ENHANCING THE EFFICIENCY OF HUMAN–COMPUTER (HUMAN-TO-MACHINE) COMMUNICATION IN NATURAL LANGUAGE IN AUTONOMOUS ROBOTIC SYSTEMS THROUGH THE ACQUISITION AND DESCRIPTION OF AGGREGATED MULTI-CHANNEL INFORMATION

Abstract

The dissertation addresses the problem of reliable voice communication between a human and an autonomous robotic system (e.g., a humanoid robot) in natural spoken interaction, specifically in scenarios involving multiple interlocutors (multi-talker), environmental disturbances, and dynamically changing speaker positions. The study assumes that speech perception is a process whose quality must be explicitly measured, diagnosed, and used within the decision-making loop of the robotic system. This stands in contrast to the classical approach, in which the acoustic path is treated as a passive channel delivering a single audio stream or a textual transcription used in a REPL-type dialogue scheme. Particular attention is devoted to the impact of sensor synchronisation and reliability (video and multichannel audio) on the correctness of beamforming path reconfiguration and, consequently, on the effectiveness of speech recognition and the fluency of the turn-taking mechanism in conversation.

The primary outcome of the dissertation is the development and validation of the **Smart Beamforming** framework architecture, designed for the acquisition, synchronisation, and fusion of multi-channel information, as well as for multichannel speech processing coupled with a supervisory control layer. An adaptive beamforming approach with parallel channels assigned to active speakers is proposed, along with a set of metrics and diagnostic modules that serve as perception-state sensors. These include output SNR measurements for interlocutors' speech channels, a Recovery Time Calculator module measuring the time required to restore quality after beam reconfiguration, and a speech recognition quality assessment module (offline via WER for utterances with reference transcriptions and online via estimates based on ASR hypothesis confidence). The Virtual Context Factory (VCF) layer aggregates information about interlocutors, stabilises context identification, supports the assignment of utterances to speakers, and estimates whether an utterance is addressed to the robot (target awareness). On this basis, supervisory control is implemented as a policy, with the SuperManager, FallbackManager, and FailsafeManager modules switching operating modes (including a REPL mode) and generating messages that correct interlocutors' behaviour, thereby closing the feedback loop between perception quality and the course of interaction.

Validation was carried out across scenarios covering both nominal and high-impact cases, including the addition of new interlocutors, partial overlap of utterances, degradation of angular separation, distance violations, degradation of the visual processing path, and overload of the number of contexts within the VCF. The studies comprised both supervised simulations and measurements performed on experimental installations involving a Pepper robot and a custom-built external microphone array (2×4 configuration). Comparative results between the baseline and full variants indicate that coupling quality metrics (SNR/RT/WER),

sensor synchronisation, and redundancy and fail-safe mechanisms stabilise system operation in multi-talker scenarios, reduce transitions to non-nominal states, and improve the functional effectiveness of interaction. Ultimately, the dissertation contributes both scientifically and practically by proposing a coherent, parameterizable model of speech perception for robotic systems, in which quality metrics serve as a fundamental element of control rather than merely tools for offline evaluation. The results of the work were implemented at Weegree sp. z o.o. sp.k. in Opole.