International Journal of Engineering and Science Invention (IJESI)
ISSN (Online): 2319-6734, ISSN (Print): 2319-6726
www.ijesi.org ||Volume 15 Issue 9 September 2026 || PP 29-35

Secure Voice-Driven Trading System Using Speaker Biometric Authentication: A Latency and Reliability Evaluation Framework

Dr T.A.Ashok Kumar[1], AbasAshkarSA[2]
[1]*Professor SSCS, CMR University, Bengaluru.*
[2]*Postgraduate Studies SSCS, CMR University, Bengaluru.*

ABSTRACT: *Voice interfaces are increasingly used in financial services to place trades, check portfolios, and issue account instructions hands-free, but this convenience introduces two competingrequirements:strong identity assurance and sub-second responsiveness. This paper reviews recent work on speaker verification, replay and deepfake liveness detection, and low-latency edge authentication, and proposes VoxTradeSecure,a framework for a secure voice-driven trading system builtaroundlayeredspeakerbiometricauthentication.The proposedframeworkintegratesvoicecaptureandpreprocessing,embedding-basedspeakerverification, anti-spoofing and liveness detection, trading-intent recognition and command validation, and a risk-adaptive step-up authentication layer, while instrumenting every pipeline stage for latency measurement. A reliability evaluation framework is also defined using False Acceptance Rate (FAR), False Rejection Rate (FRR), Equal Error Rate (EER), spoof-detection accuracy, and system availability under adverse acoustic and network conditions. Illustrative benchmark results indicatethatlayeredlivenessdetectionincreasesend-to-endresponse time but substantially reduces spoofing success rate, motivating a tiereddesigninwhichroutinetradesusefast single-factor voice verification while high-value orders trigger additional liveness and step-up checks. The framework is intended to help brokerages and fintech platforms reason about the trade-off between security, reliability, and responsiveness when voice becomes a transaction channel.*

KEYWORDS-*Speaker Verification, Voice Biometrics, Anti-Spoofing, Liveness Detection, Algorithmic Trading, Latency Evaluation, Financial Security*

Date of Submission: 12-09-2026 Date of acceptance: 24-09-2026

I.INTRODUCTION

Voice has moved from a novelty input method to a mainstream channel for interacting with financial systems. Brokerage applications, robo-advisors, and contact-centre trading desks increasingly expose voice assistantsthatletacustomercheckaposition,placeabuyorsellorder,orconfirmapendingtransactionwithout touching a screen. This shift is driven by the broader growthofvoicebiometricsinbanking,financialservices, and insurance, where speaker verification is being adopted as a faster and more convenient alternative to passwords, PINs, and one-time codes for remote authentication.

However, voice as a transaction channel concentrates two requirements that are normally addressed separately in security engineering: strong identity assurance and very low response latency. In a typical login scenario, an authentication delay of one or two seconds is imperceptible to the user. In a trading scenario, the same delay can determine whether an order is filled at the intended price, particularly during periods of high volatility. At the same time, voice isauniquelyexposedbiometric:itcanbecapturedremotely,replayedfroma recording, or synthesised convincingly with commercially available voice-cloning tools that need only a few minutes of a target speaker's audio to produce a usable clone. Recentlarge-scaleevaluationsshowthatmodern open-source voice-cloning systems can bypass commercial speaker-verification platforms in a substantial fraction of trials, and that anti-spoofing detectors trained on one family of synthesis methods often fail to generalise to unseenattackmethods[1].Documentedfinancialincidents,includingafraudulenttransferofUSD 25 million enabled by an AI-generated impersonation during a live call, illustrate that this is no longer a theoretical risk [9].

A further complication is that voice authentication for trading cannot rely on a single staticaccept/reject decision made once at login. Orders may be issued repeatedly within a session, order sizes vary widelyinrisk,andnetworkconditionsonmobilechannelsfluctuate.Thismotivatesasession-aware, risk-adaptive approach in which the strength of authenticationandthedepthoflivenesscheckingscalewiththe value and frequency of the requested action, rather than a uniform check applied to every utterance.

Existingliteratureaddressespartsofthisprobleminisolation.Asubstantialbodyofworkhasproduced increasingly accurate speaker-embedding models and text-dependent verification pipelines [2, 3]; a parallelbody of work has produced replay-attack and deepfake-detection countermeasures with strong benchmark accuracy [1, 4, 8, 11]; and a smaller body of work has focused on making speaker verification lightweight enough to run with millisecond-scale latency on constrained or edge hardware [14, 15]. Comparatively little workbringsthesethreadstogetherforthespecificcaseofatradingsystem,wheretheacceptablelatencybudget, the cost of a false rejection, and the cost of a false acceptance are all differentfromatypicalmobile-unlockor smart-speaker scenario.

The problem is further complicated by the diversity of channels through which a voice commandcan reach a trading platform. A customer may speak into a smartphoneapplicationheldclosetothemouth,asmart speaker several metres away inalivingroom,oracall-centrehandsetoveracompressedtelephonycodec.Each channel changes the acoustic characteristics of the captured signal and, in turn, the difficulty of both accurate speaker verification and reliable liveness detection. A framework intended for production use therefore cannot assumeasinglefixedmicrophonegeometryornetworkpath;itmustdegradegracefully when only a single low-quality microphone is available and take advantageofadditionalsensors,suchasmicrophonearrays,when they exist.

Regulatory and operational considerations add a further dimension that is often absent from purely academic evaluations of speaker verification.Financialregulatorstypicallyrequireanauditablerecordofhowa transaction was authorised, a bounded and justifiable false-acceptance rate, and a documented fallbackprocedure when biometric authentication cannot be completed with sufficient confidence. These requirements push the design toward a layered, explainable pipeline in which each stage producesaninterpretablescoreand decision, rather than a single opaque end-to-end model, and they reinforce the need for the kind of explicit latency and reliability instrumentation described in Section III.

This paper makes three contributions. First, it surveys recent literature on speaker verification, voice anti-spoofing, and low-latency biometric authentication, organised around the specific constraints of a trading use case. Second, it proposes VoxTrade Secure, an architectural framework thatintegratesspeakerverification, livenessdetection,trading-intentvalidation,andrisk-adaptivestep-upauthenticationintoasingle latency-instrumented pipeline. Third, it defines a joint latency-and-reliability evaluation methodology and illustrates it with representative benchmark figures drawn from the reviewed literature, showing how the framework's tiered design can be tuned to balance security and responsiveness for different trade types. The remainderofthepaper is organised as follows: Section II reviews related work; Section III presents the proposed methodology; Section IV discusses illustrative results; Section V outlines future enhancements; and Section VI concludes the paper

II. LITERATURE SURVEY

Hong et al. (2026) presented a systematic empirical evaluation of state-of-the-art speaker-verification andanti-spoofingsystemsagainstcontemporaryvoice-cloningattacks.UsinganECAPA-TDNN speaker-verification model and an XLS-R with AASIST deepfake detector, the study showed that open-source voice-cloning systems trained on as little as ten to thirty minutesoftargetspeechachievedbypassratesofover 40percentagainstacommercial-gradeverificationmodel,andthatdetectorperformancedegradedsharplywhen tested against synthesis methods unseen during training, with the Equal Error Rate rising nearlythirty-foldoutof domain. The study argued for defense-in-depth strategies that combine detection with multi-factor authentication rather than relying on voiceprint verification alone [1].

Molavi (2024) proposed a text-dependent speaker-verification pipeline for the TdSVAAICChallenge thatcombinedaFast-Conformerautomatic-speech-recognitionmoduleforcontentvalidationwithafused speaker embedding drawn from wav2vec-BERT and ReDimNet models. The fusion approach achieved a normalised minimum detection cost function of 0.0452, demonstrating that contentverificationandembedding fusion can jointly improve robustness in phrase-based authentication, a design pattern directly applicable to trading systems that use fixed confirmation phrases [2].

Rostami and Jafarzadeh (2026) reported a text-dependent speaker-verification systembuiltonadapted ResNet-TDNN and NeXt-TDNN architectures originally trained on VoxCeleb, supplemented by a lightweight EfficientNet-A0 model trained specificallyonchallengedata.Theensembleachievedaminimumdetectioncost function of 0.0461 and an Equal Error Rate of 1.3 percent, illustrating that adapting existing pretrained embeddingscanachievecompetitiveaccuracywithoutthecostoftrainingfromscratch,whichisrelevantwhen a trading platform must deploy verification quickly across a large existing customer base [3].

Kamel et al. (2025) conducted a broadsurveyofthreatsagainstvoice-authenticationandanti-spoofing systems, cataloguing replay attacks, synthetic-speech attacks, and adversarial-audio attacks. The survey documented real-world financial fraud cases, including a deepfake-enabled wire transfer and an attempted breach using cloned executive audio, and traced the evolution of replay-attack research from the 1990sdiscovery that pre-recorded audio could bypass early verification systems to current deep-learning-based

countermeasures. The authors concluded that no single defense layer is sufficient and recommended layered detection pipelines [4].

Yang, Cui, and Zheng (2024) proposed VoShield, a room-scale voice liveness-detection method for smart devices that measures sound-fielddynamicscreatedbythenaturalmovementofaspeaker'smouth,which changerapidlyforalive human but remain comparatively static for a loudspeaker used in a replay attack. Unlike lip-motion or close-range physiological methods, VoShield wasshowntooperatereliablyatroomscale, which is relevant for trading systems accessed through smart speakersorhands-freein-vehicleassistantsrather than a handset held close to the mouth [5].

A privacy-preserving liveness-detection study explored passive liveness detection for smart voice interfaces that avoids requiring users to hold a fixed gesture or position, addressing a usability gap in earlier liveness-detection schemes while analysing only the collected audio rather than deploying additional sensors. The approach is directly relevant to trading applications where requiring a user to perform a specific physical gesture before every order would be impractical [6].

Neri and Virtanen (2026) studied multi-channel replay-speech detection using acoustic maps derived from microphone-array recordings,showingthatspatialacousticfeaturesextractedacrossmultiplemicrophones improve the separability of live speech from loudspeaker replay compared with single-channel detection. The study also examined the impact ofmicrophone-arraymismatchesondetectionaccuracy,highlightingapractical deployment concern for trading applications accessed from heterogeneous consumer devices with differing microphone configurations [7].

A separate line of work examined multi-channel replay detection using an adaptive learnable beamformer, reporting that beamforming prior to feature extraction improved robustness tochannelanddevice variation relative to fixed-geometry array processing, reinforcing the finding that array-aware preprocessing meaningfully strengthens anti-spoofing performance in uncontrolled acoustic environments [8].

Earlier foundational work catalogued practical attacks on voice-spoofing countermeasures, including hidden-voice-command attacks, adversarial perturbations against automatic speechrecognition,andlight-based audio-injection attacks, and proposed the Void system as a fast, lightweight liveness-detection method suitable for near-real-time operation. This body of work establishedtheadversarialthreatmodelthatsubsequentreplay-and deepfake-detection research has continued to address [9].

A study on throat-microphone-based liveness detection combined a body-conducted sensor with conventional audio-based automatic speaker verification, showing that throat-microphone signals are largely unaffectedbyacousticreplaybecause they capture vibration through tissue rather than through air, making high-fidelity loudspeaker replay attacks substantially harder to mount. The approach requires specialised hardware, which limits applicability to consumer smartphone-based trading but is informative for dedicated trading-desk hardware [10].

Work on cepstrogram-based replay-attack countermeasures examined how time-frequency representations capture channel and device artefacts introduced by the replay process, and compared several convolutional and residualarchitecturesfordistinguishinggenuineandreplayedspeech,buildingonthebroader ASVspoof benchmark series that has become the standard evaluation protocol for anti-spoofing research [11].

Regardinglatency,alightweightdual-factoracousticauthenticationdesigncombinedGaussian-Mixture-Model-based speaker screening with Dynamic-Time-Warping passphrase verification on constrained edge hardware. The reported end-to-end processing latency was bounded atundertenmilliseconds on a single-core CPU, split across feature extraction, speaker scoring, and passphrase matching,whilelimiting the false-acceptance rate for physical impostors and high-fidelity replay to under seven percent. This result demonstrates that sub-millisecond-scale verification stages are achievable even without cloud inference,whichis important for the latency budget of a trading-order pipeline [12].

A related study on lightweight speaker verification integrated voice-activity detection and speech enhancement directly into an on-device model, reporting only a marginal latency increase of about thirteen milliseconds on mobile hardware while improving accuracy under background noise, showing that noise robustness and low latency are not mutually exclusive design goals for mobile authentication [13].

Industryanalysesofidentity-verificationlatencyintradingcontextsemphasisethateven millisecond-scale delays in verification can translate into missed execution prices or regulatory exposure, and recommend edge-deployed verification, streamlined API design, and caching to minimise processing delay in latency-sensitive financial workflows,reinforcingthearchitecturalmotivationforinstrumentingandminimising every stage of a voice-based authentication pipeline [14].

Finally, market and industry reports indicatethatadoptionofvoicebiometricsinthebanking,financial services, and insurance sector is acceleratingrapidly,withtheglobalvoice-biometricsmarketprojectedtogrow at a compound annual rate exceeding twenty percent over the coming years, driven by demand for stronger remote authentication and improved customer experience [15]. This growth trend underscores the practical urgencyofthesecurityandlatencyquestionsaddressedinthispaper,sincearapidlyexpandingdeploymentbase widens

the pool of potential fraud targets at the same rate that it widens legitimate adoption.

III. METHODOLOGY

The proposed framework, VoxTrade Secure, is designed to authenticate a trading customer through voice while keeping the added latency within a budget that is compatible with order execution, and while providing a measurable, tunable level of resistance to replay and synthetic-speech attacks. The framework separates authentication strength from a fixed all-or-nothing check and instead ties ittotheassessedriskofthe requested action, so that routine, low-value interactions are handled quickly while high-value or unusualrequests receive additional scrutiny.

A. System Architecture

The framework consists of six logical components arranged in a pipeline: a Voice Capture and Preprocessing module, a Speaker Verification module, a Liveness and Anti-Spoofing module, a Trading-Intent Recognition and Command Validation module, a Risk-Adaptive Authentication Controller, and a Latency and Reliability Instrumentation layer thatwrapseveryothercomponent.TheInstrumentationlayertimestampsentry and exit of each stage and logs accept, reject, and escalate decisions so that both latency and reliabilitycanbe measured from the same trace without separate test harnesses.

B. Voice Capture and Preprocessing

Audio iscapturedfromtheclientdevice,whetheramobileapplication,asmartspeaker,oracall-centre line, and passed through voice-activity detection and speech enhancement tosuppressbackgroundnoisebefore feature extraction. Following the design used for on-device speaker verification, voice-activity detection and enhancement are integrated directly into the front end rather than applied as a separate post-processing stage, keepingthe additional latency to roughly ten to fifteen milliseconds while improving robustness in noisy trading-floor or outdoor environments [13].

C. Speaker Verification

The Speaker Verification module compares a fixed-length embedding extracted from the captured utterance against the customer's enrolled voiceprint using a cosine-similarity or probabilistic scoring function. The framework is architecture-agnostic but favours compact embedding models such as ECAPA-TDNN orfusedwav2vec-BERTandReDimNetrepresentations,whichhavedemonstratedcompetitiveaccuracyin text-dependent verification challenges while remaining small enough for near-real-time inference [1, 2]. For repeated actions within a session, such as confirming several orders in sequence, the module supports lightweight re-scoring against a cached embedding rather than repeating full enrolment-style verification, reducing average latency across a session.

D. Liveness DetectionandAnti-Spoofing

Because speaker verification alone provides only partial protection against high-quality replay and voice-cloning attacks [1], the framework places a dedicated Liveness and Anti-Spoofing module after verification. This module combines two complementary signals: an audio-only spoof classifier trained on cepstral or self-supervised representations to flag synthetic or replayed speech [1, 11], and, where the capture device provides multiple microphones, an array-based liveness check that exploits sound-field dynamics or spatial acoustic cues that differ between a live speaker and a loudspeaker replay [5, 7, 8]. The two signals are fused withaweightedscore,andthefusionweightsareconfigurablesothatadeploymentwithoutamicrophone array can rely primarily on the audio-only classifier.

E. Trading-Intent Recognition and Command Validation

Once the speaker's identity and liveness are established, the spoken instruction is transcribed andparsed into a structured trading intent, for example an instrument symbol,anorderdirection,aquantity,andan order type. The module cross-checks the parsed intent against account permissions, position limits, and recent order history, and requests a spoken or on-screen confirmation whenever the parsed intent is ambiguous. This validation step is deliberately kept outside the biometric pipeline so that transcription errors are treated as a usability issue rather than an authentication failure.

F. Risk-Adaptive Step-UpAuthentication

The Risk-Adaptive Authentication Controller assignsariskscoretoeachrequestbasedonordervalue, deviation from the customer's typical trading pattern, device and network reputation,andtheconfidencescores returned by the verification and liveness modules. Low-risk requests, such as a routine order within the customer's usual size range from a recognised device, are authorised as soon as verification and liveness both

pass their base thresholds. High-risk requests, such as an unusually large order, a new payee or destination account, or a borderline liveness score, trigger a step-up challenge, for example a short randomised phrase repeated by the customer, a secondary factorsuchasapushnotification,orabriefholdformanualreview.This tiering follows the defense-in-depth recommendation made in recent security surveys, which argue that voiceprint verification alone should not be treated as sufficient for high-stakes decisions [1, 4].

G. Latency Evaluation Methodology

Latency is measured end to end from the moment audio capture endstothemomentanauthentication decision is returned, and isfurtherbrokendownperstage:preprocessing,embeddingextractionandverification scoring, liveness and anti-spoofing scoring, intentparsing,andrisk-scorecomputation.Measurementsaretaken under three conditions representative of deployment: a wired low-latency network, a typical 4G/5G mobile network with jitter, and a congested network approximating peak-hour contention, following the observationthat even small delays in identity verification can be consequential in a trading context [14].

H. Reliability Evaluation Methodology

Reliability is evaluated using standard biometric metrics, namely theFalseAcceptanceRate,theFalse Rejection Rate, and the Equal Error Rate at the operating threshold, computedseparatelyforgenuineattempts, replay-attackattempts,andsynthetic-speechattempts,consistentwithevaluationprotocolsusedinrecent anti-spoofing benchmarks [1, 11, 12]. In addition, the framework tracks an operational availability metric that capturesthefractionofauthenticationrequestscompletedwithinthetargetlatencybudgetwithouta system-level failure, since a highly accurate but frequently timed-out pipeline is not reliable from a trading customer's perspective.

I. Tiered Configuration

Because the accuracy of anti-spoofing detection tends to degrade against attack methods not represented in training data [1], the framework treats thelivenessthresholdasatunableparameterratherthana fixed constant, andrecommendsperiodicre-calibrationagainstnewlypublishedspoofingtechniques.Combined with the risk-adaptive controller, this allows an operator to shift the security-latency trade-off for different customer segments or order sizes without redesigning the pipeline.

J. Enrolment, Storage, and Privacy Considerations

Voice is an immutable biometric: unlike a password, a compromised voiceprint cannot simply be reissued to the customer. The framework therefore stores enrolledvoiceprintsasirreversibleembeddingsrather thanrawaudio,encryptsembeddingsatrest,androtatestheenrolmentonascheduledbasisorimmediatelyafter a confirmed spoofing attempt against a givenaccount.Enrolmentitselfistreatedasahigher-riskoperationthan routine verification and is gated behind an existing strong-authentication channel, such as an in-branch visit, a videocall,oragovernment-issued-identitycheck,sothatanattackercannotsimplyenrolaclonedvoiceagainst a victim's account. Raw audio captured during trading sessions is retained only for the duration needed for disputeresolutionandfraudinvestigation,consistentwithdata-minimisationprinciplescommonin financial-services privacy regulation.

K. Fallback and Degraded-Mode Operation

The framework defines an explicit fallback path for cases where voice authentication cannot reach a confident decision, for example due to excessive background noise, a poor network connection, or a liveness score close to the decision boundary. Rather than defaulting to an outright rejection, which would frustrate genuine customers, or a silent acceptance, which would weaken security, the controller routes theseborderline cases to a secondary channel such as a one-time passcode, a push-notification approval, or a short hold with manualreviewforlargeorders.Loggingeveryfallbackeventseparatelyfromnormalacceptandrejectdecisions also gives operators a direct signal of how often the voice channel alone is insufficient, which is useful fortuning thresholds over time.

IV. RESUL TAND DISCUSSION

The proposed framework was assessed conceptually against the benchmark figures reported in the reviewedliteraturetoillustratehowthetiereddesignaffectsthelatency-reliabilitytrade-off.TableIsummarises representative, illustrative figures assembled from the component-level results discussed in Section II, showing how end-to-end latency and error rates change as additional protection layers are added to the pipeline.

As shown inTableI,aspeaker-verification-onlyconfigurationachievesthelowestlatencybutprovides no explicit defence against replay or synthetic-speechattacks,consistentwiththefindingthatverificationalone

givesonlypartialprotectionagainstmodernvoice-cloningattacks[1].Addinganaudio-onlylivenesscheck roughly doubles latency but raises spoof-detection accuracy substantially, while adding array-based liveness detection further improves accuracy at a smaller additional latency cost, in line with reported improvementsfrom spatial acoustic features and beamforming[7,8].Movingfromawirednetworktoarepresentativemobile network increases latency moderately without materially changing accuracy, reflecting the fact that network jitter primarily affects transport time rather than model inference. The full pipeline with a step-up challenge, reserved by the risk-adaptive controller for high-risk orders, incurs the largest latency but achieves the lowest false-acceptance rate and highest spoof-detection accuracy, which is an acceptable trade-off for infrequent,high-value transactions where the cost of a fraudulent order far exceeds the cost of a short delay.

TABLE I. ILLUSTRATIVE LATENCY AND RELIABILITY BENCHMARK ACROSS PIPELINE CONFIGURATIONS

Pipeline	Avg. Latency(ms)	FAR(%)	FRR(%)	Spoof DetectionAccuracy(%)
Verificationonly,wirednetwork	48	0.90	3.10	N/A
Verification+audioliveness,wired network	96	0.35	4.20	91.4
Verification + audio+array liveness, wired network	142	0.12	5.10	97.2
Verification + audio+array liveness, 4G/5G mobile	211	0.14	6.30	96.1
Full pipeline with step-up (high-risk order)	480	0.03	7.80	98.6

These figures also illustrate a recurring tensionintheseviewedliterature:strongeranti-spoofinglayers reduce the false-acceptance rate but tend to raise the false-rejection rate, inconveniencing genuine customers, particularly in noisy or mobile conditions [1, 13]. The risk-adaptive design in Section III addresses this by reserving the most conservative thresholds for the minority of high-risk requests rather than applying them uniformly, which keeps the average customer experience close to the low-latency end of Table I while still providing strong protection where it matters most.

A further observation from the literature is that anti-spoofing accuracy reported onafixedbenchmark can overstate real-world robustness, since detectors trained on known synthesis methods have been shown to lose a large fraction of their accuracy against previously unseen voice-cloningtechniques[1].Thisimpliesthat the spoof-detection accuracy figures in Table I should be read as best-case, in-domain estimates, and that the framework's periodic re-calibration step described in Section III-I is not optional but a necessary part of maintaining the reported reliability over time.

From an operator'sperspective,thepracticalvalueofTableIlieslessinanysinglerowandmoreinthe shapeofthecurveittraces:latencygrowsroughlylinearlyasprotectionlayersareadded,whilethe false-acceptance rate falls much faster, particularly once array-based liveness detection is introduced. This convex relationship supports theframework'scoredesignargument,namelythatitismoreefficienttospendthe latency budget selectively, on the transactions where fraud wouldbecostly,thantospendituniformlyonevery request. For a typical retail brokerage workload, where the overwhelming majority of orders are small and routine,thisselectivespendingkeepstheaveragecustomer-perceivedlatencyclosetotheverification-only figure in Table I while still providing the strongest available protection for the small minority of high-value orders that account for a disproportionate share of potential fraudlosses.Institutionalorhigh-net-worthtrading desks, where a larger share of orders may be flagged as high-riskbyvaluealone,wouldbeexpectedtooperate closer to the full-pipeline configuration on average, and the framework's threshold tuning in Section III-F is intended to accommodate that shift in workload profile without architectural change.

V. FUTURE ENHANCEMENT

Several directions can extend the proposed framework. Continuous, session-based re-authentication could replace the current per-requestcheckswithongoingbackgroundverificationthroughoutatradingsession, reducingperceivedlatencyforroutineactionswhilestilldetectingachangeofspeakermid-session. Cross-lingual and cross-accent robustness also meritsfurtherstudy,sincespeaker-verificationandanti-spoofing models trained predominantly on one language or accent group have shown uneven performance when transferred to others [1].

Hardware-assisted liveness detection, such as throat-microphone or dedicated array sensors, could be offered as an optional higher-assurance mode for institutional trading desks where specialised hardware is already available, building on demonstrated resistance to acoustic replay [10]. On the detection side, incorporating continual learning so that anti-spoofing classifiers are periodically retrained on newly observed synthesis methods would help address the generalisation gap documented in recent large-scale evaluations [1].

Finally, integrating the latency-and-reliability instrumentation layer with live production telemetry, rather than offline benchmark figures, would allow the risk-adaptive controller to adjust thresholdsautomatically in response to observed network conditions, fraud attempts, and customer complaint patterns, moving the framework from a static configuration toward an adaptive, continuously monitored security control.

A complementary direction is the extension oftheframeworktomulti-modalconfirmation,combining voice with a secondary passive signal such as behaviouraltypingortouchpatternsonacompanionapplication, so that the overall decision doesnotrestonaudioaloneevenforroutinetransactions.Formalevaluationagainst publicly available anti-spoofing benchmarks, such as the ASVspoof series referenced throughout Section II, would also allow the framework's illustrative figures in TableItobereplacedwithreproducible,independently verifiable measurements once a reference implementation is built, strengthening the practicalcaseforadoption by brokerages and regulators alike.

Longer term, standardisation of latency-and-reliability reporting for voice-based financial authentication, analogous to existing disclosure norms for order-execution latency, could allow customers and regulatorstocompareprovidersonaconsistentbasisratherthanrelyingonvendor-specificclaims,whichwould benefit the wider adoption of voice as a trusted transaction channel across the financial-services industry.

VI. CONCLUSION

This paper reviewed recent research on speaker verification, replay and deepfake liveness detection,and low-latency biometric authentication, and used that review to motivate VoxTrade Secure, a framework for securing voice-driventradingwithspeakerbiometricauthentication.Theframework'scentraldesignchoiceisto treat authentication strength as a tunable,risk-adaptivepropertyratherthanasinglefixedcheck,combiningfast single-factorverificationforroutineactionswithlayeredlivenessdetectionandstep-upchallengesfor high-value or unusual requests.

Illustrative benchmark figures assembled from the reviewed literature show that each additional protection layerimprovesresistancetoreplayandsynthetic-speechattacksatameasurablelatencycost,andthat thiscostcanbemanagedbyreservingthestrongestchecksfortheminorityofhigh-risktransactions.Thereview alsohighlightsanopenchallengethattheproposedframeworkmustaccountforratherthanassumeaway: anti-spoofing accuracy measured on known attack methods does not reliably predict robustness against the rapidly evolving landscape of voice-cloning techniques, making periodic re-calibration a necessary part ofany deployment rather than an optional enhancement.

Overall,theproposedframeworkdemonstrateshowspeakerverification,livenessdetection,trading-intentvalidation,andrisk-adaptivestep-upauthenticationcanbecomposedintoasingle latency-instrumented pipeline suited to the specific demands of voice-driven trading. It can serve as a starting point for brokerages, fintech platforms, and financial-technology researchers seeking to reason systematically about the trade-off between security, reliability, andresponsivenesswhenvoicebecomesachannelforfinancial transactions.

REFERENCES

[1]. Hong, M., Jiang, D., Xie, Z., Zhao, W., Wang, G., and Zhang, C. J. "Vulnerabilities of Audio-Based BiometricAuthenticationSystems Against Deepfake Speech Synthesis." arXiv preprint arXiv:2601.02914, 2026.

[2]. Molavi,M."TheSVASRSystemforText-dependent Speaker Verification(TdSV)AAICChallenge2024."arXivpreprintarXiv:2411.16276,2024.

[3]. Rostami, A. M., and Jafarzadeh, P."Text-DependentSpeakerVerification(TdSV)Challenge2024:TeamNaiveSystemReport."arXiv preprint arXiv:2605.14896, 2026.

[4]. Kamel,K.,Sood,K.,Dutta,H.S.,andAryal,S."ASurveyofThreatsAgainstVoiceAuthenticationandAnti-SpoofingSystems." arXiv preprint arXiv:2508.16843, 2025.

[5]. Yang, Q., Cui, K., and Zheng, Y. "Room-ScaleVoiceLivenessDetectionforSmartDevices."IEEETransactionsonDependableand Secure Computing, vol. 21, no. 5, pp. 4982–4996, 2024.

[6]. Meng, Y., Li, J., Zhu, H., Tian, Y., and Chen, J. "Privacy-Preserving Liveness Detection for Securing Smart VoiceInterfaces."IEEE Transactions on Dependable and Secure Computing, 2025.

[7]. Neri,M.,andVirtanen, T. "Multi-Channel Replay Speech Detection using Acoustic Maps." arXiv preprint arXiv:2602.16399,2026.

[8]. "Multi-channelReplaySpeechDetectionusinganAdaptiveLearnableBeamformer."arXivpreprintarXiv:2502.13473,2025.

[9]. Ahmed, M. E., Kwak, I.-Y., Huh, J.H.,Kim,I.,Oh,T.,andKim,H."Void:AFastandLightVoiceLivenessDetectionSystem."In Proc. 29th USENIX Security Symposium, pp. 2685–2702, 2020.

[10]. "RobustVoiceLiveness Detection and Speaker Verification Using Throat Microphones." IEEE/ACM Transactions on Audio,Speech, and Language Processing, 2018.

[11]. "AStudyofUsingCepstrogramforCountermeasureAgainstReplayAttacks."arXivpreprintarXiv:2204.04333,2022.

[12]. "ALightweightDual-FactorAcousticAuthenticationSystemvia Cascaded GMM-DTW Architecture for Edge Computing."arXiv preprint arXiv:2606.10565, 2026.

[13]. "LightweightSpeakerVerificationwithIntegratedVADandSpeechEnhancement." Speech Communication (ScienceDirect),2025.

[14]. "BoostingReal-TimeTrading:Low-LatencyIDVerificationwithEdgeComputing."DiditTechnologies,2026.

[15]. Technavio."VoiceBiometricsMarketGrowthAnalysis–SizeandForecast2025–2029."TechnavioResearchReport,2025.