Nortal

Product specification sheet

GenAI

AI-powered voice bot overview

February 2025

Large organizations such as banks and telcos face increasing pressure to handle high call volumes for both customer support and sales.

> This challenge is present in almost all businesses that directly interact with a large number of customers over voice-based mediums. Providing personalized experience in such services is essential and requires a large number of support personnel, who are often required to speak multiple languages. Additionally, providing customer support service in the afterhours and holidays requires investments in outsourcing such services. This challenge is especially acute in regulated industries, where conversations must not only be accurate and compliant but also culturally appropriate and sensitive to local nuances.

Our AI-powered voice bot solution

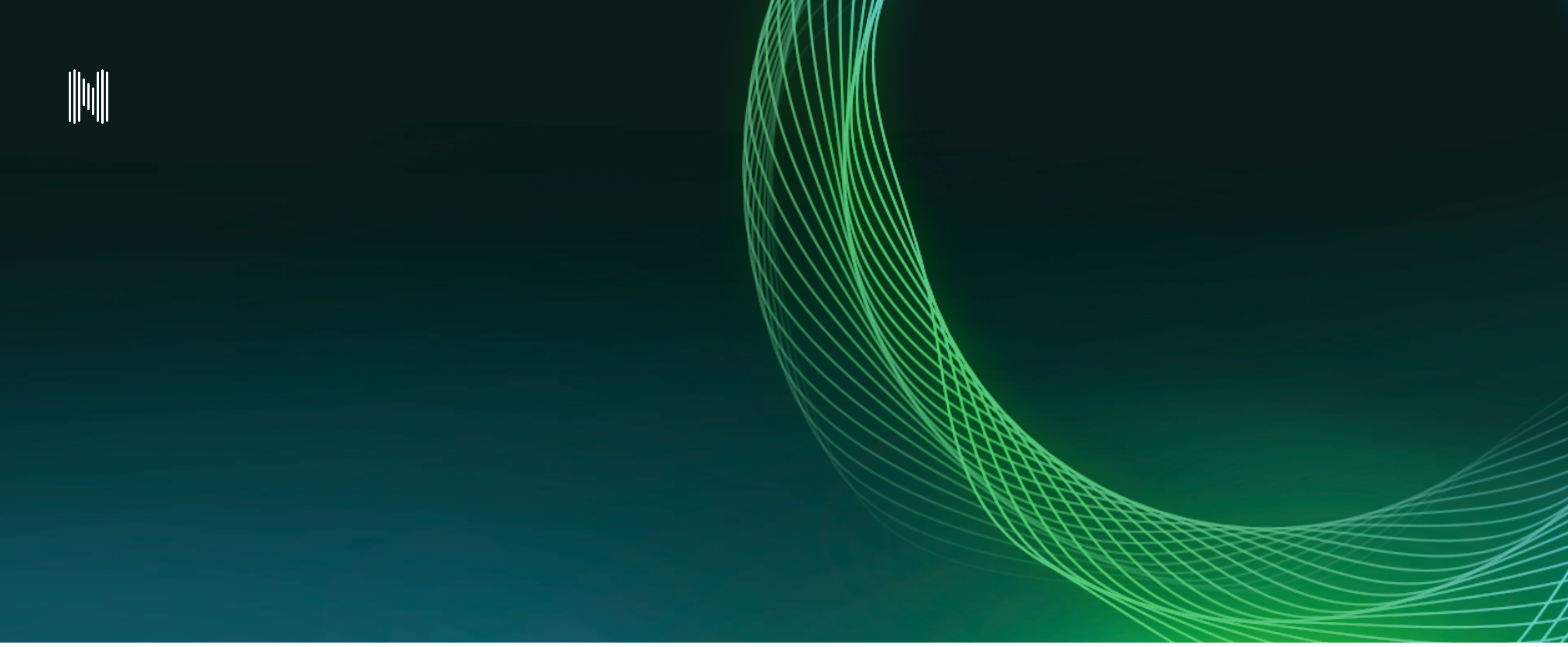
Addresses these needs by automating phone calls in multilingual setting, and allowing our client organizations to deliver seamless, human-like interactions without overburdening their support teams. It is especially valuable for the customer interaction in small languages such as Estonian.

System high-level architecture involves four major components:

1. Automatic Speech Recognition (ASR),
2. Retrieval-Augmented Generation (RAG) system for questioning-answering,
3. speech synthesis model (or Text- To- Speech, TTS) and
4. an orchestration framework.

To achieve the ambitious goal of building conversational AI systems, these technical components must come together seamlessly. ***ASR component serves*** for accurate transcription of spoken selected language into text. This requires selecting or developing an ASR AI model trained on diverse datasets to handle the nuances of selected language, including accents, dialects, and variations in pronunciation, abbreviations, etc.

Once speech is converted to text, the system must ***effectively process and respond to user queries.*** This is achieved through a ***RAG system.*** The RAG system combines two core capabilities: a retrieval module that uses semantic search to fetch relevant information from a knowledge base or document repository and a Large Language Model that produces fluent, context-aware responses in selected language.



Knowledge base can, for example, incorporate company public webpage information for the users, guides, template, any unstructured documents such PDF and DOCX documents etc. To make these document “understandable” for AI model, they are firstly split into smaller chunks, then transformed into numeric vector representations and are stored in the vector database. When user asks a question, this question is also converted to the numeric vector, and this vector is compared to the document vectors stored in the vector database. The most similar document vectors are then selected and are used as context for the answer generation. By incorporating semantic search and ranking mechanisms, the retrieval process ensures precision and relevance, while the generative model creates natural, human-like answers.

The next step is transforming the generated text responses back into speech. This is handled by a ***TTS component*** optimized for the selected. The TTS model must deliver natural-sounding speech, accurately reflecting language phonetics and tone. Accurate language representation (e.g. without foreign sounding accents) might be challenging in small languages such as Estonian.

Managing the dialogue between the system and the user requires a robust conversational flow orchestration layer, maintaining coherence and flow throughout the interaction.

Supporting these components is an end-to-end data pipeline designed to maintain system performance and adaptability. This pipeline collects and processes user interactions, enabling continuous monitoring, evaluation, and retraining of the ASR, RAG, and TTS models. This iterative approach ensures the system remains accurate and responsive to evolving user requirements.

Finally, the entire system must be integrated into a scalable and robust platform. Whether cloud-based such as Azure or AWS or deployed on-premises taking to account substantial compute power, the architecture must support modular components, enabling flexibility and future enhancements.

Challenges and implications:

Unlike chatbot interactions conversational AI requires very small latency in generating responses that is achieved by extensive optimization.

> **ASR model,** RAG LLM, **TTS model selection.** Model should be trained on selected languages for best performance. For on premises systems open source models and substantial GPU compute power should be considered.

Nortal

Challenge us, let’s create success together

Get in touch!
if you’d like to learn more or have a conversation: nortal.com

info@nortal.com