

VoicePay: A Smart Controlled Payment Application using Voice Commands

Secure, Accessible and Limit-Aware Digital Payments through Speech

Savita Biradar
Department of Computer Engineering
Zeal College of Engineering
Pune, India

Abstract—The rapid growth of digital payments has increased the need for payment methods that are fast, accessible and secure. This paper presents VoicePay, a smart controlled payment application that lets users start and approve transactions through spoken commands while keeping them in control of how, when and how much they spend. VoicePay combines automatic speech recognition, natural language understanding, speaker verification and a rule-based spending control engine on top of a standard payment gateway. Users can define per-transaction and daily limits, trusted payee lists and category budgets, and risky requests are escalated to a second authentication step. The layered design reduces typing effort, improves accessibility for visually impaired and elderly users, and limits the damage caused by fraud or accidental payments. In a reproducible simulation of 20,000 synthetic transactions, the control layer stopped 93.8 percent of simulated fraudulent requests, compared with 55.7 percent for a voice-only baseline, while adding extra authentication to 8.1 percent of legitimate payments. The paper also describes the architecture, implementation and security analysis, and outlines the evaluation of the full speech pipeline.

Keywords—voice payment; speech recognition; speaker verification; spending control; digital payments; mobile security

I. INTRODUCTION

Digital payment systems such as the Unified Payments Interface (UPI) have changed how people in India pay for goods and services [1]. Even so, most payment apps still depend on touch input: the user must unlock the phone, open the app, type an amount, choose a payee and enter a personal identification number (PIN). This process is slow in hands-busy situations and is difficult for visually impaired users, elderly users and people with limited literacy or limited typing skill.

Voice is a natural interface for these users. Advances in automatic speech recognition (ASR) [2] and speaker recognition [3] now make it practical to say, for example, “Pay 500 rupees to Rahul” and have the system understand the request. However, a voice-only payment flow brings new risks, such as misrecognized amounts, replayed or synthesized voices, and unintended payments triggered by background speech.

This paper proposes VoicePay, a smart controlled payment application. The word “controlled” is central: every voice command passes through a control engine that checks the user’s own rules (limits, trusted payees, budgets and time restrictions) before any money moves. The main contributions are: (1) a modular architecture that joins speech recognition, intent extraction, speaker verification and spending control; (2) a risk-based authentication scheme that asks for stronger proof only when a transaction is unusual; and (3) a reproducible simulation

that compares the control layer with two baselines, together with a plan for full-system evaluation.

II. RELATED WORK

Speech recognition has moved from hidden Markov models to deep neural networks and large-scale weakly supervised models that work across accents and noise conditions [2]. Text-based intent and entity extraction is commonly handled with pretrained language models such as BERT [4], and standard methods are well documented in [5].

Speaker verification has progressed from Gaussian mixture model systems [6] to neural embeddings known as x-vectors [3]. These systems are good at confirming that a voice belongs to an enrolled user, but they can be fooled by recorded or synthetic speech. The ASVspoof challenges study such spoofing attacks and countermeasures [7]. For this reason VoicePay never treats voice as the only factor for high-value payments.

Regulatory guidance on digital payment security stresses strong customer authentication, transaction limits and user alerts [8]. VoicePay follows the same direction by placing user-defined controls and step-up authentication at the center of its design instead of adding them afterward. To the best of our knowledge, few published designs combine speaker verification, user-defined spending rules and a quantitative risk score in a single voice payment pipeline, and this is the gap that VoicePay addresses.

III. PROBLEM STATEMENT AND OBJECTIVES

Existing payment apps are not designed for hands-free or low-literacy use, and existing voice assistants do not give users fine-grained control over payment risk. The objectives of VoicePay are:

- to let users complete common payments (send money, pay bills, check balance, view history) using natural spoken commands in English and Hindi;
- to verify the speaker and the intent before every transaction;
- to give users configurable controls such as daily limits, trusted payees and category budgets;
- to keep the number of steps low for routine payments while escalating risky ones; and
- to maintain a complete, tamper-evident audit trail of every voice transaction.

IV. PROPOSED SYSTEM

A. System Architecture

VoicePay follows a pipeline architecture, shown in Fig. 1. The mobile client captures speech and sends it to the backend, where the ASR, natural language understanding (NLU), verification and control modules process it in order. Only requests that pass all stages reach the payment gateway. The result is returned to the user as speech and on-screen text, so the user can confirm what happened.

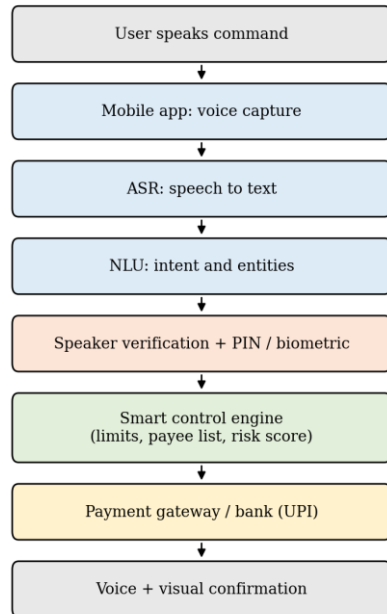


Fig. 1. Layered architecture of the VoicePay system.

B. Voice Capture and Speech Recognition

The app records audio only after a wake action (button press or wake phrase) to avoid accidental activation. Audio is denoised, converted to a standard sampling rate and passed to an ASR engine that returns a transcript with a confidence score. If the confidence is below a threshold, the system asks the user to repeat the command instead of guessing.

C. Natural Language Understanding

The transcript is mapped to an intent (for example, SEND_MONEY, PAY_BILL, CHECK_BALANCE) and to entities such as amount, payee and purpose. Spoken numbers are normalized, so “five hundred” becomes 500. Payee names are matched against the user’s contact list using phonetic and fuzzy matching, and ambiguous matches trigger a clarification question. Before any payment, the system reads back the amount and payee and waits for an explicit confirmation.

D. Authentication

VoicePay uses layered authentication. First, a speaker verification model compares the voice with the enrolled voiceprint, in the manner of [3]. Second, the transaction PIN or device biometric is required according to the risk level. A liveness check, such as asking the user to repeat a random phrase, is used to resist replay attacks [7]. Voiceprints are stored only as encrypted embeddings, never as raw audio.

E. Smart Control Engine

The control engine is what makes the application “smart” and “controlled”. It evaluates each parsed request against user rules and computes a simple risk score from factors such as amount, payee novelty, time of day, location change and speaker verification score. Depending on the outcome, the request is approved, sent for step-up authentication, or blocked. Table I lists the main controls. The risk score R is computed as in (1), where a is the amount divided by the per-transaction limit (capped at one), n is one for a new payee, t is one for an unusual hour, l is one for a change of location, and s is the speaker verification score between zero and one.

$$R = w_a a + w_n n + w_t t + w_l l + w_s (1 - s) \quad (1)$$

The weights are 0.30 for the amount, 0.25 for a new payee, 0.10 for an unusual hour, 0.10 for a change of location and 0.25 for the speaker score, so they sum to one. A request is approved if R is below 0.35, sent for step-up authentication if R lies from 0.35 to below 0.60, and blocked otherwise. Independently, any amount above the per-transaction limit always needs step-up authentication, and any amount that would exceed the daily limit is blocked.

TABLE I SMART CONTROLS IN VOICEPAY

Control	Description	Action on violation
Per-transaction limit	Maximum amount for one voice payment	Block / PIN
Daily and monthly limit	Cumulative spending cap	Block
Trusted payees	Pre-approved contacts and billers	New payee needs PIN
Category budget	Caps for food, travel, bills, etc.	Warn user
Time window	Allowed hours for voice payments	Step-up
Family / guardian mode	Parent or guardian approves large amounts	Send approval request

V. IMPLEMENTATION

A. Technology Stack

A reference implementation can be built with a cross-platform mobile client (Android or Flutter), a representational state transfer (REST) backend (for example Python with FastAPI), a relational database for users, limits and audit logs, and a bank or payment service provider application programming interface (API) for UPI transactions. ASR may use an on-device or cloud speech model, and NLU may use a fine-tuned transformer or a rule-based slot-filling framework. All communication uses Transport Layer Security (TLS), and sensitive data is encrypted at rest.

B. Transaction Workflow

A typical interaction proceeds as follows. (1) The user taps the microphone and says, “Send 300 rupees to Amit.” (2) ASR produces the transcript and NLU extracts the intent, amount 300 and payee Amit. (3) The control engine finds Amit in the trusted list and the amount below the limit, so the risk is low. (4) The app reads back the details; the user says “Yes” and the speaker verification score is accepted. (5) The gateway executes the payment, and the app announces the result and records the audit entry. If the payee were new or the amount high, the app would ask for the PIN or fingerprint before step 5.

VI. SECURITY ANALYSIS

Because voice is an open channel, VoicePay assumes that attackers may record, replay or synthesize speech, may speak near the device, or may try to confuse the recognizer. Table II summarizes the main threats and the corresponding countermeasures.

TABLE II THREATS AND COUNTERMEASURES

Threat	Countermeasure
Replay / synthetic voice	Liveness phrase, anti-spoofing model, PIN for high value
Misrecognized amount	Confidence threshold and spoken read-back confirmation
Stolen or unlocked phone	Device biometric, session timeout, speaker verification
Eavesdropping	TLS, no raw audio storage, encrypted embeddings
Accidental activation	Push-to-talk or wake phrase, explicit confirmation
Insider or log tampering	Append-only hashed audit log

No single factor is trusted completely. The combination of voice verification, user-defined limits and step-up authentication means that an attacker who defeats one layer is still constrained by the others, which is the main security argument for the controlled design.

VII. EXPERIMENTAL EVALUATION

A. Simulation Setup

The speech models themselves are not trained in this work, so we evaluate the part that is unique to VoicePay, the smart control engine, with a reproducible simulation. We generated 20,000 synthetic transactions, of which 19,400 are genuine and 600 (3 percent) are fraudulent attempts. Genuine amounts follow a log-normal distribution with a median of about 400 rupees, 20 percent go to new payees, 8 percent occur at unusual hours, and speaker scores follow a Beta(8, 2) distribution. Fraudulent attempts use larger amounts (median about 2,400 rupees), go mostly to new payees (90 percent), often occur at unusual hours (45 percent) or from a new location (40 percent), and include a strong spoofing case: 15 percent of fraudulent requests receive a high speaker score drawn from Beta(7, 2). The per-transaction limit is 5,000 rupees and the daily limit is 10,000 rupees. The random seed is fixed so that the results can be reproduced.

Two baselines are compared. The voice-only baseline accepts a payment whenever the speaker score exceeds 0.5. The limit-only baseline asks for extra authentication only when the amount exceeds the per-transaction limit. A fraudulent request is counted as stopped if it is blocked or sent for step-up authentication, because the attacker is assumed not to hold the PIN or the device biometric.

B. Results

TABLE III SIMULATION RESULTS (20,000 TRANSACTIONS)

Method	Fraud stopped	Genuine users delayed	Fraud value passed
VoicePay	93.8%	8.1%	2.9%
Voice-only	55.7%	1.9%	40.2%
Limit-only	17.5%	0.3%	56.4%

Table III shows that the layered control engine stops far more fraud than either baseline. The voice-only method fails mainly against spoofed voices and cannot use context such as payee novelty, and the limit-only method misses small and medium fraudulent payments. The price of the improvement is that about 8 percent of genuine payments (mostly those to new payees or of larger amounts) need an extra confirmation, while only 0.06 percent of genuine payments are blocked. In terms of money, only 2.9 percent of the fraudulent value passed through VoicePay, against 40.2 percent for the voice-only baseline.

C. Limitations and Plan for Full-System Tests

These results come from synthetic data with assumed distributions, so they show the behavior of the design and not performance on real users. The numbers will change with real transaction histories, real speaker verification scores and adaptive attackers. The next step is to test the complete pipeline with human participants and recorded speech using the metrics in Table IV: word error rate for speech recognition, intent accuracy for NLU, equal error rate and spoof acceptance rate for speaker verification, task completion time against a touch interface, and the System Usability Scale (SUS) for usability. Participants should include different genders, accents and age groups, and tests should be run in quiet and noisy places with informed consent and approved privacy handling.

TABLE IV PLANNED FULL-SYSTEM METRICS

Metric	Purpose	Component
Word error rate	Recognition accuracy	Speech recognition
Intent accuracy	Command understanding	NLU
Equal error rate	Speaker verification	Verification
Spoof acceptance	Replay resistance	Anti-spoofing
Task time	Speed vs. touch UI	Whole system
SUS score	Usability	Whole system

VIII. CONCLUSION AND FUTURE WORK

This paper presented VoicePay, a voice-driven payment application whose distinguishing feature is a smart control layer that enforces user-defined spending rules and risk-based authentication. By combining speech recognition, intent extraction, speaker verification and configurable controls, the system aims to make digital payments faster and more inclusive without giving up security. A simulation of 20,000 synthetic transactions showed that the control layer stopped 93.8 percent of simulated fraud against 55.7 percent for voice-only verification, at the cost of extra confirmation for about 8 percent of genuine payments. These results are indicative only and must be confirmed with real speech and users.

Future work includes support for more Indian languages and dialects, on-device processing to protect privacy, stronger anti-spoofing models, offline and feature-phone support through interactive voice response, and field trials with elderly and visually impaired users.

ACKNOWLEDGMENT

The author thanks the faculty and the department of Zeal College of Engineering, Pune, for their support during this work.

REFERENCES

- [1] National Payments Corporation of India, “Unified Payments Interface (UPI) product overview and statistics.” [Online]. Available: <https://www.npci.org.in>
- [2] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in Proc. Int. Conf. Machine Learning (ICML), 2023.
- [3] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-vectors: Robust DNN embeddings for speaker recognition,” in Proc. IEEE ICASSP, 2018, pp. 5329–5333.
- [4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. NAACL-HLT, 2019, pp. 4171–4186.
- [5] D. Jurafsky and J. H. Martin, Speech and Language Processing, 3rd ed. (draft), 2023.
- [6] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted Gaussian mixture models,” Digital Signal Process., vol. 10, no. 1–3, pp. 19–41, 2000.
- [7] Z. Wu, J. Yamagishi, T. Kinnunen, et al., “ASVspoof: The automatic speaker verification spoofing and countermeasures challenge,” IEEE J. Sel. Topics Signal Process., vol. 11, no. 4, pp. 588–604, 2017.
- [8] Reserve Bank of India, “Master Direction on Digital Payment Security Controls,” 2021.