Implementing voice moderation

Do you have audio communication on your platform? Are you worried about players abusing your audio communication and violating your community guidelines? Voice moderation is a challenge for everyone as it is a nascent concept - everything from legal compliance, policy creation, and tooling are just now emerging.

You need a partner that can provide a technology solution and strategic plan to help achieve your Trust & Safety goals, while learning and growing with you.

Spectrum has a 4+ year successful track record in providing content moderation solutions. We offer multiple behavior identification methods to solve Trust & Safety use cases in text and voice content, now and into the future.

Voice moderation has unique challenges

Accuracy is hard to achieve

Transcription to text loses context

Non-native accents & speech impairments

Cost of moderation is high

Storing & processing large audio data files is expensive

Transcription + text analysis means paying for two services

Privacy & security risks are greater

Users are accustomed to a high level of privacy in audio

Legal best practices are still being defined

How to prepare

Before you can start behavior detection in voice content, you'll need cross-team alignment on:

Question	Considerations
Do you have access to your audio data?	Is your audio service built in-house, or outsourced? If outsourced, does your partner provide direct access to raw files?
What do you want to record?	Moderating voice requires recording the audio. You must have a file to analyze; creating the file = recording voice content. When do you want to start recording the content? Will it be all content, or only a subset?
Do your privacy policies cover recording?	Users expect more privacy in voice. Proactive detection may shift the perception of privacy. Coordinate with legal and privacy to adjust terms.
What do you want to moderate for?	Trade-offs in cost and privacy must be worth the benefits. May not be worth it: everyday swear words. Worth it: worst-of-the-worst: hate speech, child grooming, etc.
What will you do when you detect problems?	Mute, warn, suspend, or ban. Will the action be applied to a single user or at a different level?
What data do you need to store to support moderation decisions?	To review and take action, then delete. To keep a trail of evidence for reporting to authorities.

Define your goals to determine a path forward

Integrating a voice moderation solution involves two components:

→ **Behavior Identification**
Which behaviors do you want to detect, and in what languages?

→ **Execution**
When do you need to detect those behaviors - in real-time or afterward?

The industry offers two approaches for voice moderation.

	Batch Processing	Client-Side SDK
Description	Introduce audio moderation with low integration effort.	Deliver real-time detection and punishment.
Behavior Identification	Full behavior library covering high-risk areas such as hate speech and child safety, across multiple languages.	Due to size constraints this approach can only cover limited behaviors, and use cases tend to focus on tonality, not safety. Typically English-only.
Execution	Batch Processing + Asynchronous Analysis	Client-Side SDK + Near-Real-Time Analysis
Use Case	Post-match analysis & player report validation: stop known bad actors from continuing behaviors.	In-game analysis: mute users in real time.
Pros	Validation & efficiency for moderators. Mute bad actors for a more engaging community. Easy integration: send all data to technology partner once; receive a response. Store evidence for compliance.	Proactive moderation with near-real-time detection and action. Prevent the very worst: self-harm, child sexual abuse material (CSAM), sexual harassment, etc. Most effective for long-term community safety.
Cons	No real-time detection.	Specific device limitations around CPU, latency, etc. limit what can be detected. More difficult integration. Potential privacy concerns: users know you're listening to live conversations.

Spectrum's solutions & timeline

Batch processing to detect an expansive list of behaviors is available now to protect your community today.
Our client-side SDK will be available soon for best-in-class proactive detection, creating enduring value for your community.

	Now	Next	Future
Behavior Identification	Direct analysis on short audio snippets to detect profanity, etc. Transcription + Text Classifiers	Detect complex behaviors directly on the raw audio	More behaviors & languages
Execution	Batch/Asynchronous (near-real-time)	Client-side SDK to perform more analysis on the device	Enhanced pre-processing
Use Case	Validate user reports to save moderators time	Proactive detection	Real-time recognition & response

How Spectrum's voice solutions are different

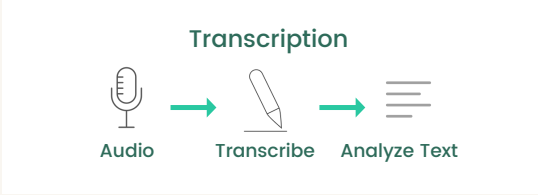
1. Direct-on-audio analysis

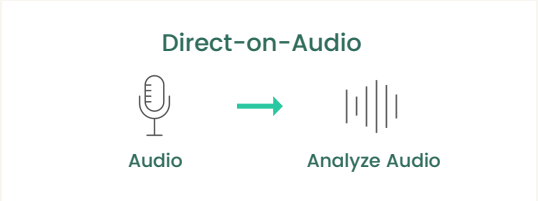
We can do transcription-based detection, but we also directly analyze short audio snippets, skipping the transcription step.

When others transcribe audio files to text and analyze the text, this results in:

- Twice the processing cost
- Transcription inaccuracies
- Slower return of analysis.

Direct-on-audio analysis considers important information that adds meaning to words: tone, pitch, range, volume, rhythm, and tempo.





2. Holistic approach to speech

We train our models on how general speech sounds, not just how individual words sound. Because we aim to learn holistic speech and don't rely on transcription, we can handle a wider range of use cases with accuracy.

Kids & Adults

Background Noise

Devices & Audio Setups

Accents & Intonations

3. Unique toolset for better detection

We've developed specialized tools that result in a better voice content moderation product. This benefits customers by detecting harmful behaviors more accurately and thoroughly in real-life voice Trust & Safety use cases.

Our toolset improves detection at every step of the process: data collection & labeling, model training, and audio-specific detection challenges.

Proprietary Tools

To collect & label audio data

Specific Labeling Lexicon

To ensure consistency

Distorts Training Data

For devices & audio setups

Introduces Noise & Variation

To train models

Multi-Faceted Feature Set

To look beyond words to tone, pitch, range, volume, rhythm, tempo

4. After we detect harmful behavior

You decide the action, we can help you implement it.

→ **Player Punishment**
Mute temporarily
Send warning
Ban player

→ **Moderator Review**
Send to your moderation queue
Moderators see the behavior: what was said & when
Further action to NCMEC, etc.

How to implement Spectrum's voice moderation

Our implementation team works with you through scoping sessions to determine technical details and timeline.

Data to send to Spectrum	
Content	Voice chats, voice messages, calls, streams etc.
Metadata	Source or game, private or public, stream, time of day, etc.
Delivery	Asynchronous or batch

Data Spectrum sends back to you	
Flagged Behavior	Which behavior was identified in the audio
Time Window	Where in the audio clip the behavior occurred
Transcription	Transcript of the occurrence for moderator review and evidence collection

Teams involved

Engineering

Integrates Spectrum's technology with your platform, including sending all data, receiving results and actioning accordingly.

+

Policy/Moderation

Advises how to interpret behavior results; ensures model performance is tuned for a positive player experience.

Major implementation steps

1. **Start sending data into Spectrum for analysis**
Engineering

2. **Greenlight model performance to start using results**
Policy/Moderation

3. **Start issuing player penalties based on model results**
Policy/Engineering