Global Research Trends In Acoustics 2019 2025

Global Research Trends in Acoustics (2019–2025)

April 30, 2025

Global Research Trends in Acoustics (2019–2025)

Introduction

Over the last five years, acoustics research has seen remarkable advances across hardware, algorithms, and applications. This report surveys key global trends from 2019–2025 in audible-frequency acoustics, focusing on: (1) transducer technologies (microphones and speakers); (2) applications of AI and machine learning in audio and speech processing; (3) speech processing advancements; (4) emerging quantum technologies for audio; and (5) other notable topics (excluding underwater, sonar, and medical acoustics). We highlight significant academic and industrial research outputs, identify leading conferences/journals and research groups (with tables provided), analyze geographic and collaborative trends, and conclude with a forward-looking outlook for 2025–2030. All findings are drawn from recent proceedings (ICASSP, Interspeech, AES, etc.), top journals (e.g., JASA, IEEE/ACM TASLP), and industry-academia initiatives.

Transducer Technologies: Microphones and Speakers (2019–2025)

- *MEMS Microphones: **Micro-Electro-Mechanical Systems (MEMS) microphones have become ubiquitous in consumer devices and continue to improve in performance. Commercial MEMS mics today achieve high signal-to-noise ratio (SNR) and wide dynamic range previously only seen in conventional electret mics. For example, Infineon’s 2022 XENSIV MEMS microphone achieved** 73 dB SNR **and** 135 dB** acoustic overload, matching studio-grade electret performance (IM73A135 | High performance

- XENSIV™ MEMS microphone with ultra high dynamic range - Infineon Technologies). Capacitive (electrostatic) MEMS mics still dominate due to CMOS-compatible fabrication and proven reliability ([

Recent Trends in Structures and Interfaces of MEMS Transducers for Audio Applications: A Review - PMC

](https://pmc.ncbi.nlm.nih.gov/articles/PMC10146864/#:~:text=Although%20MEMS%20microphones%20can%20be,or%20piezoelectric%20sensing%20solutions)). However, **piezoelectric MEMS microphones** are gaining interest for their simpler biasing (no DC bias needed) and robustness. Innovations in piezo materials, such as scandium-doped aluminum nitride (ScAlN), have enhanced piezoelectric coefficients, enabling higher sensitivity transducers. Research prototypes using ScAlN piezo MEMS structures demonstrated improved sensitivity and broad frequency response in the audible range (text{0.9}$N-Based Bimorph Piezoelectric MEMS Microphones With ...) (Electro-Acoustic Properties of Scandium-Doped Aluminum Nitride ...). MEMS arrays with multiple miniaturized mics are increasingly used for beamforming and spatial audio capture in smart speakers, AR/VR headsets, and hearing aids. Ongoing challenges include reducing self-noise and improving low-frequency response, as well as packaging techniques to reduce chip-level stress that can impact microphone sensitivity ([

Micro-Electro-Mechanical Systems Microphones: A Brief Review Emphasizing Recent Advances in Audible Spectrum Applications - PMC

](https://pmc.ncbi.nlm.nih.gov/articles/PMC10972232/#:~:text=The%20MEMS%20microphone%20is%20a,a%20comprehensive%20and%20adequate%20analysis)) ([

Micro-Electro-Mechanical Systems Microphones: A Brief Review Emphasizing Recent Advances in Audible Spectrum Applications - PMC

](https://pmc.ncbi.nlm.nih.gov/articles/PMC10972232/#:~:text=The%20microphone%2C%20a%20device%20capable,has%20expanded%20into%20various%20domains)). Recent reviews emphasize **noise reduction and directional designs** for MEMS mics, aiming for environmental noise

cancellation at the sensor level ([

Micro-Electro-Mechanical Systems Microphones: A Brief Review Emphasizing Recent Advances in Audible Spectrum Applications - PMC

](https://pmc.ncbi.nlm.nih.gov/articles/PMC10972232/#:~:text=The%20MEMS%20microphone%20is%20a,a%20comprehensive%20and%20adequate%20analysis)). Researchers are also exploring **optical and spintronic sensing** in MEMS mics: Optical MEMS microphones use a laser/photodiode to detect diaphragm vibrations, achieving very high SNR and immunity to electromagnetic interference (at the cost of complexity and power) ([

Recent Trends in Structures and Interfaces of MEMS Transducers for Audio Applications: A Review - PMC

](https://pmc.ncbi.nlm.nih.gov/articles/PMC10146864/#:~:text=instead%2C%20detect%20deflections%20induced%20by,very%20expensive%20due%20to%20the)). **Spintronic microphones** employ magnetic strain gauges on the diaphragm to dramatically boost sensitivity (gauge factors ~5000) ([

Recent Trends in Structures and Interfaces of MEMS Transducers for Audio Applications: A Review - PMC

](https://pmc.ncbi.nlm.nih.gov/articles/PMC10146864/#:~:text=A,18%E2%80%9322%20June)) ([

Recent Trends in Structures and Interfaces of MEMS Transducers for Audio Applications: A Review - PMC

](https://pmc.ncbi.nlm.nih.gov/articles/PMC10146864/#:~:text=MEMS%20microphones%20may%20rely%20on,an%20optical%20sensor%2C%20typically%20a)). These novel mechanisms remain in early research stages ([

Recent Trends in Structures and Interfaces of MEMS Transducers for Audio Applications: A Review - PMC

](https://pmc.ncbi.nlm.nih.gov/articles/PMC10146864/#:~:text=Although%20thermoacoustic%20actuation%2C%20spintronics%2C%20and,Art%20in%20Section%203)), but promise ultra-sensitive detection of sound

vibrations. In summary, 2019–2025 saw MEMS mic technology maturewith higher performance and robustness, while exploratory researchopened new transduction paradigms for the future.

- *Speaker Technologies: **Loudspeaker research has likewise pushed towards miniaturization and efficiency.** MEMS micro-speakers** have emerged as a significant trend, driven by the need for ultra-thin audio components in smartphones, earbuds, and wearables ([

Recent Trends in Structures and Interfaces of MEMS Transducers for Audio Applications: A Review - PMC

](https://pmc.ncbi.nlm.nih.gov/articles/PMC10146864/#:~:text=On%20the%20other%20hand%2C%20MEMS,acoustic%20power%20for%20a%20given)). Unlike traditional voice-coil speakers, MEMS speakers use micro-fabricated actuators (often piezoelectric or electrostatic) on silicon. In the early 2020s, several startups and projects demonstrated MEMS speakers delivering surprising output for their size. The market for MEMS speakers, though much smaller than for MEMS mics, is *“expected to be even booming”* with a **77% CAGR (2020–2026)** as these devices enter consumer products ([

Recent Trends in Structures and Interfaces of MEMS Transducers for Audio Applications: A Review - PMC

](https://pmc.ncbi.nlm.nih.gov/articles/PMC10146864/#:~:text=MEMS%20microphones%20and%20speakers%20market,forecast)). Notably, xMEMS Labs (USA) and USound (EU) introduced solid-state silicon speakers for in-ear headphones, offering high fidelity in a chip-scale package. Academic research has explored both **electromagnetic microspeakers** (with integrated coils/magnets) and **piezoelectric thin-film speakers**. A key innovation from MIT in 2022 was a **paper-thin flexible loudspeaker** that can turn any surface into a speaker using a thin piezoelectric film with micro-dome structures (Clip New Scientist Paper-thin speaker can play Queen from any surface (May 31) | MIT News | Massachusetts Institute of Technology). This film speaker operates with a fraction of the energy of a conventional speaker and produces sound with minimal

distortion, pointing to future ultra-light, conformal audio surfaces (Clip New Scientist Paper-thin speaker can play Queen from any surface (May 31) | MIT News | Massachusetts Institute of Technology). Another frontier is **thermoacoustic speakers**, which produce sound by rapidly heating and cooling a medium (often using carbon nanotube or graphene films). Thermoacoustic MEMS speakers have no moving parts and can be extremely thin; researchers see promise in them for high-frequency or niche applications ([

Recent Trends in Structures and Interfaces of MEMS Transducers for Audio Applications: A Review - PMC

](https://pmc.ncbi.nlm.nih.gov/articles/PMC10146864/#:~:text=match%20at%20L510%20Although%20thermoacoustic,reason%2C%20only%20electromagnetic%2C%20electrostatic%20and)), but efficiency and power requirements remain challenges. In summary, speaker technology is trending toward flat, thin, and efficient designs. **Acoustic metamaterials** have been applied to enhance loudspeaker output at low frequencies or to create directional sound output. For instance, engineers have designed metamaterial structures that augment bass response from small drivers or focus sound beams. A notable 2019 demonstration was a ring-shaped acoustic metamaterial "mute" that blocks 94% of incident sound while allowing airflow (BU Engineers Develop New Acoustic Metamaterial and Noise Cancellation Device | The Brink | Boston University) – a concept that could influence noise-cancelling enclosures and quiet ventilated loudspeakers.

(BU Engineers Develop New Acoustic Metamaterial and Noise Cancellation Device | The Brink | Boston University) *Figure: Boston University researchers (2019) with 3D-printed acoustic metamaterial rings that block ~94% of sound while remaining open – an example of using engineered structures to control audible noise (BU Engineers Develop New Acoustic Metamaterial and Noise Cancellation Device | The Brink | Boston University).*

- *Piezoelectric Innovations: **Piezoelectric transducers underpin many of the above developments (both microphones and**

- **microspeakers). Beyond MEMS, there have been materials breakthroughs in the past five years. Researchers are moving toward** lead-free piezoelectric materials **(due to RoHS and environmental concerns about lead-based PZT). Scandium-doped AlN, as mentioned, and other relaxor ferroelectrics are leading candidates, offering high piezoelectric constants in thin-film form (text{0.9}$N-Based Bimorph Piezoelectric MEMS Microphones With ...) (Electro-Acoustic Properties of Scandium-Doped Aluminum Nitride ...). In loudspeakers, traditional piezo buzzers have evolved into high-quality audio devices. Improved piezoelectric polymers and composites have enabled flexible actuators for wearable sound (e.g., vibration transducers in haptic audio vests). Piezoelectric micro-actuators also play a role in** haptics and ultrasonic audio** (like ultrasonic parametric speakers that generate directional audible sound). While piezoelectric transducers do not yet rival moving-coil speakers in audio output for large spaces, their efficiency and size advantages are driving new use cases, and incremental research continues to improve their acoustic quality.

Overall, transducer technology research from 2019–2025 is characterized by **miniaturization**, **new materials**, and **multi-disciplinary approaches** (mixing photonics, magnetics, metamaterials) to achieve better sound capture and reproduction. Academic groups (e.g., at University of Illinois, CEA-Leti, Tsinghua, MIT, and others) and companies (Infineon, Knowles, xMEMS, etc.) have collaboratively advanced the state of the art, bridging fundamental R&D; and product deployment.

AI and Machine Learning in Audio & Speech Processing

The past five years saw explosive growth in applying AI and deep learning to audio and speech processing. Machine learning techniques have transformed classical audio tasks, achieving unprecedented performance

in **denoising, source separation, and speech synthesis**, and enabling capabilities like **low-resource speech recognition** that were previously out of reach. Key trends include end-to-end deep learning models, self-supervised learning for data efficiency, and the integration of large-scale neural networks into real-time audio applications.

- *Audio Denoising and Speech Enhancement: **Deep learning became the dominant approach for noise reduction in audio. Traditional single-channel noise suppression (which relied on statistical noise models and spectral subtraction) has been supplanted by neural network models that learn to directly map noisy audio to clean speech. A notable milestone was the Microsoft** Deep Noise Suppression (DNS) Challenge** introduced at INTERSPEECH 2020, which open-sourced large training datasets and evaluation frameworks (The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Testing Framework, and Challenge Results - Microsoft Research) (The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Testing Framework, and Challenge Results - Microsoft Research). This catalyzed research into real-time speech enhancement. One outcome is that neural denoisers can now handle highly non-stationary noises that classical methods failed to suppress. As an MSR paper noted: *"An exciting property of deep learning–based noise suppression is that it reduces highly non-stationary noise – barking dogs, crying babies, traffic – which was not possible with earlier single-channel methods"* (Data Augmentation and Loss Normalization for Deep Noise Suppression). Modern denoisers often use recurrent or convolutional networks operating in the time-frequency domain (or even time-domain). By 2021, models like DeepMMSE, spectral gating DNNs, and generative models produced very natural results, with some approaches focusing on minimizing speech distortion (using perceptual loss functions and GANs to make speech sound clean to human listeners). These advances quickly made their way into products: popular conferencing apps and hearing aids in 2022+ began including AI-driven noise suppression that significantly improves call quality in real-world environments (suppressing keyboard clicks, sirens, etc.).

- Another development is **joint echo cancellation and noise suppression** – leveraging ML to handle echoes and noise in full-duplex communication. Challenges remain in generalization (ensuring a model trained on one noise dataset works on arbitrary real noise) and in **ultra-low-power** deployment (running complex models on tiny DSPs). Nonetheless, the collaboration between academia and industry (e.g., Microsoft's DNS challenges and ITU P.808 crowdsourced test standard ([The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Testing Framework, and Challenge Results - Microsoft Research](#)) ([The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Testing Framework, and Challenge Results - Microsoft Research](#))) has pushed denoising tech forward significantly.

- *Audio Source Separation: **The classic "cocktail party problem" of isolating individual sound sources from a mixture has seen groundbreaking progress. Deep learning methods now routinely** outperform traditional signal processing **for source separation (Recent Advances in Audio Source Separation | Frontiers Research Topic). Early in this period, researchers introduced** Conv-TasNet (2018) **– a convolutional time-domain network that for the first time surpassed the performance of ideal time-frequency masks in speech separation (Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude ... - arXiv). This was a paradigm shift: focusing on time-domain learned representations instead of spectrogram masking. Subsequently, a wave of architectures (TasNet variants, dual-path RNNs, Transformers) achieved ever-lower signal distortion and higher SI-SDR (scale-invariant signal-to-distortion ratio) in blind source separation. By 2020, deep models could separate** multiple simultaneous speakers **from single-channel audio with surprising accuracy. For example, an end-to-end separation model could demix two speakers such that automatic speech recognition (ASR) on the separated streams had under 20% word error – a feat impossible pre-DL. In music, tools like Open-Unmix, Spleeter (by Deezer, 2019),**

- **and Demucs (Facebook, 2020) brought source separation to music production, allowing splitting music into vocals, drums, bass, etc., for remixing or karaoke. These were made available as open-source libraries, demonstrating the maturity of the technology. There were also community challenges, such as the** Music Demixing Challenge 2021 **(Recent Advances in Audio Source Separation | Frontiers Research Topic), which fostered competition in music source separation. Another branch,** audio-visual separation**, leverages video cues (e.g., seeing a musician's gestures) to assist in isolating sources; this interdisciplinary approach gained traction for scenarios like separating overlapping speech by tracking who is speaking on video. In summary, by 2025 deep source separation is a robust research field with real-world applications in meeting transcription (separating speakers), content creation, and even hearing assistive devices. Current research addresses separating more than two sources, dealing with reverberation, and integrating separation directly into speech recognition systems (so-called joint separation-recognition).

- *Speech Synthesis (Text-to-Speech) and Voice Conversion: **Neural network–based speech synthesis has achieved near-human naturalness over the last five years. The era of robotic, monotonic computer speech is over, replaced by fluid and expressive synthetic voices. Building on breakthroughs like WaveNet (2016) and Tacotron (2017), researchers focused on improving quality, speed, and controllability of TTS. By 2019–2020, models such as** Tacotron 2 + WaveGlow/WaveRNN**, FastSpeech, and DeepMind's WaveNet variants were being deployed in cloud TTS services, delivering natural voices for assistants like Alexa, Siri, and Google Assistant. Key advances in this period include: *Non-autoregressive TTS* (e.g., FastSpeech 2 in 2020) for fast, parallel voice generation without compromising quality ([[PDF] fastspeech 2: fast and high-quality end-to- end text to speech - arXiv](https://arxiv.org/pdf/2006.04558#:~:text=,2s%20enjoys%20even%20faster%20inference)); *Neural vocoders* (WaveGlow, LPCNet, HiFi-GAN 2020) that

- generate high-fidelity waveform output efficiently; and *multi-speaker and cross-lingual TTS* models that can produce many different voice styles from a single model. By incorporating speaker embeddings, a single model can switch voices or even clone a person's voice with limited data. This raised concerns about **deepfake audio**, but also enabled positive applications like voice banking for people who lose the ability to speak. Another frontier has been **controllable speech synthesis** – the ability to modulate emotion, speaking style, or prosody in generated speech. Recent surveys highlight that modern TTS is evolving "beyond human-like speech to enabling *controllable* speech generation, with fine-grained control over attributes such as emotion, prosody, timbre and duration" (Towards Controllable Speech Synthesis in the Era of Large Language Models: A Survey). Techniques like style tokens, prompt-based control, and diffusion models have been introduced to let users specify if a sentence should be spoken happily, or whispering, etc. Indeed, diffusion probabilistic models (popular in image generation) have been adapted for TTS to further improve quality and provide more randomness control in speech generation. By 2025, industry TTS systems (e.g., Microsoft's Custom Neural Voice, Google's Cloud TTS) routinely achieve Mean Opinion Scores (MOS) around 4.5 out of 5 (almost indistinguishable from human recordings) for many languages. The focus has shifted to making these systems *lightweight* (so they can run on-device for real-time applications) and *data-efficient* (so that even low-resource languages can have high-quality TTS). An example of industrial-academic collaboration is the Mozilla Common Voice project and the Coqui TTS toolkit – open-source efforts that collected voices worldwide and built TTS models for dozens of languages, many of which historically lacked such technology.

- *Low-Resource and Multilingual Speech Models: **A significant trend has been tackling speech technology for low-resource languages using AI. Traditional speech recognition and synthesis required large labeled datasets, which left many languages behind. From 2019 onward, two game-changers helped address this:** self-supervised learning (SSL) **and** massively

- multilingual modeling**. In SSL, models like Facebook's wav2vec 2.0 (2020) and HuBERT (2021) are pre-trained on large amounts of unlabeled audio to learn general speech representations. These models can then be fine-tuned with minimal labeled data to achieve high speech recognition accuracy in languages with limited transcriptions. This dramatically lowered the data barrier – e.g., wav2vec 2.0 achieved excellent results on Swahili and Kyrgyz using only a few hours of transcribed audio, leveraging knowledge learned from hundreds of hours of unrelated languages. The second aspect, multilingual modeling, involves training one model on many languages, so that low-resource languages benefit from shared learning with high-resource ones. By 2021, we saw multilingual ASR models (like Google's Translatotron and Meta's XLS-R) that support dozens of languages. A pinnacle achievement came in 2023 when Meta AI announced the** Massively Multilingual Speech (MMS) project**, which expanded speech recognition and synthesis from ~100 languages to over** 1,100 languages **by training on a huge corpus of religious texts (e.g., readings of the Bible) in many tongues (Preserving the World's Language Diversity Through AI | Meta). The MMS models can also identify over** 4,000 spoken languages**, vastly more than any prior system (Preserving the World's Language Diversity Through AI | Meta). This leap was enabled by novel training techniques and data gathering, and all models were open-sourced – a significant boost for global inclusion in speech technology. Additionally,** transfer learning **and** cross-lingual transfer **have been widely researched: for example, using English speech data to help recognize accented English or related languages, and initiatives like the** BABEL program **and** AI4Bharat** (for Indian languages) created platforms for multilingual dataset development. By 2025, it is clear that the gap between high-resource and low-resource language speech tech has begun to narrow, thanks largely to AI approaches. Still, challenges like dialectal variation, code-switching, and truly low-resource languages (with <1 hour of data) are active research areas.

- *Other ML Audio Advances: **Beyond the big themes above, there are numerous other areas where AI intersected with audio.** Music generation and audio synthesis **(e.g., OpenAI's Jukebox in 2020) applied neural nets to produce music or sound effects.** Environmental sound recognition **and acoustic scene analysis have grown, with competitions like DCASE challenging teams to classify urban soundscapes or detect specific events (gunshots, glass break, etc.) using deep learning.** End-to-end audio classification **(e.g., for speaker identification, or acoustic event detection) became more accurate with architectures like CNNs on spectrograms and audio transformers.** Language processing** met speech in end-to-end spoken language understanding (SLU), where models can act directly on speech input to determine intent or extract information, skipping the intermediate transcription step. This has implications for voice assistants, making them faster and more resilient to recognition errors. Many of these innovations are being jointly pioneered by universities and industry research labs (Google Brain, Meta AI, Microsoft Research, etc.), often unveiled in forums like ICASSP and Interspeech and then quickly translated into products or open-source libraries.

In summary, AI and deep learning have revolutionized audio and speech processing in 2019–2025. The field moved from engineered features and models to data-driven end-to-end learning. Key outputs include **nearly noise-free speech in any environment**, **separation of intermixed sounds**, **human-like speech generation**, and **broad accessibility of speech tech to hundreds of languages**. This rapid progress has been enabled by large datasets, shared challenges, and interdisciplinary collaboration, positioning speech and audio AI as a cornerstone of modern human-computer interaction.

Speech Processing Advancements (Diarization, Emotion, Multilingual Models)

While the previous section covered algorithmic trends, here we highlight specific advancements in *speech processing tasks* beyond core recognition and synthesis – notably speaker diarization, paralinguistics (emotion recognition), and multilingual speech systems.

- *Speaker Diarization ("Who spoke when"): **Determining speaker segments in audio has long been important for meetings, broadcast, and forensic applications. Traditionally, diarization was done by clustering speaker embeddings (e.g., i-vectors or x-vectors) and was prone to errors, especially with overlapping speech. In the last five years,** deep learning **significantly improved diarization. One major development was the advent of** End-to-End Neural Diarization (EEND)** around 2019. Unlike clustering methods, EEND formulates diarization as a sequence labeling problem with a neural network directly labeling time frames for each speaker, using permutation-invariant training to handle the unnamed speaker outputs. Hitachi and others showed that EEND can *"explicitly handle overlapping speech"* and outperformed conventional methods on highly overlapped datasets (Multi-Talker Speech Recognition and Understanding - Research & Development : Hitachi). By introducing self-attention architectures, EEND was able to better model long-range speaker context, *"impressively reducing diarization errors"* in tests (Multi-Talker Speech Recognition and Understanding - Research & Development : Hitachi). For instance, on a telephone speech benchmark (CALLHOME), a self-attentive EEND system achieved a **10.99% diarization error rate (DER)** compared to 11.52% for the prior state-of-the-art clustering method (Multi-Talker Speech Recognition and Understanding - Research & Development : Hitachi) – a notable improvement (Multi-Talker Speech Recognition and Understanding - Research & Development : Hitachi). In parallel, enhancements to clustering-based diarization also occurred: **x-vectors** (deep speaker embeddings) became standard, and Bayesian HMM clustering (VBx) improved clustering robustness by modeling speaker turn dynamics (Bayesian HMM Based x-Vector Clustering for Speaker Diarization). The community also focused on **diarization in multispeaker,**

- **multilingual settings** – exemplified by the DISPLACE Challenge at Interspeech 2023, which targeted diarization of speakers and languages in conversation. Systems are increasingly **online (real-time)** and cope with open-set scenarios (unknown number of speakers). An exciting fusion is **ASR with diarization**: end-to-end models are being trained to jointly transcribe and diarize, aligning with how humans might directly parse “person A said ... person B said ...”. By 2025, diarization errors in clean conditions are quite low, though truly open-world, noisy scenario diarization remains a hard problem. We also see diarization technology deployed in video conferencing (identifying active speakers) and in court transcription, often aided by complementary video or microphone array data.

- *Emotion Recognition and Paralinguistics: **Recognizing a speaker’s emotional state or other paralinguistic traits (stress, intent, health) from their voice has gained attention with the rise of voice assistants and affective computing. Deep learning again drove progress here. Models using spectrogram features through CNNs or using recurrent networks on audio frames have improved accuracy on benchmark datasets (e.g., IEMOCAP, MSP-Podcast) over earlier methods. By learning high-dimensional representations, these models can detect subtle prosodic cues and voice quality features correlating with emotions like happiness, sadness, anger, or frustration. There’s also been work on** multilingual and cross-cultural emotion recognition**, since expressions of emotion in voice can vary across languages. One trend is leveraging** pre-trained speech models **(like wav2vec or speaker ID models) and fine-tuning them for emotion classification – benefiting from large-scale speech representation learning. For example, a 2023 IEEE TASLP paper introduced a non-parallel voice emotion conversion framework, indicating interest in not just recognizing but also converting or simulating emotions in speech (TASLP Volume 31 | 2023 | IEEE Signal Processing Society). Industry labs (e.g., Amazon) have started**

- **incorporating emotion detection to make AI assistants respond more empathetically; Amazon Alexa, for instance, was reported to detect user frustration and adapt its dialogue. Another aspect is** health and wellbeing applications**: identifying depression or stress from voice, which saw many research contributions (often presented in Interspeech's Computational Paralinguistics Challenge series). These tasks go beyond categorical emotion, looking at continuous measures (arousal, valence) or specific states (e.g., sleepiness, cognitive load). The global research community has grown around these topics, with dedicated workshops and increasing robustness of models to real-world conditions. We anticipate that by mid-2020s, voice emotion recognition will be a standard component in call center analytics (monitoring customer sentiment), automotive interfaces (to detect driver anger or drowsiness), and social robotics, albeit with necessary privacy and ethics considerations.

- *Multilingual and Cross-Lingual Speech Processing:** We touched on multilingual models under AI trends, but it's worth noting broader *speech processing* advancements enabling systems to handle multiple languages or switch between them. Early 2020s saw the advent of **unified multilingual speech recognition models** capable of transcribing many languages with a single model, as showcased by Meta's MMS project (covering 1,100+ languages) ([Preserving the World's Language Diversity Through AI | Meta]). This is a huge leap from the prior paradigm of one model per language. Similarly, multilingual text-to-speech models can generate speech in many languages and even perform *code-switching* (speaking two languages within one sentence) – a complex but important behavior for global applications. Additionally, **speech-to-speech translation** emerged from a research curiosity to working systems (e.g., Meta's universal translator prototype in 2022). These directly translate spoken input in one language to spoken output in another (sometimes even mimicking the original speaker's voice). While not yet mainstream, it's an active research frontier combining ASR, machine translation, and TTS. On a national level, countries with diverse languages (India, South Africa, etc.) have

- launched initiatives to create multilingual speech resources. For instance, the Indian government's "Bhashini" project and similar efforts aim to develop speech tech for all 22+ official languages, spurred by deep learning making it feasible to build acceptable systems even with relatively limited data. The focus in research has also included **language identification (LID)** from audio (discerning which language is spoken) – an essential precursor for multilingual systems to know which language model to apply. Modern LID can distinguish hundreds of languages and even detect language mixing. The global push towards multilingual models has fostered *collaboration between regions*: we see European research consortia (via Horizon2020 projects) and Asian institutions partnering to share data and models. Overall, the dream of breaking language barriers through speech technology is closer than ever, with 2019–2025 laying much of the groundwork.

In summary, speech processing R&D; in this period expanded beyond improving accuracy, to *richer understanding* of speech. Systems are now learning **who is speaking, in what language, and with what emotion or intent**, not just the literal words. These capabilities make human-machine communication more effective and natural. Academic conferences like Interspeech and IEEE ASRU featured these topics prominently, and we also saw dedicated challenges (e.g., Emotion in the Wild, VoxCeleb Speaker Recognition, DIHARD diarization) driving community progress.

Quantum Technologies in Acoustics and Audio

While quantum technology may seem far afield from everyday acoustics, the past few years have seen intriguing intersections of quantum research with audio and vibration sensing. These efforts are largely in the experimental research stage but hold potential for ultra-sensitive acoustic measurements and new signal processing paradigms.

One notable development was the creation of a **“quantum microphone”** capable of detecting individual phonons (the quantum particles of sound). In 2019, Stanford researchers built a device using superconducting quantum bits that could count phonons, essentially measuring the quietest possible sound quanta (Quantum microphone counts particles of sound | Stanford Report). This quantum microphone is far more sensitive than any conventional microphone, although it operates at cryogenic temperatures and in a lab setting. The achievement was heralded as it could enable quantum-level control of acoustic waves and advance fundamental science. Building on this, researchers are investigating **quantum sensors for audio frequencies**. A 2022 Nature paper demonstrated a **diamond nitrogen-vacancy (NV) center sensor** that achieved *“distortion-free quantum audio signal sensing”*, reconstructing audio waveforms with extremely high dynamic range by leveraging quantum phase detection (Quantum-assisted distortion-free audio signal sensing | Nature Communications). In essence, by measuring magnetic or strain fields quantum-mechanically, one can record sound without the classical distortions. The NV-center approach showed the ability to capture melodies and speech with high fidelity through a quantum heterodyne technique (Quantum-assisted distortion-free audio signal sensing | Nature Communications). These experiments indicate that quantum metrology can impact audio by offering sensors that bypass some limitations of classical transducers (e.g., the linear range limits).

Another area is **quantum computing applied to signal processing problems**. While quantum computing is still nascent, researchers have begun formulating audio signal processing tasks in quantum frameworks – for instance, using quantum algorithms for faster Fourier-like transforms or for optimization problems in audio filtering. There have been theoretical works on quantum audio watermarking and quantum-inspired filters (a novel quantum audio watermarking based on bipolar echo hiding) (Audio Compression Using Qubits and Quantum Neural Network), and even small-scale demonstrations like a quantum machine learning model for simple speech command recognition (Quantum Approaches for Dysphonia Assessment in Small Speech ...). These are very early-stage and do not

(yet) outperform classical methods, but the research is laying groundwork for future hybrid classical-quantum audio processing if and when quantum computers become more powerful.

Beyond computing, **quantum materials and phenomena** have influenced acoustics in the form of **topological acoustics** – analogous to topological insulators in electronics, researchers created structures where sound waves propagate in one-way, lossless modes around edges. Some of these works (circa 2019–2021) operate at audible frequencies, demonstrating exotic waveguides that could lead to new kinds of acoustic devices (e.g., highly robust sound wave conduits immune to scattering). While not "quantum" in operation, the design principles come from quantum topological physics concepts.

In summary, quantum technology's impact on audible acoustics during 2019–2025 is mostly through *sensing improvements* and *theoretical exploration*. A **quantum microphone sensor** can push the limits of detection (Quantum microphone counts particles of sound | Stanford Report), and **quantum-assisted audio sensing** shows promise for ultra-linear, high-resolution measurements (Quantum-assisted distortion-free audio signal sensing | Nature Communications). These could find niches in scientific measurements or perhaps high-end audio engineering (for instance, calibrating systems with quantum reference sensors). As quantum computing and sensing evolve, the convergence with acoustics is expected to grow, possibly yielding practical high-performance microphones or new signal processing algorithms leveraging quantum advantages in the coming decade.

Other Emerging Topics in Audible Acoustics (2019–2025)

Beyond the major themes above, several other noteworthy research trends emerged in audible acoustics:

- **Acoustic Metamaterials and Noise Control:** The field of acoustic metamaterials – engineered structures with properties not found in natural materials – boomed in the late 2010s. Researchers developed metamaterials for noise reduction, sound focusing, and novel acoustic lenses. We saw examples like the **open 3D-printed ring metamaterial** that cancels 94% of sound transmission (at a target frequency) while letting air flow through (BU Engineers Develop New Acoustic Metamaterial and Noise Cancellation Device | The Brink | Boston University). Such structures could revolutionize noise control in ventilation, drones, or MRI machines where traditional enclosures are impractical. Metamaterials have also been used to create ultra-thin **sound absorbers** and **barrier panels** that outperform conventional foams, and to design **acoustic holograms** that shape sound fields in desired ways. This is an active area bridging physics and engineering – conferences like the ASA (Acoustical Society of America) and journals like *Physical Review Applied* have featured many such studies. By 2025, some metamaterial concepts are moving toward commercialization (e.g., panels for HVAC noise). The **challenge remains broadening the bandwidth** of these effects, as many designs work well only in narrow frequency bands.

- **Spatial Audio and AR/VR:** With the rise of virtual and augmented reality, spatial audio (3D sound) has become crucial for immersive experiences. Research has focused on improved **HRTF (Head-Related Transfer Function) personalization**, room simulation (ambisonics, binaural rendering), and efficient encoding of spatial cues. As AES noted, *"Spatial audio is an essential technology for VR/AR, providing realism and even 'hyper-reality' for visceral immersive experiences."* (Audio for Virtual and Augmented Reality - AES). The period saw maturation of production tools: many game engines and digital audio workstations incorporated ambisonics and object-based audio workflows (Audio for Virtual and Augmented Reality - AES). Researchers also worked on **sound field synthesis** (e.g., Wave Field Synthesis, higher-order Ambisonics) to reproduce life-like sound in installations or theaters. *Binaural audio* through headphones has reached new heights

- – for instance, Dolby Atmos for headphones is now common in consumer devices, and personalized spatial audio was a selling point for Apple and Sony headphone products. Academic work on **“6DoF audio”** (where a listener can move freely and audio adjusts realistically) combined tracking, efficient rendering, and perceptual validation. Overall, AR/VR audio research is interdisciplinary, involving acoustics, signal processing, and perceptual science, and has strong industry involvement (Facebook Reality Labs, Apple’s audio teams, etc.). The consensus is that spatial audio significantly enhances user experience and is a hotbed of innovation.

- **Sound Event Detection and Acoustic Scene Analysis:** Beyond speech, there’s a growing domain of automatically recognizing sounds in our environment – like identifying if a *glass breaks* or a *dog barks* in an audio stream. The **DCASE (Detection and Classification of Acoustic Scenes and Events)** challenge has run annually, spurring advances in sound event detection, scene classification (recognizing an audio environment as “busy street”, “restaurant”, etc.), and audio captioning. Techniques here often adapt speech recognition methods (e.g., using CNNs on mel-spectrograms) to general sounds, but also need to handle very diverse noises and overlapping events. From 2019 to 2025, systems improved to reliably detect dozens of sound classes in real time, enabling smart home security devices and context-aware smartphones that can, for example, alert to alarms or distinguish noise sources. Some research has integrated **audio with video** (e.g., to detect if a *door knock* sound matches a person at the door in camera footage), reflecting a trend toward multi-modal analysis.

- **Audio for Health and Accessibility:** Audible acoustics plays a role in health monitoring (e.g., algorithms that analyze cough sounds for diagnostics, or voice analysis for Parkinson’s disease symptoms). This period included work on detecting COVID-19 from cough or breath sounds (with mixed success), and on using **hearing aid algorithms with AI** to better separate speech from noise for the hearing impaired. The latter has been boosted by deep learning noise reduction and directional microphones to create “smart hearing aids” that adapt to

- environments. *Audio-based fall detection* (listening for the sound of a fall) and *heart or lung sound analysis* (on the edge of audible range) are other topics that saw progress, often in medical or DSP-focused conferences.

- **Audio Coding and Streaming:** Even as processing gets more powerful, efficient audio coding remains important (for streaming and storage). In the early 2020s, we saw the development of **neural audio codecs**, like Qualcomm's AptX Adaptive and Google's Lyra and SoundStream. Neural codecs use autoencoders or GANs to compress audio at very low bitrates (e.g., 3 kbps) with surprisingly good quality, outperforming traditional codecs like Opus at those rates. The research community, especially at ICASSP, has been exploring improved perceptual metrics and training objectives to optimize such codecs.

- **Industry-Academia Collaborations:** Across these emerging topics, many advances were driven by collaborations and open challenges. For instance, **Amazon's Alexa Prize** challenges (focused on dialogue systems) indirectly pushed speech technology for conversational AI. **Google's AR/VR audio research** collaborations with universities advanced spatial audio (sharing datasets like the ARI HRTF set). Open-source projects (Kaldi, then K2 and lately ESPnet, SpeechBrain) led by academic-industry teams have democratized access to state-of-the-art methods, making it easier to experiment and transfer innovations across institutions. These collaborations are highlighted in the tables below and discussed further in the global trends section.

The audible acoustics field is broad; the above are just a few of the prominent "other" topics during 2019–2025. Each represents a fusion of new technology with acoustics to solve real-world problems or create new experiences. Many of these topics will likely continue to grow in importance.

Key Conferences and Journals (2019–2025)

Research findings in acoustics and audio are disseminated through a mix of specialized conferences and journals. Table 1 lists some of the **notable conferences** and **top journals** in these fields from 2019–2025, along with their focus areas.

| **Conference / Journal** | **Focus and Notability** | **Frequency/Publisher** |

|---|---|---|

| **ICASSP (Int'l Conf. on Acoustics, Speech, and Signal Processing)** | Flagship IEEE conference covering all aspects of signal processing, with large tracks on audio, speech (ASR, speech synthesis), music, and acoustics. Premier venue for new algorithms (denoising, source separation, ML for audio). E.g., Conv-TasNet and many SSL speech papers were first published here. | Annual (IEEE) |

| **INTERSPEECH** | Largest conference devoted to speech science and technology. Covers speech recognition, speaker ID, diarization, paralinguistics, spoken dialogue, etc. Often where state-of-the-art ASR and TTS works appear. E.g., many multilingual and low-resource papers, and the DNS Challenge results (The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Testing Framework, and Challenge Results - Microsoft Research), have been presented here. | Annual (ISCA) |

| **AES Convention & Conferences** | Audio Engineering Society events focusing on professional audio tech: transducers (mics, speakers), spatial audio, recording, audio DSP, and emerging tech (AR/VR audio, automotive audio). AES conferences (e.g., on **Audio for VR** in 2020) highlight applied research and industry advances in sound design (Audio for Virtual and Augmented Reality - AES). | Annual conventions; specialty conferences (AES) |

| **ASA Meetings (Acoustical Society of America)** | Broad acoustics conferences (held two times a year) covering physical acoustics, architectural acoustics, noise control, psychoacoustics, musical acoustics, etc. Important for transducer research, acoustic materials, and

fundamental acoustics. Often where metamaterial and transducer papers are presented first. | Semiannual (ASA) |

| **IEEE WASPAA (Workshop on Applications of Signal Processing to Audio and Acoustics)** | A biennial workshop (small, focused) specifically on audio signal processing. Notable for far-field speech, spatial audio, music analysis, and hearing aid algorithm research in an informal setting. | Biennial (IEEE) |

| **DCASE Workshop** | Conference/workshop for Detection and Classification of Acoustic Scenes and Events. Important venue for environmental sound recognition and audio scene analysis work (growing area in 2019–2025). | Annual (Workshop) |

| **IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)** | Top journal for audio and speech processing algorithms. Publishes extensive research articles on ASR, speech synthesis, audio analysis, etc. Many of the foundational deep learning papers for speech/audio appear here after conference. | Monthly (IEEE/ACM) (IEEE Transactions on Audio, Speech and Language Processing) |

| **The Journal of the Acoustical Society of America (JASA)** | Premier journal for acoustics broadly. Includes physical acoustics, transducer design, architectural acoustics, noise, psychoacoustics. Known for thorough experimental and theoretical papers. E.g., many MEMS microphone design studies and metamaterial evaluations are published here. | Monthly (ASA) |

| **Journal of the Audio Engineering Society (JAES)** | Key journal at the intersection of audio engineering and research. Topics range from loudspeaker design and audio electronics to spatial audio perception. Often features advances from industry R&D; (e.g., new amplifier classes, codec performance) alongside academic work. | Bi-monthly (AES) |

| **Speech Communication (Elsevier)** | An established journal focusing on speech science, including speech production/perception, phonetics, and speech technology (ASR, TTS). Good coverage of multilingual and

speech analysis research. | 8 issues/year (Elsevier) |

| **Computer Speech & Language (Elsevier)** | Another top journal for speech technology and computational linguistics related to speech. Publishes many works on ASR, spoken language understanding, and dialogue systems. | Quarterly (Elsevier) |

| **Applied Acoustics (Elsevier)** | Journal focused on applied research in acoustics: e.g., noise control, acoustic materials, transducer applications, architectural acoustics designs. Has featured many metamaterial and noise-cancelling material papers in recent years. | Monthly (Elsevier) |

- Table 1: Major conferences and journals in acoustics, audio & speech (2019–2025), with their focus areas. These venues represent the primary outlets for academic research developments discussed in this report.*

Leading Research Institutions and Organizations

Research in acoustics and audio is spread across universities, corporate labs, and national institutes worldwide. Table 2 highlights some **notable universities, labs, and companies** that have been at the forefront of the discussed areas in 2019–2025. Each has made significant contributions, from publishing influential papers to developing important technologies or collaborations.

| **Institution / Lab** | **Country** | **Notable Contributions (2019–2025)** |

|---|---|---|

| MIT (Massachusetts Institute of Technology) – Media Lab, Research Laboratory of Electronics | USA | Pioneering transducer research (e.g., paper-thin loudspeaker (Clip New Scientist Paper-thin speaker can play Queen from any surface (May 31) | MIT News | Massachusetts Institute of Technology)); spatial audio and AR (the Media Lab's Opera of the Future);

robust speech recognition research (CSAIL). |

| Stanford University (CCRMA and Applied Physics) | USA | Quantum microphone development (Quantum microphone counts particles of sound | Stanford Report); music technology at CCRMA; early work on self-supervised speech models (with Meta). |

| University of Illinois Urbana-Champaign (Coordinated Science Lab) | USA | MEMS acoustics (microphones, piezoelectric transducers); acoustic metamaterials theory and design; speech separation algorithms (Prof. P. Smaragdis's group). |

| Carnegie Mellon University (Language Technologies Institute) | USA | Speech recognition and multilingual ASR (CMU Sphinx legacy to modern end-to-end systems); speaker diarization research; social audio (e.g., CMU's work on Alexa Prize dialogue). |

| Johns Hopkins University (Center for Language and Speech Processing - CLSP) | USA | Pioneered EEND diarization with Hitachi (Multi-Talker Speech Recognition and Understanding - Research & Development : Hitachi); low-resource speech recognition (e.g., IARPA Babel program contributions); robust ASR in noisy environments. |

| Tsinghua University & Chinese Academy of Sciences (Institute of Acoustics) | China | Leading MEMS transducer research in China; major contributions to speech separation and deep learning speech translation; many papers in ICASSP and Interspeech on speech enhancement. |

| Beijing Speech and Language Technology (SLT) Center – e.g., at Univ. of Science and Technology of China (USTC) | China | Top-tier results in speech competitions (e.g., CHiME-5 Challenge winners (Multi-Talker Speech Recognition and Understanding - Research & Development : Hitachi)); advanced research in speaker recognition and speech synthesis (USTC's Voice Conversion). |

| University of Edinburgh (Centre for Speech Technology Research) | UK | Speech synthesis and voice cloning (known for the Festival and Merlin TTS systems, now deep learning based); expressive speech modeling; machine translation speech work. |

| University of Cambridge (Engineering Dept.) | UK | Automatic speech recognition (the HTK legacy; modern contributions to end-to-end ASR and adaptive noise cancellation); also strong in array signal processing (BeamformIt, etc.). |

| Fraunhofer Institutes (IDMT, IIS) | Germany | Audio coding (inventors of MP3/AAC, now neural codecs), and audio signal processing; Fraunhofer IDMT works on hearing aid algorithms and sound event detection. |

| NTT Communication Science Labs | Japan | Fundamental speech processing research (source separation, target-speaker ASR concept (Multi-Talker Speech Recognition and Understanding - Research & Development : Hitachi), voice conversion); high-quality TTS and voice morphing. |

| Hitachi R&D; (Speech Research Lab) | Japan | Collaboration with JHU on multi-talker ASR and diarization (Multi-Talker Speech Recognition and Understanding - Research & Development : Hitachi); multi-channel speech enhancement (guided source separation); applied speech technology for robotics (Multi-Talker Speech Recognition and Understanding - Research & Development : Hitachi). |

| Google (Google Research, DeepMind) | USA/Global | Major industrial player: developed WaveNet, Conformer ASR model, multilingual translation, SpeakerID in Google Assistant; conducts research published at top venues on speech and audio (often in collaboration with academics). |

| Microsoft Research (Speech and Audio groups) | USA/Global | Pioneered deep noise suppression challenge (The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Testing Framework, and Challenge Results - Microsoft Research); state-of-the-art meeting transcription (Microsoft’s AMI/ICSIsystems); cutting-edge text-to-speech (e.g., FastSpeech, VALL-E voice cloning in 2023). |

| Meta (Facebook) AI Research | USA/Global | Large-scale multilingual speech (MMS project (Preserving the World's Language Diversity Through AI | Meta)); self-supervised audio models (wav2vec series); research on far-field AR audio (Facebook Reality Labs). |

| Amazon (Alexa AI, AWS AI Labs) | USA/Global | Deployed neural speech synthesis at scale; research on wake-word detection, conversational AI; tools for Emotion AI in Alexa; sponsors academia (Alexa Prize, university collaborations on dialog and acoustics). |

| Dolby Laboratories | USA | Industry leader in audio coding and spatial audio; research and standardization of immersive audio (Dolby Atmos), dialogue enhancement, etc., often presenting in AES and IEEE. |

| National acoustics labs (e.g., NPL in UK, CNRS Orphée in France) | Various | Metrology and acoustic standards research; architectural acoustics, electro-acoustic measurements and improvements of transducer standards. These labs support fundamental measurements that back the research advancements. |

- Table 2: Selected research institutions and labs driving innovations in acoustics/audio (2019–2025). This list spans academia and industry, illustrating global contributions.* (Note: Many other groups worldwide are active; this table highlights a sample.)

Global Hotspots, Collaboration Networks, and Funding Trends

Research in acoustics and speech is a global endeavor, but certain regions have become hotspots due to concentrated expertise and investment. The **United States** and **China** stand out as two powerhouses in AI-driven speech research. The U.S. has a long legacy of speech and audio R&D; (CMU, MIT, Bell Labs tradition, etc.), and big tech companies in the U.S. (Google, Microsoft, Amazon, Apple, IBM) heavily fund internal research and academic collaborations. China, meanwhile, dramatically increased its output: since about 2017, China has led in quantity of AI publications, including speech/audio – in 2020 China surpassed the U.S. in share of top 10% most-cited AI papers (AI Report: Competition Grows Between China and the U.S.). Companies like **iFlytek, Baidu, Tencent, Alibaba** have sizable speech research divisions, and government programs (the National

AI Plan, etc.) poured resources into speech tech (for applications like voice assistants, smart homes, and surveillance). This has resulted in Chinese labs frequently topping benchmarks for speech recognition and synthesis in Mandarin and beyond.

- *Europe **remains very strong in acoustics and audio, with a collaborative flavor. Countries like the** UK, Germany, France, and Italy **host leading groups (as noted in Table 2), and the European Union has funded multi-nation projects (Horizon 2020 programs such as “METNET” for metamaterials, “ELG” – European Language Grid – for language technology, etc.). These projects encourage cross-border collaboration and sharing of resources. For example, the** European Language Grid (2019–2022) **involved partners from many EU countries building infrastructure for multilingual speech and language tools, benefiting smaller language communities.** Scandinavia** (e.g., Aalto University in Finland for audio processing, and Denmark’s acoustic clusters like Delta) also contributes significantly, especially in hearing technology and audio perception.

- *Japan and South Korea** have traditionally been strong in acoustics (transducers, audio electronics) and continue to innovate. Japan’s NTT and universities (e.g., Tokyo, Kyoto, JAIST) produce influential work in speech enhancement and synthesis. South Korea’s Samsung and academic groups (KAIST, Seoul National University) are active in speech recognition and AI music. Both countries often focus on multilingual aspects relevant to their languages and on consumer electronics integration.

- *Collaboration networks **in this field often form around shared evaluations and challenges. We’ve mentioned several: DNS Challenge (Microsoft leading but many academic participants), DCASE for sound events (led by academic consortia), the CHiME and REVERB challenges (multichannel speech in noise, with global teams competing). These efforts build an international network of researchers working together on common datasets – a form of open collaboration. Additionally, large companies**

- **routinely collaborate with academia: e.g.,** Hitachi-JHU **partnership on diarization (**[**Multi-Talker Speech Recognition and Understanding - Research & Development : Hitachi**](#)**),** Microsoft's Open Speech Research **initiative funding university projects on low-resource ASR,** Google's Faculty Research Awards **that supported work on speech separation and TTS, etc. Another collaboration vector is open-source: frameworks like** Kaldi (speech recognition toolkit)** were primarily developed by academics (Daniel Povey et al.) but saw contributions from industry and worldwide users; newer toolkits (Espresso, SpeechBrain) continue this, creating a shared platform for innovation.

In terms of **funding trends**, government funding in acoustics has been steady but somewhat eclipsed by the surge of investment in AI. Agencies like the U.S. NSF and NIH fund audio research often in context of accessibility (e.g., grants for ASR in educational tools or for improving hearing aids). Defense agencies (DARPA, IARPA) have interest in robust speech technology (the DARPA RATS program earlier in the decade on robust ASR, and IARPA's Babel for low-resource languages). From 2019 onward, **AI-related funding** clearly boosted speech: many national AI initiatives include a language technology component. For instance, India's NSF-equivalent launched "Speech Technology Mission" for Indian languages in 2020. **Private R&D; spending** via tech companies has arguably driven many breakthroughs (the budgets of a Google or Amazon research lab far exceed typical academic grants, enabling large-scale data and compute). However, academia remains crucial for foundational research and training talent; we observe that many top industry researchers in audio have roots in the labs listed in Table 2, and they maintain ties via adjunct positions or joint centers (e.g., MILA in Montreal hosts work with both university and industry participation in audio ML).

- *Countries investing significantly **include (beyond US/China)** Canada **– via CIFAR and Vector Institute programs, Canada has nurtured AI talent and some speech research (like Montreal's work on generative audio).** Australia **– with its strong cochlear implant research heritage – invests in hearing acoustics and**

- **also had a big machine learning surge (Sydney and Melbourne universities contributing to audio scene understanding).** Singapore **and** India** have growing footprints: Singapore's I^2R and universities publish in audio processing (often focusing on Asian languages and urban sound analysis), and India's IITs and IIITs increasingly publish in speech conferences with government backing to overcome language barriers.

Overall, the global landscape is one of vibrant collaboration but also healthy competition. The **"hotspots"** – Bay Area, Seattle, Cambridge (UK and MA), Beijing, etc. – draw international talent, and we've seen many researchers moving between academia and industry globally, further entwining the network. A concrete indicator: papers at ICASSP or Interspeech often have authors from multiple countries and from both universities and companies, reflecting how common cross-pollination is.

Another noteworthy trend is the **open science movement** in this field: from open datasets (LibriSpeech, CommonVoice, etc.) to open models (Kaldi, ESPnet, HuggingFace Transformers for audio), which has democratized research and allowed broader participation, including from countries with less funding. This has somewhat leveled the playing field and led to new collaborations (for example, Mozilla's Common Voice involves volunteer contributions from around the globe to collect speech data).

In conclusion, global research in acoustics (audible range) is robust and interconnected. North America, Europe, and Asia all host key centers, with China's rise in AI research making it a major contributor. Collaboration networks via challenges and open-source projects knit these regions together. Funding is strong, especially when tied to AI, but ensuring support for core acoustics (materials, transducers, fundamental psychoacoustics) remains important and is often championed by societies like ASA and AES. The next section looks ahead to how these trends might evolve in the coming five years.

Future Outlook (2025–2030)

Looking forward, several trajectories are likely based on current research:

- **Transducers and Hardware**: We anticipate **commercial proliferation of MEMS speakers** in earbuds, phones, and even flat panel TVs. By 2030, the majority of mobile devices might use solid-state microspeakers, enabling thinner form factors and new audio features (like multiple tiny speakers for true stereo on a phone). MEMS microphone performance will continue to improve; reaching 80 dB SNR or higher could happen with new materials or multi-membrane designs. **Optical microphones** might find niche use in high-end audio measurement or in extreme environments (since they are immune to EM interference). **Spintronic sensors** could mature if materials improve – perhaps seeing use in scientific instruments for vibration sensing. **Metamaterials** will likely be integrated into products: e.g., noise-cancelling ventilated panels for HVAC, metamaterial-enhanced loudspeakers that have better bass response from small enclosures, and acoustic lenses for directed audio (imagine a soundbar that can beam sound to different parts of a room without beamforming algorithms). **Quantum acoustic sensing** may remain mostly lab-bound by 2030, but if it progresses, we might see an ultra-sensitive microphone for research or a quantum-based vibration sensor used in seismology or structural monitoring (outside consumer space). Additionally, expect **greener and more robust materials** in transducers: lead-free piezos and more sustainable manufacturing for MEMS.

- **AI in Audio**: The line between “audio research” and “AI research” will blur further. **Foundation models** (very large models) that can handle audio, vision, and text together are likely to emerge – one can imagine a multimodal model that understands audio events, speech, and even music within one architecture (some early versions exist in 2023, but expect more powerful ones). **Real-time embedded AI** will be crucial: by 2030, we envision hearing aids and AR glasses with on-board neural processors running always-on sound scene analysis (to enhance

- relevant sounds and cancel noise intelligently for the wearer). **Speech recognition** might reach a plateau in accuracy for major languages (closing in on human-level performance in ideal conditions), shifting focus to **robustness** (handling heavy noise, accented speech, code-switching seamlessly) and **confidence estimation** (knowing when it's likely wrong). **Speech synthesis** will be nearly indistinguishable from human speech, including emotional nuance – a key issue will be watermarking or otherwise marking AI-generated speech to prevent misuse (an active research area likely to grow given deepfake concerns). For low-resource languages, hopefully by 2030 most languages will have at least baseline speech tech; efforts like Meta's MMS are a leap in that direction and more will follow, perhaps achieving coverage of *all* spoken languages in some form, aided by continued data gathering (e.g., via crowd-sourcing with smartphones).

- **Speech Processing and Understanding**: **Speaker diarization** might evolve into part of a broader *conversation understanding* task. By 2030, virtual meeting assistants could reliably label speakers, transcribe content, detect action items, and gauge sentiment – all in one pipeline. **Emotion recognition** and **affective computing** will likely become more integrated into customer service bots, mental health monitoring tools, and automotive systems (your car might detect you're stressed and adjust settings, for example). However, these raise privacy issues, so research on doing this *on-device* (so data doesn't leave the user's control) will be important. **Multilingual models** will continue to advance; we may see universal translators that can on-the-fly translate someone's speech into another language in their own voice with only a short delay – something that was in the realm of sci-fi, but now appears within reach due to progress in end-to-end speech translation and voice cloning.

- **Quantum and Advanced Tech**: If quantum computing becomes more practically accessible, by late 2020s we might see it applied for solving certain signal processing optimizations faster, or in designing new algorithms (quantum machine learning for audio). More tangible could be **quantum-enhanced audio hardware** – for example,

- atomic-scale vibration sensors embedded in microphones to extend dynamic range, or using quantum oscillators as extremely stable audio frequency references for calibrations.

- **Emerging Applications**: **Spatial audio** will be ubiquitous – not only in entertainment (movies, VR, gaming) but also in communications (teleconferencing with spatial cues for each participant) and perhaps in home systems (smart speakers coordinating to create immersive sound fields). **Personalized audio** (tailoring sound to an individual's hearing profile) could be a standard feature in consumer devices, thanks to quick hearing tests and adaptive filters. **Safety and accessibility** will drive innovation: e.g., acoustic sensors in smart cities for hazard detection (with privacy-preserving algorithms that only flag certain events without streaming audio), or further improvements in cochlear implants and brain-computer auditory prosthetics driven by better signal processing.

- **Global Research Community**: We expect continued strong contributions from North America and Asia, with China possibly taking a larger leadership role in publications and even standards. However, awareness of inclusivity will increase – more focus on under-represented languages and regions, partly because the tech now allows it and partly due to societal push for digital inclusion. Collaboration will remain key; perhaps new global initiatives (akin to CERN in physics) could emerge for AI and acoustics, pooling resources to tackle, say, preservation of endangered languages through speech tech, or global noise pollution reduction via acoustic research.

- **Funding and Challenges**: Research funding may shift to emphasize ethical and societal impact: projects on *AI fairness in speech* (ensuring systems work for diverse accents and demographics) and *safety* (detecting fake audio, securing voice authentication against spoofing) will be hot. Traditional acoustics (e.g., architectural acoustics, environmental noise) might leverage AI too, so expect cross-pollination. The challenge will be balancing *innovation with responsibility* – the community will likely engage more with policy, e.g., how to regulate deepfake audio or how to standardize AI hearing aid evaluations. On the

- acoustics side, climate change and sustainability might become part of the research agenda (e.g., designing soundscapes for increasingly noisy urban environments, or using bio-based materials for sound absorption).

In essence, the next five years will build on the momentum of 2019–2025. **Audio technology will become more immersive, intelligent, and inclusive**, powered by the synergy of advanced hardware and AI. Our devices will not only hear and speak with greater fidelity but also *understand* context and user needs far better. Quantum and metamaterial innovations might quietly enhance the performance behind the scenes. The global acoustics community, through shared knowledge at conferences and collaborative projects, is poised to solve long-standing challenges (like robust speech in noise, or compact high-quality speakers) and open new frontiers (like real-time universal translation, or acoustics-informed AI that can navigate complex auditory scenes as humans do). The foundation laid in the past five years gives plenty of reason for optimism about what's to come in acoustics and audio research.

References

[1] IM73A135 | High performance XENSIV™ MEMS microphone with ultra high dynamic range - Infineon Technologies: https://www.infineon.com/cms/en/product/sensor/mems-microphones/mems-microphones-for-consumer/im73a135/#:~:text=Infineon%E2%80%99s%20XENSIV%E2%84%A2%20MEMS%20analog%20microphone,benefits%20inherent%20in%20MEMS%20technology

[2] text{0.9}$N-Based Bimorph Piezoelectric MEMS Microphones With ...: https://www.researchgate.net/publication/387279659_Sc_text01Al_text09N-Based_Bimorph_Piezoelectric_MEMS_Microphones_With_Tractive_Structures#:~:text=...%20www.researchgate.net%20%20Scandium,bimorph%20obtains%20a%20higher

[3] Electro-Acoustic Properties of Scandium-Doped Aluminum Nitride ...: https://www.mdpi.com/2073-4352/12/10/1431#:~:text=,frequency%20filters%2C%20sensors%2C%20and

[4] Clip New Scientist Paper-thin speaker can play Queen from any surface (May 31) | MIT News | Massachusetts Institute of Technology: https://news.mit.edu/news-clip/new-scientist-238#:~:text=MIT%20researchers%20have%20developed%20a,attached%20to%2C%E2%80%9D%20writes%20New%20Scientist

[5] BU Engineers Develop New Acoustic Metamaterial and Noise Cancellation Device | The Brink | Boston University: https://www.bu.edu/articles/2019/making-the-world-a-lot-quieter/#:~:text=BU%20engineers%20have%20developed%20an,cancel%2094%20percent%20of%20sound

[6] BU Engineers Develop New Acoustic Metamaterial and Noise Cancellation Device | The Brink | Boston University: https://www.bu.edu/articles/2019/making-the-world-a-lot-quieter/

[7] Electro-Acoustic Properties of Scandium-Doped Aluminum Nitride ...: https://www.mdpi.com/2073-4352/12/10/1431#:~:text=Electro,frequency%20filters%2C%20sensors%2C%20and

[8] The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Testing Framework, and Challenge Results - Microsoft Research: https://www.microsoft.com/en-us/research/publication/the-interspeech-2020-deep-noise-suppression-challenge-datasets-subjective-testing-framework-and-challenge-results/#:~:text=The%20INTERSPEECH%202020%20Deep%20Noise,training%20the%20noise%20suppression%20models

[9] The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Testing Framework, and Challenge Results - Microsoft Research: https://www.microsoft.com/en-us/research/publication/the-interspeech-2020-deep-noise-suppression-challenge-datasets-subjective-testing-framework-and-challenge-results/#:~:text=and%20a%20representative%20test%20set,on%20a%20blind%20test%20set

[10] Data Augmentation and Loss Normalization for Deep Noise Suppression: https://www.microsoft.com/en-us/research/uploads/prod/2020/10/Braun-Tashev2020_Chapter_DataAugmentationAndLossNormali.pdf#:~:text=appli%02cations%20also%20targeting%20real,has%20not%20been%20possible%20so

[11] The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Testing Framework, and Challenge Results - Microsoft Research: https://www.microsoft.com/en-us/research/publication/the-interspeech-2020-deep-noise-suppression-challenge-datasets-subjective-testing-framework-and-challenge-results/#:~:text=tests%20are%20not%20scalable%20for,on%20a%20blind%20test%20set

[12] Recent Advances in Audio Source Separation | Frontiers Research Topic: https://www.frontiersin.org/research-topics/21921/recent-advances-in-audio-source-separation/magazine#:~:text=introduction%20of%20extra%20source%20information,field%20of%20audio%20source%20separation

[13] Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude ... - arXiv: https://arxiv.org/abs/1809.07454#:~:text=arXiv%20arxiv,domain%20speech%20separation

[14] Recent Advances in Audio Source Separation | Frontiers Research Topic: https://www.frontiersin.org/research-topics/21921/recent-advances-in-audio-source-separation/magazine#:~:text=28%20January%202022

[15] Towards Controllable Speech Synthesis in the Era of Large Language Models: A Survey: https://arxiv.org/html/2412.06602v1#:~:text=Text,utilizing%20natural%20language%20prompts%2C%20aiming

[16] Preserving the World's Language Diversity Through AI | Meta: https://about.fb.com/news/2023/05/ai-massively-multilingual-speech-technology/#:~:text=Massively%20Multilingual%20Speech%20,40%20times%20more%20than%20before

[17] Preserving the World's Language Diversity Through AI | Meta: https://about.fb.com/news/2023/05/ai-massively-multilingual-speech-technology/#:~:text=,that%20can%20understand%20everyone%E2%80%99s%20voice

[18] Multi-Talker Speech Recognition and Understanding - Research & Development : Hitachi: https://rd.hitachi.com/_ct/17712285#:~:text=permutation%20invariant%20training,upcoming%20IEEE%20ASRU%202019%20workshop

[19] Multi-Talker Speech Recognition and Understanding - Research & Development : Hitachi: https://rd.hitachi.com/_ct/17712285#:~:text=permutation%20invariant%20training,06247%20accepted%20at%20upcoming

[20] Multi-Talker Speech Recognition and Understanding - Research & Development : Hitachi: https://rd.hitachi.com/_ct/17712285#:~:text=the%20typical%20speaker%20diarization%20method,upcoming%20IEEE%20ASRU%202019%20workshop

[21] Multi-Talker Speech Recognition and Understanding - Research & Development : Hitachi: https://rd.hitachi.com/_ct/17712285#:~:text=similarity%20between%20two%20frames%2C%20impressively,upcoming%20IEEE%20ASRU%202019%20workshop

[22] Bayesian HMM Based x-Vector Clustering for Speaker Diarization: https://www.isca-archive.org/interspeech_2019/diez19_interspeech.html#:~:text=archive,on%20Bayesian%20Hidden%20Markov%20Models

[23] TASLP Volume 31 | 2023 | IEEE Signal Processing Society: https://signalprocessingsociety.org/publications-resources/ieeeacm-transactions-audio-speech-and-language-processing/2023/01#:~:text=Society%20signalprocessingsociety,First%2C%20we%20propose%20a

[24] Quantum microphone counts particles of sound | Stanford Report: https://news.stanford.edu/stories/2019/07/quantum-microphone-counts-particles-sound#:~:text=Report%20news,particles%20of%20sound%2C%20called%20phonons

[25] Quantum-assisted distortion-free audio signal sensing | Nature Communications: https://www.nature.com/articles/s41467-022-32150-1#:~:text=another%20important%20feature%20for%20reconstructing,measurements%20are%20required%20at%20multiple

[26] Quantum-assisted distortion-free audio signal sensing | Nature Communications: https://www.nature.com/articles/s41467-022-32150-1#:~:text=phase,bands%20within%20a%20limited%20volume

[27] a novel quantum audio watermarking based on bipolar echo hiding: https://www.researchgate.net/publication/388962982_Towards_secure_quantum_communication_a_novel_quantum_audio_watermarking_based_on_bipolar_echo_hiding#:~:text=a%20novel%20quantum%20audio%20watermarking,facilitating%20secure%20communications%20by

[28] Audio Compression Using Qubits and Quantum Neural Network: https://www.sciencedirect.com/science/article/pii/S1877050924031156#:~:text=Audio%20Compression%20Using%20Qubits%20and,embeds%20audio%20signals%20in

[29] Quantum Approaches for Dysphonia Assessment in Small Speech ...: https://arxiv.org/html/2502.08968#:~:text=Quantum%20Approaches%20for%20Dysphonia%20Assessment,QCNN%29%20or%20QNN%2C

[30] Audio for Virtual and Augmented Reality - AES: https://aes2.org/audio-topics/audio-for-virtual-and-augmented-reality-2/#:~:text=Virtual%20and%20augmented%20reality%20is,immersive%20experiences%20visceral%20and%20plausible

[31] Audio for Virtual and Augmented Reality - AES: https://aes2.org/audio-topics/audio-for-virtual-and-augmented-reality-2/#:~:text=Consequently%20it%20is%20an%20exciting,plugins%20to%20support%20VR%2FAR%20workflows

[32] IEEE Transactions on Audio, Speech and Language Processing: https://signalprocessingsociety.org/publications-resources/ieee-transactions-audio-speech-and-language-processing#:~:text=TASLPRO%20is%20dedicated%20to%20innovative,and%20language%2C%20and%20their%20applications

[33] Multi-Talker Speech Recognition and Understanding - Research & Development : Hitachi: https://rd.hitachi.com/_ct/17712285#:~:text=held%20in%202018,of%20utterances%20in%20its%20mask

[34] Multi-Talker Speech Recognition and Understanding - Research & Development : Hitachi: https://rd.hitachi.com/_ct/17712285#:~:text=Although%20the%20speech%20separation%20technology,ASR%29%20system%2C%20it%20outputs%20the

[35] Multi-Talker Speech Recognition and Understanding - Research & Development : Hitachi: https://rd.hitachi.com/_ct/17712285#:~:text=Hitachi%20and%20Johns%20Hopkins%20University,First%2C%20the

[36] Multi-Talker Speech Recognition and Understanding - Research & Development : Hitachi: https://rd.hitachi.com/_ct/17712285#:~:text=When%20Robot%20Responds%20to%20You%3F

[37] AI Report: Competition Grows Between China and the U.S.: https://hai.stanford.edu/news/ai-report-competition-grows-between-china-and-us#:~:text=AI%20Report%3A%20Competition%20Grows%20Between,