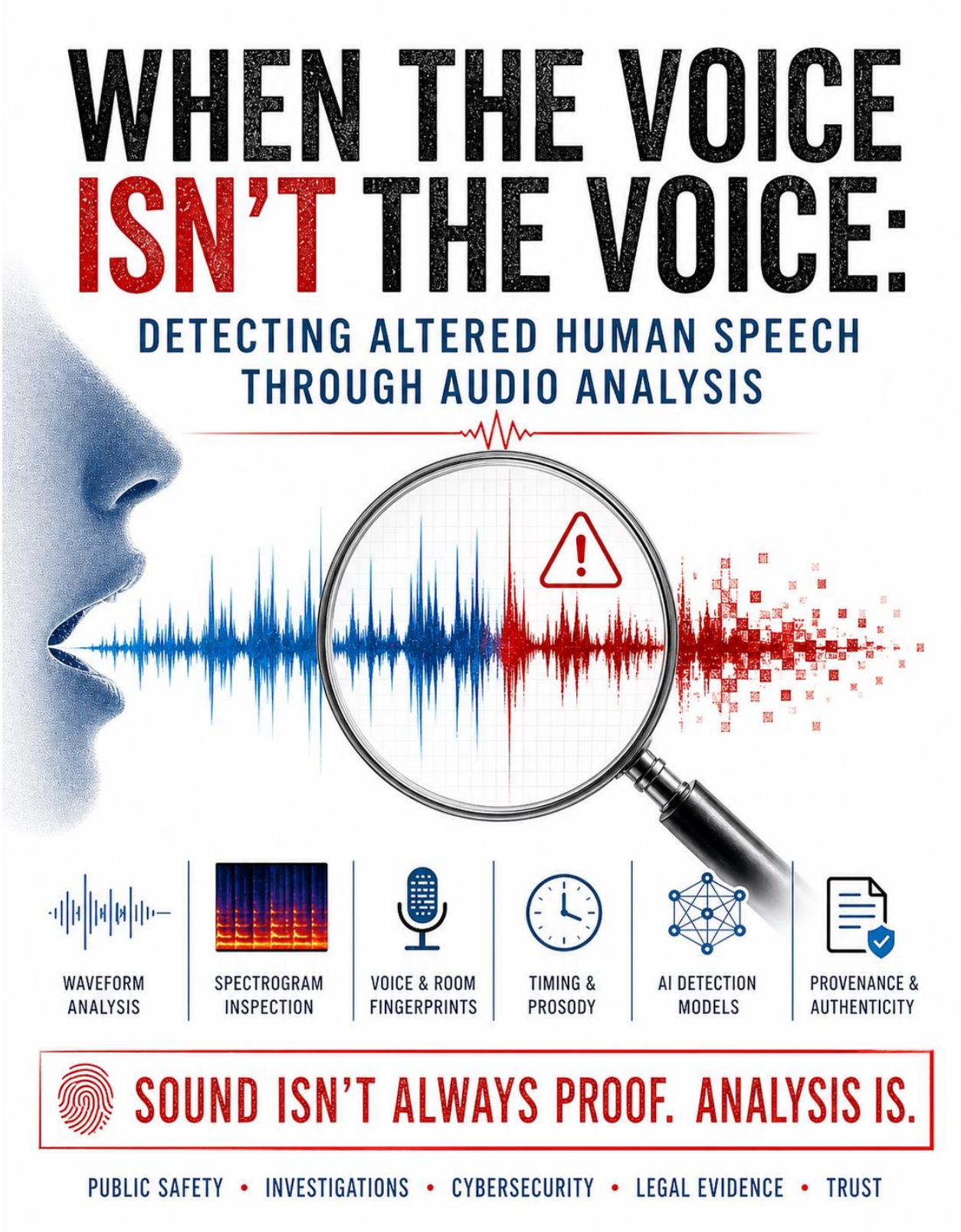

TABLE OF FIGURES

A quick note about this guide, before we dive in...

The reader acknowledges that this document is provided for informational and educational purposes only. While reasonable efforts have been made to ensure accuracy, no warranty or guarantee, express or implied, is made regarding the completeness, reliability, or accuracy of the information presented.

The content was assembled from publicly available sources, along with professional experience and general industry knowledge. Artificial intelligence tools, including those from OpenAI and others, were used to assist with content organization, analysis, and graphic generation.

Source references have been provided both inline and at the end of this document where applicable. This material does not constitute original research in all cases and should not be considered authoritative without independent verification.

The views and interpretations presented are those of the author and are intended to provide a practical, consolidated perspective on the topic.

Readers are strongly encouraged to independently validate any information that may impact operational, legal, or policy decisions. This document is not intended to provide legal, technical, or operational advice.

Mark J. Fletcher, ENP #1206
Founding Principal, FLETCH911, LLC
Fletch@Fletch911.com • (973) 826-9111

# # #

TABLE OF CONTENTS

Detecting Altered Human Speech Through Audio Analysis

We've historically treated recorded speech like a trusted witness. If someone said it, and we had the audio, that was close to the end of the argument.

Welcome to the uncomfortable little corner of the future where the voice on the recording may be real, modified, spliced, replayed, cloned, synthesized, or some lovely Frankenstein buffet of all of the above.

Today, I'd like to talk about how audio that has been altered from normal human speech may be detected through audio analysis, and why this is becoming a very real issue, not just for public safety, but also for investigations, cybersecurity, legal evidence, emergency communications, and anyone who still believes that "I heard it with my own ears" is a reliable authentication strategy.

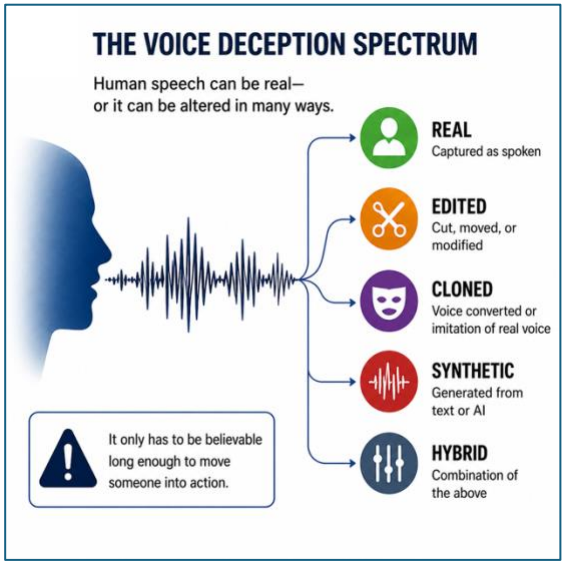

Figure 1 - The Voice Detection Spectrum

***Spoiler alert:* your ears are useful.**
They are just not enough.

For decades, voice recordings had a certain authority. The voice sounded like the person. The emotion sounded real. The cadence sounded believable. For 911 calls, they sounded urgent. And in public safety, urgency has a way of bypassing skepticism because nobody wants to be the person who hesitates when someone might be in danger.

But now we are living in a world where human speech can be edited, pitch-shifted, time-stretched, voice-converted, replayed, cloned, and generated from text. The Federal Trade Commission has warned that voice cloning can be used in family emergency scams and business impersonation schemes, including cloned voices of loved ones or company executives used to trick people into sending money or paying fraudulent invoices. The FTC also makes the larger point that this risk cannot be solved by technology alone. Their Voice Cloning Challenge states plainly that voice cloning risks "cannot be addressed by technology alone." ***(Federal Trade Commission https://www.ftc.gov/news-events/contests/ftc-voice-cloning-challenge)***

That matters. Because this is not just an audio engineering problem. It is an operational problem. It is a policy problem. It is a chain-of-custody problem. It is a training problem. And for emergency communications, it is a HUGE trust problem.

The New Problem With "I Know What I Heard"

The human voice has always carried emotional weight. A written message may be questioned. A text may be misread. But a voice? A voice feels personal. It carries panic, anger, fear, authority, grief, confusion, intoxication, hesitation, and sometimes deception.

That is exactly why fake or altered speech is so dangerous.

Figure 2 - I Know What I Heard

An altered voice does not have to be perfect. It only has to be believable long enough to move someone into action. In a public safety environment, that action could be dispatching a tactical response. In a corporate environment, it could be authorizing a transfer of funds. In a family scam, it could be sending money to someone pretending to be a child or grandchild. In a legal setting, it could influence how people interpret evidence.

And let's be honest, we already have enough trouble with people believing half the nonsense they see online. Now we get to add believable fake audio into the soup.

Fantastic. Just what the mission needed, another flavor of chaos.

The research community is taking this seriously. ASVspoof, one of the major benchmark efforts in speech spoofing and anti-spoofing research, describes ASVspoof 5 as a challenge designed to promote "the study of speech spoofing and deepfake attacks." Its 2024 challenge data included crowdsourced speech from many speakers, diverse acoustic conditions, and, for the first time, adversarial attacks. (arXiv https://arxiv.org/abs/2408.08739)

That is a polite academic way of saying: the lab problem is becoming a street problem.

First, Not All Altered Audio Is AI

Before we throw the letters "AI" at everything like a glitter-bomb at a bad craft table, we need to define the problem.

Altered speech can mean several different things.

It can be simple editing, where words are cut, moved, deleted, or rearranged. It can be pitch shifting, where a voice is made higher or lower. It can be time stretching, where speech is sped up or slowed down without changing pitch. It can be filtering, equalization, noise reduction, compression, or other processing that changes how the voice sounds. It can be replayed audio, where a recording is played through a speaker and re-recorded through another microphone. It can be voice conversion, where one person's speech is transformed to sound like someone else. Or it can be fully synthetic speech generated from text.

Figure 3 – Not all Altered Audio is AI Generated

Today, AI-generated speech is getting most of the headlines, but old-school manipulation still matters. A badly edited recording can still cause major damage if it is dropped into the right social media feed, courtroom, news cycle, or emergency response event.

In public safety, the danger is not just fake audio. The danger is fake confidence.

The Ear Is a Sensor, Not a Forensic Tool

Humans are surprisingly good at noticing when something feels wrong. A voice may sound too smooth. The timing may feel strange. The emotion may not match the words. The breathing may be odd. The pauses may feel unnatural. A synthetic voice may say the right words with the wrong soul.

But human listening has limits.

The ear hears the performance. Audio analysis examines the physics.

A forensic audio review looks at the waveform, frequency content, pitch behavior, formants, phase, noise floor, compression artifacts, microphone characteristics, background room tone, reverberation, timing, and sometimes even traces of electrical grid hum. That is a very different world than “yeah, that sounds like him.”

The point is not that every call taker, detective, attorney, or public information officer needs to become an audio engineer. The point is that agencies need to understand that audio authenticity is now a discipline, not a vibe.

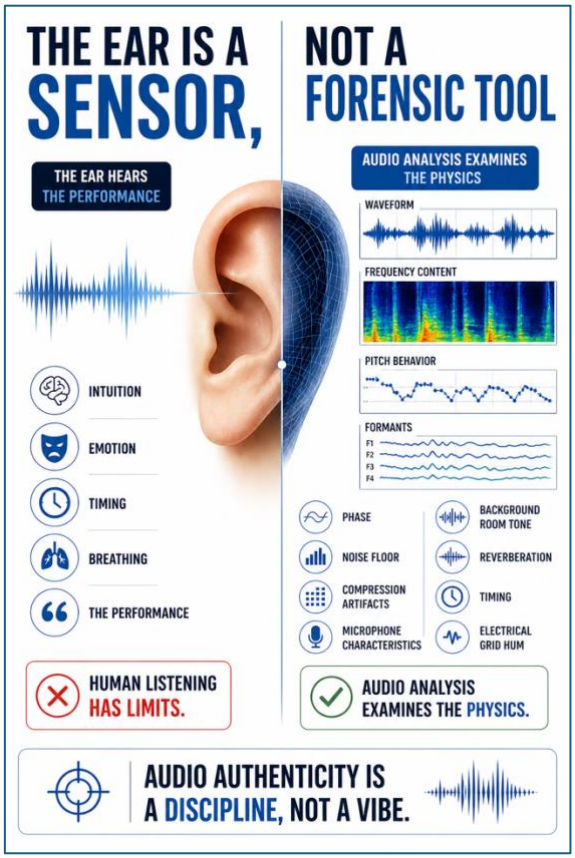

Figure 4 – The Ear is a Sensor, Not a Forensic Tool

How Audio Analysis Sees the Voice

At the most basic level, audio analysis begins with the waveform. This is the visual representation of the sound pressure changes over time. Waveform analysis can show abrupt cuts, clipping, changes in amplitude, suspicious silences, or transitions that do not match natural speech.

But the waveform is only the first layer.

A spectrogram is often more useful because it shows how energy is distributed across frequency and time. Human speech has structure. Vowels, consonants, plosives, fricatives,

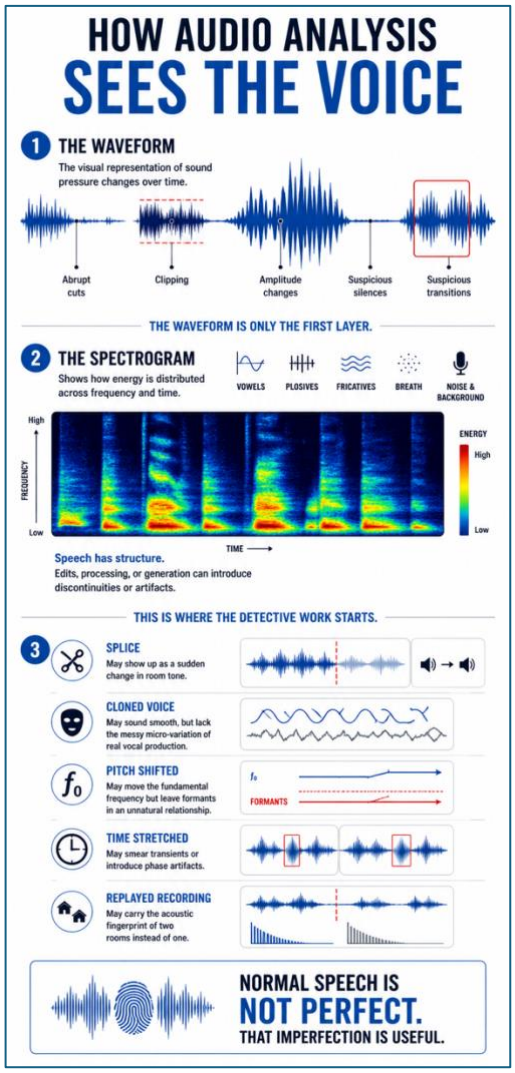

Figure 5 – How Audio Analysis Sees Voice

breath, microphone noise, and background sound all leave patterns. When speech has been edited, processed, or generated, those patterns may contain discontinuities or artifacts.

This is where the detective work starts.

A splice may show up as a sudden change in room tone. A cloned voice may sound smooth, but lack the messy micro-variation of real vocal production. A pitch-shifted voice may move the fundamental frequency but leave the formants in an unnatural relationship. A time-stretched recording may smear transients or introduce phase artifacts. A replayed recording may carry the acoustic fingerprint of two rooms instead of one.

Normal speech is not perfect. That imperfection is useful.

Pitch, Formants, and the Biology of Speech

Human speech is a biological process before it is an audio file.

Your vocal folds vibrate and produce a fundamental frequency, often referred to as F0. Your mouth, throat, tongue, nasal cavity, and vocal tract shape that sound into resonances called formants. Those formants are part of what makes vowels sound like vowels and speakers sound like themselves.

A simple voice changer may shift pitch up or down. But if the formants do not move in a biologically plausible way, the result can sound unnatural. That is why some altered voices have that "chipmunk," "robot," or "masked villain with a coupon-code microphone" quality.

More advanced voice conversion systems are harder to detect because they can manipulate more than pitch. They may preserve linguistic content while transforming speaker identity. A 2024 survey on audio anti-spoofing explains that text-to-speech systems generate speech from text, while voice conversion attacks alter original speech to mimic a target speaker while preserving the linguistic information. (arXiv - https://arxiv.org/html/2404.13914v1)

That distinction matters. In one case, the words may have never been spoken by anyone. In the other, someone spoke the words, but the voice identity may have been transformed. Operationally, those are very different problems.

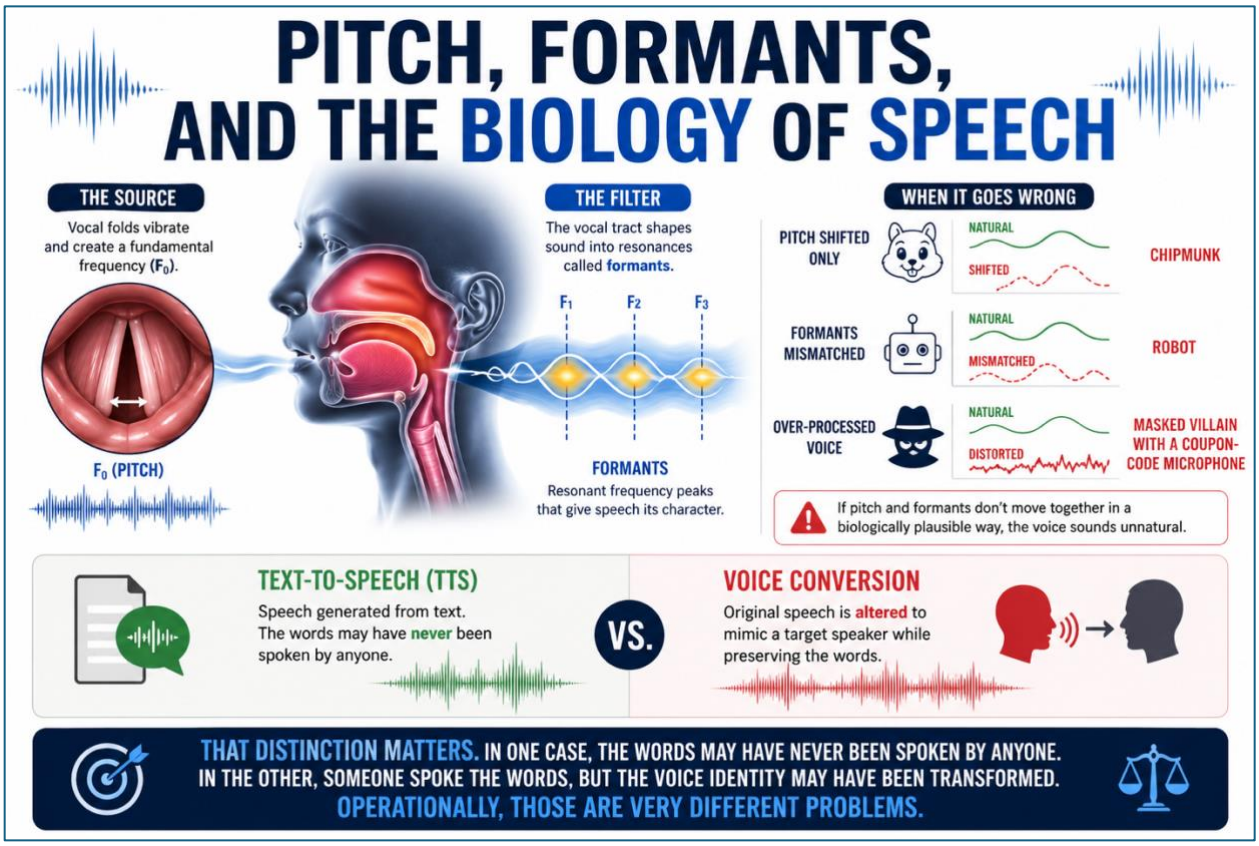

Figure 6 - Pitch, Formants, and the Biology of Speech

Timing Is a Tell

Human speech has rhythm, and not the kind you can fix with a metronome.

We breathe. We hesitate. We swallow. We stress certain words. We speed up when anxious. We pause in strange places when thinking. We interrupt ourselves. We trip over words. We recover. We drag vowels. We clip consonants. We do all the little messy things that make speech human.

Synthetic speech has improved dramatically, but timing remains one of the areas where detection can still find clues. Prosody, which includes rhythm, intonation, stress, and timing, can reveal whether speech behaves like natural human speech or like speech assembled by a system trying very hard to pass as human.

This is especially important in emergency calling.

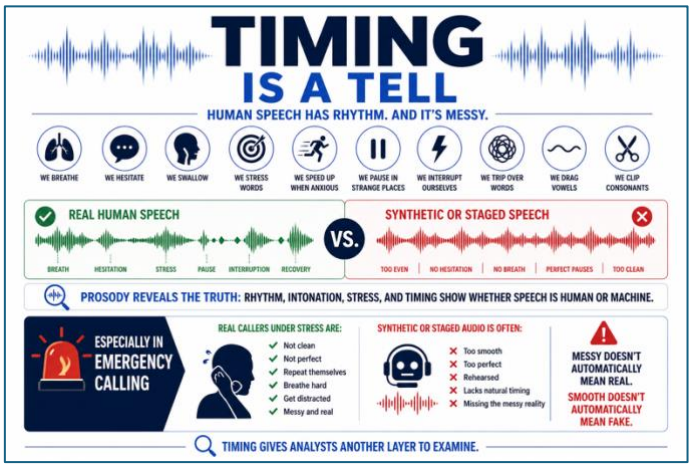

Figure 7 – Timing is a Tell

A real caller under stress may not sound clean. They may not produce a perfect narrative. They may not speak in complete sentences. They may repeat themselves. They may breathe hard. They may be distracted by the scene around them. Synthetic or staged audio may preserve dramatic language but fail to reproduce the timing and cognitive messiness of a real emergency.

Now, to be clear, messy does not automatically mean real, and smooth does not automatically mean fake. Some people are calm under pressure. Some recordings are cleaned up by systems. Some calls are compressed by networks. But timing gives analysts another layer to examine.

The Room Has a Voice Too

One of the biggest mistakes people make is thinking that altered audio detection is only about the speaker.

It is not.

The room has a voice. The microphone has a voice. The codec has a voice. The network path has a voice. The background noise has a voice.

Real recordings usually have continuity. There is a consistent noise floor. There may be HVAC hum, radio hiss, keyboard clicks, room reflections, street noise, fan noise, body-camera handling noise, VoIP compression, or headset artifacts. When a phrase is inserted from somewhere else, the voice may match, but the environment may not.

That is where forensic analysis can get interesting.

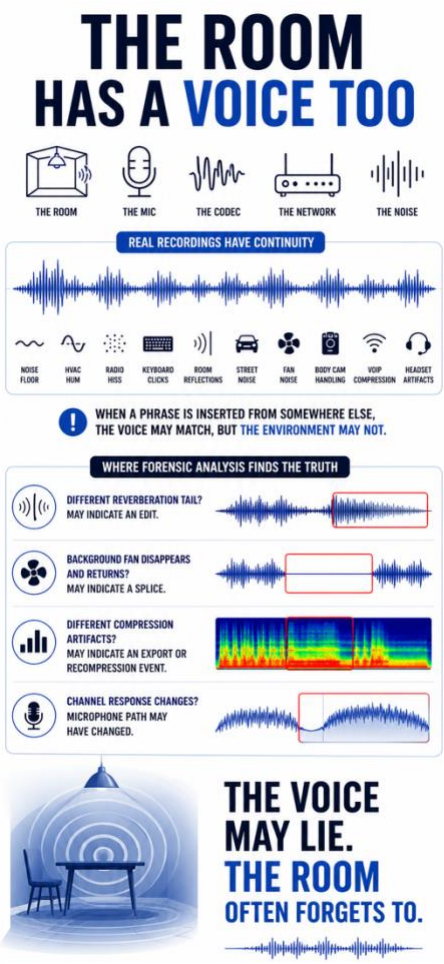

Figure 8 - The Room Has a Voice Too

If one sentence has a different reverberation tail, that may indicate an edit. If a background fan disappears for two seconds and then returns, that may indicate a splice. If the speech

carries different compression artifacts from the rest of the file, that may indicate an export or recompression event. If the channel response changes, the microphone path may have changed.

The voice may lie. The room often forgets to.

Time Stretching, Pitch Shifting, and Phase Artifacts

Some altered audio is created through time and pitch manipulation. This can be done for legitimate reasons, such as audio production, accessibility, editing, or broadcast timing. But the same tools can also be used to disguise a voice or alter meaning.

A phase vocoder is one common class of method used for time stretching and pitch manipulation. The basic idea is that the audio is broken into short time-frequency frames, processed, and then reconstructed. Carnegie Mellon's phase vocoder tutorial describes the process as analyzing "short grains of sound," adjusting frequency component phase, and combining the output back into a sound file. (Wikipedia)

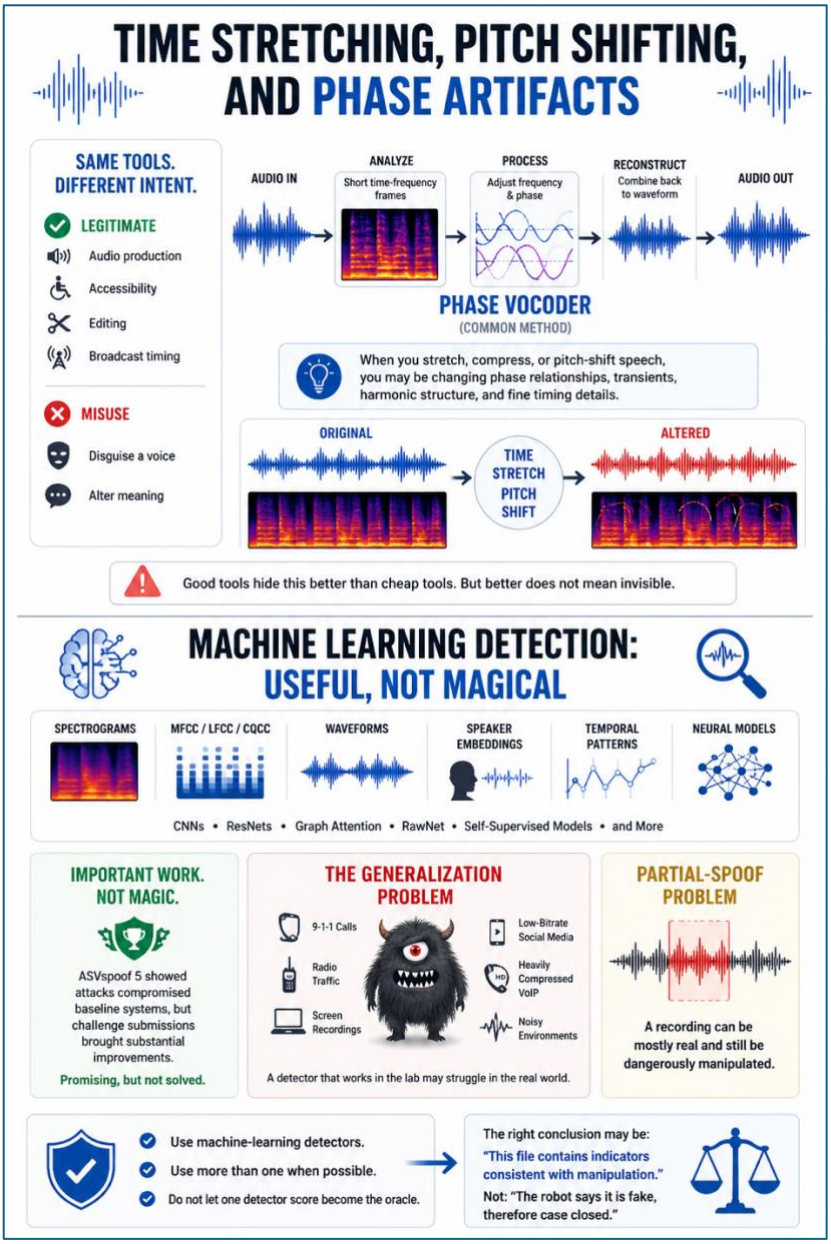

Figure 9 - Time Stretching, Pitch Shifting, and Phase Artifacts

The details get mathematical quickly, but the important point is simple: when you stretch, compress, or pitch-shift speech, you are not just changing how it sounds to the ear. You may be changing the phase relationships, transients, harmonic structure, and fine timing details inside the signal.

Good tools hide this better than cheap tools. But better does not mean invisible.

Machine Learning Detection: Useful, Not Magical

Modern audio deepfake detection increasingly uses machine learning. Systems may analyze spectrograms, cepstral features such as MFCC, LFCC, or CQCC, raw waveforms, speaker embeddings, temporal patterns, and other features. Neural models may include convolutional networks, residual networks, graph attention models, RawNet-style models, or self-supervised speech models.

This is important work. It is also not magic.

ASVspoof 5 reported that attacks significantly compromised baseline systems, while challenge submissions brought substantial improvements. That is encouraging, but it also shows why detection cannot be treated as solved. (arXiv - https://arxiv.org/abs/2408.08739)

The generalization problem is the big, ugly monster in the room. A detector trained on one type of fake audio may not perform as well on another. A detector that performs well on clean lab data may struggle with a noisy 9-1-1 call, a radio transmission, a screen recording, a low-bitrate social media clip, or a heavily compressed VoIP recording. The 2024 anti-spoofing survey notes that spoofed audio can include entire fake clips as well as "partial spoofed audio," where only portions of a clip are fake. (arXiv - https://arxiv.org/html/2404.13914v1)

That partial-spoof problem matters. A recording can be mostly real and still be dangerously manipulated.

So yes, use machine-learning detectors. Use more than one when possible. But do not let a single detector score serve as the oracle.

The right conclusion may be, "This file contains indicators consistent with manipulation," not "The robot says it is fake, therefore case closed."

Electric Network Frequency: Background Hum That Helps

Electric Network Frequency analysis, often called ENF analysis, is one of the more interesting forensic techniques. The basic concept is that audio recordings may capture faint

hum from the electrical power grid. Because grid frequency fluctuates slightly over time, those fluctuations can sometimes be compared to reference data to help verify when a recording was made or whether it was altered.

Recent ENF research describes using the unique Electric Network Frequency signal to identify tampered audio files.

(PMC https://pmc.ncbi.nlm.nih.gov/articles/PMC10458025)

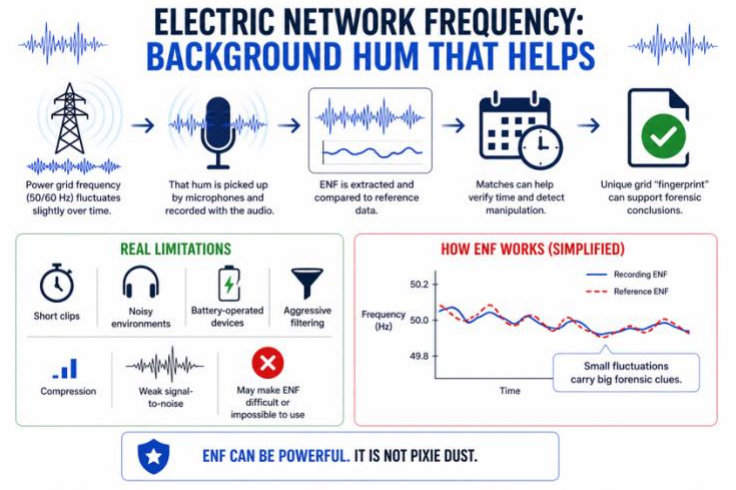

Figure 10 - Electric Network Frequency: Background Hum That Helps

That sounds like spy-movie magic, but it has real limitations. ENF is only useful if the recording actually captured a usable power-grid trace. Short clips, noisy environments, battery-operated devices, aggressive filtering, compression, or weak signal-to-noise conditions can make the technique difficult or impossible to use. In other words, ENF can be powerful. It is not pixie dust.

Provenance: The Best Evidence May Start Before the Analysis

Audio authenticity is not only about what is inside the signal. It is also about where the file came from.

Who recorded it? On what device? Was the original file preserved? Was it exported from an app? Was it sent through a social media platform? Was it compressed? Was it converted? Who had access to it? Was the chain of custody documented?

This is where provenance becomes important.

The Coalition for Content Provenance and Authenticity, or C2PA, provides a standard intended to help establish the origin and edit history of digital content. C2PA describes Content Credentials as functioning "like a nutrition label for digital content." (C2PA - https://c2pa.org/)

That is a useful analogy. A nutrition label does not tell you whether dinner tastes good. It tells you what went into it. In the same way, provenance data does not automatically tell you whether audio is truthful, but it can help establish origin, edits, and handling history.

The C2PA specification defines provenance as understanding the history of an asset and authenticity as a set of facts that can be cryptographically verified as not having been tampered with. It also recognizes composed assets, such as a video created by importing existing video clips and audio segments. (C2PA Spec - https://spec.c2pa.org/specifications/specifications/2.4/specs/C2PA_Specification.html)

For public safety and investigations, this points to a future where authenticity begins at capture, not after a controversy goes viral.

Public Safety Use Cases: Where This Gets Real

This subject is not academic for emergency communications.

Imagine a swatting call using a cloned voice of a student, parent, employee, public official, or known victim. Imagine a threat call that appears to come from someone local but was generated from text. Imagine an edited emergency recording circulating online without context. Imagine a synthetic voice impersonating a police chief, school administrator, mayor, sheriff, ECC director, hospital executive, or IT leader.

The public safety mission depends on timely decisions under uncertainty. That will not change. But audio manipulation increases the cost of assuming too much too quickly.

That does not mean call takers should become suspicious of every caller. That would be operationally impossible and ethically dangerous. Emergency calls still need to be handled as emergencies.

But agencies do need escalation pathways for suspicious audio, especially when recordings are used as evidence, public claims, media clips, threat intelligence, or operational triggers outside the immediate call-handling moment.

The first response may still be, "Send help."

The follow-up must increasingly be, "Preserve the original, verify the source, and analyze before declaring certainty."

A Practical Audio Authenticity Workflow

A reasonable workflow does not need to start with a million-dollar lab. It starts with discipline.

1. First, preserve the original file. Do not enhance it, normalize it, convert it, denoise it, upload it, clip it, or run it through five free tools from the internet before someone qualified sees it. Every conversion can destroy evidence or create artifacts that confuse the analysis.
2. Second, document where the recording came from. Device, platform, timestamp, file format, transfer method, and chain of custody all matter.
3. Third, make a working copy. Analysis should happen on the copy, not the original.
4. Fourth, segment the recording. Identify speakers, pauses, sudden transitions, background changes, and suspicious regions.
5. Fifth, examine the signal. Look at waveform, spectrogram, pitch, formants, timing, noise floor, and compression behavior.
6. Sixth, evaluate environmental consistency. Does the room tone match? Does the background noise continue naturally? Does reverberation change? Does the microphone path remain consistent?
7. Seventh, use machine-learning detectors as supporting tools. Do not rely on a single score.
8. Eighth, compare against known authentic samples if speaker identity matters. But remember that speaker comparison and fake detection are related, not identical.
9. Ninth, corroborate externally. Call logs, CAD records, device logs, access control logs, phone records, video, metadata, dispatch timestamps, and witness accounts may be just as important as the audio itself.
10. Finally, report in confidence levels. Avoid absolute statements unless the evidence truly supports them.

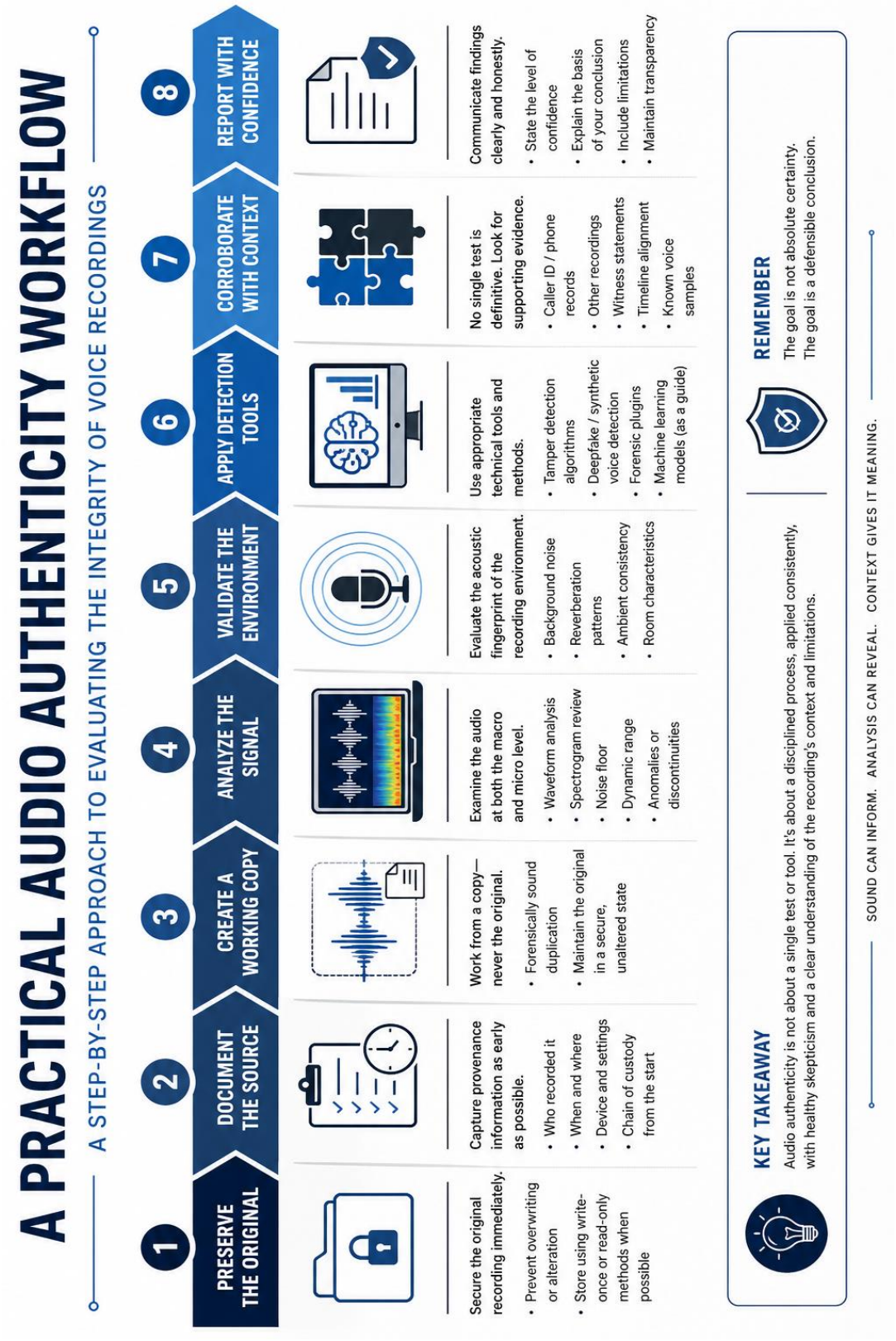

Figure 11 - A Practical Workflow for PSAPs

Figure 12 - What Agencies Should Do NOW

What Agencies Should Do Now

What agencies should decide now how they will handle suspected altered audio. That policy should address preservation, chain of custody, when to involve forensic experts, how to document analysis, when to notify leadership, how to coordinate with legal counsel, and how to communicate uncertainty to the public.

Training should include basic awareness. Not everyone needs to know how to calculate cepstral coefficients. But supervisors, investigators, records custodians, public information officers, ECC leadership, and legal teams should understand the categories of manipulation and the importance of preserving originals.

Cybersecurity teams should also be part of the conversation. Voice cloning is not just a media problem. It is an impersonation problem. The FTC has warned that scammers may clone a CEO or executive voice to trick employees into sending money or paying fake invoices. (Consumer Advice https://consumer.ftc.gov/consumer-alerts/2023/11/announcing-ftcs-voice-cloning-challenge)

The same concept applies to public safety operations. If someone can impersonate leadership convincingly, your verification process cannot depend only on "that sounded like the chief."

Use call-back procedures. Use known numbers. Use multi-factor verification for unusual requests. Use pre-established authentication phrases or procedures where appropriate. And yes, write it down before the circus arrives.

What Detection Cannot Do

This is where we need to stay honest.

Audio analysis cannot always prove who created a fake. It cannot always prove intent. It cannot always recover the original. It cannot always separate poor recording quality from manipulation. It cannot always detect a sophisticated fake, especially if the clip is short, noisy, compressed, or stripped of context.

And machine-learning detectors are not immune to failure.

A clean result does not always prove authenticity. A suspicious result does not always prove fraud. The job is to reduce uncertainty with evidence, not replace one assumption with another.

That is a lesson public safety already understands. One sensor is information. Multiple independent signals create confidence.

The Bottom Line

The human voice used to be one of the most trusted signals in communication. Now it is becoming another data source that must be validated.

That does not mean every recording is fake. It does not mean every emergency call is suspicious. It does not mean technology has ruined trust forever,

Figure 13- - What Detection Cannot Do

It means important audio deserves a process.

Preserve the original. Understand the signal. Analyze the voice, the room, the timing, the channel, and the provenance. Use machine-learning tools, but do not worship them. Corroborate with operational records. Train people before the crisis. And for the love of all things mission-critical, do not let “it sounded real” become your entire authentication strategy.

The voice may sound human. The words may sound urgent. The emotion may sound real.

But in the AI era, sound alone is no longer proof.

- Follow Me on Social Media @Fletch911
- Check out http://911tips.com/ every Monday, Wednesday, and Friday
- My Blogs are published at http://Fletch.TV

References and Resource Links

ASVspoof 5: Crowdsourced Speech Data, Deepfakes, and Adversarial Attacks at Scale
Research challenge paper describing ASVspoof 5, speech spoofing, deepfake attacks, crowdsourced data, and adversarial attacks. (arXiv - https://arxiv.org/abs/2408.08739)

Audio Anti-Spoofing Detection: A Survey
Survey covering text-to-speech, voice conversion, anti-spoofing countermeasures, partial spoofed audio, and detection methods. (arXiv - https://arxiv.org/html/2404.13914v1)

Federal Trade Commission, The FTC Voice Cloning Challenge
FTC challenge page describing multidisciplinary responses to malicious voice cloning and the limits of technology-only solutions. (Federal Trade Commission - https://www.ftc.gov/news-events/contests/ftc-voice-cloning-challenge)

Federal Trade Commission, Announcing the FTC's Voice Cloning Challenge
Consumer warning on voice cloning, family emergency scams, and business impersonation risks. (Consumer Advice - https://consumer.ftc.gov/consumer-alerts/2023/11/announcing-ftcs-voice-cloning-challenge)

Federal Trade Commission, Fighting Back Against Harmful Voice Cloning
FTC consumer guidance on voice cloning scams and verification through known contact methods. (Consumer Advice - https://consumer.ftc.gov/consumer-alerts/2024/04/fighting-back-against-harmful-voice-cloning)

NIST, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile
NIST cross-sector guidance for managing generative AI risks. (NIST - https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-generative-artificial-intelligence)

C2PA, Verifying Media Content Sources
Overview of the C2PA standard and Content Credentials for digital provenance. (C2PA - https://c2pa.org/)

C2PA Technical Specification
Technical definitions for provenance, authenticity, claims, manifests, content bindings, and composed assets. (C2PA Spec - https://spec.c2pa.org/specifications/specifications/2.4/specs/C2PA_Specification.html)

Electric Network Frequency-Based Audio Tampering Research
Research on using Electric Network Frequency signals for audio tampering detection. (PMC - https://pmc.ncbi.nlm.nih.gov/articles/PMC10458025)

Phase Vocoder and Time/Pitch Modification Background
Technical background on phase-vocoder-style time stretching and pitch shifting. (Wikipedia - https://en.wikipedia.org/wiki/Audio_time_stretching_and_pitch_scaling)