NFON Intelligent Assistant

Call Transcription & Summarization – AI Transparency Technical Note

May 2025 – V1.0 – Final

Contents

Introduction

NFON is committed to using artificial intelligence (AI) responsibly to improve the accessibility and usability of business communications. With **Call Transcription and Summarization**, users gain greater visibility into the content of their calls through **automatic, AI-generated transcripts and post-call summaries**.
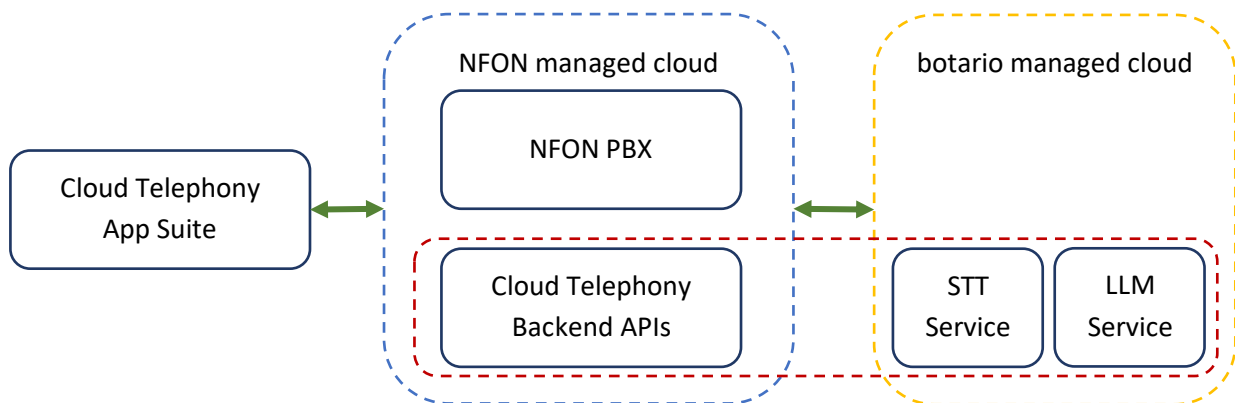
This feature has been developed in line with NFON's privacy and security principles, ensuring transparency, control, and compliance with relevant data protection regulations. This document outlines the design, functionality, and safeguards of the Call Transcription and Summarization feature.

Feature Overview

Call Transcription and Summarization is a **subscription-based feature** available for users of NFON's cloud telephony service. Once purchased, administrators can assign the feature to individual users and their associated extensions.

After all call participants have **agreed to the terms and conditions of the transcription**, the process is triggered automatically during a call and operates in the background in real time. Once the call ends, the system initiates a summarization task based **solely on the call's transcript**. Both the full transcript and the summary are made available to the user in the call history section of the NFON client interface. Access is **limited strictly** to the user who participated in the call.

The feature is designed to operate independently of call recording. Audio is not retained unless other recording services have been separately enabled.

- 🟢 HTTPS secured connections
- 🔴 Stateless operations: no customer data is persisted in the botario STT Service

Model Overview

Model Architecture

Call transcription is performed using the same **Whisper-based Speech-to-Text (STT) model** as described in the Voicemail Transcriptions feature. This model, managed and optimized by botario (an NFON company), operates within a dedicated, GPU-based cloud environment **located in the EU**.

The summarization task is executed **post-call** using a large language model (LLM) — **Gemma-3-27B-IT** — also hosted on the same botario-managed GPU cluster. The summary is generated **based exclusively on the transcript of the call**, without additional external data sources.

The processing environment includes multiple servers for redundancy and performance, as well as a load balancing infrastructure. All data is **processed in the EU and stored within Germany**.

Usage Guidelines

Users are notified about the availability of transcripts and summaries only after a call concludes and the summarization task is complete. These results are accessible within the NFON client under the relevant **call history entry**.

No other parties have access to another user's call summaries or transcripts. The design ensures that only the directly involved user can view the generated content.

Model Inputs and Outputs

The input to the system is the **audio stream** of a live call, which is securely transmitted from the NFON PBX to the botario STT service using **TLS 1.3 encryption**. The **output** of this step is the textual transcription of the call.

This transcription serves as the **input** for the second step — summarization — which produces a structured **text summary** of the call's content. Both the transcript and the summary are then stored within NFON's cloud infrastructure and linked to the specific call history item.

Data Sources for Training and Evaluation

The STT model is based on **OpenAI's Whisper** and a **fine-tuned Wav2Vec model**, trained using publicly available datasets and additional data gathered through a collaboration with the ITSC (Institute for Technical Communication Systems). No customer data is used for training or fine-tuning.

For the summarization component, the **Gemma-3-27B-IT** model was trained on a broad corpus that, while not explicitly detailed, is understood to include **publicly available and synthetic datasets**. NFON does not use customer data for training or tuning this model.

Model Evaluation and Performance

NFON / botario conducts **regular assessments and evaluations** of the AI models powering Call Transcription & Summarization to ensure high performance, accuracy, and reliability. This includes **testing and quality assurance processes**, where **human oversight** plays a key role in validating model behavior, identifying potential areas for improvement, and maintaining consistency across different speech patterns and accents.

Safety and Ethical Considerations

The summarization feature does not replace human judgment and is designed to **support** rather than automate decision-making. Both the transcription and summary are **neutral outputs** reflecting the content of the original conversation.

Efforts are taken to ensure that the models operate fairly across different speech patterns, accents, and content types. **No automated decisions** are made based on the content of a call, and the user remains in control of how the transcription and summary are used.

Privacy and Security

All processing is performed within NFON's infrastructure and its trusted, **EU-based** partners. Audio data is transmitted securely and not stored unless separately configured via other services.

The transcription and summary outputs are securely stored in the NFON cloud and are accessible **only to the user who participated in the call**. This ensures confidentiality and aligns with the principles of data minimization and access control.

References

- **NFON GTC & SLA** (see Service Description Cloud Telephony)
- **NFON Privacy Policy**
- **NFON Trust Center**
- **OpenAI Whisper**
- **Gemma 3 Technical Report**