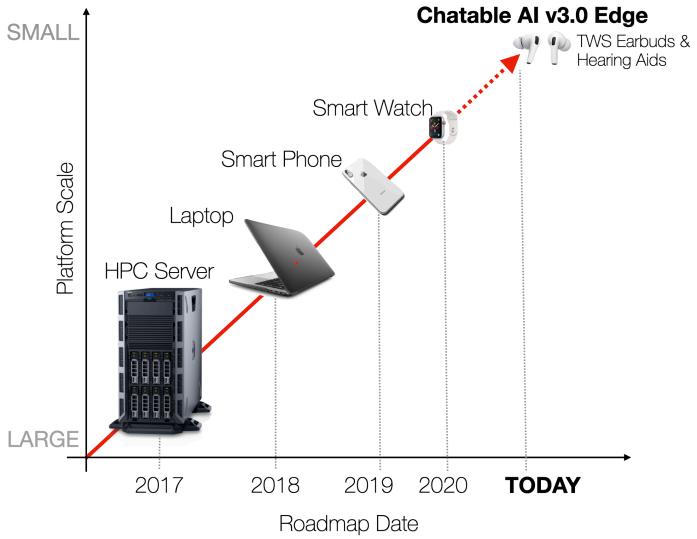

Figure 1. Getting on-chip: The Chatable AI R&D journey
Taking our groundbreaking Conversation Enhancement AI from large scale HPC servers down to on-chip TWS and Hearing Aid applications of today.

Chatable AI v3.0 Edge is our latest generation AI - the worlds first on-chip zero-latency Conversational Enhancement AI for TWS Earbuds and Hearing Aids. This new class of DNN technology is a product of our proprietary neuroscience-led AI and comes from reverse engineering the human auditory brain and how its active neural circuits function when listening to speech.

Designed specifically for next-generation AI-oriented TWS Earbud and Hearing Aid processors, Chatable AI v3.0 Edge is an advanced Deep Neural Network architecture featuring over one hundred million AI calculations per second.

Chatable AI v3.0 Edge: The Journey to On-Chip

Fig. 1 shows various platform milestones on the Chatable AI roadmap. Our groundbreaking Chatable AI was first developed in the lab on large-scale high performance computing servers. As efficiency was improved, subsequent generations of the AI progressed to laptop, smartphone and then smart watch platforms. Each subsequent milestone leap in platform scale requires an order of magnitude increase in efficiency (reduction in processing and RAM requirements) in order to eventually meet the stringent requirements of on-chip processing for TWS earbud and Hearing Aids.

Breaking Through the 6ms Latency Barrier

The total time it takes for sound from the microphone to be processed and delivered to the receiver (loudspeaker) is known as the roundtrip latency. Hearing industry research shows that in order that sound be processed quickly enough that it is perceived as having no delay/latency, the roundtrip latency must be 6ms or less.

For Inline DNN direct processing of audio, the roundtrip latency is a combination of algorithmic latency (of the DNN itself), the time it takes for the processing to be executed, and any additional processing overheads associated with the operating system.

Thanks to our groundbreaking advances in AI, and because it is deployed directly on-chip, Chatable AI v3.0 Edge is the worlds first Conversational Enhancement AI to break through the latency barrier of 6ms-or-less roundtrip latency, making it the first AI with imperceptible latency. This means that our Inline direct-sound processing architecture provides peerless, vivid sound quality without latency.

This leap in AI latency performance is a result of our neuroscience-led AI approach and was made possible by our fundamental advances in both AI and auditory neuroscience. Additionally, the move to on-chip deployment allows us to preserve the vanishingly low latency of the AI by removing overheads associated with operating system and/or tethered smartdevices. Using latest generation AI-capable chips allows our breakthrough zero-latency algorithms to provide a listening experience that has never been heard before.

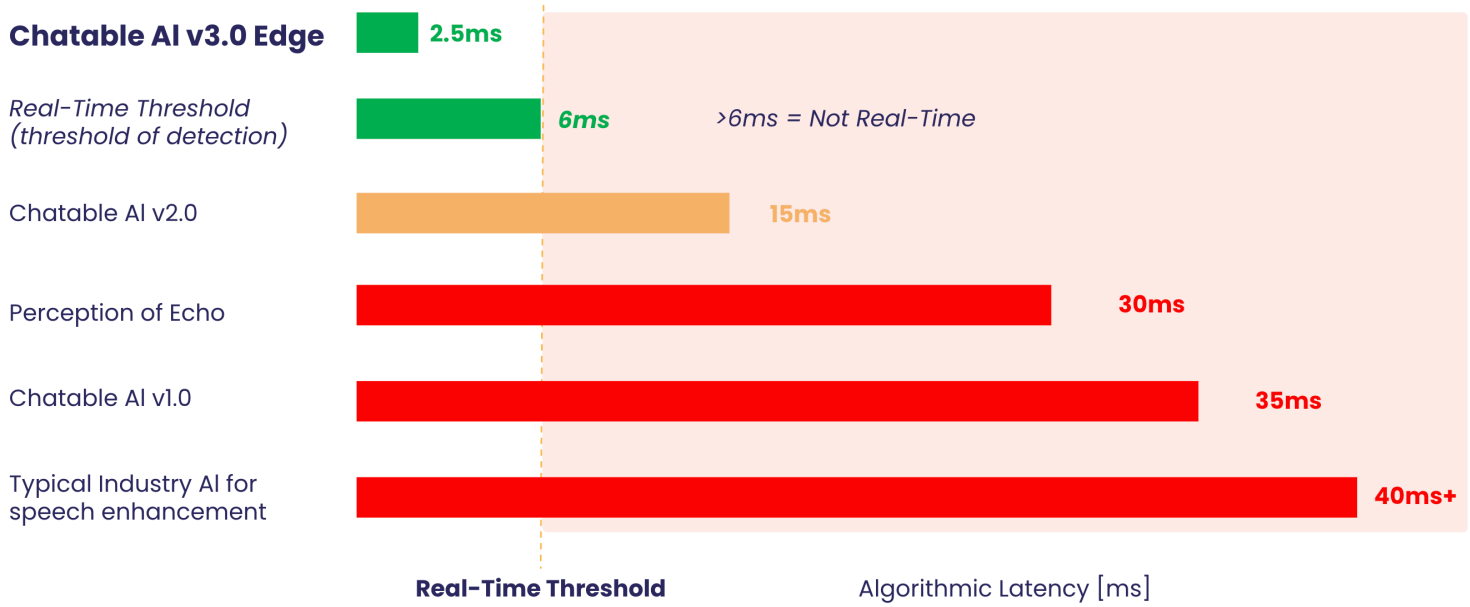

Figure 2. Breaking through the 6ms latency barrier
Groundbreaking latency of Chatable AI v3.0 Edge compared with various references

Scientifically Proven Conversation Enhancement

Our groundbreaking AI has been through rigorous, state of the art scientific testing and validation processes both in the lab and out in the real-world. The first and only Conversation Enhancement AI that is directly and scientifically validated for enhancement of real conversation in the real world.

In extensive listening tests featuring over 600 people with diverse hearing profiles, ranging from normal hearing to mild, moderate, severe and profound hearing impairment, listening-oriented neuro-diverse conditions including Autism, ADHD, Dyslexia and Auditory Processing Disorder, the majority (90%) reported that the Chatable AI substantially enhanced their conversation.

A New AI Experience

The unique combination of advanced on-chip Deep Learning AI for Conversation Enhancement and undetectable latency provide a vivid new AI experience for users, where conversation 'pops out'. For the first time, super-human hearing powered by AI is possible for Hearing Aids and TWS Earbuds and the experience provided by Chatable AI v3.0 Edge on-chip technology is designed to make conversation a fun, new AI experience.

Further Information

The authors may be contacted as shown above for further information on compatible chips, platform overheads and customisation options.

About the authors

Dr Andrew Simpson, Founder & CTO @Chatable

Andrew Simpson holds a PhD in auditory neural signal processing and has published widely on artificial intelligence and hearing topics. Working at the intersection of auditory neuroscience and artificial intelligence, Andrew is pioneering a unique neuroscience-led AI approach to hearing aid technology.

Prior to Chatable, Andrew held Research Fellow and Research Associate positions at the University of Surrey, Queen Mary University of London and University College London, where he worked on topics as diverse as adaptive hearing, neural coding of speech and artificial intelligence for voice separation. He has published more than 45 academic papers and has over 400 academic citations.

Giles H Tongue, Founder & CEO @Chatable

Giles has been a tech innovator for the last decade, playing a leadership role in high-growth consumer electronics businesses tech21, Impact Tech Labs and most recently wearable startup NURVV, a revolutionary running wearable featuring state of the art sensors and algorithms, electronics, software, manufacturing and app development.

Giles is now CEO and Co-Founder of Chatable, leading an intrepid and interdisciplinary team of technologists out of the lab and into the world of consumer electronics to usher in a new AI-driven era.