AIR GAP VOICE

The Hidden Cost of Cloud Dictation

April 2026

PUBLIC RELEASE

Airgap Voice is a speech-to-text application for macOS and iOS that processes all audio locally on the user's device. No audio, transcription text, or telemetry data is ever transmitted over a network. This document describes the technical architecture for security architects, compliance officers, and technical evaluators who need to understand how Airgap Voice achieves data sovereignty guarantees that cloud-based transcription services cannot provide.

Author: Marvin F. L. Hansen
Contact: marvin.hansen@airgapvoice.com
Web: www.airgapvoice.com

Table of Contents

1. Executive Summary

Every organisation that uses cloud-based dictation services is making an implicit decision: that the convenience of server-side speech recognition outweighs the risk of transmitting spoken words — often containing the organisation's most sensitive information — to infrastructure it does not own, cannot audit, and may not even be able to locate.

This paper examines what that decision actually costs. Not in licensing fees, which are straightforward, but in regulatory exposure, supply chain risk, compliance overhead, and the erosion of client confidence that follows when an organisation cannot definitively answer the question: "Who else heard what I just said?"

The cost is not hypothetical. Organisations subject to GDPR, HIPAA, ITAR, SOX, and attorney-client privilege face concrete obligations around the processing of spoken data. Cloud dictation creates data flows that intersect with all of these frameworks, often in ways that procurement teams do not fully map at the time of purchase.

A category of alternatives now exists: local-first transcription, where speech recognition runs entirely on the user's device with no network transmission. This approach does not merely reduce risk; it eliminates entire categories of exposure that cloud architectures cannot address regardless of encryption, contractual protections, or vendor assurances.

> **The purpose of this paper is not to argue that cloud services are inherently unacceptable. It is to ensure that organisations making this decision understand its full scope, and that they are aware of alternatives that did not exist five years ago.**

2. The Architecture of Exposure

2.1 How Cloud Dictation Works

The basic architecture of cloud-based speech recognition has remained consistent across vendors for over a decade. Audio is captured on the user's device, compressed, transmitted over a network connection to vendor-operated servers, processed by machine learning models that require significant compute resources, and the resulting text is returned to the client application.

This architecture exists for a practical reason: until recently, the computational demands of accurate speech recognition exceeded what consumer and enterprise hardware could deliver in real time. That constraint drove a cloud-first design, and it worked. But the data flow it creates has consequences that extend well beyond the technical domain.

2.2 The Data Path

When a user dictates into a cloud-connected application, the audio data traverses a path that typically includes:

- Capture on the local device microphone
- Encoding and compression (often to a lossy format, but still containing the full speech signal)
- Transmission over TLS to a vendor API endpoint
- Routing through vendor load balancers and potentially multiple data centres
- Processing by inference servers, which may be operated by the vendor or by subprocessors

- Temporary (or not-so-temporary) storage of the audio and resulting transcript
- Return of the transcript to the client

Each hop in this chain represents a point where the audio data exists outside the user's control. The security of the overall system is bounded by the weakest link in this chain, not by the strongest.

2.3 Retention Policies

Vendor retention policies for audio data vary widely and are often written with more flexibility than customers realise. Common patterns include:

- Audio retained for "service improvement" with opt-out mechanisms that may not apply to all tiers or regions
- Transcripts stored separately from audio, with different retention windows for each
- Metadata (timestamps, device identifiers, usage patterns) retained indefinitely even when audio is deleted
- Backup and disaster recovery systems that may retain copies beyond the stated retention period
- Retention terms that change with updated terms of service, sometimes without explicit notification

The critical question is not what the vendor's current policy says. It is whether the customer has any independent means of verifying compliance with that policy. In almost all cases, the answer is no.

2.4 The Training Data Question

Machine learning models improve with data. This creates a structural incentive for vendors to use customer audio as training data. While major vendors have responded to public pressure by offering opt-out mechanisms, several uncomfortable realities remain:

- Opt-out is not the same as opt-in. Default configurations in many services include data contribution for model improvement.
- Enterprise agreements may override default policies, but verifying that override in practice — across all data pipelines — is non-trivial.
- Once audio has been used to train a model, the contribution cannot be meaningfully reversed. Model unlearning is an active research area, not a production capability.
- The boundary between "service improvement" and "model training" is often blurred in vendor documentation.

For organisations handling sensitive information, the question is straightforward: can you accept the possibility that fragments of your spoken data influence a model that serves your competitors?

3. Regulatory Exposure

Cloud dictation intersects with multiple regulatory frameworks. The following is not legal advice; it is a map of the terrain that legal and compliance teams should evaluate.

3.1 GDPR: Audio as Personal Data

Under the General Data Protection Regulation, voice recordings constitute personal data. A speaker's voice is biometric data under Article 9, placing it in the "special categories" that require explicit consent or another Article 9 basis for processing. Key implications for cloud dictation:

- Cross-border transfer: If the vendor processes audio outside the EEA, a valid transfer mechanism (SCCs, adequacy decision, or BCRs) must be in place. Post-Schrems II, this requires a transfer impact assessment for each destination country.
- Data processing agreements: The organisation must have a DPA with the vendor that specifies processing purposes, duration, and deletion obligations.
- Right to erasure: Article 17 requests may require the vendor to delete audio data, but verifying deletion across backup systems and training pipelines is operationally difficult.
- Data protection impact assessment: High-risk processing (which large-scale voice processing likely qualifies as) requires a DPIA under Article 35.

3.2 HIPAA: Protected Health Information

Healthcare professionals routinely dictate clinical notes, diagnoses, treatment plans, and patient-identifiable information. When dictation flows through a cloud service, the audio constitutes protected health information (PHI) under HIPAA.

- A Business Associate Agreement (BAA) with the dictation vendor is required. Not all vendors offer BAAs, and those that do may limit covered services.
- The Minimum Necessary Rule requires that PHI disclosures be limited to the minimum necessary for the purpose. Transmitting full audio recordings (which may contain incidental PHI beyond the intended dictation) raises questions about compliance with this principle.
- Breach notification obligations under the HITECH Act apply if the vendor experiences a security incident affecting PHI. The covered entity bears notification responsibility regardless of where the breach occurred in the supply chain.

3.3 ITAR/EAR: Export-Controlled Information

Engineers, programme managers, and analysts in defence and aerospace routinely dictate notes, reports, and communications that contain information controlled under the International Traffic in Arms Regulations (ITAR) or Export Administration Regulations (EAR).

- Transmitting ITAR-controlled technical data to a server located outside the United States, or accessible by non-U.S. persons, constitutes an export. This applies regardless of encryption.
- Cloud vendors that use global infrastructure may route data through non-U.S. facilities or employ non-U.S. nationals in operations roles, potentially triggering deemed export violations.
- Penalties for ITAR violations are severe: up to $1 million per violation and potential debarment from government contracting.

3.4 SOX: Financial Data Integrity

Executives and financial professionals who dictate communications related to financial reporting, earnings, or material non-public information create records that intersect with Sarbanes-Oxley requirements.

- Section 302 requires CEO/CFO certification of internal controls over financial reporting. Cloud dictation of financial data introduces a control point outside the organisation's direct management.
- Section 404 requires assessment of internal controls. Auditors may question whether transmitting financial dictation to third-party servers constitutes an adequate control environment.

3.5 Attorney-Client Privilege

The implications of cloud dictation for attorney-client privilege warrant dedicated analysis and are addressed in a separate paper. The core concern is straightforward: privilege may be waived if communications are disclosed to third parties, and the question of whether a cloud vendor constitutes a "third party" for privilege purposes is jurisdiction-dependent and actively evolving.

Regulation	Data Type at Risk	Key Obligation	Cloud Dictation Concern
GDPR	Voice (biometric)	Lawful basis, transfer mechanism	Cross-border audio transmission
HIPAA	PHI in clinical dictation	BAA, minimum necessary	Audio exceeds minimum necessary
ITAR/EAR	Controlled technical data	No unauthorised export	Server location, personnel access
SOX	MNPI, financial data	Internal control integrity	Third-party processing of sensitive financial data
Privilege	Attorney-client communications	No third-party disclosure	Vendor as potential third party

4. The Supply Chain Risk

4.1 Vendor Dependency

Cloud dictation creates an operational dependency on a vendor's continued willingness and ability to provide the service on acceptable terms. This dependency manifests in several ways:

- API changes: Vendors periodically deprecate API versions, modify response formats, or alter supported languages. Each change requires integration work on the customer side.
- Pricing changes: Cloud speech APIs are typically priced per audio-hour. Vendors can and do adjust pricing, and enterprise discount agreements have fixed terms.
- Acquisition risk: If the vendor is acquired, the acquiring entity's data practices, jurisdiction, and strategic priorities may differ materially from the original vendor's.

- Discontinuation: Vendors discontinue products. A cloud dictation dependency becomes a migration project when the service is sunset.

4.2 Subprocessor Chains

Most cloud dictation vendors do not operate entirely self-contained infrastructure. Audio data may be processed by or transit through:

- Infrastructure providers (IaaS) who host the vendor's compute and storage
- Content delivery networks that route API traffic
- Third-party machine learning platforms used for model inference
- Analytics and monitoring services that may have access to metadata
- Support and operations subcontractors who may access data during incident response

Each subprocessor extends the trust boundary. Under GDPR, each must be disclosed and covered by appropriate contractual protections. In practice, customers rarely have visibility into the full subprocessor chain, and changes to that chain may occur with minimal notice.

4.3 Jurisdiction Risk

Where audio is processed matters, legally, not just technically. A vendor may be headquartered in one jurisdiction, process data in a second, store backups in a third, and employ operations staff in a fourth. Each jurisdiction's laws may grant government access to data under different conditions and with different transparency obligations.

The U.S. CLOUD Act, for example, enables U.S. law enforcement to compel disclosure of data held by U.S.-headquartered companies regardless of where the data is physically stored. For non-U.S. organisations using U.S.-based dictation vendors, this creates exposure that contractual terms cannot eliminate.

4.4 Business Continuity

Cloud dictation fails when the network fails. This is not a theoretical concern:

- Internet connectivity outages — whether local, regional, or vendor-side — render cloud dictation unavailable.
- Vendor service incidents affect all customers simultaneously. Major cloud provider outages regularly impact thousands of organisations for hours or days.
- Bandwidth constraints in field deployments, secure facilities, or areas with limited connectivity may make cloud dictation impractical or unreliable.
- Air-gapped environments, common in defence, intelligence, and critical infrastructure, cannot use cloud dictation at all.

5. The Hidden Costs

The direct cost of cloud dictation — licensing fees, per-minute API charges — is visible in procurement budgets. The indirect costs are often larger but distributed across departments in ways that make them difficult to aggregate.

5.1 Compliance Overhead

Every cloud service that processes sensitive data generates compliance work:

- Annual vendor risk assessments, including review of SOC 2 reports, penetration test summaries, and data handling documentation
- Ongoing monitoring of vendor subprocessor lists and policy changes
- Internal documentation of data flows for audit and regulatory purposes
- Employee training on acceptable use: what can and cannot be dictated through the cloud service
- Periodic re-evaluation of the vendor's compliance posture as regulations evolve

These activities consume security, legal, and IT staff hours. For organisations with multiple cloud vendors — which is nearly all of them — the cumulative burden is substantial.

5.2 Legal Review

Cloud dictation deployments require legal review of:

- Data Processing Agreements (DPAs) for GDPR compliance
- Business Associate Agreements (BAAs) for HIPAA-covered entities
- Master Service Agreements and their data protection addenda
- Terms of service changes, which may alter data handling practices
- Acceptable use policies that may grant the vendor rights to aggregated or anonymised data

Legal review at enterprise rates is expensive. More importantly, it creates a recurring obligation: agreements must be re-reviewed when terms change, and vendors update terms regularly.

5.3 Incident Response

When a cloud vendor experiences a security incident, every customer must determine whether their data was affected. This analysis requires:

- Coordination with the vendor's incident response team, often with limited information and high latency
- Internal assessment of what data was exposed, which is difficult when the organisation cannot independently audit the vendor's systems
- Potential breach notification obligations under GDPR (72 hours), HIPAA (60 days), and state breach notification laws
- Client and stakeholder communication, even when the organisation's own systems were not compromised

The reputational cost of notifying clients that their dictated words may have been exposed in a vendor breach is difficult to quantify but real. "Our vendor was breached" does not inspire confidence.

5.4 Reputational Risk

Clients and counterparties increasingly ask pointed questions about data handling. Law firm clients want to know how their communications are protected. Healthcare patients have growing awareness of data privacy. Government contractors face explicit security requirements from contracting officers.

> **The answer "we send your dictated words to a third-party cloud service" is becoming harder to deliver with confidence. The answer "everything is processed on-device and nothing leaves this machine" is qualitatively different and verifiable.**

Cost Category	Cloud Dictation	Local-First Dictation
Licensing / API fees	Recurring, usage-based	One-time or subscription
Compliance overhead	Ongoing vendor assessments	Minimal: no third-party data flow
Legal review	DPAs, BAAs, TOS changes	Standard software license review
Incident response	Vendor breach creates obligations	No vendor data exposure possible
Reputational risk	Client concern about third-party processing	Verifiable on-device processing
Business continuity	Dependent on network + vendor uptime	Functions offline

6. The Local-First Alternative

6.1 On-Device Processing as a Category

Local-first transcription is not a single product; it is an architectural approach in which speech recognition runs entirely on the user's hardware. No audio is transmitted to external servers. No network connection is required for operation. The processing happens where the speaking happens: on the device.

Several products now implement this approach with varying degrees of completeness. The common thread is a fundamental architectural decision: the audio never leaves the device. This is not a feature that can be bolted onto a cloud architecture; it requires building the system from the ground up with local processing as the foundation.

6.2 Why Now

The viability of local speech recognition has changed dramatically in recent years due to convergent advances in hardware and software:

- Neural processing hardware: Modern CPUs, GPUs, and dedicated neural engines provide sufficient throughput for real-time speech recognition without cloud offloading.
- Model efficiency: Research in model compression, quantisation, and architecture optimisation has produced speech recognition models that achieve production-quality accuracy within the memory and compute constraints of consumer hardware.

- On-device memory: Current devices ship with sufficient RAM to hold speech recognition models in memory alongside normal workloads.

Five years ago, local transcription was a compromise: it worked, but accuracy lagged behind cloud services. That gap has closed. For many languages and use cases, on-device models now deliver accuracy that meets or exceeds the cloud services that preceded them.

6.3 The Verification Advantage

The most significant advantage of local-first transcription is not performance; it is verifiability. When a vendor claims that data is processed locally, the user can independently verify that claim using standard operating system tools.

On macOS, for example, a user or IT administrator can use built-in tools such as Activity Monitor, lsof, and tcpdump to confirm that an application makes zero network connections during dictation. This is not a matter of trusting the vendor's documentation; it is observable, reproducible fact. This verification capability inverts the trust model: instead of relying on a vendor's privacy policy and hoping for consistent implementation, the user can confirm the behaviour directly.

6.4 Eliminating Categories of Risk

Local-first processing does not merely reduce the risks described in this paper. It eliminates entire categories:

Risk Category	Cloud Dictation	Local-First Dictation
Data in transit	Audio traverses network	No network transmission
Vendor data breach	Audio exposed if vendor compromised	No vendor holds audio data
Subprocessor risk	Multiple third parties in chain	No third parties involved
Cross-border transfer	Audio may cross jurisdictions	Data never leaves device
Retention policy compliance	Depends on vendor implementation	Audio exists only in volatile memory
Training data contribution	Possible depending on terms	Impossible: no data leaves device
Government access (CLOUD Act)	Vendor may be compelled	No vendor holds data to disclose

> **There is an important distinction between managing risk and eliminating it. Cloud security is fundamentally about risk management. Local-first processing eliminates the underlying exposure. You cannot intercept data that was never transmitted.**

7. Evaluating a Local Transcription Solution

Not all products that claim local processing deliver it completely. The following guidance is intended for procurement, security, and IT teams evaluating local transcription solutions.

7.1 Questions Procurement Teams Should Ask

- Does the application require an internet connection for installation, activation, or any aspect of operation?
- Are speech recognition models bundled with the application, or downloaded from vendor servers after installation?
- Does the application transmit any data (audio, transcripts, usage telemetry, crash reports, or license checks) to external servers?
- Is the application code signed and notarised by the platform vendor (e.g., Apple)?
- Does the application operate within the platform's security sandbox (e.g., macOS App Sandbox)?
- How is transcribed text delivered to the target application? (Clipboard access vs. direct text insertion via accessibility APIs carries different security implications.)
- Where is audio data stored during processing? (Volatile memory only vs. disk-based temporary files.)
- What is the vendor's business model? (Subscription, one-time purchase, or data monetisation.)

7.2 Red Flags in Vendor Claims

Security and procurement teams should be alert to claims that sound reassuring but do not address the fundamental data flow question:

"We encrypt all data in transit."

Encryption protects data from interception during transmission. It does not prevent the vendor from accessing the data at the endpoint. If audio reaches a vendor server, encryption in transit is necessary but not sufficient.

"We do not store audio data."

This claim is difficult to verify and may have exceptions for debugging, quality assurance, or abuse detection. Ask for specifics about transient processing, logging, and backup systems.

"We are SOC 2 Type II certified."

SOC 2 certification confirms that controls exist and are operating effectively. It does not confirm that those controls prevent all unauthorised access, and it does not eliminate the fundamental exposure of transmitting data to a third party.

"We comply with GDPR."

Compliance is a continuous obligation, not a state. It depends on the vendor's ongoing practices, and the customer retains responsibility as data controller regardless of the vendor's claims.

7.3 Verification Checklist

Verification Item	Method	Expected Result
No network activity during dictation	tcpdump, Wireshark, or Little Snitch during active use	Zero packets transmitted to external hosts
No network activity at idle	Monitor network connections with lsof -i while app is running	No established connections to external servers
App Sandbox enforcement	Check code signature: codesign -dvvv	Sandbox entitlement present
Hardened Runtime	Check code signature flags	Hardened Runtime flag enabled
Code signing and notarisation	spctl --assess --type exec	Application accepted by Gatekeeper
No disk-based audio storage	Monitor file system activity with fs_usage during dictation	No audio files written to disk
Models bundled locally	Check application bundle contents; test in airplane mode	Full functionality without network

8. Conclusion

Cloud dictation was the right answer when local hardware could not run speech recognition at production quality. That constraint no longer holds. The question facing organisations today is not whether local transcription is technically viable; it is whether the accumulated risk, cost, and complexity of cloud dictation remain justified when a local alternative exists.

For organisations handling regulated data — healthcare records, export-controlled technical information, financial disclosures, privileged communications — the answer increasingly favours local processing. Not because cloud vendors are negligent, but because the architecture itself creates exposures that no amount of contractual protection can fully address.

The hidden cost of cloud dictation is not the API bill. It is the compliance overhead, the legal review, the incident response planning, the client confidence questions, and the persistent uncertainty about where spoken words end up after they leave the device. Local-first transcription does not merely reduce these costs. It eliminates the conditions that create them.

> **Organizations evaluating dictation solutions should demand verifiable claims. Encryption in transit, SOC 2 reports, and privacy policies are necessary components of cloud security, but they are not equivalent to the simple, independently verifiable assertion: nothing left this machine. The technology exists today to make that assertion true.**

This white paper is provided for informational purposes only and does not constitute legal advice. Organizations should consult qualified legal counsel regarding their specific regulatory obligations.