Assignment 6: KPIs for AI Implementation

AI Voice Agent for Telecom Customer Support

Ramesh Maharaddi

CONTEXT & AI SYSTEM DESCRIPTION

Organization Profile

- **Sector:** Telecommunications
- **Company Type:** Regional Mobile Network Operator (MNO)
- **Department:** Customer Service Operations
- **Company Size:** Mid-sized telco serving 2.5M subscribers across urban and rural markets

AI System Implementation

System Name: Intelligent Voice Assistant for Customer Support (IVA-CS)

Purpose: Deploy an AI-powered voice agent to handle inbound customer service calls 24/7, automating routine inquiries and freeing human agents for complex issue resolution.

Key Capabilities:

- Understand natural language customer inquiries (no rigid menu systems)
- Handle common requests: bill inquiries, plan changes, network outage checks, SIM replacements, connectivity troubleshooting
- Access real-time customer data from CRM, billing systems, and network status dashboards
- Intelligently escalate complex issues to human agents with full context
- Provide support across multiple languages (English, Spanish, local dialects)
- Detect customer sentiment and adapt tone/response accordingly
- Handle multi-turn conversations with context awareness

Integration Points:

- Telephony/CCaaS platform (Asterisk/Broadsoft)
- Customer Relationship Management (CRM) system
- Billing system (BSS - Business Support System)
- Network Operations (OSS - Operations Support System)
- Ticketing and knowledge base systems

Expected Business Impact:

- Reduce inbound call volume handled by human agents by 40-50%
- Decrease average call wait time from 8 minutes to <1 minute for automated calls
- Improve First Contact Resolution (FCR) from 72% to 85%+
- Reduce operational costs per interaction by 65-75%
- Enable 24/7 support availability without additional staff

KEY PERFORMANCE INDICATORS (KPIs)

KPI #1: Containment Rate (First Call Resolution - FCR)

Category: Operational / Effectiveness

Definition: The percentage of customer calls successfully resolved by the AI voice agent without requiring escalation to a human agent or follow-up contact.

Formula: (Number of calls fully resolved by AI / Total calls handled by AI) × 100

Target: 65% in weeks 1-4 (pilot validation) → 75% by week 8 → 85% by week 12 (production readiness)

Why This Matters:

- Directly measures agent effectiveness at accomplishing its primary task
- High FCR reduces overall call volume, freeing human agents for complex issues
- Reduces customer frustration from repeated calls for the same issue
- Core driver of cost savings per interaction
- Aligns with industry benchmarks (Vodafone TOBi achieves >75% automation; ING Turkey handles 50%+ without escalation)

How to Measure:

- Automated logging of call outcomes in CCaaS platform
- Tag escalations with reason (customer requested human agent, agent recognition failure, out-of-scope query)
- Weekly dashboard reporting by inquiry type (billing, plan changes, outage checks, etc.)
- Compare FCR rates across language groups and time periods

Implementation Challenges & Mitigation:

- *Challenge:* Complex billing issues may require human review → *Mitigation:* Set realistic baseline targets; segment FCR by inquiry complexity
- *Challenge:* Customer preference for human agents early in deployment → *Mitigation:* Monitor over 12 weeks as trust builds
- *Challenge:* New agents may over-escalate to avoid errors → *Mitigation:* Use rules-based guardrails rather than pure ML

KPI #2: Average Call Handle Time (AHT) / Time to Resolution

Category: Efficiency

Definition: Average duration (in seconds or minutes) of customer calls handled by the AI voice agent, from connection to call completion (including any escalation setup).

Formula: Total call duration (all AI-handled calls) / Number of calls handled by AI

Target: <3 minutes for automated calls (vs. 7-10 minutes for human agents) by week 6; steady-state <2.5 minutes by week 12

Why This Matters:

- Measures operational efficiency and resource utilization
- Shorter resolution time improves customer satisfaction (reduced wait, faster problem-solving)
- Industry data shows AI agents close calls 15-35% sooner than human agents
- Enables the system to handle more concurrent calls (cost per interaction optimization)
- Indicates conversation naturalness—overly slow or repetitive agent interactions drive longer calls

How to Measure:

- CCaaS platform automatically logs call start/end times
- Dashboard segmentation by call type (allows comparison across inquiry categories)
- Track both median AHT (typical call) and 95th percentile (slower calls that may indicate problems)
- Weekly trend analysis with moving average to detect drift

Implementation Challenges & Mitigation:

- *Challenge:* Rushing calls may hurt resolution quality (reducing FCR) → *Mitigation:* Monitor AHT alongside FCR; don't optimize time at the expense of resolution
- *Challenge:* Complex calls may skew average → *Mitigation:* Segment metrics by inquiry type; compare within categories
- *Challenge:* Additional greeting/confirmation steps may slow early-stage calls → *Mitigation:* Expect initial AHT higher, gradually optimize as agent learns

---

KPI #3: Customer Satisfaction Score (CSAT) / Bot Experience Score (BES)

Category: User Experience / Human Factors

Definition: Direct measurement of customer satisfaction with the AI voice agent interaction, capturing both transaction success and interaction quality (tone, clarity, helpfulness, naturalness of conversation).

Formula (Simple CSAT):

- Post-call SMS/IVR survey: "Rate your experience 1-5"

- CSAT (%) = (Number of responses 4-5 / Total responses) × 100

Formula (Bot Experience Score - BES, advanced):

- Automated sentiment analysis across all conversations
- Deductions for negative signals: frustration cues, repeated requests, unnatural pauses, request for human agent
- Scale: 0-100, with conversation score determined by negative signal frequency
- Aggregate across all conversations for period to get overall BES

Target:

- Week 1-4: 70% CSAT (4.0/5.0 average) — pilot validation, expectations being set
- Week 5-8: 78% CSAT (4.2/5.0 average) — improved after model refinement
- Week 9-12: 85%+ CSAT (4.4/5.0 average) — production readiness; target parity with human agent scores (86-88%)

Why This Matters:

- Ensures technology is meeting customer emotional and practical needs, not just technical resolution
- High CSAT directly correlates with customer loyalty and reduced churn
- Identifies gaps between task completion and customer perception (e.g., agent solved problem but tone was cold)
- Detects model drift early—declining CSAT before operational metrics show problems
- Voice agent tone, pacing, and empathy are competitive differentiators vs. traditional IVR systems

How to Measure:

- **Post-Call Survey:** Automated SMS or IVR prompt "How would you rate this call?" immediately after completion. 15-20% response rate typical.
- **Sentiment Analysis:** Continuous NLP analysis of call transcripts for frustration markers (raised voice, negative language, empathy gaps)
- **NPS Alternative:** Quarterly deeper survey: "Would you recommend this support experience to others?" (0-10 scale)
- **Segmentation:** Break down by inquiry type, customer tenure, language group, time of day

Implementation Challenges & Mitigation:

- *Challenge:* Low survey response rate (15-20%) may not represent all calls → *Mitigation:* Combine with automated sentiment analysis
- *Challenge:* Customers may rate based on outcome alone, not interaction quality → *Mitigation:* Surveys can ask separate questions: "Did we solve your issue?" vs. "How natural/helpful was interaction?"
- *Challenge:* Negative sentiment may occur despite successful resolution (e.g., customer frustrated before call) → *Mitigation:* Distinguish pre-call frustration from agent-caused frustration using early/late call sentiment

KPI #4: Cost Per Interaction (CPI) / Cost Savings

Category: Financial / Business Impact

Definition: Total operational cost to handle one customer interaction via AI voice agent, compared to human agent baseline. Includes infrastructure, licensing, compute, maintenance, training, and escalation overhead.

Formula:

- **AI CPI:** (Total weekly AI operations cost / Number of calls handled by AI)
  - Includes: cloud infrastructure, voice platform licensing, model inference costs, escalation to human agent, oversight/monitoring
- **Human CPI:** (Agent salary + overhead + tools / calls handled per agent per day) ≈ $4-8 per call for mid-tier telecom
- **Savings Ratio:** (Human CPI - AI CPI) / Human CPI × 100

Target:

- Week 1-4: $1.50-1.80 per AI call (pilot, learning phase, more manual oversight)
- Week 5-8: $1.20-1.40 per AI call (efficiency gains, less manual intervention)
- Week 9-12: $1.00-1.20 per AI call (mature stage, ~65-75% cost reduction vs. human agent)

Baseline Human Agent Cost: $4.50-6.00 per call (includes salary $28/hr, overhead, tools, 120-150 calls/agent/day)

Expected ROI: 12-week pilot investment (~$80-120K in setup, training, infrastructure) recouped in month 2-3; positive ROI by month 5 for full rollout

Why This Matters:

- Quantifies business value in terms leadership understands and cares about
- Justifies continued investment and expansion to other departments
- Identifies inflection points (where AI cost efficiency surpasses human agent model)
- Drives decisions on automation scope (e.g., which call types to automate based on CPI impact)
- Critical for last-mile integration success—if costs don't decrease despite deployment, adoption will stall

How to Measure:

- **Infrastructure costs:** Cloud provider billing (AWS/Azure ASR, LLM inference, storage, bandwidth)
- **Licensing:** Voice platform costs (Twiilio, Vonage, custom), LLM API usage (OpenAI, Anthropic, local models)
- **Labor for oversight:** QA reviewers, prompt engineers, escalation handlers (partial, as they handle complex calls)
- **Maintenance & monitoring:** Engineering support, monitoring tools, model retraining labor
- **Escalation overhead:** Cost of transferring calls to human agents when AI needs help

- Compare to baseline human agent cost per interaction

Implementation Challenges & Mitigation:

- *Challenge:* Escalated calls to human agents add cost (may not achieve 65-75% savings until FCR improves) → *Mitigation:* Model FCR improvement into cost projections; expect higher CPI early
- *Challenge:* Difficult to allocate shared infrastructure costs → *Mitigation:* Use activity-based costing; allocate based on call volume and compute time
- *Challenge:* Hidden costs may emerge (regulatory compliance reviews, extra monitoring) → *Mitigation:* Track all costs in cost accounting system; update weekly

---

KPI #5: Escalation Rate & Escalation Quality

Category: Learning & Adaptability / Governance

Definition:

- **Escalation Rate:** Percentage of calls transferred to human agents (inverse of FCR; should be ~15-35% early, declining to 10-15% by week 12)
- **Escalation Quality:** Percentage of escalated calls that human agents successfully resolve in follow-up (measures agent handoff quality)

Formula:

- **Escalation Rate:** (Calls escalated to human agents / Total calls handled) × 100
- **Escalation Quality:** (Escalated calls resolved by human agent on first contact / Total escalated calls) × 100

Target:

- **Escalation Rate:** 35% week 1 → 25% week 4 → 15-20% week 12 (mature state; 80-85% self-sufficient)
- **Escalation Quality:** >85% of escalated calls resolved by human on first contact (indicating good context handoff)

Why This Matters:

- Shows whether AI is improving (learning) over time or stagnating
- Identifies which issue types AI struggles with (helps prioritize retraining)
- Escalation quality ensures poor handoffs don't degrade overall customer experience
- Declining escalation rate is a leading indicator of agent maturation
- Maps to sustainability—if escalations don't decrease, deployment stalls and ROI suffers

How to Measure:

- CCaaS logs all escalation triggers (customer request, agent recognition failure, confidence threshold breach, governance guardrail)

- Tag escalations by reason category (OOB = Out of Bounds, NLU Failure = intent not recognized, Policy = sensitive/compliance issue)
- Track human agent first-contact resolution on escalated calls via post-call survey
- Weekly dashboard showing escalation trends and top escalation reasons

Implementation Challenges & Mitigation:

- *Challenge:* High initial escalation rate (35%) may demoralize the team expecting immediate ROI → *Mitigation:* Set clear expectation that week 1-2 are baseline; improvement is the goal
- *Challenge:* Escalations due to customer preference ("I want a human") skews metric → *Mitigation:* Separate voluntary escalations from system escalations
- *Challenge:* If escalated calls aren't resolved, the problem may be human handoff, not AI training → *Mitigation:* Measure escalation quality separately; human teams may need coaching

ALIGNMENT WITH BUSINESS GOALS

Business Goal	KPI(s) Measuring Success	Target Impact
Reduce operational costs	CPI, FCR	65-75% cost reduction per interaction
Improve customer satisfaction	CSAT, BES, AHT	85%+ CSAT; faster resolution time
Increase 24/7 availability	FCR (availability metric)	Calls answered immediately, no wait times
Free up agent capacity for complex issues	Escalation Rate, Escalation Quality	80-85% of calls handled without human; escalated calls highly contextual
Enable future automation	Escalation Rate, learning signals	Data foundation to automate more use cases (billing disputes, account recovery, etc.)