TECHNOLOGIE MULTIMEDIALNE - ETI PG - 4 SEMESTR

Interfejsy Multimodalne

Od klawiatury i myszy do systemow rozumiejacych Twoj wzrok, głos, gesty i emocje

mgr inż. Szymon Zaporowski

ROZWÓJ INTERFEJSOW

Od Liczydła do Agenta AI

~2500 p.n.e.
Liczydło
Pierwszy interfejs. Dotyk + wzrok.

lata 60.
CLI
Klawiatura, Komendy tekstowe.

lata 80.
GUI + Mysz
Xerox->Apple->Windows. Ikony, metafora.

lata 90.
Mowa + Pismo
Pierwsze ASR, OCR. DataGlove - gesty.

Trend: Kazda dekada przybliża interfejs do naturalnej komunikacji człowiek-człowiek → od znaków binarnych do głosu, gestów i emocji jednocześnie.

ROZWÓJ INTERFEJSOW

Od Liczydła do Agenta AI

2000-10

Multi-Touch

iPhone 2007.
Kinect, Siri.

2010-20

VR / AR

Oculus,
HoloLens.
Eye-tracking.

2020 ->

Agenci LLM

GPT-4o, Gemini.
Text+obraz+audio.

Trend: Kazda dekada przybliża interfejs do naturalnej komunikacji człowiek-człowiek → od znaków binarnych do głosu, gestów i emocji jednocześnie.

TERMINOLOGIA

Czym jest modalność i jak ją nazwiemy?

EWOLUCJA NAZEWNICTWA

II Woj. Swiat.	MMI	Man-Machine Interaction
lata 70.	HCI	Human-Computer Interaction
lata 80.	HMC	Human-Machine Communication
lata 90.	PUI	Perceptual User Interface
2000+	NIS	Natural Interactive Systems

Nazwy się zmieniają, problem pozostaje bez zmian:
Jak sprawić, by maszyna rozumiała człowieka w sposób dla nas naturalny?

TERMINOLOGIA

Czym jest modalność i jak ją nazwiemy?

CZYM JEST MODALNOŚĆ?

Modalność = sposób przekazywania i odbierania informacji

Wejście (Input)

Jezyk naturalny, gesty, mowa ciała, mimika, pismo, wzrok (gaze), fale mózgowe (EEG), reakcje skóry (EMG)

Wyjście (Output)

Ekran/projektor (wzrok), głośnik (słuch), siła zwrotna/tekstura (dotyk), skafander, zapach

Klucz: Modalności ≠ tylko zmysly. To kanały komunikacyjne - jeden człowiek używa kilku jednocześnie (mówisz + gestykulujesz + patrzysz).
Dobry interfejs potrafi je scalic.

KATALOG MODALNOSCI

Co człowiek daje komputerowi - i co dostaje z powrotem?

INPUT -- Czlowiek -> Komputer

Ręce / ramiona	Klawiatura, mysz, joystick, DataGlove, ekran dotykowy, kamera (gest)
Mimika twarzy	Kamera + detekcja 68 punktów twarzy; analiza emocji (FACS+CNN)
Jezyk ciała	Kamera + estymacja pozy (MediaPipe, OpenPose); czujniki pozycji IMU
Wzrok (gaze)	Kamera + IR diody -> eye-tracker (Tobii, Apple Vision Pro)
Skóra / mięśnie	Czujniki EMG (aktywność mięśni); GSR (Galvanic Skin Response - stres)
Fale mózgowe	EEG (delta, alfa, beta, gamma); BCI - sterowanie myślami

KATALOG MODALNOSCI

Co człowiek daje komputerowi - i co dostaje z powrotem?

OUTPUT -- Komputer -> Człowiek

Wzrok	Monitor, projektor, AR/VR display, HUD, hologram
Słuch	Głosniki stereo/surround, spatial audio (ambisonia), TTS
Dotyk/Haptyka	Siła zwrotna (force feedback), wibracje, tekstura, rękawice haptyczne
Ciało	Skafander (exosuit) - symulacja ruchu, grawitacji, temperatury
Węch (rzadkie)	Interfejs olfaktoryczny - lawenda, cytrus; dyfuzja zimnego powietrza

KLASYFIKACJA

Jednomodalne vs. Multimodalne - Co i Kiedy?

Kryterium	Jednomodalne	Multimodalne
Kanały	1 (np. tylko glos)	2+ jednoczesnie
Odporność	Niska -- 1 szum = awaria	Wysoka -- redundancja
Naturalność	Sztuczna (uczysz sie sys.)	Naturalna (jak z człowiekiem)
Złożoność	Niska	Wysoka (sync, fuzja)
Przykład	CLI, ASR, OCR	Apple Vision Pro, Whisper+gesty

Zasada uzupełniania (Complementarity): Mówisz **'weź to'** -- system nie wie, co oznacza "***to***".
Dodajesz wskazanie gestem - kontekst jest jasny.
Żadna modalność sama nie wystarczy.

KLASYFIKACJA

Jednomodalne vs. Multimodalne - Co i Kiedy?

JAK MODALNOŚCI WSPÓŁPRACUJĄ?

Rownoważność	Dwie modalności robią to samo - użytkownik wybiera wygodniejszą (glos LUB dotyk)
Redundancja	Obie przekazują to samo -> wyższa pewność systemu (gesty + mowa = ta sama komenda)
Uzupełnianie	Kazda niesie inną informację. Razem dają pełny obraz (mowa 'co to?' + wzrok na obiekt)
Specjalizacja	Kazda modalność do innego zadania: głos-> komendy, wzrok->cel, haptyka->potwierdzenie
Transfer	Informacja przekształcana między kanałami: tekst -> TTS; obraz -> opis głosowy

INTERFEJSY JEDNOMODALNE

Skad Wyszliśmy? - Przegląd Interfejsów Unimodalnych

Skad Wyszliśmy? - Przegląd Interfejsów Unimodalnych

Pismo odreczne

OCR / HWR

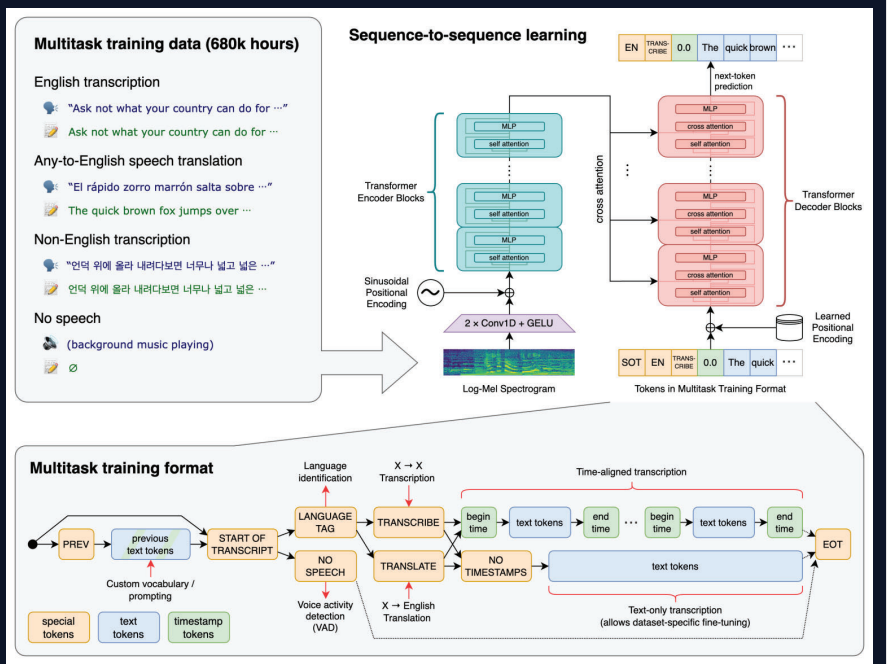
Unistroke/Graffiti (Palm 3Com) -- zdefiniowane symbole.
Apple Newton (1993) -- komputer uczy się pisma użytkownika.
Problem: roznorodność stylisów, nierozłączność liter.

Skad Wyszliśmy? - Przegląd Interfejsów Unimodalnych

Rozpoznawanie mowy

ASR

Podejście 1: uzytkownik uczy komputer słow (speaker-dependent)
Podejście 2: gotowy słownik + adaptacja do głosu.
Problem: mowa ciagła, halasy w tle, akcent.
Dziś: Whisper <6% WER.

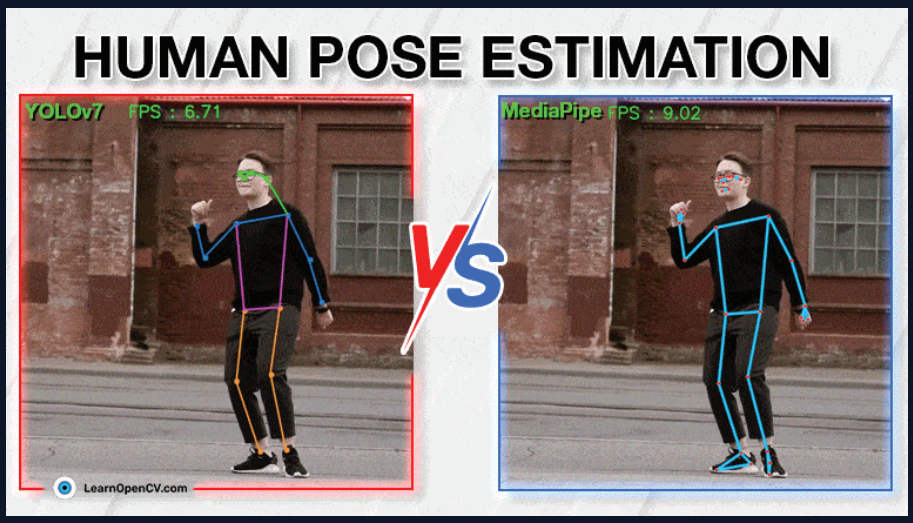


Skad Wyszliśmy? - Przegląd Interfejsów Unimodalnych

Rozpoznawanie gestów

Gesture

DataGlove (~1987) -- czujniki zgięciowe, ~10 pozycji.
Camera-based: Leap Motion, MediaPipe -- 21 keypoints per ręka ok. 60 fps.
Problem: oświetlenie, okluzja, rożnice kulturowe



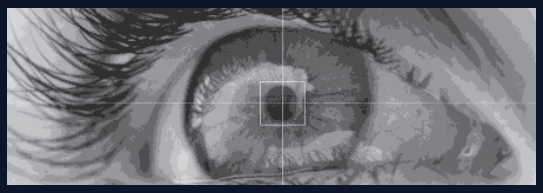
INTERFEJSY JEDNOMODALNE

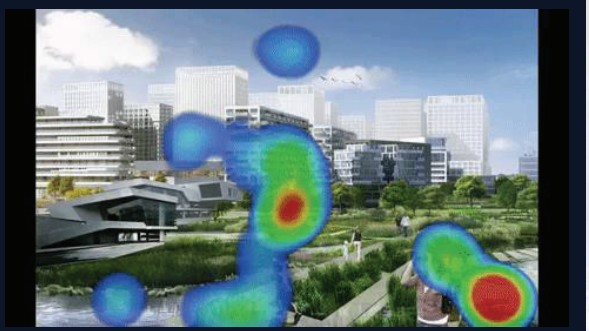
Skad Wyszliśmy? - Przegląd Interfejsów Unimodalnych

Eye Gaze Tracking

Gaze

4 diody IR -> refleksy rogówkowe -> kamera śledzi oko.
Szacuję punkt fiksacji niezależnie dla każdego oka.
Dzis: Tobii +/-0.4 stopnia, Vision Pro -- zero kalibracji.





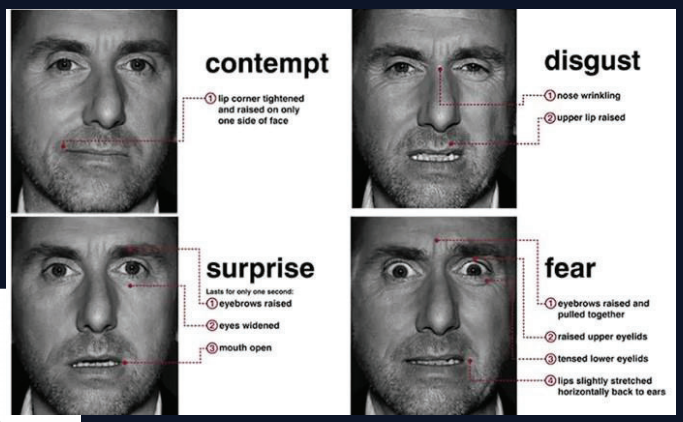
Technologie Multimedialne · ETI PG · 4 semestr
INTERFEJSY MULTIMODALNE
B-10

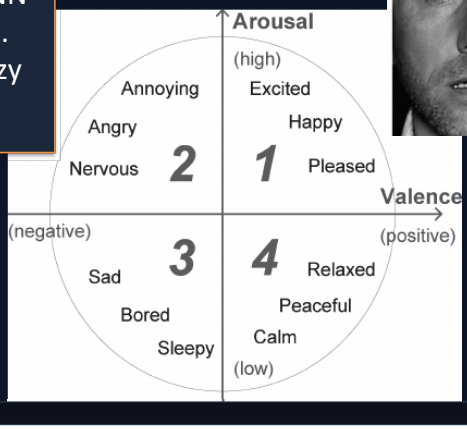
Skad Wyszliśmy? - Przegląd Interfejsów Unimodalnych

Mimika / Emocje

Facial

Avatar SONY (1994): model twarzy z 500 wielokątow, 26 wyrazów.
Dziś: Action Units (FACS) + CNN -> 7 emocji + valence/arousal.
Problem: emocje rożne między kulturami



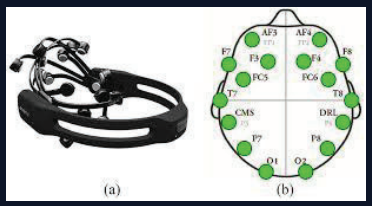


Skad Wyszliśmy? - Przegląd Interfejsów Unimodalnych

Biofeedback / BCI

EEG/EMG

Czujniki EMG (mięsnie), EEG (rytmy: delta/alfa/beta/gamma), GSR.
Neurofeedback: użytkownik świadomie kontroluje swój stan mózgu.
Dzis: OpenBCI, Emotiv -- BCI dla graczy, rehabilitacji, para-sportu



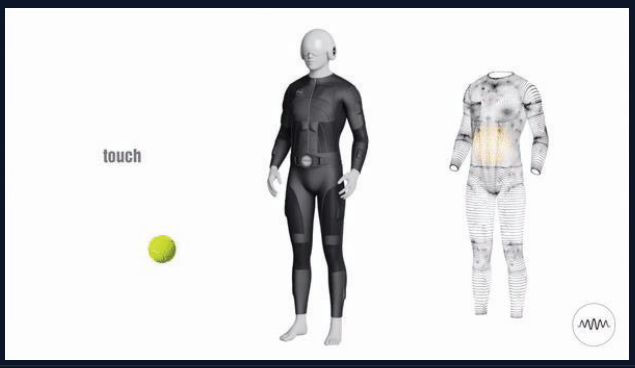
Kiedy dodajemy dotyk, węch i ciało do interfejsu

Interfejsy Haptyczne Dotyk + Sila

Stymulacja kanałów kinestetycznych (ruch/siła) i dotykowych (tekstura/wibracja).

Cecha wyróżniajaca: jednoczesna wymiana informacji
-- pętla zamknięta (system reaguje na ruch użytkownika i użytkownik czuje reakcję systemu).

Dziś: DualSense PS5, rękawice HaptX,
skafandry bHaptics, rękawice chirurgiczne VR.



PIERWSZE SYSTEMY MULTIMODALNE

Kiedy dodajemy dotyk, węch i ciało do interfejsu

Interfejs Węchowy Olfaktoryczny

Emisja zapachów sterowana komputerowo --
dostosowanie środowiska do stanu użytkownika.
Projekt KSMM PG: interfejs dla dzieci z ADHD.
Lawenda (uspokajanie), cytrus (koncentracja).
Metoda: dyfuzja zimnego powietrza.
Problem: zapach nie da się 'oczyścić' szybko
z pomieszczenia; mieszanie komplikuje system.

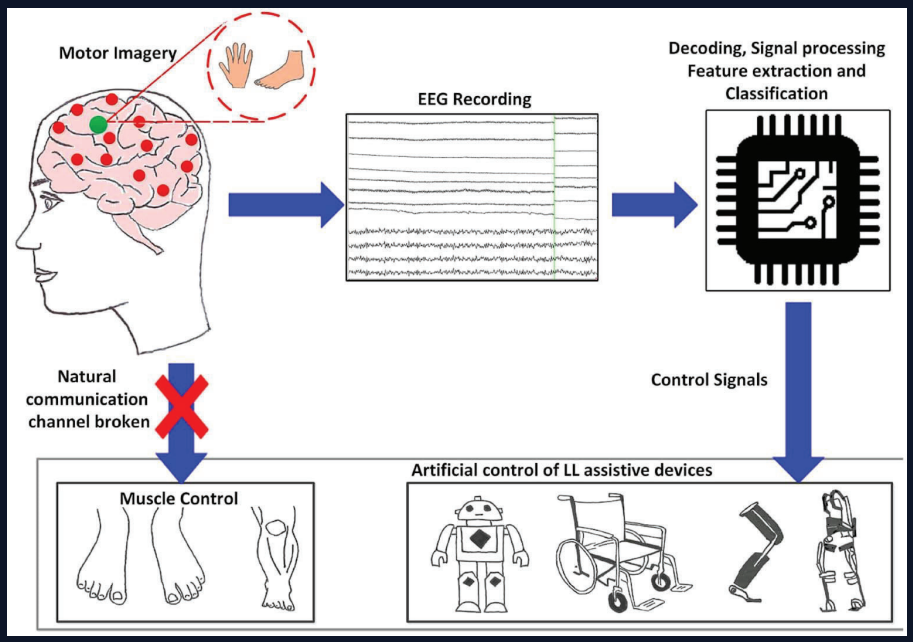
Technologie Multimedialne · ETI PG · 4 semestr INTERFEJSY MULTIMODALNE B-11

Kiedy dodajemy dotyk, węch i ciało do interfejsu

Biofeedback / BCI Neurokontrola

Sygały biometryczne jako kanał wejścia:
EEG, EMG, GSR, EKG, akcelerometr.
Pipeline EEG: Elektrody -> filtracja adaptacyjna (artefakty) -> filtr pasmowy (delta/alfa/beta) -> Decyzja.
Synchronizacja półkul: system mierzy dominujący rytm i daje wizualny feedback -> użytkownik kontroluje swój stan mózgu.

Dziś: rehab. po udarze, protezy (motor BCI),
wykrywanie zmęczenia pilota, gry neurofeedback.



Strategie Łączenia Modalności (Fusion Strategies)

1 Early Fusion (Fuzja Wczesna)

Audio (48 kHz) + Video (30 FPS) → połączenie ZANIM nastąpi ekstrakcja cech → sieć neuronowa → decyzja

✓ **Zaleta:** Sieć uczy się współzależności między modalnościami od razu

✗ **Wada:** Wymaga synchronizacji (jitter < 50 ms), szum w jednym = szum w drugim

2 Late Fusion (Fuzja Późna)

Audio → Sieć 1 → wynik | Video → Sieć 2 → wynik | Połącz wyniki → decyzja

✓ **Zaleta:** Modularna, tolerancja na szum, niezależne kanały

✗ **Wada:** Mniej uczy się zależności między modalnościami

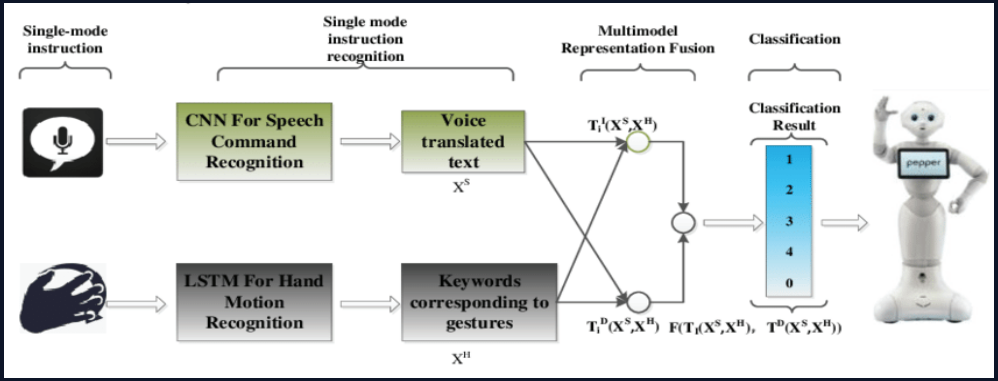
FUZJA MODALNOŚCI

Strategie Łączenia Modalności (Fusion Strategies)

3 Hybrid Fusion (Fuzja Hybrydowa)

Wczesna fuzja dla szybkich współzależności (lip-sync) + Późna fuzja dla odporności na szum

✓ **Najlepsza praktyka** dla systemów świata rzeczywistego



PRZYKŁADY

Praktyczne Przykłady: Multimodalność Wokół Nas

Smartphone (Codzienny)

- **INPUT:** Dotyk, głos, gesty
- **OUTPUT:** Ekran, dźwięk, wibracja
- **Fusion:** Pozwala „mówić i dotykać" równocześnie

Gaming VR (Immersive)

- **INPUT:** Wzrok 3D, ruch głowy, ręce, mowa
- **OUTPUT:** 4K stereo, dźwięk przestrzenny, kombinezon haptyczny
- **Fusion:** Naturalne „bycie" w wirtualnym świecie

Accessibility (Inkluzja)

- **INPUT:** Eye-gaze + EEG + EMG
- **OUTPUT:** Mowa syntetyczna + wskaźniki wizualne
- **Fusion:** Pozwala osobom sparaliżowanym sterować interfejsem

Robot Mobilny (Autonomiczny)

- **INPUT:** LiDAR, kamera, mikrofon, zderzak
- **OUTPUT:** Ruch, dźwięk, LED, gesty
- **Fusion:** Redundancja (jeśli LiDAR zawiedzie, kamera pomaga)

ARCHITEKTURY MULTIMODALNE ML

CLIP · Flamingo · GPT-4o — jak modele widzą i słyszą

CLIP (OpenAI, 2021)

- Podwójny koder: obraz (ViT) + tekst (Transformer)
- Trening: 400M par obraz-opis (contrastive loss)
- Embedding do wspólnej przestrzeni 512-D
- Zero-shot: pasuje opis do obrazu bez fine-tuningu
- Podstawa dla DALL·E 2, Stable Diffusion

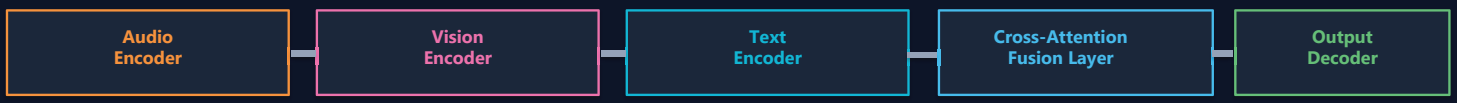
Flamingo (DeepMind, 2022)

- Vision encoder (NFNet) zamrożony
- Bramkowany mechanizm uwagi krzyżowej: obraz ↔ LLM warstwy
- Few-shot: wystarczą 4 przykłady w prompcie
- Odpowiada na pytania o sekwencje obrazów
- Poprzednik rodziny Gemini

GPT-4o (OpenAI, 2024)

- End-to-end: audio+obraz+tekst → jeden model
- Brak osobnych modułów ASR/OCR — współdzielone wagi
- Latencja głosowa ~320 ms (vs 2,8 s GPT-4 Voice)
- Rozumie emocje w głosie (tembr, intonacja)
- Nowy paradygmat: modalności uczą się wzajemnie

SCHEMAT MECHANIZMU UWAGI KRZYŻOWEJ (Hybrid Fusion)



Uwaga krzyżowa pozwala każdemu koderowi 'pytać' inne modalności o kontekst — np. dekoder mowy może uwzględnić obraz kamery.

WYZWANIA INŻYNIERSKIE

Latencja · Kalibracja · Niezbalansowane dane

LATENCY BUDGET — BUDŻET OPÓŹNIEŃ

Haptyka / Force-feedback	< 10 ms
Audio / TTS response	< 20 ms
Eye-gaze cursor update	< 50 ms
ASR (rozpozn. mowy)	< 150 ms
Visual AR overlay	< 20 ms

⚠ Latencja haptyczna > 10 ms: mózg wykrywa brak spójności dotyk–obraz (rubber hand illusion się rozpada).

🔑 Reguła praktyczna: najpierw zmierz latencję end-to-end całego pipeline'u ZANIM dodasz kolejną modalność — każdy nowy sensor to potencjalny bottleneck.

Kalibracja Cross-Modalna

- Różne sensory mają różne częstotliwości próbkowania (EEG 256 Hz, kamera 30 fps, mikrofon 48 kHz)
- Timestamp synchronisation: PTPv2 / NTP < 1 ms drift
- Kalibracja przestrzenna (eye-tracker ↔ kamera RGB)
- Online re-calibration gdy użytkownik zmienia pozycję
- Drift IMU: błąd nakładający się w czasie bez korekty

Data Imbalance

- Dane głosowe: miliardy próbek (Common Voice, LibriSpeech)
- Dane EEG: tysiące prób od kilkudziesięciu osób
- Dane haptyczne: niemal brak publicznych zbiorów
- Strategie: SMOTE, synthetic data (GAN/VAE), transfer learning z domenowo bliskich zbiorów
- Ryzyko: model 'ignoruje' rzadką modalność → bias

PRYWATNOŚĆ I ETYKA

BCI · Eye-Gaze · FACS · GDPR — kto jest właścicielem twoich danych?

BCI / EEG — Dane Mózgowe

- Sygnał EEG ujawnia: stan skupienia, emocje, intencje ruchu, predyspozycje neurologiczne, ślady chorób (epilepsja, ADHD)
- Neurodata = najbardziej intymny typ danych biometrycznych
- Neuralink, Emotiv, OpenBCI — brak dedykowanych regulacji w UE
- 2024: Chile pierwsze państwo z konstytucyjną ochroną neurodata
- ❓ Czy pracodawca może żądać BCI 'dla sprawdzenia produktywności'?

Eye-Gaze — Okno na Myśli

- Śledzi co czytasz, co omijasz, gdzie się wahasz
- Ujawnia: zainteresowania, preferencje seksualne, kłamstwa, problemy z czytaniem (dysleksja), choroby oczu
- Apple Vision Pro zbiera gaze 24/7 podczas użytkowania
- Tobii: polityka 'gaze data never leaves device' — audytowalna?
- ❓ GDPR Art. 9: dane biometryczne = kategoria szczególna

FACS / Rozpozn. Emocji

- Pracodawcy używają FACS w rozmowach rekrutacyjnych (HireVue)
- Emotion AI może profilować bez wiedzy i zgody użytkownika
- Problem międzykulturowy: ten sam wyraz twarzy ≠ ta sama emocja
- EU AI Act (2024): zakaz real-time biometric surveillance w przestrzeni publ.
- Zakaz klasyfikacji emocji w miejscach pracy i szkołach (Annex III)
- ❓ Co z systemami antyzmęczeniowymi dla kierowców?

GDPR + AI Act — Ramy Prawne

- GDPR Art. 4(14): dane biometryczne = dane osobowe szczególnej kategorii
- Przetwarzanie wymaga: wyraźnej zgody LUB ważnego interesu publ.
- Prawo do usunięcia: 'zapomnij mój głos/gesty/EEG'
- EU AI Act (sierpień 2024): wysokie ryzyko → audyt, przejrzystość
- Właściciel danych EEG: de iure użytkownik, de facto - kto ma serwer
- ❓ Open question: czy myśl to dane osobowe?

UX I COGNITIVE LOAD

Kiedy WIĘCEJ modalności jest GORZEJ? — zasada 'right modality for right task'

TEORIA: COGNITIVE LOAD (Sweller, 1988)

- **Intrinsic load** — złożoność zadania (stała, nie możesz zmienić)
- **Extraneous load** — nieistotne info w interfejsie (TU możesz pomagać)
- **Germane load** — obciążenie budującym się schematem poznawczym

⚠ Dodanie nowej modalności zwiększa extraneous load jeśli użytkownik musi się jej uczyć lub jeśli kanały sobie przeszkadzają (attention bottleneck).

Bolt (1980): "Put That There"

Pierwsze formalne badanie nad multimodalnością. Użytkownik łączył mowę ("Put that") z gestem wskazania ("there") → precyzja i naturalność wyższa niż każda z modalności osobno. Zasada: modalność dobieraj do charakteru zadania (wskazywanie=gest, nazewnictwo=głos).

✕ KIEDY WIĘCEJ MODALNOŚCI = GORZEJ?

Redundancja bez wartości

System potwierdza każdą akcję dźwiękiem + wibracją + komunikatem. Użytkownik ignoruje sygnały — 'alarm fatigue' (jak w ICU).

Modalności konkurują o uwagę

Pilot słucha instrukcji głosowych I czyta HUD I reaguje na haptykę — attention bottleneck, błędy krytyczne (CAST, 2017).

Brak spójności (inkongruencja)

Obraz w lewo, dźwięk dochodzi z prawej — mózg traci 20–40 ms na rekalibrację czasową (efekt McGurk). Jakość spada.

Krzywa uczenia się nowej modalności

EEG BCI wymaga tygodni treningu. Wprowadzony w trybie 'zawsze włączony' dezorientuje zamiast pomagać.

🔑 Złota zasada: "A modality should be added only when it reduces the overall cognitive load for the primary task, not when it technically can be added."