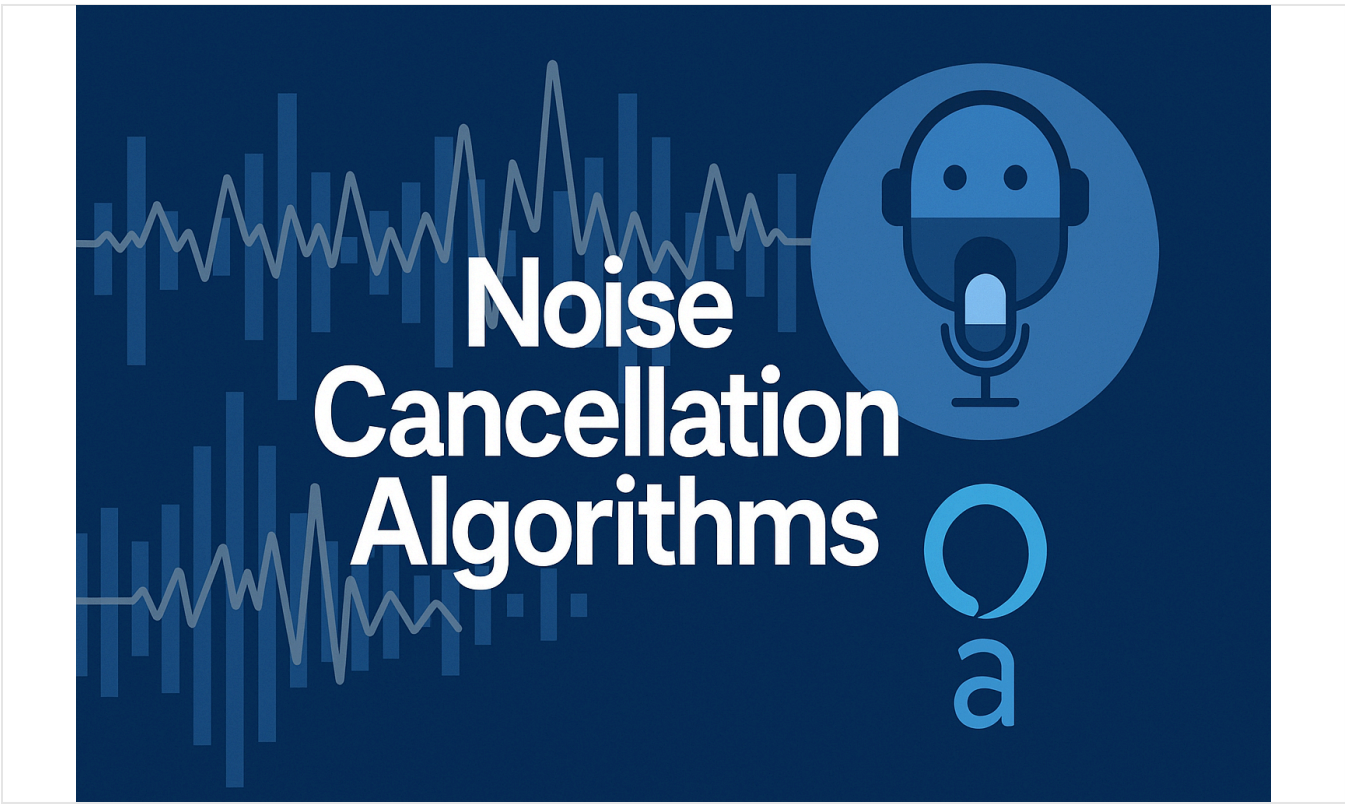
Noise Cancellation in Voice Bot Audio Pre-Processing

Published August 5, 2025 35 min read



State-of-the-Art Noise Cancellation in Voice Bot Audio Pre-Processing

Introduction

Voice-enabled bots and virtual assistants must contend with a variety of audio impurities in real-world environments, from background noise and reverberation to echoes of their own playback. Before speech-to-text (STT) conversion, an **audio pre-processing pipeline** is employed to enhance the speech signal and suppress undesired noise. This report surveys the state-of-the-art

techniques for real-time noise cancellation at this front-end stage. We cover both traditional digital signal processing (DSP) algorithms and modern AI-based methods for **noise suppression**, **acoustic echo cancellation (AEC)**, **dereverberation**, and general **voice enhancement**. We discuss open-source solutions (e.g. RNNoise, DeepFilterNet) and proprietary SDKs (NVIDIA Riva/Maxine, Google's pipeline, Microsoft's real-time enhancements, etc.), neural network architectures (RNNs, CNNs, Transformers, hybrid DSP-neural approaches), integration with automatic speech recognition (ASR) systems, as well as performance metrics, latency constraints, and robustness under challenging acoustic conditions.

Noise, Echo, and Reverberation Challenges

Background Noise (stationary hum or non-stationary sounds) can significantly degrade the intelligibility of speech and the accuracy of automatic speech recognition (ASR). **Room Reverberation** (multipath echoes of the speaker's voice) blurs temporal detail, while **Acoustic Echo** (the playback of a bot's own voice or audio picked up by its microphone) can completely overwhelm the desired speech input. In voice bots – especially far-field devices like smart speakers or conferencing systems – these effects are pronounced. The goal of the pre-processing stage is thus to suppress background noise and reverberation, and cancel any playback echo, **without distorting the user's voice**, all in **real time**.

Traditional Signal Processing Approaches

Early noise reduction in speech relied on DSP techniques such as spectral subtraction, Wiener filtering, and statistical model-based estimators (e.g. Ephraim-Malah filters). These typically follow the classic structure of **Voice Activity Detection (VAD)** → **Noise Spectrum Estimation** → **Spectral Filtering**, as illustrated below. The VAD detects when speech is present, the noise estimator updates a noise profile during non-speech segments, and a filter (subtractive or gain-based) attenuates frequencies dominated by noise. Such methods assume noise is relatively stationary; they struggle with transient or overlapping noises. They also involve many heuristic tuning parameters to avoid musical noise artifacts or speech distortion – a process described as "50% science, 50% art".

Conceptual diagram of a traditional noise suppression pipeline with VAD, noise spectrum estimation, and spectral subtraction.

For **echo cancellation**, classical acoustic echo cancellers use adaptive filters (e.g. normalized least-mean-square) to model the room impulse response between the loudspeaker (playing the bot's voice) and the microphone, subtracting the echoed signal. They often include double-talk detection logic (to handle when user speech and echo occur together) and nonlinear residual echo suppressors. **Dereverberation** methods in DSP include techniques like the weighted prediction error (WPE) algorithm, which uses linear prediction across frames to reduce late reverberation. Multi-microphone arrays further enable beamforming approaches – steering a spatial beam toward the user while nulling other directions. Beamformers such as delay-and-sum or MVDR improve signal-to-noise ratio by exploiting spatial diversity. These traditional approaches are effective under certain assumptions (linear echo paths, stationary noise, known array geometry) but often degrade in complex, dynamic acoustic scenarios.

Emergence of AI-Based Noise Suppression

In the last decade, **deep learning** revolutionized speech enhancement. Instead of relying on hand-tuned spectral rules, neural networks can learn to directly map noisy audio to clean speech. Early works (e.g. Xu et al. 2015) trained deep neural networks to predict a time-frequency mask (e.g. a ratio mask) to suppress noise. The typical pipeline involves mixing clean speech with noises to create a large synthetic dataset, training a model to output a mask or enhanced signal given noisy input, and then applying that model in real time. Unlike traditional methods, these learned models can handle highly non-stationary noises (typing keyboards, barking dogs, sirens, etc.) which "you cannot estimate in speech pauses" using classic noise profiling. By 2020, major tech platforms began deploying AI noise suppression: e.g. Google Meet's cloud denoiser and Microsoft Teams' new noise suppression both leverage supervised deep learning on extensive noise datasets rather than simple spectral subtraction.

RNNoise (2017) was a seminal open-source project that proved a small deep learning model could run in real-time on CPUs for noise suppression. RNNoise, by Jean-Marc Valin, uses a **recurrent neural network (RNN)** with Gated Recurrent Units (GRUs) to predict gains for 22 frequency bands (derived from a Bark-scale decomposition of the audio) instead of manipulating every FFT bin. This dramatically reduces model outputs and complexity, acting like a fast adaptive multi-band noise gate. The approach is **hybrid DSP/NN**: standard signal processing computes spectral features (and even basic VAD and spectral subtraction as a baseline), and the neural net learns the complex part – the "tricky tuning" of spectral gain for each band over time. The RNN (a GRU in this case) retains temporal context to distinguish speech from noise across frames. RNNoise runs with just ~10 ms

algorithmic latency and minimal CPU, yet yields quality better than traditional suppressors. It demonstrated that **real-time** deep learning noise suppression was feasible without a GPU – inspiring a wave of subsequent designs.

Since RNNoise, architectures have diversified. Many models use **convolutional neural networks (CNNs)** or **convolutional recurrent networks (CRNs)** to exploit local spectro-temporal structure. Others use full **U-Net architectures** (encoder-decoder CNNs) operating on spectrograms or even time-domain waveforms. For example, Microsoft's **Deep Noise Suppression (DNS)** research introduced models like **NSNet** (an LSTM-based suppressor) and later **GAN-based** and **Transformer-based** enhancers through global challenges. However, models with tens of millions of parameters are impractical for low-latency use without acceleration. Thus, there's intense focus on efficient designs: **DeepFilterNet** (Schröter et al. 2022) is one such state-of-the-art framework that marries perceptually motivated DSP with neural networks. It uses a two-stage process: first a network predicts **ERB-scaled spectral gains** to clean the spectral envelope (similar to RNNoise's band gains), then a second stage applies **deep filtering** to enhance periodic components of speech. By using separable convolutions, grouped layers, and exploiting human perception (e.g. coarse envelope vs fine harmonic structure), DeepFilterNet achieves excellent quality with low complexity. Its authors report outperforming prior complex mask-based methods at a fraction of the computational cost.

Another notable approach is Amazon's **PercepNet** (Valin et al., Interspeech 2020), which powers Amazon Chime's **Voice Focus** feature (Source: amazon.science). PercepNet explicitly incorporates perceptual considerations: for voiced speech (vowels with clear pitch), it applies a **pitch-tuned comb filter** to strip out noise between harmonics. By accurately tracking the fundamental frequency (via Viterbi decoding on frame-wise autocorrelation) even in noisy conditions, PercepNet uses classic DSP to remove a large portion of noise from periodic speech components. The neural network then has a lighter task – mostly handling unvoiced sounds and fine adjustments. This "do as little as possible with the DNN" philosophy yields a small model that runs in real time on a CPU using <5% of one core, yet achieved **state-of-the-art** results in the 2020 DNS Challenge real-time track. In fact, PercepNet ranked 2nd place (real-time category) in that challenge using only ~4% CPU, while a heavier Amazon model "PoCoNet" took 1st in offline category by utilizing a GPU for more intensive processing (including full complex spectral modeling) (Source: amazon.science). This illustrates a key trade-off: real-time systems favor lightweight or hybrid models, whereas offline/batch enhancement can afford large models that even estimate phase for maximum quality.

Today's cutting-edge noise suppressors often use **complex spectral masks or complex neural networks** (operating on both magnitude and phase of the STFT). Examples include deep complex U-Nets and Phase-aware GAN models that improve speech naturalness by correcting phase distortions – though these are typically heavier. We also see research into ** Transformer-based** models for speech enhancement. Transformers can model long-range dependencies (useful for consistent noise suppression over time), but their high computational cost has so far limited their real-time use. Instead, many real-time solutions stick to RNNs or TCNs (temporal convolutional networks) with carefully constrained sizes.

Importantly, all these models are trained on huge datasets of noisy speech. Microsoft, for instance, released an open dataset for training non-stationary noise suppression models (hundreds of noise types combined with thousands of speakers). Training is supervised, with the target being either the original clean speech or an oracle spectral mask. Because it's infeasible to get paired "clean/noisy" recordings in every environment, datasets are often synthetically generated by mixing clean speech corpora with recorded noise clips at various signal-to-noise ratios. This yields models that learn a general notion of "speech vs noise" that generalizes to new voices and noises not seen in training. The downside is that training requires significant compute (Microsoft's team trained on Azure GPU clusters for days to tune their model), but once trained, the model can run inference efficiently on consumer devices.

Acoustic Echo Cancellation and Dereverberation with AI

While noise suppression got an early head start with deep learning, **acoustic echo cancellation (AEC)** is now also being tackled with neural networks. Traditional AEC can struggle during **double-talk** (when far-end echo and near-end speech overlap) or when the echo path changes (e.g. moving speakers). Interspeech/ICASSP challenges in 2021–2023 have spurred hybrid and neural AEC solutions. A common approach is a two-stage system: a linear adaptive filter first performs coarse echo removal (modeling the bulk of the echo), then a neural **residual echo suppressor** eliminates any remaining echo while preserving the user's speech. Deep **complex-valued networks** (which handle magnitude and phase) have been used to improve AEC, as echoes carry distinct phase structure. For example, one challenge entry used a *Deep Complex Convolutional Recurrent Network* to jointly cancel echo and noise, outperforming traditional methods in the blind test set. These learned echo cancellers can adapt to nonlinearities or odd speaker distortion that classical linear filters cannot handle. However, like noise models, they must be small enough for real-time; thus researchers focus on pruning and quantizing models for deployment.

Dereverberation – removing room reverberation – has seen both classic and learned methods. Classic WPE is still a solid baseline for multi-mic dereverb in ASR pipelines. On the learning side, neural networks (often similar to noise suppression models) can be trained to predict the anechoic speech from reverberant input. Some enhancement models implicitly reduce reverberation as part of overall speech cleanup. Notably, PercepNet's design to suppress "noise and reverberation" (Source: amazon.science) suggests it treats late reverberation as another form of interference to attenuate. Indeed, **reverberation** appears as a smeared, noise-like component in the spectrogram, and DNNs can learn to reduce that smearing to some extent. There are also specialized models and datasets focusing on far-field reverberant speech enhancement, sometimes using room impulse simulations for training data. In practice, many voice bot devices rely on *acoustic design* (e.g. microphone arrays with beamforming) and on *downstream ASR robustness* to handle reverberation, since fully removing reverberation without harming speech can be challenging. Nonetheless, including a dereverb module or model before ASR has shown to improve word error rates for far-field speech recognition, especially in highly echoey environments.

Integration with ASR in Voice Bots

In a real-time voice bot, these pre-processing components operate *in tandem with* the ASR module. The typical processing chain is: **Microphone array capture → AEC → Beamforming/noise suppression → Voice activity detection → ASR decoding**. The noise suppression and echo cancellation dramatically improve the quality of audio that the speech recognizer sees, thus boosting accuracy. For instance, Amazon Alexa and Google Assistant devices use multi-microphone beamforming plus echo cancellation to achieve far-field voice recognition even while music is playing. As Sonos researchers note, a well-designed noise reduction front-end **"will enhance the performance of the smart speaker (wake word detection, word error rate, etc) by cleaning the speech signal."**.

One integration approach is to treat enhancement as an independent front-end module. For example, Zoom and WebEx allow a user to enable noise suppression which then filters the audio before any speech recognition or transmission. This modular approach is straightforward: the ASR system is largely agnostic to how the audio was enhanced (other than enjoying a higher SNR). Modern cloud ASR APIs (Google, Amazon, Microsoft) typically assume the user or device will perform some noise mitigation locally; indeed NVIDIA's Riva ASR pipeline offers a configurable **"background noise removal"** component in front of the speech recognizer.

Another approach is **joint modeling** or tightly coupled front-ends. Some research trains speech enhancement and ASR together, in a multi-task or end-to-end fashion. The enhancement network can be optimized to produce output that maximizes ASR accuracy (instead of, or in addition to, sounding good to human ears). This can sometimes yield even better recognition in noise, since the front-end learns to preserve just the information needed by the ASR. However, this is mostly seen in research prototypes (e.g. joint SE+ASR training in end-to-end models) and not yet the norm in deployed systems, due to training complexity and the desire to have a single enhancement module usable for both human listening and ASR.

A key point is that ASR engines themselves have become more noise-robust thanks to training on noisy, far-field data (so-called multi-condition training). For instance, Google's Recurrent Neural Network Transducer (RNNT) ASR model is trained on a wide range of noise types and reverberations. This means it can tolerate moderate noise without a pre-processor. Nonetheless, front-end noise suppression is still valuable in low-SNR scenarios. It *prevents* the ASR from mis-triggering on noise or missing words due to heavy noise masking. Google has reported that using a learned front-end denoiser can improve their speech recognition results, especially in extreme noise cases. In some cases, Google's production speech pipeline (for example in telephony or meeting settings) has run a server-side denoising model on the audio before feeding it to the recognizer – essentially inserting their AI "denoiser" as a preprocessing stage in the cloud.

There are trade-offs to consider: an overly aggressive noise canceller might **remove parts of speech** (e.g. truncate fricatives or low-energy phonemes thinking they are noise) and *worsen* ASR accuracy. Thus, designers tune these systems to balance noise removal with speech preservation. In fact, Microsoft explicitly does *not* filter out certain sounds like **music or laughter** in their dataset/model, treating them as desirable signals to pass through, because completely silencing those could be detrimental or unnatural in a conversation. The end goal, as Microsoft's team said, isn't zero noise at all costs – it's **making speech dominant and clear** while keeping the speech content intact. When integration is done well, users get the benefit of cleaner audio and the ASR gets a higher fidelity signal, all with minimal added latency.

Speaking of latency: integration requires that the enhancement step operate in streaming fashion. Typically these models work on short frames (e.g. 20 ms frames with maybe 10 ms overlap). The **processing latency** must be low (on the order of a few milliseconds beyond the frame size). As a rule of thumb, the **total end-to-end latency** for conversation should stay < 200 ms to avoid talk-over issues. If a noise suppressor added, say, 100 ms of algorithmic delay, it would severely impair interactivity. Thus, designers constrain frame lookahead and often align the enhancement frame size to the ASR's frame size or the audio packet size to avoid additional buffering delays.

RNNoise/PercepNet used 20 ms frames with 50% overlap, adding only ~10 ms extra delay from frame overlap processing. Microsoft's noise suppression in Teams was designed to process each 10–20 ms frame within that frame time on the client device. In cloud scenarios like Google Meet, the challenge was to run the DNN on TPUs fast enough that the network round-trip (capturing audio, sending to cloud, processing, returning audio) didn't exceed a few tens of milliseconds. Google achieved this by heavy optimization for their TPU inference and by doing noise suppression in parallel with other audio processing in Meet's servers. In summary, integration with ASR demands that noise cancellation be *streaming and low-latency*, often running in parallel on a separate thread or hardware (DSP or GPU) so as not to slow down the overall pipeline.

Neural Network Architectures and Hybrid Systems

A variety of neural network architectures are employed in these audio front-ends:

- **Recurrent Networks (RNNs)**: As discussed, gated RNNs like LSTM and GRU are popular for their ability to model temporal sequences. RNNoise's use of GRUs is a prime example, chosen for low compute and ability to remember noise characteristics over time. Many early speech enhancement models (e.g. DeepXi, Kagami) used LSTMs to estimate spectral masks frame by frame. Bidirectional RNNs (BLSTMs) can leverage future context but induce too much latency for real-time, so most real-time systems use unidirectional RNNs (only past context) or limited lookahead. RNNs remain popular in industry implementations due to their parameter efficiency and temporal smoothing capabilities.

- **Convolutional Networks (CNNs)**: CNNs (often in time-frequency domain) can exploit local structure in spectrograms. For example, a stack of convolution layers can act on a spectrogram chunk (covering a few frames of context) to predict a mask. CNNs are highly parallelizable on GPUs, which is great for cloud-based enhancement. NVIDIA's earlier neural noise reduction experiments used feed-forward and convolutional layers but found that to get high quality, the models became large, thus they pivoted to more efficient recurrent or hybrid schemes for real-time use (Source: developer.nvidia.com). Still, many challenge-winning models (e.g. CRUSE – a Convolutional Recurrent U-Net, or DCCRN) use CNN encoders/decoders combined with some recurrent units, marrying the strengths of both.

- **Transformers**: Transformer-based models (with self-attention) have been proposed for speech enhancement, benefiting from their ability to capture long-range dependencies. A Transformer could, in theory, learn the difference between speech and noise by attending over an entire utterance. In practice, their computation and memory footprint is a hurdle for real-time,

especially on edge devices. Some recent research (2022–2023) on streaming Transformers uses chunk-wise self-attention to limit latency. As of 2025, Transformers are more common in non-real-time enhancement or as a small component in a larger network. For instance, a *dual-path Transformer network* was tried in one of Microsoft's DNS challenge baselines, but simpler CNN-RNN hybrids still often win on the latency/quality trade-off. It will be interesting to watch if efficient attention mechanisms (or smaller "Tiny Transformers") find their way into mobile-friendly noise suppressors in the future.

- **Autoencoders and Generative Models**: Beyond discriminative models that directly predict clean speech, generative approaches like VAEs and diffusion models are emerging. A notable example is Facebook's **EnCodec**/"SoundStream" which included a noise reduction module as part of a neural codec pipeline. These models attempt to *generate* clean speech given noisy input. Diffusion models have shown excellent speech enhancement quality in offline tests, but they currently require iterative refinement steps that are too slow for real-time. Research is ongoing to distill such models for faster inference. GANs (Generative Adversarial Networks) have been used to make enhanced speech sound more natural (reducing the muffled or robotic artifacts sometimes heard with purely supervised models). For instance, Qualcomm's **AIMobile** GAN-based noise suppression targeted real-time mobile hardware. GAN training can make a model prioritize perceptual quality by fooling a discriminator that checks if output sounds real. However, stability and complexity are concerns, and many production systems still rely on straightforward L1/L2 loss trained models with carefully tuned post-filters.

- **Hybrid DSP-Neural Systems**: Many state-of-the-art systems integrate neural networks with traditional DSP in a complementary way. We've seen how RNNoise and PercepNet combine classic VAD, filtering, and pitch tracking with a neural mask estimator. Such hybrid systems leverage domain knowledge (e.g. human auditory perception, known physics of echos) to simplify the learning task. Another example is **beamforming+DNN**: use beamforming to get an initial spatially filtered signal, then apply a DNN to do residual noise removal. This is used in some far-field ASR front-ends, where a neural post-filter enhances the beamformer output (sometimes called a "neural beamformer"). Microsoft's recent research on a *directional ASR* model combined a spatial filter with a neural mask estimator to suppress competing talkers. Overall, hybrids are attractive because they allow small networks to focus on what's hard to solve analytically, while leaving easier or well-defined tasks to DSP. As Jean-Marc Valin put it, *"keep all the basic signal processing that's needed anyway, but let the neural network learn all the tricky parts that require endless tweaking"*. This philosophy underpins many production designs, ensuring efficiency and robustness.

Performance Benchmarks and Robustness

Objective metrics for speech enhancement include **PESQ** (Perceptual Evaluation of Speech Quality), **STOI** (Speech Intelligibility Index), and derived MOS (Mean Opinion Score) estimates. In the deep noise suppression literature, it's common to report improvements in PESQ (where 4.5 is clean speech) or STOI (0-1). For instance, one 2018 experiment by a startup (2Hz) cited an average MOS increase of 1.4 points with their deep learning model, whereas traditional single-mic DSP often *degrades* MOS in very noisy conditions. The lack of standardized benchmarks was an issue a few years ago, which prompted companies like Microsoft to release datasets and host challenges to let the community measure models on common data. The DNS Challenge (Interspeech/ICASSP) and recent AEC Challenge have provided such common evaluation: models are ranked by subjective quality ratings (crowd-sourced MOS) as well as objective scores. These benchmarks indicate that current top models can produce **remarkably high quality**: DNS Challenge winners often achieve MOS scores approaching or even exceeding 4.0 on noisy speech (where the noisy input might have MOS ~2.5). In terms of **intelligibility**, many models can bring a STOI of e.g. 0.6 (for noisy speech) up to 0.9+, which is near ceiling for intelligibility.

Latency is as critical a metric as quality for real-time voice bots. The processing algorithmic delay typically must be < 20 ms (some set 20 ms as an upper limit for the suppressor itself). Many frameworks use 10 ms frames with a lookahead of a few milliseconds. Microsoft's Teams noise canceler processes 20 ms frames entirely on the client side to avoid network delay. Google Meet's cloud approach had to use powerful TPUs to ensure the server-side model introduced minimal delay; they succeeded such that users did not notice added lag in meetings. When designing these systems, engineers consider the *entire pipeline latency*: audio buffers, codec frame size (if encoding is used), network hop (if applicable), etc., to budget how much time the noise suppression can take. They also offload heavy computation to hardware accelerators (GPUs, TPUs, NPUs) to keep CPU load low and latency consistent.

Robustness is another crucial consideration. A noise canceller should handle a wide range of acoustic conditions: from low SNR (~0 dB) factory noise to echoey chambers, from single stationary hums to multiple competing speakers. The best AI models tend to be very **generalized**, thanks to training on diverse data. For example, Microsoft's model was trained on 100+ noise types and thousands of speakers, and they even stress-tested with *unseen* noise types to ensure the model didn't overfit to specific training noises. Still, edge cases exist. Extremely low SNR (< -5 dB) might result in speech being lost. Certain noises that resemble speech (e.g. a radio/TV in background, or another person talking) are tricky – the model might mistakenly keep another voice thinking it's the

user, or conversely, remove parts of the target speech thinking it's noise. Some companies explicitly choose not to remove *music* or *laughter* and the like, as it can be contextually important (or in the case of music, users might prefer it not be completely stripped out in a call).

Another robustness aspect is **artifact avoidance**. Early neural suppressors sometimes introduced warbling or "robotic" artifacts, especially under high suppression levels. NVIDIA's RTX Voice (a GPU-based noise removal for streamers) initially had reports of the voice sounding a bit artificial or "underwater" when it aggressively filtered complex noise. Ongoing work in loss functions (perceptual loss, etc.) and post-processing (like adding a tiny bit of the noise back to avoid dead-silence feeling) aim to make the output sound natural. The Amazon PercepNet approach of *retaining some noise* intentionally was to avoid an unnaturally sterile sound and to prevent phase inconsistencies (Source: amazon.science). Subjective user testing is invaluable here – sometimes a small residual noise is preferable to a distortion of speech.

Far-field and multi-mic scenarios deserve special mention for robustness. Voice bots in smart speakers or conferencing phones often use multiple mics spaced apart. This allows **spatial noise reduction** (beamforming), which can dramatically improve quality by ~5-10 dB before the signal even hits the neural suppressor. Many commercial solutions combine beamforming with DNNs: e.g. a fixed or adaptive beamformer outputs a primary signal and possibly some reference signals; a neural network then processes them. In fact, one common strategy in industry is a **neural beamformer** such as the GSC (Generalized Sidelobe Canceller) with a DNN-based blocking matrix or mask estimator. These advanced methods can jointly handle directional interference and ambient noise. Far-field robustness also means the system deals with more reverberation (since the mic is not close to the mouth). Some products include **dereverberation filters** or train their DNN to handle reverberation explicitly. Microsoft's latest Azure Cognitive Services include a pre-processing stack that does AEC, noise suppression, and dereverberation, providing cleaner audio for their cloud ASR (particularly targeting meeting scenarios with speakerphones).

Finally, a practical performance factor is **computational load**. On a smartphone or embedded device, the CPU and battery impact are important. Many AI noise suppression solutions are optimized for specific hardware: Qualcomm's Snapdragon chips have Hexagon DSP blocks that can run noise reduction at low power; Apple's Neural Engine likely powers their AirPods Pro adaptive noise reduction. NVIDIA's Maxine audio effects leverage GPU acceleration to run multiple enhancements in parallel in the cloud. Metrics like CPU usage (percentage of a core) or memory footprint (model size in MB) are closely watched. Amazon's PercepNet bragged it used only 4% of a CPU core; Mozilla's RNNoise can run on a Raspberry Pi or even microcontrollers for simple use

cases. As AI hardware becomes more common even in low-power devices, the feasibility of running more complex models at the edge increases, but currently most deployments still carefully balance complexity to meet real-time constraints without draining resources.

Real-World Deployments and Examples

All major players in voice technology have adopted these noise cancellation techniques in their products:

- **Google**: In 2020, Google rolled out an AI-based noise suppressor in **Google Meet** for G Suite users. This "denoiser" runs in Google's cloud, on TPUs in their datacenters. It was developed by the team from Google's 2017 acquisition of Limes Audio, combining their audio DSP expertise with Google's AI prowess. The model can intelligently filter out typing sounds, paper rustling, door slams, dog barks, etc., while keeping speech. Notably, it operates on the server side: Meet sends audio to the cloud, cleans it, then sends it to meeting participants. This showcases Google's confidence in their network and TPU latency to handle real-time. Google has also integrated noise suppression in Android (for instance in Gboard's voice input and phone calls) but details are less public. For ASR, Google's assistant and voice search rely more on robust modeling than on heavy front-end processing (since they expect users on phones may not have advanced algorithms locally). However, Google's **Pixel phones** do include multiple microphones and use classic noise reduction during calls (the Pixel 2016 had dual microphones specifically for noise canceling in phone calls). For far-field devices like **Google Home/Nest** smart speakers, Google uses beamforming and presumably some denoising to achieve good recognition in noisy rooms (though the specific algorithms are not publicly detailed). In summary, Google applies noise cancellation in communication (Meet) and likely in devices, often with AI models. Their **RNNT ASR pipeline** has been documented to include robust features; a Google research paper in 2014 already showed using a deep autoencoder to preprocess noisy features improved an RNN-T's recognition accuracy. We can infer that lessons from such research have informed Google's production pipelines.

- **Microsoft**: Microsoft has invested heavily in real-time voice enhancement for **Microsoft Teams** and related products. In late 2020, they announced an AI-based **Real-Time Noise Suppression** in Teams, which shipped soon after. This feature runs locally in the Teams app (Windows, etc.), using a DNN to suppress non-stationary noises like typing and crunching that previous Skype-era noise reducers couldn't handle. Microsoft's team (led by Robert Aichner) described building a large training set and carefully optimizing the model to run on typical CPUs. They decided to

do it on the **edge (client-side)** for both privacy and latency reasons – sending every call's audio to the cloud for cleaning would add delay and cost, and peer-to-peer calls in Teams might not even touch a server. The model is likely based on their DNS challenge work (possibly an improved version of NSNet2). Microsoft also integrated AI noise removal in **Surface devices and Windows** (the "Voice Focus" feature uses AI to filter background noise in any app, if using a supported device). Additionally, Microsoft Azure offers **Cognitive Services** for speech that include noise suppression and echo cancellation for developers to preprocess audio. On the far-field front, Microsoft's **Xbox Kinect** and **HoloLens** had advanced multi-mic processing; more recently their **Teams Rooms** devices use AI-based audio processing for echo/noise in conference rooms. Microsoft's research and patents in this area are extensive, including a patent (CN109841206A) on adapting a deep neural network for speech enhancement in a recognition system. In short, Microsoft uses these techniques across enterprise and consumer scenarios, often branding it as making calls sound more professional by "AI filtering out" unwanted sounds.

- **Amazon**: Amazon's use of noise cancellation spans both **communication** (AWS Chime/Voice Focus) and **voice assistants** (Alexa). We've detailed Amazon Chime's **Voice Focus**, which uses PercepNet to handle noise and reverb in meetings in real-time (Source: amazon.science). This is offered to AWS developers as well (Amazon has an SDK so others can use Voice Focus in their own apps). On the Alexa side, far-field speech recognition is a core competency: Echo devices have a 7-microphone array and perform beamforming, noise reduction, and AEC to hear "Alexa" wake word even during loud music playback. While Amazon hasn't published all details, they did present technologies like **PoCoNet** (likely a powerful DNN for speech enhancement possibly used offline for improving training data or for super-resolution). Alexa's wake word engine and ASR certainly benefit from the front-end: array processing algorithms (some originated from the DARPA EAR program and Amazon's own research) ensure that even if you shout to Alexa from across a noisy room, it can isolate your voice. Amazon's lab 126 and Alexa AI teams have numerous papers on multi-microphone speech enhancement and acoustic modeling for far-field. Additionally, Amazon's *Transcribe* API (ASR service) has an option for a **"channel synthesis" and noise reduction** when multiple mics or a stereo signal is provided, indicating they apply some form of noise suppression in the cloud ASR pipeline as needed.

- **NVIDIA**: NVIDIA provides **Riva**, an SDK for speech AI, and **Maxine**, a cloud-AI suite for audio/video enhancements. As part of Maxine, NVIDIA has **Background Noise Removal** and **Room Echo Removal** features that developers can integrate. These likely leverage models similar to those in RTX Voice/NVIDIA Broadcast (which are GPU-accelerated noise removal tools originally made for gamers and streamers). **RTX Voice**, launched in 2020, showcased an AI model that could strip out virtually all background sounds (keyboard, fans, etc.) from a live

microphone feed, using the Tensor cores on an NVIDIA GPU. It was essentially a high-end real-time noise suppressor, reportedly using a model trained on many noise types. It impressed users by even removing music or other voices to isolate the main speaker. RTX Voice has since been integrated and rebranded as part of **NVIDIA Broadcast** (for consumer) and as **Maxine Audio Effects** for developers. The technology was also offered in NVIDIA's collaboration with enterprise partners – for instance, **Cisco Webex** integrated NVIDIA's noise removal for users on NVIDIA GPUs as an option. NVIDIA Riva, while primarily focusing on ASR/TTS, can incorporate these enhancements so that an end-to-end pipeline might be: audio -> noise removal -> ASR (Conformer or Transformer-based) -> text. In fact, NVIDIA's documentation mentions the use of a "neural VAD to filter out noise" and the availability of pre-trained models for noise removal in the Riva framework. Real deployment example: **Tencent Meeting** (Chinese Zoom equivalent) uses NVIDIA Maxine in the cloud for noise cancellation and other effects. **PolyAI**, a startup building voice agents, uses NVIDIA Riva and specifically notes noise-optimized speech pipelines for robust understanding. NVIDIA's competitive edge is having GPU acceleration which allows running heavier models in real time, so they often tout being able to chain multiple enhancements (noise removal + dereverb + super-resolution) for "studio quality" audio in live calls.

- **Zoom**: At the start of the pandemic, Zoom famously licensed technology from **Krisp**, an AI noise suppression startup, to provide noise cancellation in Zoom calls. Krisp's technology – based on deep learning – could remove background voices and noises on-the-fly and was packaged as an SDK. By late 2020, Zoom introduced its own built-in **"Background Noise Suppression"** with levels (Low, Medium, High) in the settings. It's believed that Zoom's native solution is influenced by or built on the techniques pioneered by Krisp (Zoom hired some folks in that area). Zoom's noise suppression includes **echo cancellation, noise reduction and gain control** for Zoom Rooms. Users have noted that on the highest suppression setting, Zoom will aggressively filter even music and could create a slight lisp on speech, suggesting a strong ML model at work. Zoom allows "Original Sound" mode (which disables noise suppression) for musicians, implying that their default noise suppression would cut out music otherwise. In summary, Zoom leveraged state-of-the-art AI noise removal (first via Krisp, then in-house) to improve call quality for millions of users. This was a key differentiator during the work-from-home boom, in competition with Google Meet and Microsoft Teams, which were also rolling out AI noise canceling.

- **Cisco/Webex**: Cisco acquired a startup called **BabbleLabs** in 2020 specifically for its AI noise removal tech (Source: rtcweb.in)(Source: rtcweb.in). BabbleLabs had developed a deep learning based noise elimination that could be implemented on-device or cloud. After acquisition, Cisco integrated it into **Webex Meetings** as "Speech Enhancement" that removes

background noise for meeting participants. Cisco's marketing showed examples of eliminating construction noise and dog barks using the BabbleLabs AI. Now, every Webex user benefits from that whenever they enable noise removal (and by default in many cases). This shows how valuable the technology was considered – Cisco chose acquisition to keep pace with competitors in the collaboration space.

- **Discord**: The popular gaming chat app **Discord** partnered with Krisp as well in 2020 to add a noise suppression option (so that gamers' noisy backgrounds or clacky keyboards don't annoy their teammates). This brought AI noise cancellation to millions of users rapidly. Discord later made this available on mobile too, using the mobile-optimized version of Krisp's library. This is an example of a direct product integration of a third-party AI SDK.

- **Others**: Virtually every player in voice comms has moved to adopt these technologies. **Apple**, known for its integrated hardware-software approach, uses advanced audio processing in AirPods and iPhones. While much of Apple's work is not public, the latest AirPods Pro use computational algorithms for **Adaptive ANC and noise reduction** that likely involve ML (as hinted by the audioXpress article on Apple leveraging AI for audio personalization and real-time enhancement). Apple's FaceTime and phone call noise reduction (like "Voice Isolation" mode introduced in iOS 15) are AI-based, focusing on your voice and cutting background sounds using on-device neural nets. **Meta (Facebook)** uses AI noise suppression in Messenger and Portal devices. **Tencent** and **Alibaba** have their own versions for their communication platforms. Even niche areas like call centers (e.g. Cisco's contact center solutions, Amazon Connect) use AI to reduce customer and agent background noise for clearer calls – sometimes termed "voice optimization".

In the automotive realm, companies are looking at AI noise suppression for in-car voice assistants (to deal with road noise). And startups are innovating too: e.g. **Sanas** is a company doing real-time voice enhancement including noise reduction and even accent conversion, targeting call centers. **Altered** and others offer voice changers with built-in noise cleanup for streamers.

In summary, what was once a novel lab experiment (neural network noise removal) is now ubiquitous. From videoconferencing to smart homes to mobile voice interfaces, state-of-the-art noise cancellation – powered by advanced signal processing and AI – has become an expected feature for any voice-driven technology.

Conclusion

The audio pre-processing stage of voice bots has undergone a remarkable transformation with the infusion of AI. Classic DSP algorithms (spectral subtraction, Wiener filters, beamformers, etc.) laid the groundwork in earlier decades, but they are being augmented or overtaken by **deep learning models** that can suppress noise, cancel echo, and reduce reverberation with a new level of effectiveness. The state-of-the-art solutions today often blend the best of both worlds: **neural networks guided by signal processing principles**, tuned for real-time performance. Techniques like RNNoise and PercepNet pioneered the hybrid approach, achieving strong noise reduction on tiny models by leveraging human speech characteristics. Building on that, larger but more optimized models like DeepFilterNet and others push toward near-clean speech even in very noisy environments – all while respecting the stringent latency and compute limits of live voice interaction.

We have seen how industry leaders deploy these advances: reducing background clatter in conference calls, enabling far-field devices to wake up only to actual human commands, and ensuring AI assistants hear us accurately regardless of the acoustic mess around us. **Performance benchmarks** indicate that current systems can drastically improve perceptual quality (MOS) and intelligibility (STOI) of noisy speech, though careful evaluation and tuning are ongoing to avoid artifacts. **Latency and resource usage** remain top considerations – the best algorithm is useless if it introduces too much delay or drains a device's battery. Therefore, research continues into model compression, efficient architectures, and hardware acceleration to make **"AI noise cancellation" ubiquitous and invisible** in operation.

Looking ahead, we anticipate even more integration of these front-ends with core ASR and **adaptive behavior** based on context. For example, a voice bot might dynamically adjust its noise suppression aggressiveness based on detection of whether the user is speaking or whether it's playing music itself. The field is also expanding beyond noise into general **voice enhancement** – not just removing undesirable sounds, but actively improving the desired voice (making it sound as if on a high-quality mic). NVIDIA's "Studio Voice" and others hint at this, where the AI can make a cheap laptop mic sound like a studio recording by dereverb, EQ, and noise cleanup. In telecommunications, we see the concept of **"HD voice"** being taken further with AI, to deliver crystal-clear, noise-free calls that far exceed traditional phone audio quality.

In conclusion, **noise cancellation for voice bots** has reached a sophisticated state of the art, combining decades of signal processing know-how with cutting-edge deep learning. The result is a new generation of voice interfaces that can operate robustly in everyday environments – whether it's a busy home, a noisy street, or a reverberant office – thus bringing us ever closer to seamless,

ubiquitous voice interaction. As one audio industry leader aptly stated, *"AI is no longer an edge case for audio. It's the new baseline."* The technologies described in this report are prime examples of that baseline in action, ensuring voice bots hear us clearly and we hear them, no matter the acoustic challenges in between.

References: *(The report content is based on information from numerous sources, including academic papers, industry whitepapers, and technical blogs by experts at Amazon, Microsoft, NVIDIA, and others. In-line citations in the format source † lines correspond to the specific sources and line ranges from which the information was drawn.)*

Tags: audio pre-processing, noise suppression, acoustic echo cancellation, voice enhancement, digital signal processing, neural networks, automatic speech recognition, voice bots

---

About ClearlyIP

ClearlyIP Inc. — Company Profile (June 2025)

---

1. Who they are

ClearlyIP is a privately-held unified-communications (UC) vendor headquartered in Appleton, Wisconsin, with additional offices in Canada and a globally distributed workforce. Founded in 2019 by veteran FreePBX/Asterisk contributors, the firm follows a "build-and-buy" growth strategy, combining in-house R&D with targeted acquisitions (e.g., the 2023 purchase of Voneto's EPlatform UCaaS). Its mission is to "design and develop the world's most respected VoIP brand" by delivering secure, modern, cloud-first communications that reduce cost and boost collaboration, while its vision focuses on unlocking the full potential of open-source VoIP for organisations of every size. The leadership team collectively brings more than 300 years of telecom experience.

---

2. Product portfolio

- **Cloud Solutions** – Including *Clearly Cloud* (flagship UCaaS), **SIP Trunking**, **SendFax.to** cloud fax, **ClusterPBX OEM**, **Business Connect** managed cloud PBX, and **EPlatform** multitenant UCaaS. These provide fully hosted voice, video, chat and collaboration with 100+ features, per-seat licensing, geo-redundant PoPs, built-in call-recording and mobile/desktop apps.

- **On-Site Phone Systems** – Including CIP PBX appliances (FreePBX pre-installed), ClusterPBX Enterprise, and Business Connect (on-prem variant). These offer local survivability for compliance-sensitive sites; appliances start at 25 extensions and scale into HA clusters.
- **IP Phones & Softphones** – Including CIP SIP Desk-phone Series (CIP-25x/27x/28x), fully white-label branding kit, and *Clearly Anywhere* softphone (iOS, Android, desktop). Features zero-touch provisioning via Cloud Device Manager or FreePBX "Clearly Devices" module; Opus, HD-voice, BLF-rich colour LCDs.
- **VoIP Gateways** – Including Analog FXS/FXO models, VoIP Fail-Over Gateway, POTS Replacement (for copper sun-set), and 2-port T1/E1 digital gateway. These bridge legacy endpoints or PSTN circuits to SIP; fail-over models keep 911 active during WAN outages.
- **Emergency Alert Systems** – Including **CodeX** room-status dashboard, **Panic Button**, and **Silent Intercom**. This K-12-focused mass-notification suite integrates with CIP PBX or third-party FreePBX for Alyssa's-Law compliance.
- **Hospitality** – Including **ComXchange** PBX plus PMS integrations, hardware & software assurance plans. Replaces aging Mitel/NEC hotel PBXs; supports guest-room phones, 911 localisation, check-in/out APIs.
- **Device & System Management** – Including **Cloud Device Manager** and **Update Control (Mirror)**. Provides multi-vendor auto-provisioning, firmware management, and secure FreePBX mirror updates.
- **XCast Suite** – Including Hosted PBX, SIP trunking, carrier/call-centre solutions, SOHO plans, and XCL mobile app. Delivers value-oriented, high-volume VoIP from ClearlyIP's carrier network.

---

3. Services

- **Telecom Consulting & Custom Development** – FreePBX/Asterisk architecture reviews, mergers & acquisitions diligence, bespoke application builds and Tier-3 support.
- **Regulatory Compliance** – E911 planning plus **Kari's Law**, **Ray Baum's Act** and **Alyssa's Law** solutions; automated dispatchable location tagging.
- **STIR/SHAKEN Certificate Management** – Signing services for Originating Service Providers, helping customers combat robocalling and maintain full attestation.
- **Attestation Lookup Tool** – Free web utility to identify a telephone number's service-provider code and SHAKEN attestation rating.
- **FreePBX® Training** – Three-day administrator boot camps (remote or on-site) covering installation, security hardening and troubleshooting.
- **Partner & OEM Programs** – Wholesale SIP trunk bundles, white-label device programs, and ClusterPBX OEM licensing.

---

4. Executive management (June 2025)

- **CEO & Co-Founder: Tony Lewis** – Former CEO of Schmooze Com (FreePBX sponsor); drives vision, acquisitions and channel network.
- **CFO & Co-Founder: Luke Duquaine** – Ex-Sangoma software engineer; oversees finance, international operations and supply-chain.
- **CTO & Co-Founder: Bryan Walters** – Long-time Asterisk contributor; leads product security and cloud architecture.
- **Chief Revenue Officer: Preston McNair** – 25+ years in channel development at Sangoma & Hargray; owns sales, marketing and partner success.
- **Chief Hospitality Strategist: Doug Schwartz** – Former 360 Networks CEO; guides hotel vertical strategy and PMS integrations.
- **Chief Business Development Officer: Bob Webb** – 30+ years telco experience (Nsight/Cellcom); cultivates ILEC/CLEC alliances for Clearly Cloud.
- **Chief Product Officer: Corey McFadden** – Founder of Voneto; architect of EPlatform UCaaS, now shapes ClearlyIP product roadmap.
- **VP Support Services: Lorne Gaetz** (appointed Jul 2024) – Former Sangoma FreePBX lead; builds 24×7 global support organisation.
- **VP Channel Sales: Tracy Liu** (appointed Jun 2024) – Channel-program veteran; expands MSP/VAR ecosystem worldwide.

---

5. Differentiators

- **Open-Source DNA:** Deep roots in the FreePBX/Asterisk community allow rapid feature releases and robust interoperability.
- **White-Label Flexibility:** Brandable phones and ClusterPBX OEM let carriers and MSPs present a fully bespoke UCaaS stack.
- **End-to-End Stack:** From hardware endpoints to cloud, gateways and compliance services, ClearlyIP owns every layer, simplifying procurement and support.
- **Education & Safety Focus:** Panic Button, CodeX and e911 tool-sets position the firm strongly in K-12 and public-sector markets.

In summary

ClearlyIP delivers a comprehensive, modular UC ecosystem—cloud, on-prem and hybrid—backed by a management team with decades of open-source telephony pedigree. Its blend of carrier-grade infrastructure, white-label flexibility and vertical-specific solutions (hospitality, education, emergency-

compliance) makes it a compelling option for ITSPs, MSPs and multi-site enterprises seeking modern, secure and cost-effective communications.

DISCLAIMER