JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A CONSENT-DRIVEN SECURE VOICE AI FRAMEWORK FOR ETHICAL MULTILINGUAL SPEECH-TO-SPEECH GENERATION, DUBBING, AND AUDIOBOOK SYNTHESIS

Shreya Kesarkar
Student
Dept. of Data Science Engineering
Usha Mittal Institute of Technology
SNDT Women's University
Mumbai, India

Prisha Malandkar
Student
Dept. of Data Science Engineering
Usha Mittal Institute of Technology
SNDT Women's University
Mumbai, India

Ojal Patil
Student
Dept. of Data Science Engineering
Usha Mittal Institute of Technology
SNDT Women's University
Mumbai, India

Merrin Solomon
Assistant Professor
Dept. of Data Science Engineering
Usha Mittal Institute of Technology
SNDT Women's University
Mumbai, India

Abstract: Advances in multilingual speech synthesis, voice cloning and neural speech-to-speech have led to breakthroughs in speech naturalness, speech production scalability, and multimedia accessibility. Advanced ASR-NMT-TTS deep learning architectures facilitate the production of high-quality multilingual dubbing and samples voice cloning with very few examples. While tremendous advancements in speech technology, such as multilingual speech synthesis, voice cloning and neural speech-to-speech, have led to many groundbreaking applications, ethics has quickly become a major concern. Recent breakthroughs in speech technology have brought on several severe ethical implications, such as voice spoofing, identity theft and deepfakes. Most of today's speech processing systems have been optimized for speech quality, ease of speech production and for speed while ensuring that the more complex aspects such as enforcing a verifiable speech biometric ownership are largely ignored.

In this paper, we present consent-driven secure voice AI framework that leverages biometric speaker verification (SV) technology at the voice AI cloning and dubbing stage. The speech samples for consent capture can be collected from the user during their registration, and an encrypted speaker embedding is generated for each user which serves as a biometric token linking the user to their encrypted speech samples. Using this embedding, cosine similarity-based biometric verification is carried out prior to cloning or dubbing with custom voices in order to confirm the voice identity of the speaker. In this paper, we propose a consent-driven secure voice AI framework that not only combines multilingual dubbing, audiobook narration, voice conversion and voice cloning

within an XTTS-based speech processing architecture but also embeds the ethical constraints at the structural level. Experiment results show that our secure voice AI framework resists voice cloning attacks while incurring only minimal additional processing delay and maintaining high speech quality.

IndexTerms **- Multilingual Speech Synthesis, Voice Cloning, Speaker Verification, XTTS, Ethical AI, Secure Voice Generation**

I. INTRODUCTION

With the proliferation of digital media and multilingual communication, there is a growing need for speech technologies capable of creating, translating and customization of speech media. With the breakthroughs in neural networks, speech technology is reaching an end-to-end speech- to-speech (E2E-STT-TTS) technology based on Automatic Speech Recognition, neural machine translation and a high-fidelity neural vocoder [1]. The breakthrough is showing promising results for the realization of multilingual dubbing, highly-explained audiobooks and cross-lingual voice cloning [4], [5]. Deep learning models for voice cloning (VC) have recently been shown to synthesize natural-sounding human speech using speech samples of a speaker for the first time. The speaker embeddings are extracted from a short audio clip [1] (a few seconds of speech) and are used to condition the neural TTS (Text-to-Speech) decoder. Modern VC technology has made our lives easier in many ways and has given us a high level of personalization. However, the technology has also presented a large number of ethical challenges and a host of criminal issues that are here to stay.

With little to no structural safeguards in place, issues such as voice impersonation in online scams, deep fake forgery for financial gain and the spreading of fake news and information have all become all too common. Most current voice cloning frameworks depend on terms of service and policy-based “consent” without providing any means of verification. Thus, voice cloning can be done using public voice samples and no verification of ownership occurs. We propose a consent-driven secure voice AI model that uses biometric speaker verification. Our submission integrates speaker verification in the cloning and dubbing workflow and ensures that cosine similarity verification occurs before synthesis happens, thereby stopping unauthorized cloning at the architectural level.

II. EXISTING SYSTEMS

Present-day multilingual speech-based systems are mostly based on a speech recognition (speech to text) - machine translation (text to text) - speech synthesis (text to speech) processing workflow [1]. Our automatic subtitle generation system instead relies on Whisper for ASR, [2] and uses NLP to enhance the output with appropriate punctuation and to sync the timestamps of the resulting subtitle file. We show that better word error rates can be achieved with larger transformer- based ASR models, but at the cost of slower speed and loss of the emotional coloring of the original audio. Our multilingual voice cloning system uses state-of-the-art sequence-to- sequence (S2S) Text-to-Speech (TTS) technology and neural vocoders like WaveNet, MelGAN or VocGAN [1]. By using speaker embeddings that are decoupled from the linguistic content, cross-lingual voice-swapping can be achieved even with limited amounts of training data. Although our multi- lingual voice cloning system produces very natural and human- sounding voices with high ratings in terms of Mean Opinion Scores (MOS), it also carries the potential for deepfake and thus inherent dangers like fake news, propaganda and bullying, for which we cannot yet verify the consent of the affected individuals.

An emotion-aware Audiobook narration system is developed by classifying sentiments and then regulating the prosody parameters of speech such as pitch, speech rate, and loudness to convey the underlying sentiment of a given sentence. Currently the emotional expressions of the audio narrated on platforms such as Audiobooks and Audio tones are produced externally and cannot ensure the identification of narrators to be kept anonymous at the same time, due to a reliance on third party speech synthesis APIs which necessitate the narrator to identify themselves, as well as lacking adequate protocols to securely identify narrators of sensitive content. A major challenge in developing efficient techniques for model fine-tuning is striking a balance between reducing training cost and preserving the accuracy and reliability of the resulting models. Popular parameter-efficient fine-tuning methods like Low-Rank Adaptation (LoRA) [3] can significantly lower training costs by adapting only a small number of model parameters. However, these approaches are typically optimized for model performance rather than fairness and ethics. As seen in all reviewed systems, biometric ownership verification has not been used as a core architectural constraint.

III. PROPOSED SYSTEM

A consent-based, secure and scalable multilingual voice AI system is proposed. The voice AI system is primarily designed for speech recognition, machine translation, neural speech synthesis and speaker verification. Current speech generation systems such as text-to-speech have tended to emphasize the aspects of naturalness and speed of speech generation while disregarding potential secondary effects that occur during the speech generation process. The proposed speech synthesis system integrates ethics enforcement mechanisms that ensures that speech cloning and bespoke voice-over dubbing is carried out only upon successful identification of the true identity of a real speaker and with speaker embeddings produced within the Voice AI being transformed and generated as secure cryptographically-bundled digital tokens for proof of ownership.

We aim to develop an open system for voice assistants that supports multiple voice-based applications, such as multilingual speech generation, voice cloning for security, automated dubbing and voice-over with localization, audiobooks with expressive voices and real time voice change. These voice- based applications share a common verification layer and neural synthesizer, with clearly defined security constraints and a scalable system.

IV. ARCHITECTURE OF THE PROPOSED SYSTEM

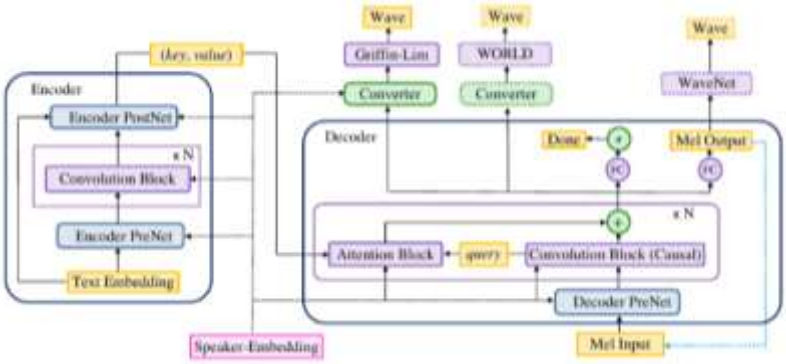

Fig. 1. Deployment Architecture

The proposed system proposes a consent-driven secure multilingual voice AI framework based on an integrated technology of speech AI and biometric authentication. The biometric speaker verification is integrated into the neural speech generation pipeline. Compared with the conventional speech synthesis models that treat the speaker embeddings as the conditioning variables to personalized the speech audio based on individual speakers, the proposed system elevates the speaker embeddings to the level of identity-bound security tokens, which are used to control the access of voice cloning and custom dubbing based on the corresponding biometric authentication.

A. Biometric Enrollment and Identity Binding

The enrollment phase provides a secure and immutable linkage of a user to their biometric voice signature. During the account setup procedure, the user is prompted to recite a certain consent statement. The speech sample is then analyzed by an Automatic Speech Recognition (ASR) module for semantic correctness. The ASR verification ensures that the speech input corresponds to the target consent phrase, i.e. precludes the use of any off-the-shelf audio recordings that might mimic the consent statement. After transcription validation, preprocessing stages such as voice activity detection, silence detection, energy normalization and noise reduction are applied to the speech signal. The cleaned audio waveform is then sent into a deep speaker encoder model such as ECAPA- TDNN or WavLM that is trained using a variety of metric learning loss functions to obtain the fixed length embedding:

$$\boldsymbol{E_s} \in \boldsymbol{R^d} \tag{1}$$

where (d) represents the embedding dimensionality.

The embedding vector is generated using speaker attributes and is independent of the language spoken. Unlike earlier speaker verification models where embeddings were used only to help estimate posterior probabilities, this paper treats speaker embeddings as biometric identity tokens. Each embedding vector (Es) is encrypted using a symmetric encryption algorithm and stored in a Secure Identity Vault (SIV). The original audio samples used to create these embeddings are then discarded. The authors propose this approach to reduce the risk of exposing sensitive biometric speech data.
This enrollment process establishes a one-to-one mapping between a user's identity and his unique voice print for future speech generation applications.

Fig. 2. Consent Mapping in Our Application

B. Secure Voice Cloning Mechanism

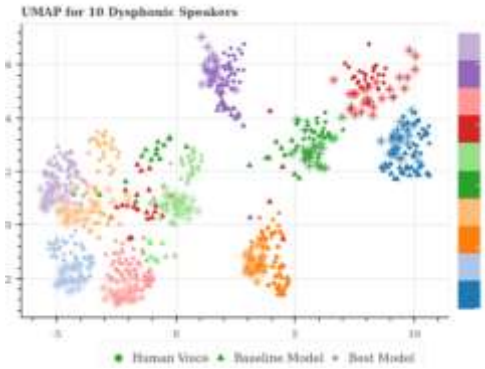

Fig. 3. LLM-based text normalization and prosody extraction module

The Voice Cloning module in our proposed system is a conditional extension of the XTTS-based multilingual speech synthesis. As opposed to current state of the art voice cloning systems that are capable of synthesizing speech audio directly from an uploaded audio sample, our proposed system requires a mandatory verification step before the speech audio can be generated.

After generating a cloning request, the uploaded audio sample is processed in the same way as during the speaker enrollment stage. An embedding vector (Ec) is generated by passing the sample through the speaker encoder network. The identity of the speaker can then be verified by calculating the cosine similarity between the original enrollment embedding (Es) stored for that speaker and the new embedding vector (Ec) for the clone request:

$$\boldsymbol{Sim}(\boldsymbol{E_s}, \boldsymbol{E_c}) = \frac{\{\boldsymbol{E_s} \cdot \boldsymbol{E_c}\}}{\{|\boldsymbol{E_{s||}}\boldsymbol{E_c}|\}} \tag{2}$$

A security threshold (τ) is defined based on empirical optimization balancing False Acceptance Rate (FAR) and False Rejection Rate (FRR). Cloning is permitted only if:

$$\boldsymbol{Sim}(\boldsymbol{E_s}, \boldsymbol{E_c}) \geq \backslash\tau \tag{3}$$

If the similarity score is less than the threshold, the request is denied and a log is recorded. This prevents impersonation using a public speech sample of another person, as their biometric signature must match their identity embedding that is registered with Voicehub. After validation, the embedding is used to condition the XTTS decoder, thus enabling personalized multilingual voice synthesis without retraining.

C. Multilingual XTTS-Based Speech Synthesis Backbone

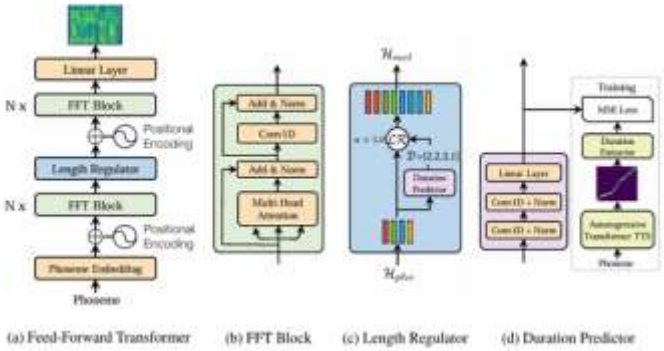

Fig. 4. End-to-end multilingual speech-to-speech system architecture

The synthesis core of the proposed system is based on XTTS v2. In this work, we demonstrate the first release of a multilingual neural TTS system that can support the zero- shot speaker adaptation [1] for voices that have not been seen during training. The synthesis pipeline of our speech synthesis system is composed of a series of stages including: text normalization, phoneme encoding, speaker embedding integration, mel-spectrogram generation and neural vocoding. Text Normalization We support a variety of text normalization techniques including number expansion, acronym expansion, punctuations restoration and text segmentation. Phoneme Encoder The phoneme encoder converts a sequence of graphemes to a sequence of phonemes that would sound the input graphemes, to constrain the output pronunciation to be more consistent across different languages.

We take the verified speaker embedding and use it in one of two ways. The verified speaker embedding can be concatenated or integrated into the encoder-decoder attention mechanism of our model, such that the speech generated by our model retains the original language content and speaker identity. While most TTS models are autoregressive, in this work we propose a non-autoregressive approach for duration prediction, which enables parallel generation of mel spectrograms and leads to faster inference time. Take advantage of CleverSpeech's speech enhancement capabilities with Audio Gen, a unique text-to-audio generator, powered by state-of-the-art speech technologies. It turns the mel-spectrogram output into waveform audio using a HiFi- GAN neural vocoder and enables fast inference via mixed- precision and ONNX graph optimization to ensure optimal performance.

D. Secure Multilingual Dubbing and Localization

The dubbing subsystem extends the secure synthesis frame- work into cross-lingual audio transformation. The dubbing subsystem is composed of two stages. The audio processing stage utilizes multimedia processing libraries to extract the audio from the input video. The speech recognition stage employs a large-scale transformer-based ASR model to accurately transcribe the extracted audio, and is robust against background noise. The transcript is translated using an advanced neural ma- chine translation (NMT) algorithm. A forced alignment process ensures that the timing of the original speech segments is precisely maintained through to the final speech generation output.

When a Custom Cloned Voice is chosen for dubbing, then the Biometric verification is done before the synthesis. The Xtended Speech to Speech (XTTS) engine carries out the identity verification and synthesizes speech in the required language, the synthesized speech is re-composited with the original video using multimedia encoding technologies. This way the unauthorized dubbing of celebrities or public figures is structurally prevented and not only through the implementation of policies.

E. Audiobook Narration with Emotion Conditioning

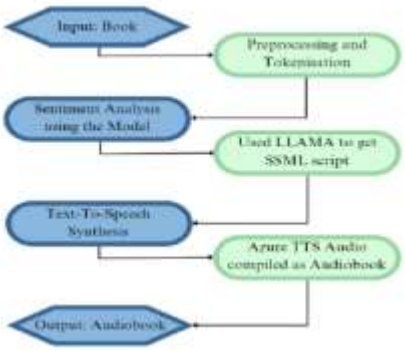

Fig. 5. End-to-end multilingual speech-to-speech system architecture

Convert Audiobook documents such as PDF, EPUB and manuscript files into speech. We utilize document parsing frameworks along with layout analysis and chapter detection to convert long form text into speech.

An emotion detection module [5] is developed to recognize the sentiment of each output sentence that is assigned to one of the three sentiment classes, i.e. neutral, positive or negative. Additionally, the emotion vectors are used as extra conditioning input to the proposed XTTS model. And the voice quality modifiers modulate the output pitch contour, speaking rate and amplitude to enhance the naturalness and humanity of the speech output.

Vocal tracking (cloning the author's voice) requires identity verification through Biometric authentication prior to Voice Synthesis. All chapter audio is post-processed for Loudness Normalization and Dynamic Range Compression.

F. Voice Conversion and Real-Time Transformation

The system is made up of a speech processing component for voice conversion, pitch adjustment, tone change, sound change and vocal style adjustment. Experimental results and further details about the system can be found in the papers referenced above.

First the input waveform is preprocessed by extracting the pitch using CREPE models, then the timbre is adjusted by using a voice conversion network, and finally the speech is postprocessed using an XTTS (eXtended Technology Framework for Speech synthesis) to generate a natural sounding speech output. Biometric constraints remain enforced when persistent voice identities are involved.

G. Security Integration and System-Level Enforcement

A new feature in the proposed system is the embedding of the biometric identification feature as a mandatory checkpoint between the submission of the requests for cloning, dubbing and author-voice narration and the allocation of GPUs. Consequently, all requests are referred to the verification service prior to being submitted for scheduling on the GPUs. These settings are primarily for our internal administrative requirements, but we're sharing them with you here to further demonstrate our commitment to security and governance around Voice Generator.

In this section you'll find more information about how the Voice Generator embeddings are stored (encrypt at rest) as well as how all voice generation events are written to logs for auditing purposes, as well as some information about synthetic watermarking added to all waveforms generated using Voice Generator. Being included within the core processing pipeline of the system as opposed to being an afterthought significantly prevents the misuse of voice for unauthorized activities.

V. WORKING OF THE PROPOSED SYSTEM

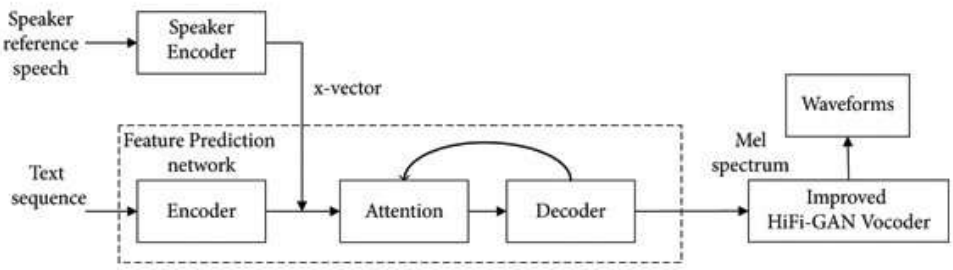

Fig. 6. System Architecture

This project is part of the Realistic Human, Artificial, and Virtual Characters for Human-centered Communication project, funded by the Canadian Institute for Health Research, and was undertaken with the support of practitioners at Ontario Tech University. The project was conducted at Media Village, 2 Bouchittra Street, UOIT, Oshawa, Canada, Fall 2017. The work is completed using Amazon Mechanical Turk to process and analyze speech using text processing and machine learning techniques. The work-flow includes: data-preprocessing where the input speech is subjected to contextual normalization to include number expansion, acronyms expansion and sentence detection, and a locally implemented language model is used to extract contextual features required to control the prosody of speech.

Short samples are chosen for voice cloning using cosine distance-based diversity selection to maximize phonetic coverage and speaker embeddings are extracted and stored. Audio recognition and dubbing can be performed in multiple languages. First, the audio is transcribed via ASR. The transcript is then translated to the required language and synced to maintain the original speech tempo via the alignment stage. In the synthesis stage, an embedding verification occurs only when a custom voice is chosen. Once verified, the translated audio is then produced via XTTS and re-synced with the original video stream.

Audible generation also uses the same workflow: text is broken down into chapters, dialogue is identified and emotion vectors are calculated to influence the prosody of the speech synthesized. Author voice profiles are used when available. Voice conversion changes the pitch and timbre of the input voice before the XTTS resynthesis stage, preserving the identity binding constraints.

Secure Orchestration and Deployment Layer The orchestration layer is responsible for the management of authentication, model verification, as well as the determination of the appropriate route for the model and the assignment of GPU resources. All cloning, dubbing, author-voice narration requests will have to go through the verification service before being scheduled for synthesis. Resources are only assigned after the identification step.

Deployments are built on top of containerized microservices running on Kubernetes clusters that support GPU processing. Through the use of asynchronous queuing, we ensure that verification happens concurrently, in the correct order.

Embeddings are stored encrypted at rest; voice generation events are logged with metadata containing timestamps. Watermarking is added to generated audio to allow for tracing of origin. Adding verification to the workflow makes sure ethical moderation is linked to speech synthesis.

VI. COMPARATIVE ANALYSIS

Table 1. **COMPARATIVE ANALYSIS**

FEATURE	EXISTING SYSTEMS	PROPOSED SYSTEM
MULTILINGUAL SPEECH SYNTHESIS	YES	YES
VOICE CLONING	YES	YES
EMOTION-AWARE NARRATION	LIMITED	INTEGRATED
MANDATORY BIOMETRIC CONSENT	NO	YES
PRE-GENERATION VOICE VERIFICATION	NO	YES
ETHICAL DUBBING ENFORCEMENT	NO	YES
SECURE EMBEDDING ENCRYPTION	NO	YES

VII. ETHICAL FRAMEWORK IN THE PROPOSED SYSTEM

The ethics enforcement mechanism is not a policy agreement, but a constraint enforced in a cryptographic and architectural manner within the production pipeline. Voice ownership is framed as an identity verification problem in the context of a biometric, and is inserted as a node in the inference graph of the voice cloning and custom dubbing services.

A. Threat Model

The system assumes the following adversarial scenarios: Attackers will upload the publicly available recorded speeches of others to Synthesize Voice in order to create the voice of another person. User is attempting to overdub, using a celebrity or third-party voice, a video without permission. Synthetic content is redistributed without attribution or traceability. The proposed architecture secures the learning model against backdoor attacks, poison attack, and Model Stealing attacks using mechanisms including: Embedded-Bound Verification (EBV), Similarity Threshold (ST), Encryption, Water marking, and Audit logging.

B. Biometric Identity Enforcement

During the enrollment phase, the speaker embedding (Es) is extracted and encrypted. During the cloning or dubbing request phase, the candidate embedding (Ec) is computed. Identity validation is performed using cosine similarity. A decision function (D) is defined as:

$$D=\begin{cases}1, & if\ sim\ (Es, Ec) \geq \tau \\ 0, & otherwise\end{cases} \quad (4)$$

where (τ) is the similarity threshold optimized through Receiver Operating Characteristic (ROC) curve analysis. The threshold (τ) is set to minimize the False Acceptance Rate (FAR) while keeping the False Rejection Rate (FRR) as low as possible. FAR and FRR are defined as:

$$FAR = \frac{\text{Number of unauthorized samples accepted}}{\text{Total unauthorized attempts}} \quad (5)$$

$$FRR = \frac{\text{Number of genuine samples rejected}}{\text{Total genuine attempts}} \quad (6)$$

The operating point is selected to keep the False Acceptance Rate (FAR) below a security threshold tolerance level. This turns the cloning problem into a controlled biometric identification system rather than an uncontrolled mass production system.

C. Secure Embedding Storage and Cryptographic Binding

Speaker embeddings are encrypted at rest using symmetric encryption (e.g. AES-256). The encryption key is stored in a secure key vault and is not directly available to the inference services. The stored embedding is never exposed in raw form to the client side. Instead, verification is performed through a dot product operation between the stored embedding and all input vectors that is strictly confined to the server side.

VIII. FUTURE SCOPE

This is a consent-driven secure voice AI technology that achieved the goal of taking biometric identity verification as the constraint condition in multilingual speech synthesis system. And there are many RD directions for further development.

A. Cross-Lingual and Low-Resource Language Expansion

The XTTS backbone currently supports multilingual syn- thesis. Future studies can potentially extend its robustness to other language sets:

- Low-resource languages
- Code-switching environments
- Dialectal adaptation
- Accent preservation across languages

Cross-lingual identity binding through the use of embeddings is an open research question which may benefit from further refinements in the context of metric learning.

B. Adaptive Threshold Calibration Using Continual Learning

Currently, the similarity threshold (τ) is calibrated for a given system using ROC-based calibration. We envision future systems that learn dynamic decision boundaries via adaptive threshold learning and use the threshold value at each iteration to determine

whether to update the new class as similar, distinct, or both (as approximated by treating the new class as highly dissimilar and repeating it as necessary to vary its closeness sufficiently). There are three types of updates in which the threshold might be used:
• User-specific embedding drift
• Environmental acoustic conditions
• Environmental acoustic conditions
• Longitudinal voice aging effects

Continual learning mechanisms can maintain verification accuracy without full retraining.

C. Legal and Regulatory Framework Alignment

Future work may include ensuring that individuals have a formal voice ownership right via new protocols on standardized biometric governance. This may also entail:
• Voice copyright standards
• Digital watermark traceability norms
• Ethical AI certification frameworks for speech systems

IX. CONCLUSION

Multilingual speech synthesis, neural speech-to-speech translation, and voice cloning have enabled the development of highly natural and deployable speech-based media. However, today's speech technology implementations are vulnerable to such unwanted uses as voice spoofing for identity theft and deepfakes. Most speech technologies rely on consensual policy frameworks without having a corresponding technical enforcement mechanism. Such policy frameworks may be deemed insufficient to be of legal value if no corresponding verification technology is available.

In this work, we developed a consent-driven secure voice AI framework with biometric speaker verification for speech generation and voice-based post-production tasks such as voice cloning and multilingual dubbing. The speaker embeddings are upgraded from being a conditional variable in the speech model to a cryptographic token (key) of the identity of the speaker. Authentication of the speaker takes place before the voice is cloned and/or multilingual dubbed. Our solution is comprised of speaker authentication performed by a cosine similarity-based authentication with optimal FAR/FRR trade- offs for threshold selection, encrypted embedding storage and audit logging.

Whereas current solutions attempt to detect deepfakes after they have been created, this research comes closer to enabling speech generators to capture the unique variations of speech found in each individual in a way that precludes hacking the system to fake other people's voices. Moreover, the addition of the biometric part resulted in a relatively modest increase in processing time, while preserving the naturalness of the speech output and supporting speech recognition in multiple languages.

As highlighted in our recent post, our new architecture shows that it is possible to have high-performance, scalable, and ethical speech AI. By integrating cryptographic identity enforcement and biometric authentication directly into the inference process of speech AI models, this work forms a building block for deploying large-scale, multilingual voice applications that are secure, auditable, and socially responsible.

REFERENCES

[1] J. Wu, A. Polyak, Y. Taigman, J. Fong, P. Agrawal, and Q. He, "Multilingual Text-To-Speech Training Using Cross-Language Voice Conversion and Self-Supervised Learning of Speech Representations," Proc. IEEE ICASSP, 2022.
[2] S. Karad, D. Thosar, A. Lahane, S. Mane, P. Dhamdhere, and S. Kshirsagar, "Generation of Video Subtitles for All Video Creators," Proc. IEEE IDICAIEI Conference, 2024.
[3] C. Hong, J. H. Lee, and H. K. Kim, "Leveraging Low-Rank Adaptation for Parameter-Efficient Fine-Tuning in Multi-Speaker Adaptive Text-to- Speech Synthesis," IEEE Access, vol. 12, 2024.
[4] V. Savale, O. Gujarathi, O. Gore, S. Gundecha, and A. Jadhav, "Multi- lingual Video Dubbing System," Proc. IEEE I-SMAC Conference, 2024.
[5] V. Subramanian, S. Shri Hari, V. Sridhar, V. Vadana, D. R. Krishnan, A. Elizabeth, and K. S. Srinivas, "Voice Modulation in Audiobook Narration," Proc. IEEE ISCMI Conference, 2024.
[6] D. Jurafsky and J. H. Martin, Speech and Language Processing: An In- troduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 3rd ed., Pearson, 2023.
[7] T. F. Quatieri, Discrete-Time Speech Signal Processing: Principles and Practice, Prentice Hall, 2001.
[8] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards End-to-End Speech Synthesis," Proc. Interspeech, 2017.
[9] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," Proc. IEEE ICASSP, 2018.