Songjam  
A Cryptographic Voice Verification Network

Adam Place

2025-06-03

Abstract

A fully encrypted, privacy-preserving voice verification network could effectively eliminate voice-based deepfake fraud without requiring users to compromise their security or personal data. Zero-knowledge proofs provide part of the solution, but a fundamental challenge remains: it is difficult to verify with absolute certainty that a given voice truly belongs to a specific individual, especially as voice synthesis technologies become increasingly sophisticated. This work introduces a solution to the verification problem by leveraging cryptoeconomics. The network rewards users for submitting accurate credentials, while imposing penalties on those who attempt to provide false information, leveraging the security of a Proof-of-Stake (PoS) consensus mechanism, while aligning user incentives with the integrity of the network. Cryptographic keys to voice biometrics and other sensitive data are stored in Trusted Execution Environments (TEEs) providing hardware-backed isolation. Access to these keys is granted to users through a Soulbound Token (SBT), a non-transferable, non-replicable digital asset permanently tied to an individual’s digital identity (Buterin, Weyl, and Ohlhaver 2022). Over time, the user builds-up a personalised voice model, with each new audio sample organized and stored within a WavRAG structure—an Audio-Integrated Retrieval Augmented Generation framework, enabling efficient and context-rich retrieval of voice data (Chen et al. 2025).

# Introduction

The rise of deepfakes has led to a growing concern about the falibility of humans to verify the authenticity of voice communication. High profile cases such as the 2024 deepfake attack on engineering firm Arup, which led to a twenty-five million dollar loss, highlight the importance of secure and reliable enterprise verification systems.

While a plethora of fradulent online videos featuring deepfakes of celebrities promoting supposed investment opportunities, and the rise of romance or 'pig butchering' scams which collectively siphon billions of dollars from unsuspecting victims every year illustrate the nessecity of consumer or citizen protections. The sub-150ms latency required for natural conversation in VoIP calls leaves minimal time for analysis on standard hardware, making real-time detection of deepfakes challenging. While the proprietary nature of platforms such as Zoom, Teams and SIP providers, hinders the development of universal detection frameworks.

What is needed is a privacy-preserving verification system leveraging zero-knowledge proofs and selective disclosure mechanisms, where only cryptographically attested credentials (e.g., identity, access rights) are validated—without exposing extraneous user data. This approach ensures networks authenticate voice participants a priori through mathematical guarantees (e.g., zk-SNARKs), verifying the existence of valid credentials rather than analyzing raw biometrics or personal details. By limiting verification to key attributes (e.g., "member of authorised group" or "registered device") and eliminating unnecessary data collection, such systems support user control over personal information while thwarting deepfake fraud and metadata misuse. This paper proposes a novel cryptoeconomic framework for voice verification, in which users are incentivised to provide accurate and verifiable credentials, while malicious actors face penalties—akin to PoS slashing mechanisms—for fraudulent voice attribution attempts.

# The Proof-of-Stake Security Guarantee?

Much has been written about the security of the PoS consensus mechanism, with proponents arguing that its economic incentives guarantee security, while opponents contend that the consolidation of staked token supply will lead to centralisation. Nevertheless, the economic incentives and penalties underpinning Ethereum’s Proof-of-Stake mechanism have proven effective, driving its adoption and establishing it as the preferred consensus model for securing one of the world’s largest blockchain networks—demonstrated by the very small number of validator slashing events recorded to date (with only around 0.045% of validators slashed since Ethereum staking began).



Ethereum’s PoS economics rewards honest validators and penalises malicious ones

Ethereum’s ability to ensure security through staked ETH is attributable to the distribution and value it accumulated over several years as a Proof-of-Work (PoW) network. A new network, even built with the same consensus mechanism, would not be able to guarantee security with the same level of confidence. Taking the PoS principle to Voice Verification, to validate and secure the authenticity of voice biometrics from day one requires a different approach. A solution is to introduce an AI biosignal verification analysis step, in which the user must provide a voice sample accompanied by the physiological motions necessary for human speech production. Projects such as OriginStory propose a microphone that includes a sensor to measure these movements (Federal Trade Commission 2024), however a laryngeal mounted Piezoelectric transducer would be a viable alternative that can capture vocal fold vibrations (70-3000 Hz) through neck skin and would be broadly accessible and cost effective for a global network of users. The emphasis is on open-source hardware and a ResNet-18 signal processing architecture (Liu et al. 2025), which prepares the data with preliminary "authentic/synthetic" labels. Such process is a pre-requisite for the user to be able to submit a voice sample to the network, and is a key component of the voice verification system. To enhance security and incentivise honest participation, the staking mechanism requires users to regularly re-verify and stake SANG tokens as collateral, in addition to successfully completing the preliminary AI biosignal verification analysis.

The security contribution () combines an AI verification confidence score (, on a 0–1 scale), user-staked tokens (), a stake elasticity parameter ( for superlinear scaling), a time decay constant (), and the time since last verification refresh (), creating a time-adjusted stake mechanism that prioritizes recent biometric confirmations.

# Sybil Resistance

Pairing PoS and physical biosignal analysis alone is not enough to protect the network against sybil attacks, in which a single attacker creates and operates multiple fake identities.



An attacker can recruit humans who are not participating in a network to act as Sybils

Soul Bound Tokens (SBTs) are proposed as a potential decentralised identity system, that would be used to ensure individual credentials are not transferable or replicable, and may therefore prove critical in sybil resistance. However before community recovery mecanisms are broadly adopted by wallet issuers, SBTs are not a complete solution. In the interim the network may leverage a revocable proto-SBT such as ERC-5484, which maintains the non-transferable and non-replicable properties of an SBT, but allows the token to be burnt under certain conditions (Proposals 2022). Sybil resistance must therefore be managed through a layered combination of social and technical mechanisms, including zero-knowledge identity proofs, integrating existing social media account validation, and the use of a reputation system, which can include on-chain activity, repeated engagement and other social signals.

A voice verification network should also support a pseudonymous economy, whereby users can port their authenticated voice biometrics to a pseudonymous identity they control. In fact, in order to bootstrap the network, it may not be necessary for voice biometrics to be coupled with a real identity at all. Such a 'training wheels' phase—liberated from the need for real identities can enable adoption and stress testing while community wallet recovery mechanisms mature.

# Privacy

Current internet architecture limits privacy, often relying on centralised servers that process and store user data (Li and Yang 2021). Privacy for a Voice Verification Network can be maintained by encrypting all biometric and audio data prior to storage in a decentralised cloud network such as IPFS (IPFS 2023).



A wallet signature derives a fernet symmetric encryption key (Fernet 2010) from a Trusted Execution Enviornment (TEE) network such as Phala (Network 2023), making the data undecipherable without key access.



A Content Identifier (CID) is generated when the data is stored and this is deployed into a proto-SBT contract issued to the signing wallet. Only this proto-SBT holding wallet can derive the decryption key from the TEE network.



Self-custody solution for voice data management

# Intellectual Property, Sovereignty and Custody

The unique physiological properties of a voice cannot be understood as Intellectual Property (IP) in a traditional sense, and are not protected by existing IP laws. Though legislation such as the Ensuring Likeness Voice Image Security (ELVIS) Act (Legislature 2024) has been signed into law in Tennessee, US, and bills are proposed in the US and other jurisictions, the ELVIS Act is merely an expansion of *publicity* rights of the voice as "a sound attributable to a particular individual, regardless of whether it contains the actual voice or a simulation" and only applies to unathorised *commercial* use. This makes sense, as the voice is a natural biological property of an individual, it is not a product of human invention, however when commercialised a voice being purveyed in a new work can have attributes that allow it to fit into an IP regime, therefore elements of proposed legislation are concerned with making AI voice models licensable. The No Artificial Intelligence Fake Replicas And Unauthorized Duplications (No AI FRAUD) Act (Representatives 2024) proposed in the US House of Representatives would "recognize that every individual has a property right in their own likeness and voice, akin to other forms of IP rights" and the Nuture Originals, Foster Art and Keep Entertainment Safe (NO FAKES) Act (Senate and Representatives 2024) introduced in both the US Senate and House of Representatives would grant "each individual or right holder the right to authorize the use of their voice or visual likeness in a digital replica, which the bill states is a property right." Though the scope of these proposals targets broad protection for individuals, as well as outlining the commercial terms for licensing, in practice enforcement is likely to be difficult for anyone other than the corporations that rely on IP as a product.

A voice verification network should exist independently of any IP regime, and not be encumbered by politically or commercially vulnerable legislation. A voice verification network should be sovereign, enabling the user to own and control their data, and to enable it to be used as they see fit. This work proposes cryptographic, rather than legal controls, to maintain the integrity of the network, and to ensure that the user’s data is not used for purposes other than those intended. It however may be utilised by nation states to uphold their laws and protect their citizens from deepfake fraud. Voice data custody solutions may be necessary to support a broader rollout, and data custodians may serve as a bridge between the network and voice owner.

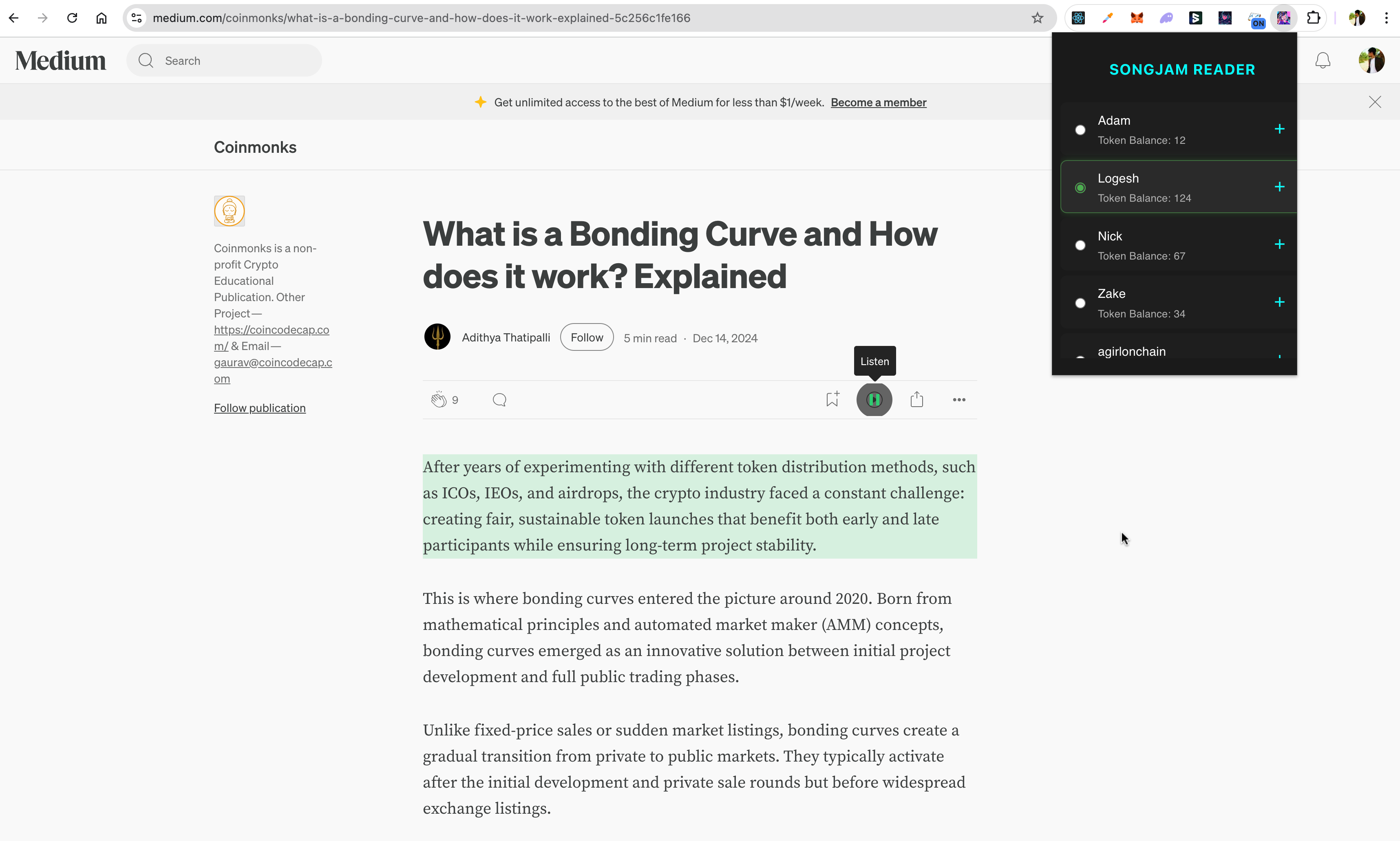


Custodial solution for voice data management

Custodial solutions are often used in the cryptocurrency space, while self-custody is available to users who wish to administer their cryptographic assets independently of any third party.

# AI Voice Agents

Network participants are able to provide thier voice to an AI voice agent, which can be used to provide a voice-based service. The voice model can be provided 'as-is' or in combination with other voices and sources to create unique voices. For example, a voice agent could be deployed in a reader app, a chatbot, or a voice assistant, and paid for with royalties issued in a native token representing the individual voice. Such an AI voice agent marketplace would provide an incentive for users to provide their voice data to the network, and to maintain it over time.



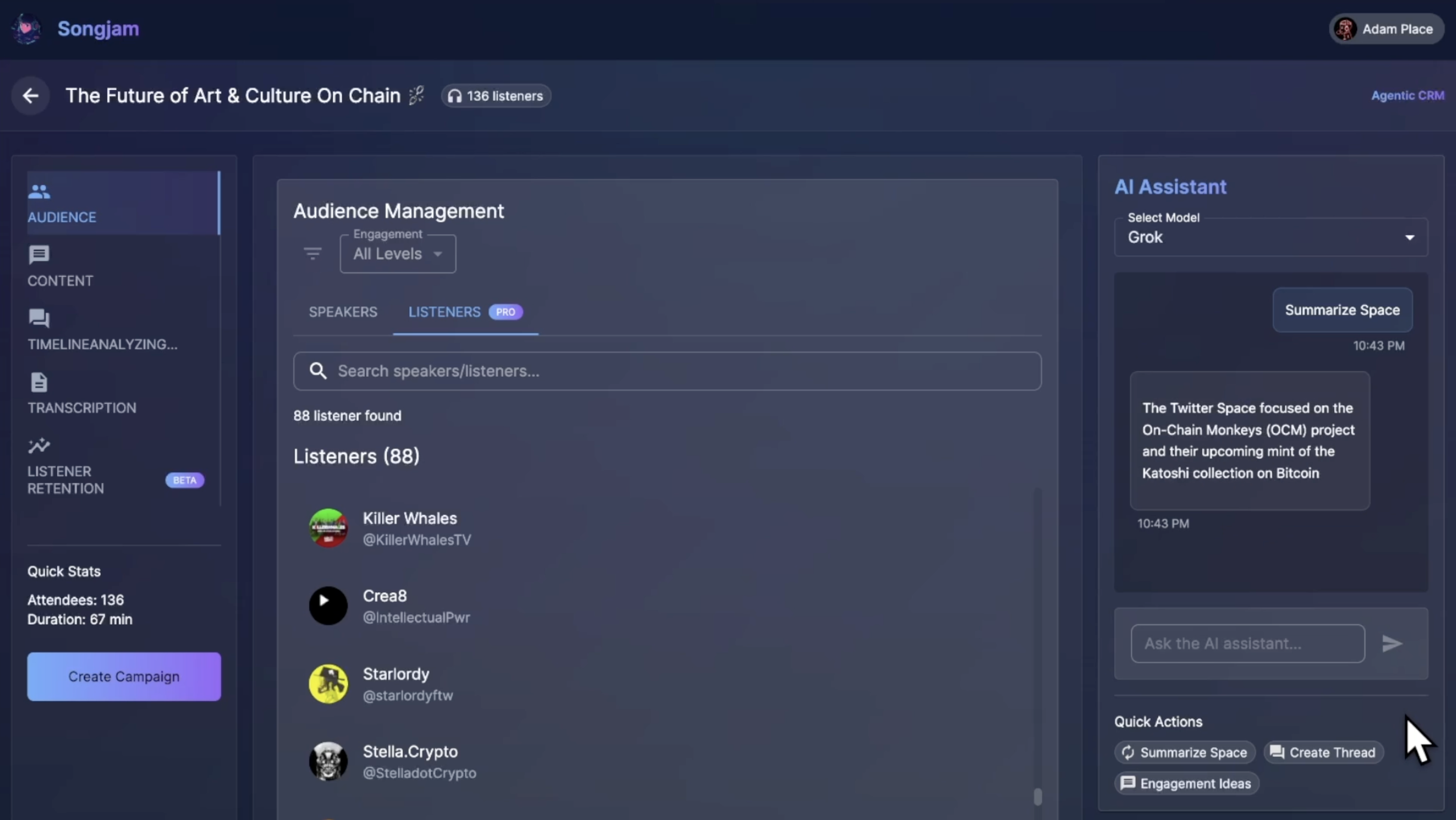
Voice agents deployed in a reader browser extension, displaying remaining token balance

Users stake corresponding tokens to enable use of an agent, which are burnt based on usage:

Where is the remaining stake at time , is the number of tokens spoken by the AI model for utterance , and is the total number of tokens in the model’s vocabulary. This ensures that one staked token is burned for each token spoken by the AI voice agent, with partial tokens being accumulated until they sum to a whole token.

Another early use-case sees users spinning-up thier own voice agent once they have recorded a sufficient amount of voice data, and deploying it inside an X (formerly Twitter) Space—interactive live audio conversations which have become popular forums for the cryptocurrency community. By participating in these X Spaces with AI voice agents trained on their voice data and personality, they can grow their X following, expanding their reach and earning potential.

An AI voice agent marketplace can be introduced through the Songjam agentic CRM, an exisiting platform designed to support growth and anlytics for X users by automating processes such as direct messaging, tracking listener data and Automatic Speech Recognition (ASR) with an emphasis on X Spaces.



Songjam agentic CRM interface for managing X Spaces

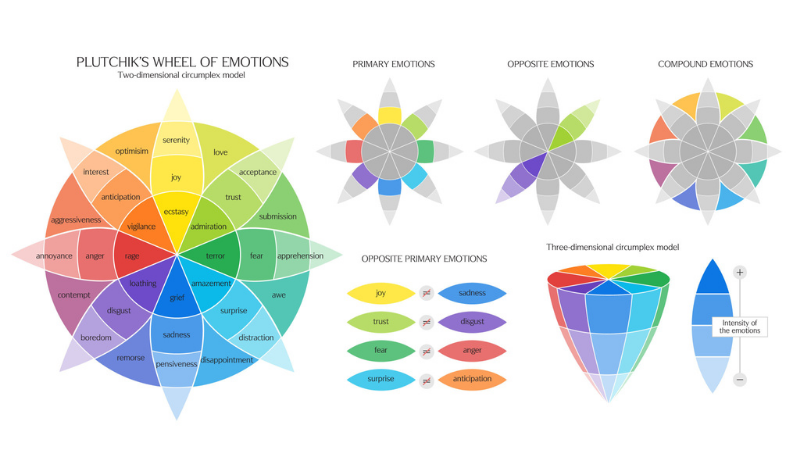
Focusing on an X Spaces native sandbox can encourage users to spin up creative AI voice agents with customisable personalities, using their own voice or a combination of available voices to create unique voice agents with royalties being split between the original voice owner(s) and the agent owner. As mentioned in section 3, such a 'training wheels' phase can support early network growth, helping to ensure the network is robust, while providing a platform for users to experiment within the voice agent ecosystem. There may be several iterations of the voice agent marketplace and its underlying protocol that must be rolled out and stress-tested at scale before a fully decentralised, cryptographically secure voice verification network can reach maturity.

# Data

Retrieval Augmented Generation (RAG) is a data standard which supports context-rich retrieval for AI models such as Large Language Models (LLMs). It has become a key component of the AI ecosystem, giving LLMs the ability to access and utilise external data sources to enhance their responses. WavRAG (Chen et al. 2025) is a framework that integrates audio data into RAG pipelines without requiring an intermediate ASR step, therefore reducing computational overhead, preserving acoustic information and leveraging a unified multimodal knowledge base ideal for voice data. Despite these advantages, WavRAG is not yet widely adopted, and the extent to which RAG can contribute to the acoustic aspects such as intonation, expressiveness and other voice characteristics essential to AI voice agents has yet to be extensively investigated.

Furthermore, there are limited high-quality emotion datasets available for the training of voice models, and many researchers are resorting to synthetic data generation due to the cost of collecting and annotating real data. Model collapse is predicted when a model is trained on recursive synthetic data (Shumailov et al. 2024), in other words when a model is trained on data itself generated by a model. Though there is limited research into voice model degeneration due to these causes, why risk it when X Spaces presents a goldmine of real-world voice data for training and testing?

Incentivising users to annotate and own their voice data can help to address these issues, and can be supported by the network through a voice data marketplace. Users may not charge the same price for their voice data as they would for their time, particularly when captured in a public forum they participate in for their own professional or personal benefit.



Plutchik’s Wheel of Emotions

A speech database can draw from psychologist Robert Plutchik’s Wheel of Emotions (Plutchik 1980), which is a model of human emotions based on the theory of psychotherapy. The wheel is divided into 8 primary emotions, and subsequent sets of secondary and tertirary emotions which represent different levels of intensity, as well as compound emotions. X Space data is recorded and catagorised by the speaker, or other participants in the space including listeners. Metadata is enriched with further context and this forms the foundation of a structured speech dataset.



Process of annotating voice data using Plutchik’s Wheel of Emotions

Making such data available can be defined by the proto-SBT holder, they may wish to keep the data private, give it away for free, monetise it by selling it to be used for training an AI model, or use it for their own voice agent exclusively. Programmable terms can be used to define the conditions under which the data can be used, and the royalties that are due to the voice owner(s). Compiling such a structured dataset in a WavRAG pipeline and enabling the network to define its own terms would be a step towards a sovereign open-source AI voice ecosystem.

# Conclusion

This work proposed a cryptographically secure voice verification network, and a framework for the development of an open-source AI voice ecosystem, including a voice agent and voice data marketplace. Powered by cryptoeconomics and open-source hardware (**originstory2024?**), with market demand the network can be bootstrapped rapidly, and scaled to support the needs of the global population. Leveraging a Proof-of-Stake consensus mechanism, the network is secured through its participants. The network aims for a future where voice data is owned by the individual, and can be used as they see fit. Nevertheless, the network may be used in a custodial manner, and data custodians may administer voice data for the benefit of indiviudal voice owners. This architecture establishes a durable paradigm for voice data security, ensuring persistent user control and safeguarding individual voice assets against emerging threats in an adversarial digital landscape.

Buterin, Vitalik, E. Glen Weyl, and Puja Ohlhaver. 2022. “Decentralized Society: Finding Web3’s Soul.” *Available at SSRN 4105763*.

Chen, Yifu, Shengpeng Ji, Haoxiao Wang, Ziqing Wang, Siyu Chen, Jinzheng He, Jin Xu, and Zhou Zhao. 2025. “WavRAG: Audio-Integrated Retrieval Augmented Generation for Spoken Dialogue Models.” *arXiv Preprint arXiv:2502.14727*. <https://doi.org/10.48550/arXiv.2502.14727>.

Federal Trade Commission. 2024. “OriginStory: A Framework for Verifying the Authenticity of Digital Content.” *FTC Abstract*. <https://www.ftc.gov/system/files/ftc_gov/pdf/OriginStory-Abstract.pdf>.

Fernet, Fernet. 2010. “Fernet (Symmetric Encryption).” *Python Cryptography*.

IPFS. 2023. “IPFS: InterPlanetary File System.” *Protocol Labs*.

Legislature, Tennessee State. 2024. “Ensuring Likeness Voice Image Security (ELVIS) Act.” *Public Chapter No. 926*.

Li, Hui, and Xin Yang. 2021. “Co-Governed Sovereignty Network: Legal Basis and Its Prototype Applications with MIN Architecture.” *ResearchGate*. <https://www.researchgate.net/publication/353484102_Co-governed_Sovereignty_Network_Legal_Basis_and_Its_Prototype_Applications_with_MIN_Architecture>.

Liu, Tao, Mingyang Zhang, Zhihao Li, Hanjie Dou, Wangyang Zhang, Jiaqian Yang, Pengfan Wu, Dongxiao Li, and Xiaojing Mu. 2025. “Machine Learning-Assisted Wearable Sensing Systems for Speech Recognition and Interaction.” *Nature Communications* 16 (2363). <https://doi.org/10.1038/s41467-025-57629-5>.

Network, Phala. 2023. “Phala Network: A Privacy-Preserving Cloud Computing Service.” *Technical Documentation*.

Plutchik, Robert. 1980. “A General Psychoevolutionary Theory of Emotion.” *Theories of Emotion* 1: 3–33.

Proposals, Ethereum Improvement. 2022. “ERC-5484: Consensual Soulbound Tokens.” *EIP-5484*.

Representatives, US House of. 2024. “No Artificial Intelligence Fake Replicas and Unauthorized Duplications (No AI FRAUD) Act.” *H.R. 6943*.

Senate, US, and House of Representatives. 2024. “Nuture Originals, Foster Art and Keep Entertainment Safe (NO FAKES) Act.” *S. 1271 and H.R. 6943*.

Shumailov, Ilia, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. “AI Models Collapse When Trained on Recursively Generated Data.” *Nature* 631: 755–59. <https://doi.org/10.1038/s41586-024-07566-y>.