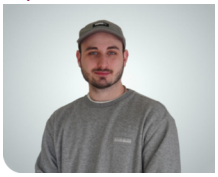
Text-to-Speech (TTS) in Public Transportation

Diplomanden



Baris Catan

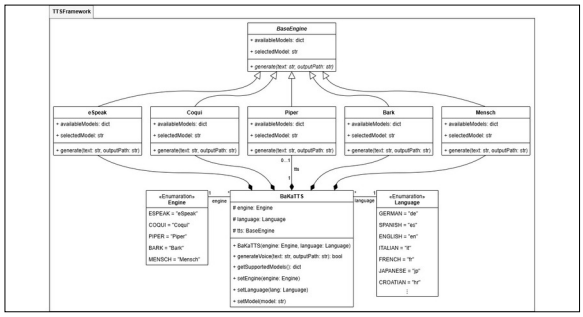
Kai Erdin

Ziel der Arbeit: Die Firma ErvoCom entwickelt Kommunikationssysteme für den öffentlichen Verkehr. Um manuelle Durchsagen zu automatisieren und zugleich eine konsistente Sprachqualität sicherzustellen, soll eine Text-to-Speech-(TTS)-Lösung auf einem Embedded-Linux-System evaluiert und realisiert werden. Ziel dieser Arbeit war es, geeignete TTS-Engines zu untersuchen, ein modulares Framework zu erstellen, die Engines hinsichtlich Ressourcenbedarf (CPU, RAM) sowie der Qualität der erzeugten Sprache objektiv und subjektiv zu bewerten und schliesslich einen Prototypen zu entwickeln. Dieser überträgt Texte via Client-GUI über MQTT an einen TTS-Server und gibt die generierte Sprache über einen VoIP-SIP-Audiostream an einen netzwerkbasierten Audioverstärker aus.

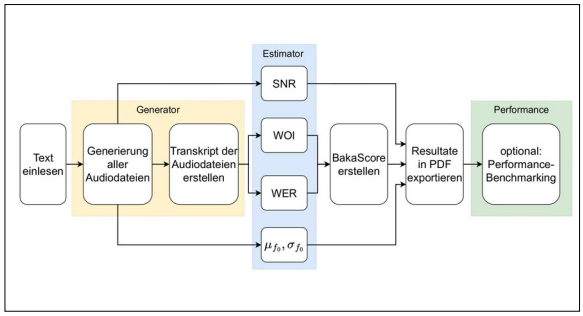
Vorgehen / Technologien: Zu Beginn wurden acht Open-Source-TTS-Engines recherchiert und hinsichtlich Sprachunterstützung, Audioqualität und Systemanforderungen untersucht. Vier Engines (eSpeak, Piper, Bark und Coqui) erfüllten die Anforderungen (lokale Ausführbarkeit, mehrsprachige Unterstützung) und wurden in ein eigens entwickeltes Python-Package integriert. Dies ermöglicht über eine einheitliche Schnittstelle den flexiblen Austausch von Engines und Sprachmodellen. Zur Bewertung wurde ein ausgeklügeltes Benchmarking-System entworfen. Insgesamt wurden 298 verschiedene Sprachmodelle getestet. Zur Beurteilung der Verständlichkeit wurden objektive Metriken wie Word Error Rate (WER) und Word of Interest (WOI) entwickelt und als "BaKaScore" zusammengefasst. Die Prosodie wurde anhand der Veränderung der Tonlage analysiert, während die Audioqualität zusätzlich durch die Signal-to-Noise Ratio (SNR) bewertet wurde. Über eine benutzerfreundliche GUI wurden zudem subjektive Bewertungen mittels Mean Opinion Score (MOS) mit Proband:innen durchgeführt. Die Ergebnisse wurden in der Arbeit analysiert und diskutiert. Abschliessend wurde ein funktionaler Prototyp realisiert, der die gesamte TTS-Kette - von der Texteingabe über die MQTT-Kommunikation bis zur Sprachausgabe via SIP - erfolgreich demonstriert.

Fazit: Das entwickelte System bietet eine skalierbare und wiederverwendbare Grundlage zur objektiven und subjektiven Bewertung von TTS-Engines. Durch die Kombination aus modularer Architektur, systematischer Bewertungsmethodik und funktionsfähigem Prototyp wurde gezeigt, dass sich synthetische Sprachsysteme effektiv in bestehende Kommunikationslösungen integrieren lassen. Die Resultate liefern eine fundierte Entscheidungsbasis für den praktischen Einsatz im Bereich automatisierter Durchsagen im öffentlichen Verkehr.

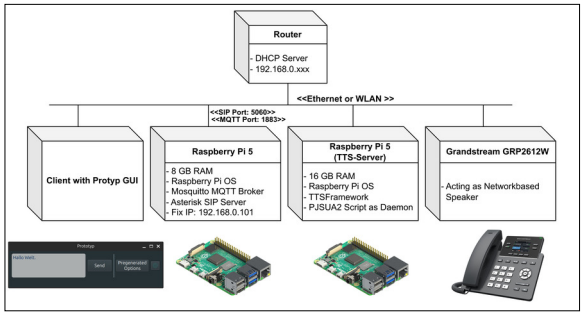
TTSFramework, das verschiedene Engines einheitlich vereint
Eigene Darstellung



Übersicht der Benchmarking Suite
Eigene Darstellung



Anwendung des TTSFrameworks in einem Prototypen, welches ein Kommunikationssystem im öffentlichen Verkehr simuliert
Eigene Darstellung



Referenten
Prof. Reto Bonderer, Ramon Moscatelli

Korreferent
Urs Reidt, Bonaduz, GR

Themengebiet
Embedded Software Engineering

Projektpartner
Ervocom Engineering AG, Feusisberg, SZ