AI-driven interactions powered by high-SNR MEMS microphones

by Dr. Gunar Lorenz, Infineon Technologies

Introduction

At Infineon, we have long believed that superior audio solutions are essential to enhancing the user experience offered by consumer devices. We are proud of our unwavering commitment to innovation, which has resulted in remarkable advances in active noise cancellation, transparent hearing, studio recording, audio zoom, and other related technologies. As a leading supplier of XENSIV™ MEMS microphones, we have focused our resources on improving the audio quality of MEMS microphones, enabling superior experiences in a wide range of consumer devices, including TWS and over-ear headphones, laptops, tablets, conferencing systems, smartphones, smart speakers, hearing aids, and even cars.

Today, we live in an exciting era where AI is revolutionizing everyday life and tools like ChatGPT are redefining productivity through intuitive text and voice interactions. As AI-based systems continue to advance, as traditional business models, beliefs, and assumptions are being challenged. What is the role of voice in the emerging AI ecosystem? Do we, as business leaders, need to rethink our beliefs? Will the rise of generative AI diminish the importance of high-quality audio input, or will high-quality audio input become essential for the widespread adoption of AI-based services and personal assistants?

AI – from helpful assistant to best friend

It is natural for humans to adapt their responses not only to the content of a question, but also to the form in which it is asked. The human voice provides a variety of cues that can be used to determine the age, gender, social and cultural background, and emotional state of the questioner. In addition, recognizing the setting (e.g., airport, office, transportation, or physical activity such as running) can help determine the questioner's intent and adjust the answer or conversation accordingly.

Despite significant advances in AI capabilities, there is still a perception that AI-based assistance lacks the ability to correctly predict the intent of a human question or how a particular message will be interpreted. To improve human-machine interaction, three key factors should be considered in the rhetorical choices an AI makes: knowledge of the listener, the listener's emotional state, and the environmental context.

In many cases, the received audio signal alone is sufficient to extract useful information and adapt an appropriate response. Consider, for example, a phone call or audio conference with people you have never met. More importantly, consider how a person's perception of another person evolves and changes after repeated conversations without ever having the opportunity to interact in person.

Recent research suggests that even small changes in an AI's linguistic response style result in a noticeable shift in the AI's perceived social competence and personality. It is reasonable to hypothesize that with the right level of acoustic input, future AI systems will be able to function as effective companions, exhibiting behaviors of a human friend, such as inquiring and truly listening to answers, or alternatively, simply listening and withholding judgment when appropriate.

How do humans experience audio signals?

As with any verbal communication, an audio message employs language and words to convey thoughts, feelings, and ideas. Furthermore, other elements of communication such as pitch, speed, volume, and background noise can affect the overall perception of the message.

From a scientific perspective, the human ear is capable of perceiving audio signals based on two key factors: frequency and sound pressure level. Sound pressure level (SPL) is quantified in decibels (dB_{SPL}), which indicates the amplitude of the sound pressure oscillating around the ambient atmospheric pressure. An SPL of 100 dB_{SPL} is comparable to the very loud noise from a lawn mower or helicopter.

The lowest point in the SPL range (0 dB) is associated with a sound pressure oscillation of 20 µPa, which represents the hearing threshold at 1 kHz of a young, healthy individual with optimal hearing. All human sounds related to speech fall into the frequency band of 100 Hz and 8 kHz. The corresponding human hearing threshold according to ISO 226:2023 is shown in Figure 1.

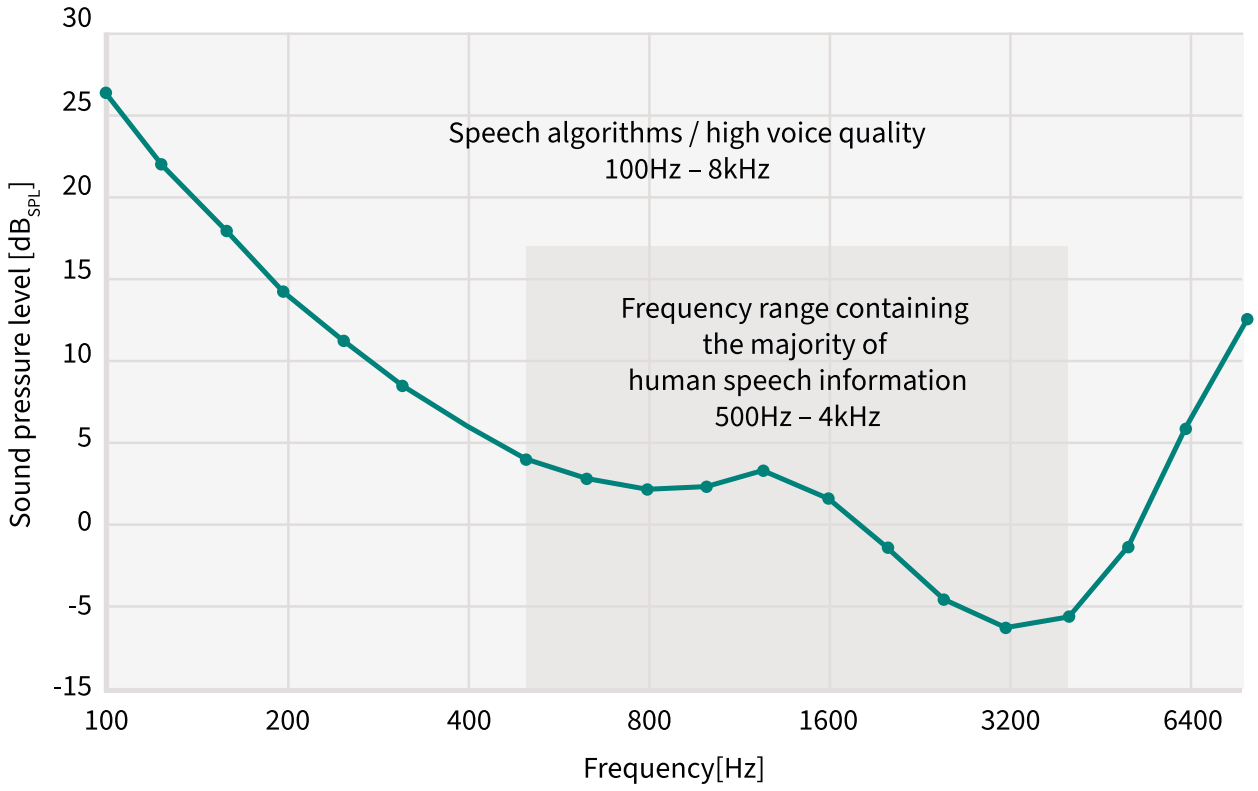

Figure 1: Threshold of hearing: The level of sound at which a person gives 50% of correct detection responses on repeated trials, according to ISO 226:2023.

As illustrated in Figure 1, the human ear is particularly sensitive to frequencies within the range of 500 Hz to 6 kHz. Any issues with the relative frequency balance at those frequencies can have a significant impact on the perceived quality of voices and instruments. The frequencies between 500 Hz and 4 kHz contain the majority of the information in human speech that affects speech intelligibility.

Specifically, the frequencies around 2 kHz are of particular importance. Frequencies from 5 kHz to 10 kHz are important for music. These frequencies add "life" and "brightness" to sound. However, these frequencies contain relatively little speech information, only sibilance, which is the hissing sound at the beginning of words like "ship" "chip", and "zip". Reducing sibilance to around 6 to 8 kHz can have a detrimental effect on speech intelligibility.

As most of us are aware, the human hearing threshold declines with age, as illustrated in Figure 2.

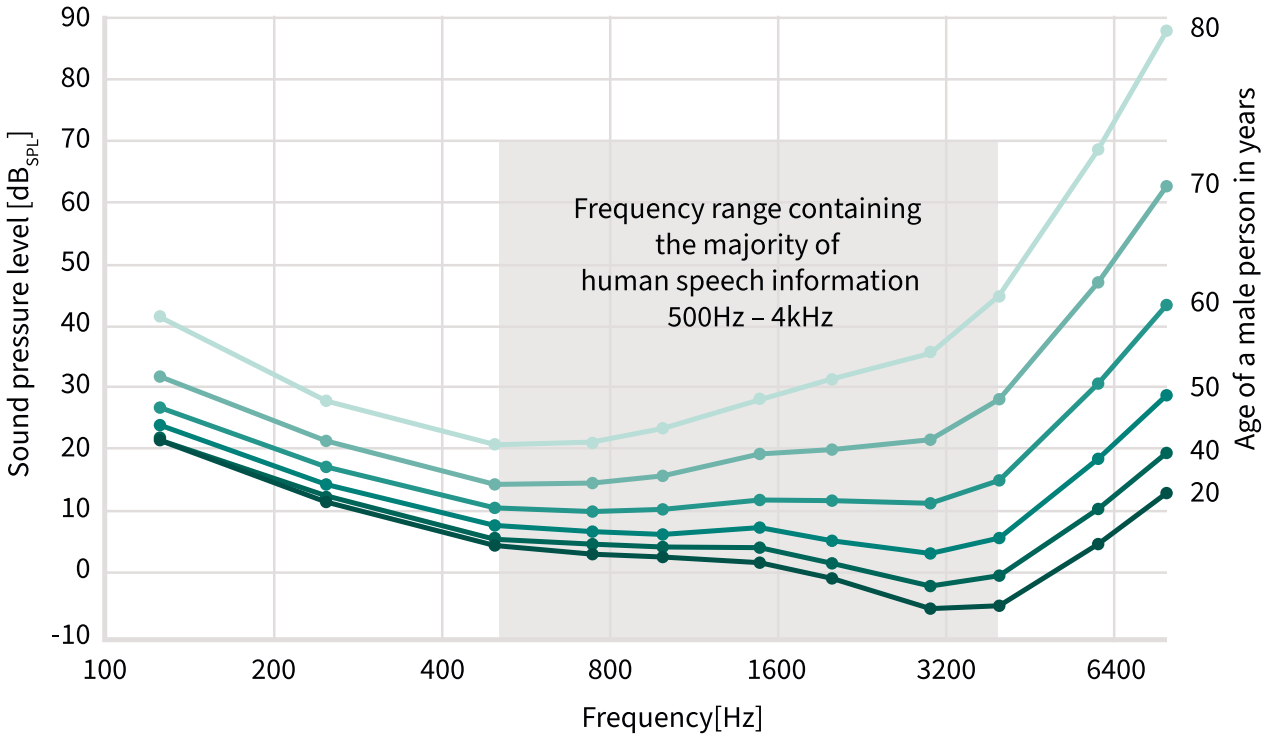

Figure 2: This graph shows the degradation of the hearing threshold of ontologically normal male persons of various ages under monaural earphone listening conditions. Please note that a similar graph exists for female persons, which shows a slightly lower hearing degradation over age (ISO 7029:2017).

It is important to note that even mild hearing loss, which is experienced by most individuals between the ages of 40 and 50, can have significant impacts on an individual's life. For instance, an individual with mild hearing loss may face challenges in following a group conversation in a noisy environment. Additionally, they may miss important auditory cues, such as warning signals or alarms.

Is the current audio hardware sufficient for future AI generations?

Now that we have a better understanding of how humans perceive audio signals, let's revisit the initial question regarding the audio input quality needed for current and future AI to perform at a level indistinguishable from humans today.

As is the case with most consumer devices currently on the market, audio signals are recorded using MEMS microphones. MEMS microphones are the primary audio capture technology for AI-powered personal assistants, which are now becoming available on the market.

The quality of the audio recordings produced by a MEMS microphone depends on its dynamic range. The upper limit of the dynamic range is defined by the acoustic overload point (AOP), which defines the microphone's distortion performance at high SPLs. The self-noise of the microphone limits its dynamic range at the lower end of the spectrum. The traditional measure of microphone self-noise is the signal-to-noise ratio (SNR), which defines the ratio between the self-noise of a microphone and the desired signal it captures. However, the SNR figure is somewhat misleading for the purposes of our discussion, as its definition implies how humans perceive an audio signal using A-weighting.

If the intended recipient of the recorded signal is an AI, a related microphone parameter, called equivalent noise level (ENL), is a more appropriate way to specify performance, as it ignores the human perception element of recorded sound. ENL refers to the signal produced by the microphone in the absence of an external sound source. ENL indicates the sound pressure level that will create the same voltage as the self-noise from the microphone.

The microphone's ENL across frequency can be considered the closest match to a microphone's hearing threshold. It should be noted that this is a highly simplified assumption, as there are usually numerous other components in the audio chain, including sound channels, additional water protection, and the audio processing chain.

Please refer to Figure 3 for a visual representation of the ENL curves of two MEMS microphones in comparison with the human hearing threshold.

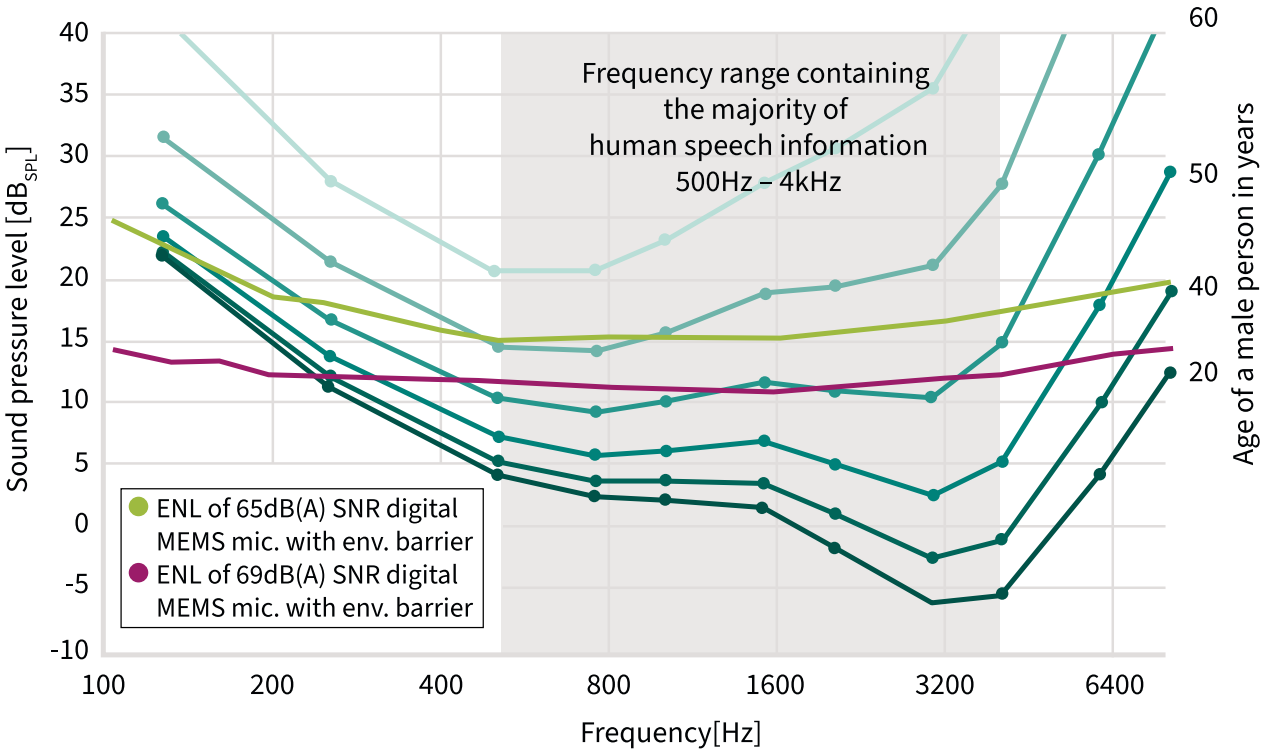

Figure 3: Acoustic 1/3 octave self-noise (ENL) of a mid-range and a high-end MEMS microphone compared to the hearing threshold of a typical male person.

The red line depicts the ENL curve of a 65 dB(A) SNR microphone with an integrated environmental barrier for dust protection. The corresponding MEMS microphone is currently used in several high-end smartphones from a variety of vendors.

The purple line below illustrates the ENL curve of Infineon's latest high-end digital microphone, which features an innovative environmental barrier that protects against particles and humidity. This microphone represents the current state of the art and was released only this year in a high-end tablet. We anticipate that microphones with comparable performance will be available in high-end smartphones by the end of the year. It is worth noting that reducing the microphone self-noise by 5 to 10 dB$_{SPL}$ represents a significant achievement, particularly when considering the logarithmic scale of sound pressure.

While Infineon has made notable progress in reducing the self-noise of high-end MEMS microphones, there is still a significant gap in the microphone's ability to discern low sound pressure levels when compared to the human ear. In particular, the 2 kHz range is of paramount importance for ensuring high levels of sound intelligibility for human listeners. There is a discrepancy of over 12 dB$_{SPL}$ between the capabilities of a young person and Infineon's state-of-the-art microphone. In comparison to the microphones currently used in high-end phones, there is a significantly higher discrepancy of 17 dBSPL. Once again, it should be noted that this assessment only considers the self-noise of the MEMS microphone and does not take into account the additional noise sources of the audio chain, which will further reduce the overall performance.

The current limitations of MEMS microphone technology are most apparent in the frequency range that contains the majority of human speech information (500–4 kHz). Even the most sophisticated MEMS microphones on the market are only capable of understanding sound at a level comparable to that of a 60-year-old person. Based on the available data, it is reasonable to expect that AI-based virtual assistants using the latest MEMS microphone technology will experience hearing impairments similar to those of elderly people, particularly in situations where they are required to follow conversations in noisy environments or from long distances.

Summary and outlook

The rapid progress in AI will not slow down, but rather accelerate the trend towards higher-SNR MEMS microphones. While the latest MEMS microphones cannot yet match the audio quality of the human ear, Infineon's progress in reducing self-noise benefits existing and future AI. Further improving the audio chain will be key to enhancing AI capabilities such as environmental classification, context understanding, emotional awareness, speaker identification, and multi-speaker diarization. With better audio input, AI will be able to interact with humans in a way that matches or even rivals the best of human behavior.

In addition, improved levels of human-machine interaction will enable new AI-based use cases and services. For example, imagine a future version of Microsoft's Copilot that not only summarizes a Teams meeting, but also provides an overall assessment of the mood of the conversation. Future AI might be able to highlight or rank the importance of the action items discussed, based solely on human speech and audio. There is also the potential to add AI-based coaching capabilities that give the user helpful advice on how to better steer future conversations in a desired direction.

Imagine AI-based first-round interviews with new job candidates, or the ability to identify speakers based on audio alone, with a level of security sufficient for online shopping.

All of this is likely just a small sample of what can be expected from future AIs with hearing capabilities that match or exceed those of humans. With our enhanced MEMS microphone solution, we are proud to be part of this exciting journey here at Infineon.

Published by
Infineon Technologies AG
Am Campeon 1-15, 85579 Neubiberg
Germany

Public

Version: V1.0_EN
Date: 10/2024

Stay connected!

Scan QR code and explore offering
www.infineon.com

Please note!
This Document is for information purposes only and any information given herein shall in no event be regarded as a warranty, guarantee or description of any functionality, conditions and/or quality of our products or any suitability for a particular purpose. With regard to the technical specifications of our products, we kindly ask you to refer to the relevant product data sheets provided by us. Our customers and their technical departments are required to evaluate the suitability of our products for the intended application.

We reserve the right to change this document and/or the information given herein at any time.

Additional information
For further information on technologies, our products, the application of our products, delivery terms and conditions and/or prices, please contact your nearest Infineon Technologies office (www.infineon.com).

Warnings
Due to technical requirements, our products may contain dangerous substances. For information on the types in question, please contact your nearest Infineon Technologies office.

Except as otherwise explicitly approved by us in a written document signed by authorized representatives of Infineon Technologies, our products may not be used in any life-endangering applications, including but not limited to medical, nuclear, military, life-critical or any other applications where a failure of the product or any consequences of the use thereof can result in personal injury.