

ZWIĘKSZANIE EFEKTYWNOŚCI KOMUNIKACJI CZŁOWIEK-KOMPUTER (HUMAN-TO-MACHINE) W JĘZYKU NATURALNYM W AUTONOMICZNYCH SYSTEMACH ROBOTYCZNYCH POPRZECZ AKWIZYJCJĘ I DESKRYPCJĘ ZAGREGOWANEJ INFORMACJI WIELOKANAŁOWEJ

Streszczenie

Rozprawa dotyczy problemu niezawodnej komunikacji głosowej człowieka z autonomicznym systemem robotycznym (np. robotem humanoidalnym) w warunkach naturalnej interakcji głosowej, w szczególności w scenariuszach z wieloma interlokutorami (multi-talker), zakłóceniami pochodzącymi ze środowiska oraz dynamiczną zmianą położenia interlokutorów. W pracy przyjęto, że percepcja mowy jest procesem, którego jakość musi być jawnie mierzona, diagnozowana i wykorzystywana w pętli decyzyjnej systemu robotycznego. Co jest w przeciwieństwie do klasycznego podejścia, w którym tor akustyczny jest traktowany jako pasywny tor dostarczający pojedynczy strumień audio lub transkrypcję tekstową wykorzystaną w schemacie dialogowym typu REPL. W pracy szczególną uwagę poświęcono wpływowi synchronizacji i wiarygodności sensorów (wideo i wielokanałowego audio) na poprawność rekonfiguracji toru beamformingu, a w konsekwencji na skuteczność rozpoznawania mowy i płynność mechanizmu turn-taking w rozmowie.

Głównym wynikiem pracy jest opracowanie i walidacja architektury frameworka **Smart Beamforming**, który służy do akwizycji, synchronizacji i fuzji informacji wielokanałowej oraz wielokanałowego przetwarzania mowy, sprzęgniętego z warstwą sterowania nadzorczego. Zaproponowano adaptacyjny beamforming z równoległymi kanałami przypisanymi do aktywnych rozmówców, a także zestaw metryk i modułów diagnostycznych pełniących rolę czujników stanu percepcji, w tym pomiar SNR na wyjściu kanałów mowy interlokutorów, moduł Recovery Time Calculator mierzący czas powrotu jakości po rekonfiguracji wiązek oraz moduł oceny jakości rozpoznawania mowy (offline przez WER dla wypowiedzi z referencją i online przez estymaty oparte na pewności hipotezy ASR). Warstwa Virtual Context Factory (VCF) agreguje informację o rozmówcach, stabilizuje identyfikację kontekstów, wspiera przypisanie wypowiedzi do mówców oraz estymuje adresowanie wypowiedzi do robota (target awareness). Na tej podstawie zaprojektowano sterowanie w postaci polityki nadzorczej, realizowanej przez moduły SuperManager, FallbackManager i FailsafeManager, które przełączają tryby pracy (w tym w tryb REPL) oraz generują komunikaty korygujące zachowanie interlokutorów, domykając pętlę sprzężenia zwrotnego pomiędzy jakością percepcji a przebiegiem interakcji.

Walidację przeprowadzono w scenariuszach obejmujących przypadki nominalne i krytyczne, które uwzględniały dołączanie kolejnych rozmówców, częściowe nakładanie wypowiedzi, pogorszenie separacji kątowej, naruszenia odległości, degradację toru wizyjnego oraz przeciążenie liczby kontekstów w VCF. Badania obejmowały zarówno symulacje kontrolowane, jak i pomiary na stanowiskach badawczych z robotem Pepper oraz zewnętrzną macierzą mikrofonową własnej konstrukcji (układ 2x4). Wyniki porównawcze wariantu bazowego (*baseline*) i wariantu pełnego (*full*) wskazują, że sprzężenie metryk jakości (SNR/RT/WER), synchronizacja sensorów oraz mechanizmy redundancji i fail-safe stabilizują pracę systemu w scenariuszach multi-talker, ograniczają liczbę przejść do stanów

nienominalnych i poprawiają funkcjonalną skuteczność interakcji. Ostatecznie praca wnosi wkład zarówno naukowy, jak i aplikacyjny, proponując spójny, parametryzowalny model percepcji mowy przez systemy robotyczne, w którym metryki jakościowe są integralnym elementem sterowania, a nie jedynie narzędziem ewaluacji offline. Wdrożenie wyników pracy dokonano w firmie Weegree sp. z o.o. sp.k w Opolu.