Synchronizace mezi jazykem a gesty během interakce s robotem

Karina Zamrazilová, Michal Vavrečka, Sofiia Ostapenko, Gabriela Šejnová, Júlia Skovierová

Český institut informatiky, robotiky a kybernetiky, ČVUT v Praze, Česká republika
michal.vavrecka@cvut.cz

Abstrakt

Cílem naší studie je ověření faktu, zda jsou deklarativní gesta během člověk–robot interakce (HRI) časově sladěna s příslušnými částmi jazykových instrukcí. Navrhli jsme proto VR experiment, ve kterém 26 participantů učilo humanoidního robota pomocí gest a řeči. Získaná data ($n = 1126$) ukazují, že gesta zaostávají za řečí: nástup gesta průměrně o 0.56 ± 1.30 s ($p < 0.001$), vrchol o 0.66 ± 1.25 s ($p < 0.001$), avšak celková délka gesta trvala déle než řeč. Analýza distribucí dat navíc odhalila, že přestože se signály liší v absolutním načasování, existuje mezi jejich klíčovými body korelace: nástup $r = 0{,}644$ ($p < 0{,}001$), vrchol $r = 0{,}646$ ($p < 0{,}001$). Výsledky naznačují, že lidé koordinují gesta s řečí spíše relačně než absolutně, což je důležité pro návrh multimodálních HRI rozhraní.

1 Úvod

Klíčovými modalitami lidské komunikace jsou řeč a gesta. Není však jasné, zda lidé gesta s řečí přirozeně synchronizují, nebo zda mezi modalitami existuje časová asynchronie. Cílem studie je analýza vztahu mezi začátkem a vrcholem deiktických gest a odpovídajícími prvky řeči. Zaměřili jsme se proto na přesné časové body: nástup gesta, vrchol gesta, začátek řečové instrukce a vyřčení klíčového slova.

Během výzkumu jsme testovali tyto hypotézy:

- **H1a**: Existuje statisticky významný rozdíl mezi začátkem gesta a začátkem řeči.
- **H1b**:Existuje statisticky významný rozdíl mezi časem vrcholu gesta a časem odpovídající jazykové části.
- **H2**: Existuje statisticky významný rozdíl mezi celkovou délkou trvání gesta a délkou trvání řečové instrukce.

2 Související práce

Současné přístupy v robotice se často zaměřují jen na jednu modalitu nebo používají různé modality k popisu jednotlivých částí sdělení, ale jen v silně strukturovaných scénářích. Pro přirozenější interakci je nutné efektivní rozpoznání spojení více modalit s přesnou detekcí lidského záměru (Vanc a spol., 2024). Holzapfel a spol. (2004) zjistili, že gesta často předcházejí řeči o cca 0,3 s, zatímco Matuszek et al. upozorňují, že striktní synchronizace není nutná, avšak flexibilní načasování stále umožňuje efektivní komunikaci (Matuszek a spol., 2014).

Někteří výzkumníci použili VR systémy pro zkoumání těchto modalit, např. Wu a spol. (2021) vyvinuli systém Ges-THOR pro interakci pomocí gest v prostředí simulovaného učení. Tyto studie potvrzují potenciál VR pro zkoumání HRI, ale nezaměřují se na časové sladění řeči a gest. Naše práce na těchto poznatcích staví a využívá VR pro přesné sledování časování gest s ohledem na řeč.

3 Metodologie

3.1 Prostředí a průběh experimentu

Studie probíhala ve virtuálním prostředí vytvořeném pomocí simulátoru iGibson (Li a spol., 2021), propojeného s rozhraním OpenVR a headsetem HTC Vive Pro Eye. Účastníci (n = 26) komunikovali s humanoidním robotem Fetch, kterého učili rozpoznávat objekty a jejich vlastnosti (barva, tvar, velikost) pomocí řečových instrukcí a deiktických gest. Scénář zahrnoval 16 úloh s pěti objekty v pseudonáhodném rozmístění. Robot poskytoval zpětnou vazbu v reálném čase a zaznamenával multimodální vstupy.

3.2 Vzorek

Výzkumu se zúčastnilo 19 žen a 7 mužů (průměrný věk 26 let), všichni byli praváci a mluvili plynně anglicky.

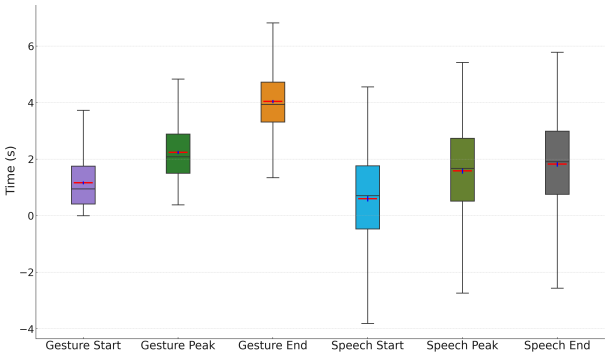
3.3 Záznam a analýza dat

Během experimentálních úloh byly zaznamenávány pohyby hlavy, trupu a obou rukou (vzorkovací frekvence 30 Hz), dále zvukový záznam z vestavěného mikrofonu headsetu a videozáznam z perspektivy účastníka. Pro každý experimentální pokus byly pohybové trajektorie anotovány na tři klíčové události: začátek, vrchol a ukončení gesta. Paralelně byly z audiozáznamů automaticky extrahovány jazykové události, zejména moment vyslovení klíčového slova nebo požadované vlastnosti

cílového objektu, pomocí nástroje WhisperX, který poskytuje přesné časové značky jednotlivých komponent řeči. Celkem bylo identifikováno 1126 validních párů gesto–řeč. Následná statistická analýza zahrnovala výpočet časových rozdílů mezi začátky a vrcholy obou modalit, analýzu trvání jednotlivých signálů a výpočet korelací za účelem zjištění míry časové synchronizace.

4 Výsledky

K ověření všech hypotéz byly použity Wilcoxonovy párové testy a výsledné p-hodnoty byly upraveny Bonferroniho korekcí. V rámci H1a byl identifikován statisticky významný rozdíl mezi začátky řeči a gest ($W = 348\,495$, $p < 0{,}001$), přičemž řeč typicky předcházela gestu. Průměrný čas zahájení řeči byl $0{,}60\,\mathrm{s}$ ($\sigma^2 = 2{,}66$), zatímco u gest činil $1{,}16\,\mathrm{s}$ ($\sigma^2 = 0{,}79$), s rozdílem $\Delta T = 0{,}56\,\mathrm{s}$ ($\sigma^2 = 1{,}69$). Spearmanova korelace mezi začátky obou modalit dosahovala $\rho = 0{,}644$ ($p < 0{,}001$), což naznačuje silnou, ale ne dokonalou časovou souvislost. Hypotéza H1b, zaměřená na vrcholy gest a řeči, odhalila rovněž významný rozdíl ($W = 287\,006$, $p < 0{,}001$); vrcholy gest následovaly po výslovnosti klíčového slova (střední hodnota pro řeč: $M = 1{,}58\,\mathrm{s}$, $\sigma^2 = 2{,}56$; pro gesto: $M = 2{,}24\,\mathrm{s}$, $\sigma^2 = 0{,}92$; $\Delta T = 0{,}66\,\mathrm{s}$), s korelací $\rho = 0{,}646$ ($p < 0{,}001$). Rozpětí rozdílů se pohybovalo od $-1{,}10$ do $6{,}26\,\mathrm{s}$ a směrodatná odchylka činila $1{,}25\,\mathrm{s}$, což ukazuje na variabilitu strategií mezi anticipací a reakcí. V rámci H2 byl zaznamenán statisticky významný rozdíl v délce trvání obou modalit ($W = 48\,672$, $p < 0{,}001$), přičemž gesta trvala výrazně déle ($M = 2{,}01\,\mathrm{s}$, $\sigma^2 = 0{,}71$) než řeč ($M = 1{,}03\,\mathrm{s}$, $\sigma^2 = 0{,}30$; $\Delta T = 0{,}98\,\mathrm{s}$). Gesta tedy nejen začínají a kulminují později, ale i přetrvávají déle, což potvrzuje jejich podpůrnou a doplňující roli v komunikaci. Na obr. 1 lze vidět celkový čas v trvání gest a řeči. Zároveň ilustruje významné rozdíly, které jsme zaznamenali. Tato zjištění mohou mít zásadní dopady na návrh budoucích systémů v rámci HRI.



Obr. 1: Vizualizace průměrných reakčních časů pro jednotlivé modality.

5 Diskuze a závěr

Studie detailně analyzovala časové souvislosti mezi řečí a deklarativními gesty v prostředí HRI pomocí přesného sledování pohybu ve VR. Na základě výše provedených statistických testů nemůže vyvrátit H1a, tudíž můžeme říci, že začátek gest se významně liší od začátku řeči, což vyvrací např. zjištění (Holzapfel a spol., 2004), kde gesta řeč předcházela. Rozpor může souviset s rozdílnou náročností úloh – naše scénáře vyžadovaly vyšší míru kognitivního zapojení a často vedly k "just-in-time" použití gest jako doplňku po verbální instrukci, podobně jako ve studii u Matuszek a spol. (2014). H1b nelze také vyvrátit, poněvadž se ukázalo, že vrcholy gest přicházejí až po výslovnosti klíčových slov, na to navazuje i H2, kterou rovněž nelze vyvrátit, protože gesta trvají déle než odpovídající řeč. To naznačuje, že gesta slouží nejen k okamžité ilustraci, ale i jako vizuální opora sdělení během celé komunikační sekvence.

Poděkování Tento projekt byl podpořen Grantovou agenturou České republiky, grant č. 23-04080L. Tato práce byla podpořena Ministerstvem školství, mládeže a tělovýchovy České republiky prostřednictvím e-INFRA CZ (ID:90254)

Reference

Holzapfel, H., Nickel, K. a Stiefelhagen, R. (2004). Implementation and evaluation of a constraint-based multimodal fusion system for speech and 3d pointing gestures. V *Proceedings of the 6th international conference on Multimodal interfaces*, str. 175–182.

Li, C., Xia, F. a spol. (2021). igibson 2.0: Object-centric simulation for robot learning of everyday household tasks. *arXiv preprint arXiv:2108.03272*.

Matuszek, C., Bo, L., Zettlemoyer, L. a Fox, D. (2014). Learning from unscripted deictic gesture and language for human-robot interactions. V *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 28.

Vanc, P., Skoviera, R. a Stepanova, K. (2024). Tell and show: Combining multiple modalities to communicate manipulation tasks to a robot. *arXiv*.

Wu, Q., Wu, C. J., Zhu, Y. a Joo, J. (2021). Communicative learning with natural gestures for embodied navigation agents with human-in-the-scene. V *2021 (IROS)*, str. 4095–4102, Prague, Czech Republic.