Whitepaper


# Webex AI Codec

## The AI-powered codec for modern communications



webex by CISCO

# Contents

# Introduction

Delivering clear speech is essential for effective modern communication systems. This holds true for various scenarios, including videoconferencing meetings, one-to-one calls, and the deployment of voice assistants, translation services, or text-to-speech to enhance accessibility and inclusion.

However, factors such as background noise, reverberation, voice capture and sound reproduction quality, and network impairments can degrade speech quality and hinder understanding.

At Webex, we understand the pivotal role of clear speech in creating an inclusive future for all. That's why we are committed to providing exceptional, reliable, high-fidelity communication systems to our customers. Through significant investments in AI-based speech enhancement technologies like Background Noise Removal and Optimize for my/all voice(s), we have greatly enhanced the user experience on Webex.

Today, we introduced a groundbreaking technology: Webex AI Codec. This AI-based speech codec achieves outstanding quality at remarkably low bitrates and demonstrates robustness against network impairments. The result is a remarkably resilient communication system that enables noise-free, high-fidelity, and crystal-clear voice communication.

The future of speech communication has entered a new era, and Webex is at the forefront, pioneering this transformative journey.

# Fundamentals of modern digital voice communications

Modern digital voice communication systems rely on two essential building blocks: audio/speech codecs and the ability to mitigate packet loss.

## Audio/speech codecs

Audio/speech codecs are integral to digital communication systems. Transmitting a digital representation of audio/speech would require a large number of bits, rendering network transmission impractical. To address this, audio/speech codecs reduce the number of bits needed. The term “codec” combines “coder” and “decoder:” the coder (or encoder) compresses the audio waveform to a specific bitrate, while the decoder reconstructs the waveform at the receiving end. The objective of audio/speech coding is to compress an audio stream captured by a microphone within a given bitrate budget, allowing for reconstruction at the receiving end with audio quality as close as possible to the original.

In real-time communication systems, the captured audio is typically divided into frames, compressed by the codec, and packetized before transmission over a network connection. The successful delivery of these “audio packets” depends on the stability and reliability of the network connection – which cannot always be guaranteed.

## Mitigation of packet loss

Packet loss refers to the loss of packets due to network impairments. For instance, when the network connection is partially interrupted, the audio at the receiving end may experience “breaks”, impacting intelligibility - and therefore productivity.

In a voice communication system, mechanisms should be in place to handle such losses – examples include, but are not limited to: transmission of lost packets on demand, preventative transmission of redundant audio frames, and Packet Loss Concealment (PLC) techniques.

Each technique has its own pros and cons. For example, retransmission of lost packets is bandwidth efficient (good) at the cost of increased latency (bad). Conversely, duplicate transmission of packets is less bandwidth efficient (bad) but results in lower latency (good).

On the other hand, PLC aims at masking the audio that is not received while waiting for new packets to arrive, e.g., by replaying the audio from the last successfully received packets, replacing lost packets with silence, or reconstructing plausible “filler” audio based on typical speech patterns.

Regardless of the technique being used to mitigate the effects of packet losses, there are inherent trade-offs between bandwidth usage, latency, computational cost, and audio quality. Plus, in case of a significant number of lost packets, all the techniques mentioned above may fail, and a good portion of the conversation may be permanently lost.

If we were able to achieve excellent audio quality with fewer bits (thus increasing bandwidth efficiency), we could overcome the drawbacks of duplicate transmission or retransmission of packets (by transmitting more packets withing a given bitrate budget). Or, in cases with significant packet loss, a Generative AI-based PLC could (re)create the lost audio based on the context. As it turns out, AI-powered technologies for audio coding and mitigation of packet loss effects have the potential to solve most of the limitations of existing techniques. This is discussed in the next sections in more details.

## Limitations of existing technology

Speech codecs have played a pivotal role in the development of communication systems for several decades.

Audio coding is based on the concept of simplifying the audio stream while retaining virtually all information, reducing the required bandwidth for transmission while retaining sufficient audio quality for understanding speech or enjoying music. This process, known as "lossy audio compression," is akin to image or video compression techniques.

The field of audio/speech coding draws upon various disciplines, including information theory, psychoacoustics, neuroscience, and digital signal processing. These scientific advancements have become increasingly significant with the rise of digital technologies and the need to exchange digital media over limited bandwidth channels, particularly through the internet. Consequently, audio compression became imperative. This led to the development of audio coding technologies and formats like MP3, AAC, Vorbis, FLAC, Opus for music.

In real-time communication systems, customers expect to have clear video and audio regardless of their location, whether it be at home, in the office, or on the go. However, delivering high-quality media experiences requires a significant amount of bandwidth, which is a limited resource that needs to be utilized efficiently. This is why it is crucial to have audio and video codecs that can optimize the use of bandwidth by providing excellent quality with lower bitrates and minimal latency.

While traditional audio codecs have served us well over the years, they have limitations when it comes to effectively compressing audio content while maintaining excellent quality. Reducing the bitrate budget significantly impacts audio quality, rendering the reconstructed audio unintelligible or unpleasant to listen to. Several studies, such as those on Opus and EVS, support this notion [2], [3], [4].

To overcome these limitations, we have witnessed the emergence of AI-based audio codecs in recent years. Groundbreaking research by industry leaders such as Google [3], Amazon [5], Microsoft [6], and Meta [2] have demonstrated that AI-based coding can deliver exceptional quality even at extremely low bitrates, going as low as 1.6 to 6 kbps. This advancement is a game-changer for real-time voice communication.

However, a powerful low-bitrate codec alone is not enough to implement a robust voice communication system. To ensure outstanding audio quality and provide an exceptional user experience, the system must also address challenges such as noise removal, reverberation reduction, correction of microphone artifacts, and resilience against network impairments, among others. Furthermore, speech plays a vital role in various applications such as transcription, translation, and voice biometrics. As customers increasingly incorporate multiple technologies into their workflows, it becomes essential to take a comprehensive approach to handling speech-related tasks.

For instance, imagine a scenario where a customer contacts a customer service agent from a noisy environment. Currently, the customer would have to find a quiet location (to minimize background noise) with good network coverage. This common scenario involves compromises on both ends and does not provide a seamless experience.

Wouldn't it be great if there was a technology that could simultaneously:

1. Remove background noise at both ends of the communication.

2. Ensure flawless audio transmission on any network?

Having such a technology would greatly enhance the overall communication experience, eliminating the trade-offs and providing a seamless solution for customers and agents alike.

To address these requirements comprehensively, we have developed Webex AI Codec—a single, end-to-end solution for speech enhancement and highly resilient communications.

Webex AI Codec combines cutting-edge technologies to deliver remarkable audio quality and ensure reliable communication in various scenarios.

# Webex AI Codec: The AI-powered codec for modern communications

Webex AI Codec is a novel AI-based speech codec designed to address the challenges of real-time communication systems by providing high quality low bitrate audio compression, resilience to network impairments, and audio enhancements. By leveraging the excellent speech coding performance at low bitrates, the Webex AI Codec can be used to implement a communication system that is highly resilient to network impairments. The fundamental idea is to encode the current speech frame at a low bitrate (e.g., 6 kbps) as well as copies of the previous frames, coded at an even lower bitrates (e.g., 1 kbps).

As shown in Figure 1, Webex AI Codec comprises an encoder, vector quantizer (VQ) and audio decoder, which is trained end-to-end using thousands of hours of speech and background noises.

## Encoder

Our encoder is a causal convolutional network with zero algorithm latency. The audio encoder takes raw audio input and extract a feature vector embedding for every frame of audio. The encoder can also remove noise, reverberation and other artifacts from the input stream, Purified speech can enable denser encoding.

## Vector quantizer

Vector quantization (VQ) is a well-known quantization technique that has been widely used in various areas, including audio/video coding, clustering, and similarity-based vector searching. Residual VQ [6] also known as multi-stage VQ [7], employs multiple VQ layers, each of which takes the residual signal from the previous layer to further quantize in a sequential manner as shown in Figure 1. In the Webex AI Codec system, we use the Residual VQ to compress the embeddings even further before transmission, aiming to transmit content with minimal bit usage.

## Decoder

The decoder architecture follows a similar design as in the encoder, though in a 'mirroring' manner. The decoder reconstructs audio from the received speech vectors including compensating for lost audio frames.

## Webex AI Codec

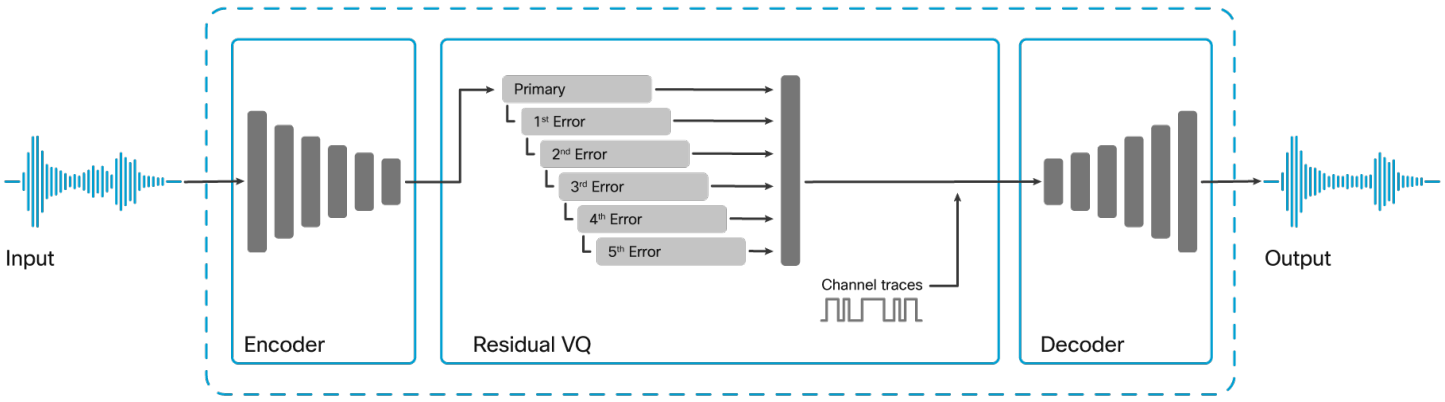


**Figure 1:** Webex AI Codec Core Technology

### Neural codec training

To train the neural codec system, we introduce various artifacts into clean speech signals, including background noise, reverberation, band limitation, packet losses, and more. We trained the codec on millions of hours of unique noisy speech composed from more than 10,000 hours of clean speech and noise samples. The large scale of training data ensures our model broad generalization across speech and artifacts. Our training process involves incorporating both generative and adversarial loss functions into the model. As a result of this training, the audio encoder takes raw audio input and leverages a deep neural network to extract a comprehensive set of features that encapsulate intricate speech and background noise characteristics jointly or separately. The extracted speech features represent both the speech semantic as well as speech stationary attributes such as volume, pitch modulation, accent nuances, and more.

This represents a departure from conventional audio codecs that rely on manually designed features, as the neural encoder learns and refines its feature extraction process from extensive and diverse datasets, resulting in a more versatile and generalized representation.

## Performance validation

At Cisco, we have developed a crowd-sourced MUSHRA (Multi Stimuli with Hidden Reference and Anchor) listening test framework using Qualtrics, an online survey platform, and Amazon MTurk. This tool allows us to conduct MUSHRA tests more frequently with the desired number of participants. However, this approach can sometimes lead to inconsistent results. To address these discrepancies, we've implemented both real-time and post-test screening processes.

This thorough screening process is essential to ensure that we obtain consistent and reproducible results, especially when conducting multiple test iterations.

To achieve this, we have developed a two-step screening process for qualifying listeners and their scores, as well as validating those scores.

When designing a listening test in Qualtrics, specific rules are imposed, which may disqualify listeners and exclude their scores from our post-screening step.

### Testing methodology

In the first step, listeners are subjected to a hearing ability test. If they pass this test, they proceed to a 'training test,' which provides an example similar to what they will encounter in the actual listening test. During the listening test, we closely monitor their scores, and they may be disqualified if their answers are inconsistent.

### Test results

Figure 2 shows the results of the listening tests comparing Webex AI Codec operating at 1 kbps and 6 kbps with Opus 16 kbps. Most noticeably, Webex AI Codec at 6 kbps outperforms Opus at 16 kbps. Webex AI Codec at 1 kbps provides excellent quality.

Note that, results are combined by normalizing all scores to common "anchor" and "hidden reference" scores.

## Optimal quality audio at a fraction of the bandwidth

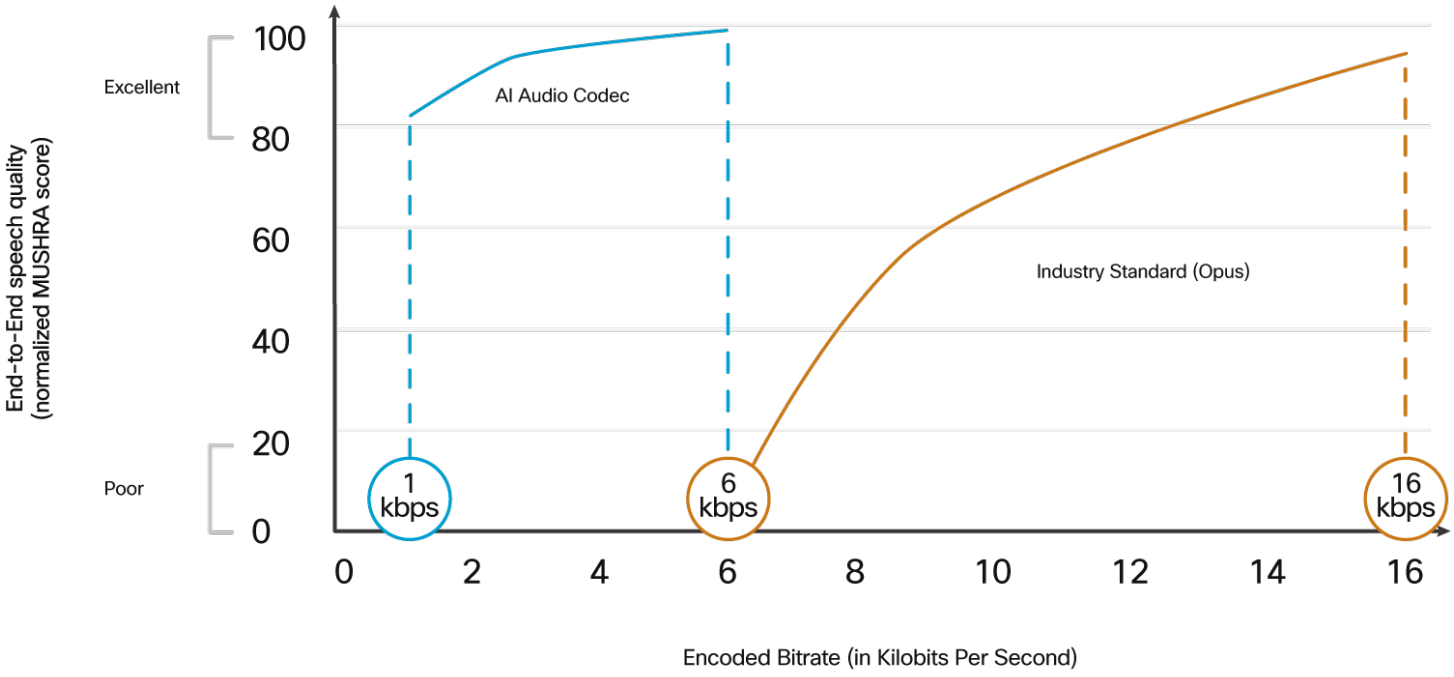


**Figure 2:** Webex AI Codec performance

# 94%

**Less bandwidth usage than industry standard codec at comparable call quality**

# >80 points

**Webex AI Codec MUSHRA score at 1, 3, and 6 kbps.**

# Webex AI Codec impact / applications across the Webex Platform

Webex AI Codec's versatility make it well-suited for all Webex use cases, including the Webex Suite (Meetings, Calling, and Webinars) and Webex Contact Center. It seamlessly integrates with various Webex-enabled endpoints such as the Webex App, Room Systems, IP Phones, and Desk Series devices, among others.

**Table 1.** Features and benefits

| WEBEX PRODUCT | WEBEX AI CODEC APPLICATIONS |
|---|---|
| **Webex Calling** | • Background noise removal<br>• Optimize for my / all voices<br>• Noise removal and HD Voice for external callers |
| **Webex Meetings** | • Background noise removal<br>• Optimize for my / all voices<br>• Improved audio quality in low bandwidth<br>• Increased accuracy of captions and translations |
| **Webex Contact Center** | • Background noise removal<br>• Real-time transcription<br>• Conversation summaries<br>• Optimize for my voice |
| **Cisco IP Phones** | • Background Noise removal<br>• Optimize for my voice<br>• Noise removal and HD Voice for external callers |

**Table 1.** Features and benefits

| WEBEX PRODUCT | WEBEX AI CODEC APPLICATIONS |
|---|---|
| **Cisco Room Series** | • Noise removal |
| **Desk Devices** | • Background noise removal<br>• Optimize for my voice<br>• Noise removal and HD Voice for external callers |

Moreover, the benefits of Webex AI Codec extend beyond Webex. Other applications like Text-to-Speech (TTS), Automatic Speech Recognition (ASR), language translation, and voice biometrics can also leverage its capabilities. The Compact Speech Representation, which is the speech format generated by the Webex AI Codec encoder, can be easily integrated with other systems. This enables low-latency and highly accurate interaction between services that rely on speech information, facilitating efficient and seamless communication between different platforms.

# Cisco's commitment to responsible AI

As an AI-based innovation, Webex AI Codec adheres to Cisco's Responsible AI principles, which include Transparency, Fairness, Accountability, Privacy, Security, and Reliability. These principles guide every stage of the Webex AI Codec product lifecycle, from requirements gathering and model evaluation to deployment and documentation.

**Table 1.** Cisco's Responsible AI Principles

| CISCO'S RESPONSIBLE AI PRINCIPLES | |
|---|---|
| **Transparency** | AI relies on large datasets and advanced algorithms. Often, it's not clear to users when and how AI is involved in decision-making. As transparency is one of our Trust Principles and core to this framework, we inform customers when AI is being used to make decisions that affect them in material and consequential ways. Customers and users can then inform us of their concerns or let us know when they disagree with decisions. By keeping communications channels open, we intend to build, maintain, and grow the trust that our customers, users, employees, and other stakeholders place in our AI offerings.<br><br>*Cisco's goal is to provide clarity and consistency in informing users when AI is employed in our technologies; the intent of the AI; the model class; the data demographics; and the security, privacy, and human rights controls applied to the model in a manner that is accessible, transparent, and understandable. We also share how to get more information about our use of AI.* |
| **Fairness** | AI creates the potential for harmful human bias to become ingrained or amplified by technological systems. At the same time, it presents an opportunity to better understand and mitigate harmful bias and discriminatory results in decision making and to create technology that promotes inclusion. Achieving better decisions requires assurance that the training data represents the demographics of individuals or groups across the full spectrum of diversity to which AI will be applied.<br><br>*Cisco strives to identify and remediate any harmful bias within our algorithms, training data, and applications that are directly involved in consequential decisions; that is, decisions that could have a legal or human rights impact on individuals or groups. As an integral component of our responsible AI framework, we have also developed mechanisms for our customers to provide feedback and raise any concerns for review and action by our Incident Response Team. We regularly update these practices to reflect the latest technological advancements, including those in AI.* |
| **Accountability** | Accountability for AI solutions and the teams that develop them is essential to responsible development and operations throughout the AI lifecycle. AI tools often have more than one application, including unintended use cases and uses that might not have been foreseeable at the time of development. Companies that develop, deploy, and use AI solutions must take responsibility for their work in this area by implementing appropriate governance and controls to ensure that their AI solutions operate as intended and to help prevent inappropriate use.<br><br>*Cisco Public 3 Cisco is committed to upholding and respecting the human rights of all people, as articulated in our Global Human Rights Policy. The Cisco Responsible AI Framework requires teams to account for privacy, security, and human rights impacts from the very beginning of development through the end of the AI lifecycle. Accountability measures include requiring documentation of AI use cases, conducting impact assessments, and oversight provided by a group of cross-functional leaders.* |

**Table 1.** Cisco's Responsible AI Principles

| CISCO'S RESPONSIBLE AI PRINCIPLES | |
|---|---|
| **Privacy** | Applications of AI often use personal data that could impact individual privacy and civil liberties if not managed properly. When AI uses personal data or makes decisions for or about a person, privacy controls must be designed into the supporting technology to assure that personal data usage is permitted, purpose-aligned, proportional, and fair. Those controls must be maintained throughout the data and solution's lifecycle.<br><br>*Cisco has built privacy engineering practices into the Cisco Secure Development Lifecycle (CSDL). These practices help ensure that we design, build, and operate privacy-enhancing features, functionality, and processes into our product and service offerings. When processing personal information, Cisco is committed to following the principles set forth in our Global Personal Data Protection and Privacy Policy, which aligns with applicable international privacy laws and standards.* |
| **Security** | AI systems must be resilient and protected from malicious actors using similar secure development lifecycle controls as standard software development. Protection against security threats includes testing the resilience of AI systems for conventional as well as adversarial machine-learning attacks; sharing information about vulnerabilities and cyber-attacks; and protecting the privacy, integrity, and confidentiality of personal data.<br><br>*Cisco builds AI technologies using leading security practices, drawing on our secure development lifecycle to maximize resilience and trustworthiness. To meet the unique characteristics of AI, Cisco has added specific security controls for AI that improve attack resiliency, data protection, privacy, threat modeling, monitoring, and third-party compliance* |
| **Reliability** | Across a range of AI applications, the efficacy of AI solutions is measured by how reliably that solution produces a desired output based on the data set on which it has been trained, and the data from which it continuously learns. One of the key offerings of AI solutions is increased accuracy, which can only be achieved if the solutions are systematically tested for and engineered to produce replicable results.<br><br>*Cisco prioritizes innovation, and we design and test AI systems and their components for reliability. As part of our responsible AI assessment, we review AI-based solutions for embedding controls in their lifecycle to maintain consistency of purpose and intent when operating in varying conditions and use cases. Where we identify that an AI solution has potential impacts on safety, we impose additional integrity controls.* |

To ensure transparency and accountability, we plan on publishing more details regarding the Responsible AI assessment of Webex AI Codec as it is deployed across the range of Webex products. This will provide a deeper understanding of the technology and its adherence to responsible AI practices, fostering trust and confidence among our users.

# Conclusion

The primary objective of a communication system is to ensure clear speech by addressing issues such as noise, reverberation, microphone/headset artifacts, and providing robustness against network impairments.

To tackle these challenges, Cisco has developed Webex AI Codec, an AI-based speech codec that integrates noise removal and other speech enhancement capabilities. Webex AI Codec achieves near-transparent quality at bitrates as low as of 1 kbps. This allows for the implementation of highly resilient communication systems.

Webex AI Codec incorporates a unique approach where encoded packets include speech from the current frame as well as copies of previous frames, coded at a lower bitrate. This design enables the reconstruction of audio even in the presence of significant packet loss, minimizing any loss of information during meetings, calls, and other interactions.

Currently, Webex AI Codec is undergoing trials in Webex Meetings with multiple participants and selected types of Webex Calling calls.

With this innovative solution, Cisco leverages the power of AI to address fundamental challenges in speech communication, delivering reliable and high-quality technology that facilitates collaboration. Moreover, Cisco ensures that any AI innovation, including Webex AI Codec, adheres to the principles of Responsible AI.

Webex AI Codec represents the latest advancement in the quest to power an inclusive future for all, providing a vital building block to high-quality communications in diverse settings.

**References:**


1. A. Rämö and H. Toukomaa, “Subjective quality evaluation of the 3GPP EVS codec,” 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 2015, pp. 5157-5161, doi: 10.1109/ICASSP.2015.7178954.
2. J.-M. Valin, U. Isik, P. Smaragdis, A. Krishnaswamy, “Neural speech synthesis on a shoestring: Improving the efficiency of LPCNET,” Preprint, February 2022
3. J. Dani and S. Srinivasan, “Satin: Microsoft’s latest AI-powered audio codec for real-time communications,” Microsoft Teams Blog webpage, February 2021, URL: https://techcommunity.microsoft.com/t5/microsoft-teams-blog/satin-microsoft-s-latest-ai-powered-audio-codec-for-real-time/ba-p/2141382
4. N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, M. Tagliasacchi, “SoundStream: An End-to-End Neural Audio Codec,” Preprint, 2021, https://doi.org/10.48550/arXiv.2107.03312
5. A. Défossez, J. Copet, G. Synnaeve, Y. Adi, “High Fidelity Neural Audio Compression,” Preprint, October 2022, https://doi.org/10.48550/arXiv.2210.13438
6. A. Vasuki and P. Vanathi, “A review of vector quantization techniques,” IEEE Potentials, July 2006, doi: 10.1109/MP.2006.1664069
7. R. Gray, “Vector quantization,” in IEEE ASSP Magazine, vol. 1, no. 2, pp. 4-29, April 1984, doi: 10.1109/MASSP.1984.1162229


**For more information**
Please visit webex.com/products/collaboration-ai.html

January 2024