Speaker Diarization: The Evolution of AI Understanding 'Who Spoke When?' with High SNR MEMS Microphones

by Dr. Gunar Lorenz, Infineon Technologies

Better audio input is essential for making AI suitable for everyday use. In noisy environments, AI can more easily recognize different speakers in a conversation if audio is recorded with high SNR (Signal-to-Noise Ratio) MEMS microphones. Improving audio input reduces misunderstandings and enhances the quality of translations. AI applications using high SNR MEMS microphones revolutionize how devices interact with the world by enabling advanced environment recognition, speaker recognition and sound classification.

As a parent of bilingual children, I have always been excited by the potential of AI to break down language barriers in both personal and professional settings. Seeing all the AI hype from the major phone manufacturers last year, I was eager to evaluate the latest advances in live translation during my subsequent business trip to Asia. Armed with the latest flagship smartphone, I tried to follow a Korean conversation and watched as the English text appeared on the screen with a slight delay.

I was impressed by the speed of the translations, however on occasion the meaning of the original text was not fully captured. The real challenge arose when additional people joined the Korean conversation, as the English translation appeared on the screen without any indication of who the speaker was. The multiple and sometimes overlapping speakers, in combination with the small delay between the spoken and the translated word, made it difficult to follow the conversation.
I tested several live translation apps, but unfortunately, none of them added IDs or labels to indicate who was speaking.

In AI terminology, understanding "who spoke when" is called speaker diarization. Without it, live translation feels like reading a Shakespeare play without knowing who's who – you can't really understand the meaning. Have a look at the text on the left:

What an AI gets without diarization:

Have not saints' lips, and holy palmers too?- Ay, pilgrim, lips that they must use in prayer. O, then, dear saint, let lips do what hands do; Saints do not move, though grant for prayers' sake. Then move not, while my prayer's effect I take. Thus from my lips, by yours, my sin is purged. Then have my lips the sin that they have took. Sin from thy lips? O trespass sweetly urged! Give me my sin again. You kiss by the book.

What an AI gets with diarization:

Speaker 1: Have not saints lips, and holy palmers too?
Speaker 2: Ay, pilgrim, lips that they must use in prayer.
Speaker 1: O, then, dear saint, let lips do what hands do;They pray, grant thou, lest faith turn to despair.
Speaker 2: Saints do not move, though grant for prayers' sake.
Speaker 1: Then move not, while my prayer's effect I take. Thus from my lips, by yours, my sin is purged.
Speaker 2: Then have my lips the sin that they have took.
Speaker 1: Sin from thy lips? O trespass sweetly urged!Give me my sin again.
Speaker 2: You kiss by the book.

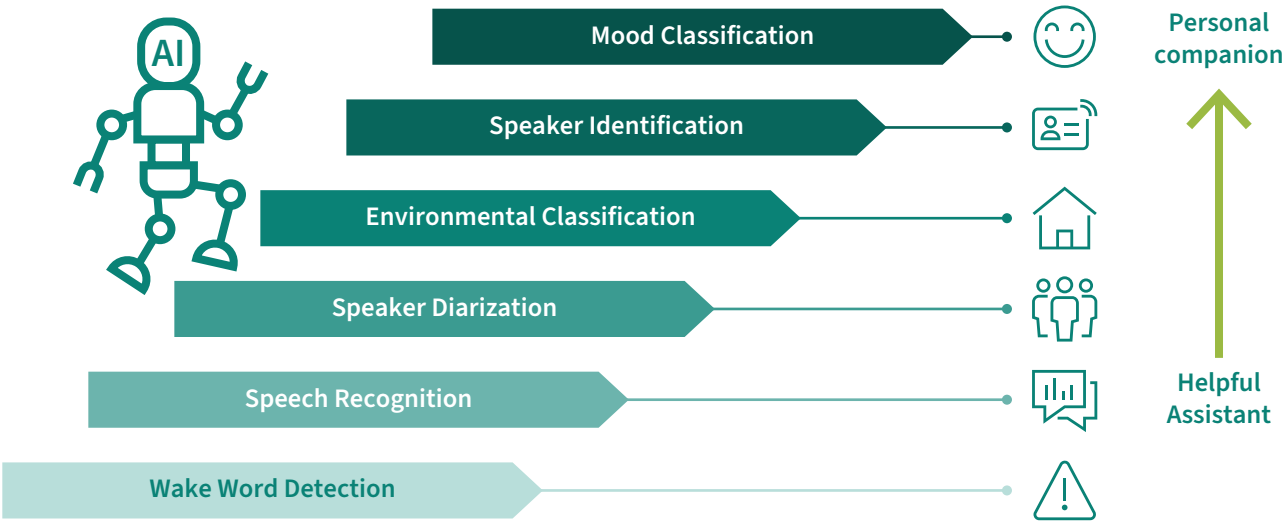
Shakespeare's original intend:

Romeo: Have not saints lips, and holy palmers too?
Juliet: Ay, pilgrim, lips that they must use in prayer.
Romeo: O, then, dear saint, let lips do what hands do;They pray, grant thou, lest faith turn to despair.
Juliet: Saints do not move, though grant for prayers' sake.
Romeo: Then move not, while my prayer's effect I take. Thus from my lips, by yours, my sin is purged.
Juliet: Then have my lips the sin that they have took.
Romeo: Sin from thy lips? O trespass sweetly urged!Give me my sin again.
Juliet: You kiss by the book.

That's pretty much what an AI would get from a smartphone recording of a beautifully acted Shakespeare play. With proper diarization (text in the middle), the AI gets a clear indication that the recorded audio contains a dialog involving two people. Based on the content plus the fact that two people are talking might reveal the romantic nature of the conversation.

What is still missing is, even with proper diarization, is the sex and the age of the two people involved.

We see AI's ability to interpret audio as a set of skills, with "Wake Word Detection" being the most basic AI ability to interpret input audio. If we were to visualise the AI's audio skills on a staircase, "Speaker Diarization" would be at level 3, just after "Speech Recognition":



Identifying Romeo and Juliet in the recorded dialogue would require speaker identification. Speaker identification would provide Shakespeare's original intention on the right-hand side of the dialog box on page 1. Even if Romeo and Juliet were correctly identified by the AI, all the romance and beauty of the actor's tone and performance would be lost. At least that's how it seemed to me at school. It wasn't until I saw the dialogue between Leonardo DiCaprio and Claire Danes in Peter Martin's adaptation of Romeo + Juliet that I began to understand the depth and beauty of the dialogue.

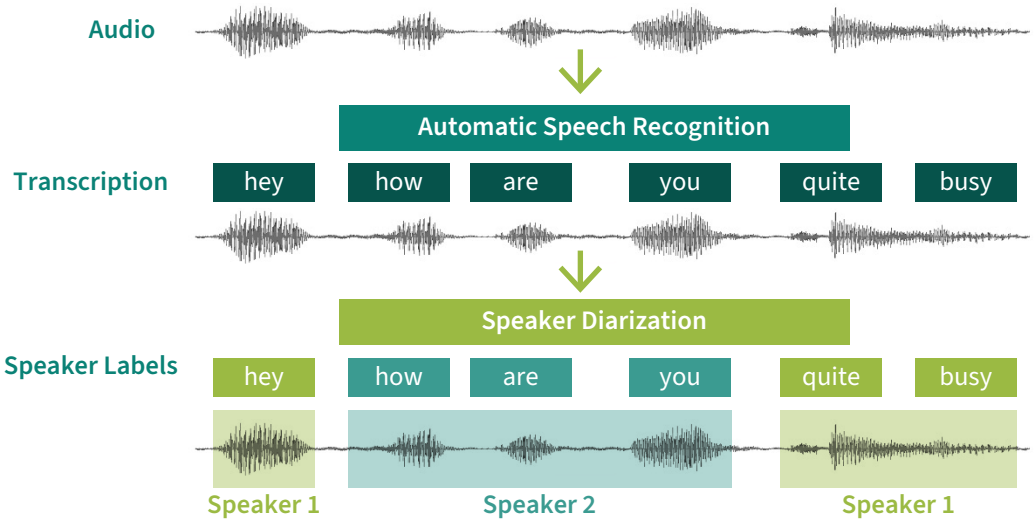
Research has shown that most communication is non-verbal, including tone of voice (38%) and body language (55%). In one of my recent articles argued that by improving AI's listening skills, we can turn machines into digital companions or colleagues. These machines will be able to understand conversations and create summaries of recorded interactions.

At Infineon, we believe that high-quality audio is key to a better user experience. As a leading supplier of high SNR MEMS, we're curious to know if the same applies to AI. We wanted to see how improving "listening" would affect the AI's ability to tell different speakers apart. A recent publication from Syntiant confirmed our internal measurements on the high SNR microphone benefits for wake word detection and voice command recognition.

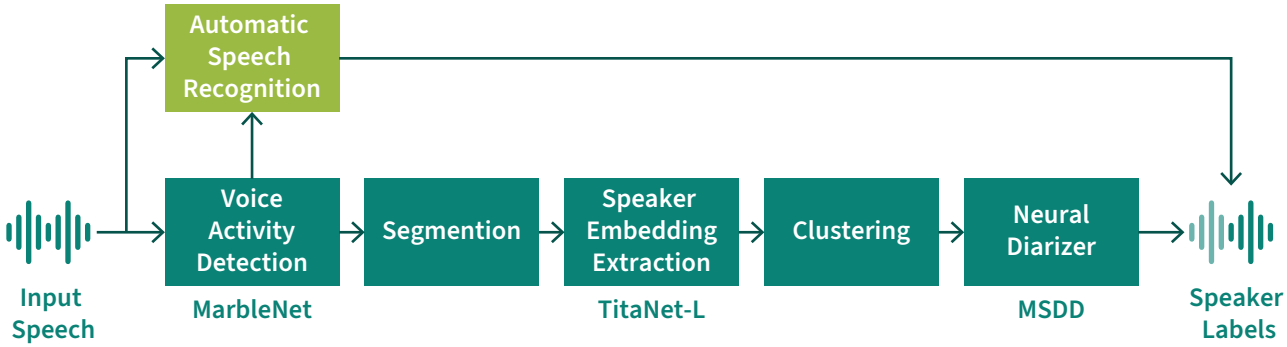
To get a better understanding of the latest AI capabilities in speaker diarization, we chose a publicly available AI tool that could measure the error in identifying speakers based on an audio file of a conversation. As with previous tests, it's challenging to keep up with the frequent releases and updates of new AI tools. At the end we decided to use NVIDIA's NeMo Framework. NeMo proved to be particularly useful in getting real measurement data to quantify what "listening better" means.

On the NVIDIA website it says: “Speaker diarization is the process of segmenting audio recordings by speaker labels and aims to answer the question “who spoke when?”. Speaker diarization makes a clear distinction when it is compared with speech recognition.

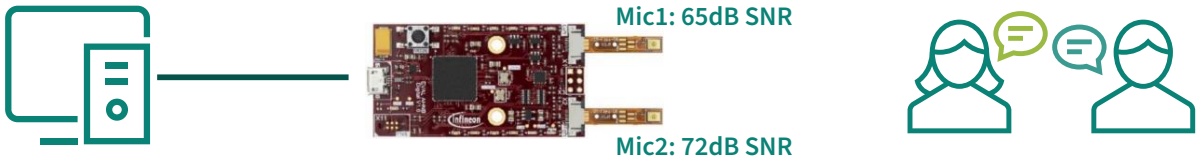
As shown in the figure below, before we perform speaker diarization, we know “what is spoken” yet we do not know “who spoke it”. Therefore, speaker diarization is an essential feature for a speech recognition system to enrich the transcription with speaker labels.



Live translation with speaker diarization requires multiple components in the audio processing chain starting with “voice activity detection” as shown in the block diagram from NVIDIA website:

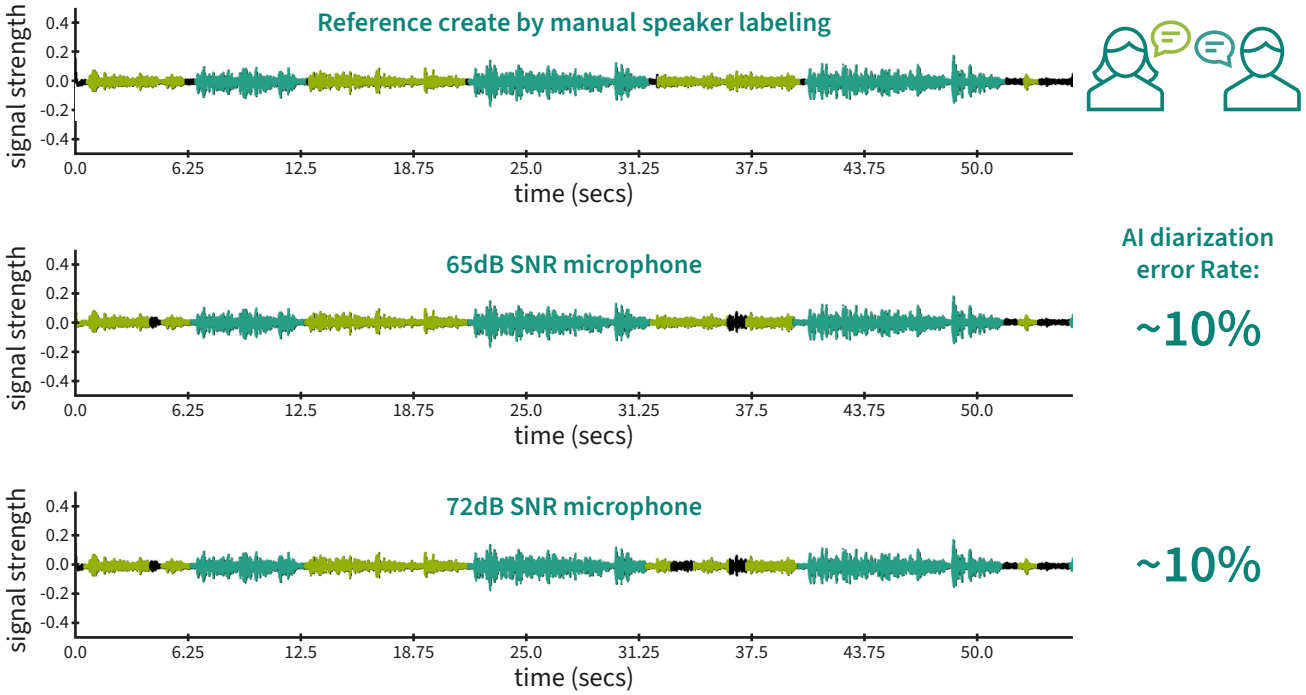


For our first test setup we used Infineon’s microphone evaluation board Audiohub Nano which allows to record audio with two different MEMS microphones at the same time.



The first microphone is roughly the same as one can find in most high-end mobile phones. It has a signal-to-noise ratio of 65dB(A). The microphones signal-to-noise-ratio (SNR), measures the microphone's ability to distinguish sounds from background noise, comparing the strength of the desired sound (e.g., a voice) to the microphones self-noise. The higher SNR the better the audio quality of the recorded signal. As a high-end reference for our test, we used our latest digital XENSIV™ microphone IM72D128V with an SNR of 72dB(A).

The recorded audio streams of the two microphones were analyzed with NVIDIAs speaker diarization system NeMo. The figure shows the audio signals of the recorded conversation.



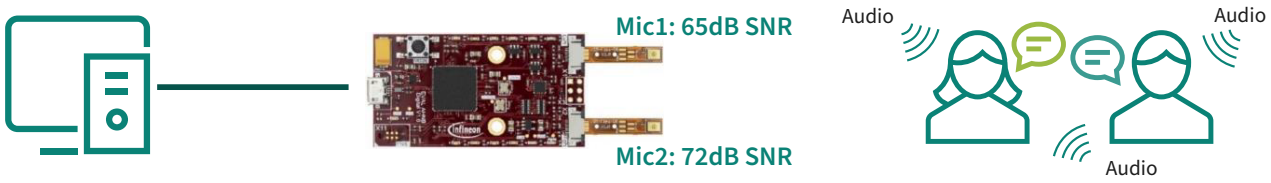
The two graphs show the reference with the audio sequence of the first speaker in green and the audio sequence of the second speaker in turquese. The reference was given to NeMo by manual labeling. Each time a specific button was pressed at each turn of the speaker.

The graph in the middle shows the NeMo's automatic speaker labeling using a standard smart phone microphone. Whereas the bottom graph shows NeMo's speaker labeling using the high SNR microphone.

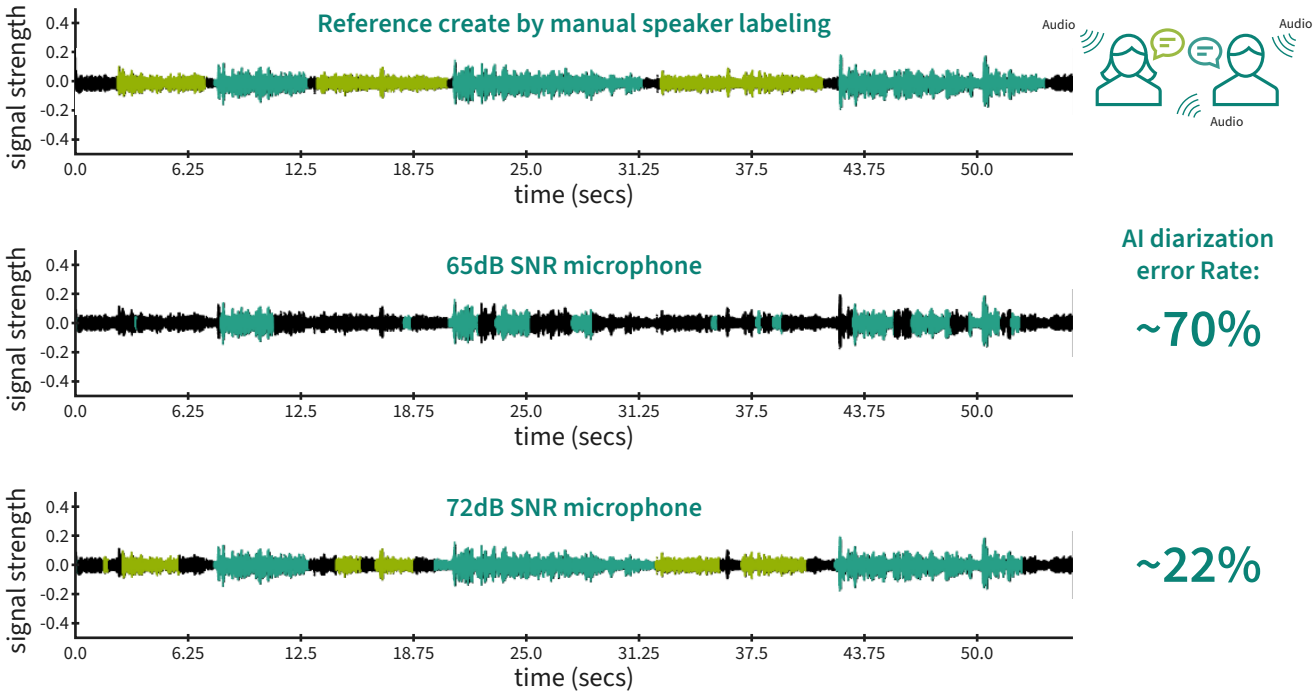
Both graphs show how the used AI identified correctly speaker 1 and 2 except for a few "black" sequences in the middle which could not be attribute to either or both speakers. For both the high and the low-end microphone recording NeMo calculated a fairly low diarization error of about 10%. Thus, regardless of the used microphone NeMo could successfully identify both speakers 90% of the time.

It should be noted that our two test speakers didn't make it particularly hard for NeMo by leaving a small little pause before starting to speak. Both never spoke at the same time.

For the second test we added background noise. Performing the same conversation in our very busy Infineon cafeteria.



The figure shows again three audio sequences. On the top once again the reference with the corresponding speaker sequence colored in green for speaker 1 and light green for speaker 2.



With background noise, NeMo struggled to even detect the second speaker when the recording was done with the medium SNR microphone. The diarization error rate showed up as around 70%.

Using a high SNR microphone in the same situation the diarization success rate was dramatically improved. What still seems to work fine in a quite environment, starts to become a challenge in the noisy environments.

The results made us think about elderly people with a mild hearing loss, which face challenges in following a group conversation in a noisy environment.

If we get older, our hearing gradually declines, often making it harder to hear high-pitched sounds or to understand speech in noisy environments. AI's using low SNR MEMS microphones seem to struggle with similar limitations.

When compared to human hearing, most currently used MEMS microphones perform at a level that's roughly comparable to the hearing abilities of a 60-year-old man, especially when it comes to detecting subtle details in speech. Learn more.

Summary and outlook

Despite the rapid progress in AI technology, there is still a long way to go for AI to truly “understand” human conversations. Mastering speech-to-text is only the first step; the next critical step towards human-like understanding of human conversations is mastering speaker diarization. Our own experiments have shown that better audio input will be crucial in making AI suitable for everyday use. High SNR (signal-to-noise ratio) audio input plays a significant role in this advancement, as a higher SNR means clearer audio, allowing AI to process speech more accurately. This is especially important in noisy environments, as the AI can more easily tell the difference between different speakers, reducing misunderstandings and improving the quality of translations. In the future, AI applications using MEMS/high SNR microphones will change how devices interact with the world by enabling advanced environment recognition, and sound classification.

Each innovation in MEMS microphone technology brings us closer to AI that can naturally listen, understand, and empathise, making it an integral part of daily life.

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

Public

Date: 03/2025

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.