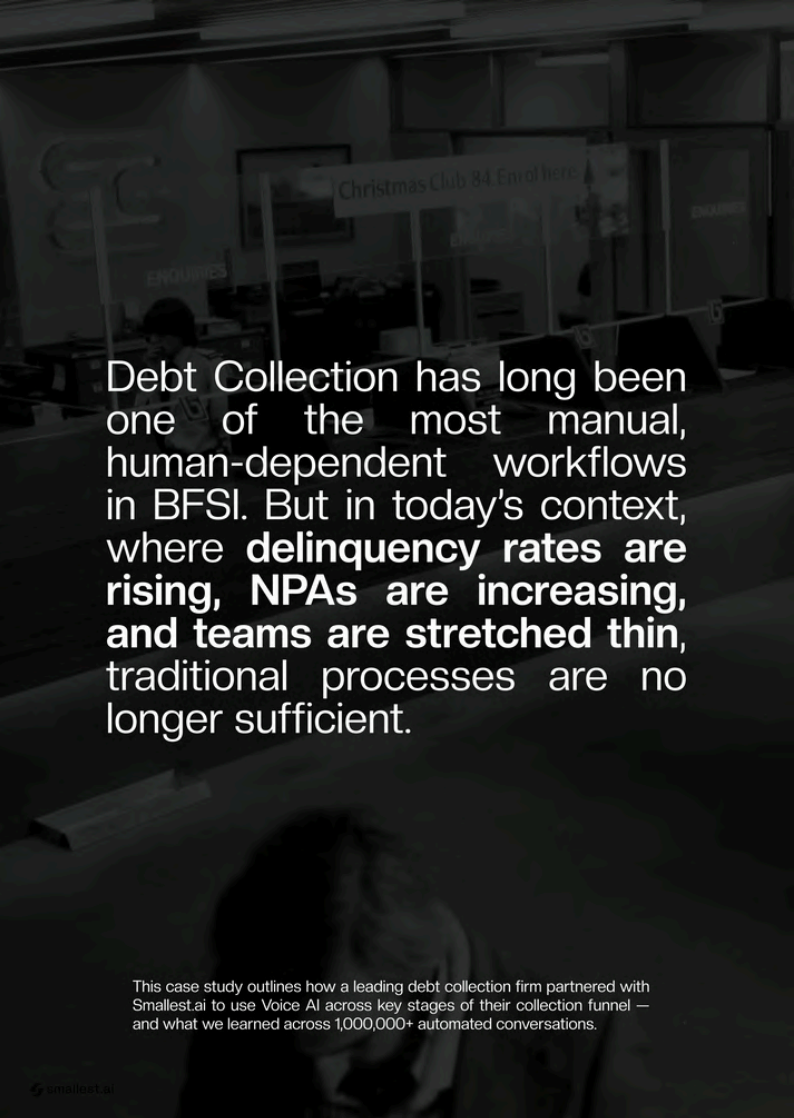
Case Study

BFSI

How Voice Agents are Modernizing BFSI Operations

smallest.ai

16192 Coastal Highway
Lewes, DE 19958
https://smallest.ai



Debt Collection has long been one of the most manual, human-dependent workflows in BFSI. But in today's context, where **delinquency rates are rising, NPAs are increasing, and teams are stretched thin**, traditional processes are no longer sufficient.

This case study outlines how a leading debt collection firm partnered with Smallest.ai to use Voice AI across key stages of their collection funnel — and what we learned across 1,000,000+ automated conversations.

The Problem

Rising Delinquency, Fragmented Teams & Cost Pressures

Banks are seeing a spike in **early-stage defaults**, growing **non-performing assets (NPAs)**, and **higher cost-to-collect** than ever before.

What makes the problem worse? **Fragmentation.**

Collections are not a singular process; they are divided into stages, each with its own workflows and teams.

Despite having structured teams across pre-due, post-due, and recovery buckets, **execution is breaking down at the frontlines.** Here are some problems BFSI ops leaders shared with us:

- **Follow-ups don't happen on time**
- **Promise-to-pay commitments aren't always tracked**
- **Connectivity rates are dropping due to spam-flagged numbers**
- **Agent memory + manual notes = inconsistent customer experience**
- **Training new reps takes 4–6 weeks with attrition being a constant problem.**

The Result?

Revenue gets delayed, escalations rise, and high-value accounts start slipping through. And this isn't just a people problem - it's a system problem.

A Secondary Problem

Why Debt Collection is Riddled with Leakages?

Debt collection is still largely manual and that's where most of the leakage happens. Customers don't pay on time for a range of reasons.

1. Forgetfulness
2. Dispute over amount or billing
3. Financial stress
4. Lack of urgency

But often, the real issue lies within the collection process itself. Where human-led collections break down:

1. Unstructured Calls
2. Memory/Notes for Context
3. Human Error
4. High Churn Rate
5. Limited after-hour contact
6. Emotional Fatigue
7. Missed payment follow-ups

These errors create a chain reaction — missed payments, escalations, customer dissatisfaction.

- Agents forget prior commitments made by customer
- No proactive reminders before due dates
- Call notes are incomplete or inconsistently logged
- Follow-ups delayed or never made
- Conversations lack personalisation or continuity

This is the domino effect of inefficiency- and it's one of the **biggest untapped opportunities in BFSI automation today**.

According to a study conducted by TI3.co

The Solution

How AI Voice Agents Create Leverage?

We worked with the bank's recovery teams to automate first-touch and follow-up calls using Smallest.ai Voice Agents - **multilingual, human-sounding, real-time voice automation with memory, retry logic, and CRM integration.**

Deployment Approach

Our deployment begins with a focused **Proof of Concept (POC)**- where we **identify a target use-case** (for example, DPD 0–30), **time-box it**, and **execute at speed** to prove both value and fit. This helps validate impact early, build stakeholder confidence, and quickly surface learnings from real conversations.

Once the POC delivers results, the system moves to **full-scale deployment** - where workflows are adapted to handle larger volumes, multiple languages, and layered compliance rules.

But performance doesn't stop at launch. **Our agents learn post-deployment** - identifying points where conversations stall and adjusting accordingly, much like a human agent who improves with experience. These learnings are continuously shared with stakeholders, so the improvement roadmap stays aligned with business goals.

Our workflows are flexible enough to handle **complex enterprise-specific use-cases**, and our enterprise onboarding team brings deep cross-industry expertise.

Why this matters for BFSI?

AI Voice Agents Increasing Operational Efficiencies

62%
reduction in operational cost

50%
rise in collection rates

1.7x
higher satisfaction by AI-Driven CX*

30%
cost-savings on service and support**

*as per Capgemini 2023 Report
**Juniper Research Report

Voice Agent Demo

Watch this Voice Agent in action and see it engage real customers with clarity, empathy, and results.

	Human Agent	Voice Agent
Daily Calls	300-400	+3000
Accounts Managed	~1000/mo	+10,000/mo
Connectivity Rate	~50%	60-65%
Promise-to-Pay Rate	22%	34%
Disposition Accuracy	~70%	>90%
Delinquency Reduction	-	up to 18%

Voice Agents at Work

How AI Voice Agents Helped Leading Debt Collector Partners

Smallest.ai, in collaboration with its channel partners like Knowl.ai, Convin, Plivo, Rezolv, Skit, BNY Mellon and 10+ global enterprises, has been able to collate insights from over 1 million calls per month. For the purpose of this case-study, we shall focus on **outbound call from the early (close to DPD 0) to the DPD 30 delinquency buckets.**

Use Case	Debt Collection
Collections Amount	1.8 Billion USD
Connected Call per month	+1,000,000
Peak concurrency	>1000 Calls
Cost Savings vs Human	60%
Collections Rate	50%
Effective Call Pickup Rate	85%
Average Call Duration	~1.5 mins

At peak capacity, our Voice Agents were running over 1,000,000 calls every week.

What We Learned from 1,000,000 calls?

1. Not all customers respond the same

DPD 0-30 few days overdue	90% pickup rate	High Adherence to PTP
DPD 30+ 2-3 weeks overdue	60-70% pickup rate	Need to make multiple attempts

Voice AI worked especially well in the DPD 0-30 window by helping reduce the load on human agents and letting them focus on tougher late-stage accounts. In fact, call volumes for human teams went down by 80%, cutting costs without dropping performance.

2. Tier 1 vs Tier 2 cities behave differently

Tier 1	Lower connectivity rate	Clearer Conversation
Tier 2	Heavier Accents	50% drop in call duration

To fix this, we trained the AI to adjust tone, language, and pace to match each region better.

What We Learned from 1,000,000 calls?

3. Retries Work - Just Once Isn't Enough

+3 retries in 12 hours increases pickup rates by 50%

Voice Agents handle retries without sounding robotic & repetitive.

4. Number Rotation - Avoid Spam Labels

Even with great scripts and voices, calls won't connect if numbers are flagged as spam

We used smart number rotation to switch out numbers getting marked as spam (on apps like Truecaller). This simple step improved pickup rates by over 20% and helped maintain trust with customers.

Smallest.ai has worked directly and indirectly with banks like HDFC, IDFC and Kotak to gain valuable insights in making Voice Agents for various use-cases in the banking process.

Our Secret Sauce

Lightning V2

World's Fastest TTS Model

At the time of the case study, we deployed our first-generation voice agents. Since then, we've launched Lightning V2, delivering even more powerful capabilities.

Key Upgrades:

- **Hyper-realistic voices** (MOS: 4.33/5)
- Support for **16+ regional & global languages** (Spanish, Hindi, Arabic, Russian, French)
- **100ms latency** for real-time dialogue
- Accurate handling of **amounts, due dates, acronyms**

Benchmark comparisons show that **Lightning V2 outperforms competitors** like ElevenLabs, Cartesia, and Bulbul in clarity, speed, and contextual accuracy, which are essential for debt collection conversations.

P90 Latency Comparison (ms):

90th percentile latency - includes all network and serving latency

Our Secret Sauce

Electron

Small Model. Enterprise Muscle.

Alongside Lightning V2, we've also introduced Electron, our proprietary series of Small Language Models (SLMs) designed for high-stakes enterprise workflows such as collections.

What Makes Electron Different?

- Inspired by decoder-only architectures like LLAMA
- **Trained on over 1 million multilingual conversational examples**, synthetically created and refined using **human-in-the-loop automations** for accuracy
- Designed for **low-latency, high-volume** customer interactions, with a **Time to First Byte (TTFB) of 45ms** compared to GPT-4.1 Mini's 450ms.
- Robust instruction-following, conversational goal adherence, and strict guardrail execution
- **Built to learn post-deployment** via enterprise-specific adapters like human agents improving with on-ground experience
- Continuously identifies failure points in real conversations and corrects them over time
- Customises itself for complex use-cases without compromising tone, compliance, or outcome alignment

Recover More, Spend Less

smallest.ai's Voice Agents can help you:

- **Recover up to 3x more** in missed payments
- **Operate at 1/5th the cost** of human-led efforts
- Get **24/7 scalable coverage** with 100% audit trails
- Improve **compliance, accuracy, and customer sentiment**

Rezolv skit.ai CONVIN Celetel mosaic WELLNESS plivo make my trip

Ready to Transform Your Collection Operations?

Email us at **support@smallest.ai** and schedule a demo on how smallest.ai's Voice Agent's can transform your operations.

We've been featured on

Blogs

- Introducing Lightning V2: World's Fastest TTS Model
- Benchmark Report - Smallest.ai vs ElevenLabs
- Benchmark Report - Smallest.ai vs Cartesia