ROBOTYKA AFEKTYWNA

Wykład 5: Generowanie ekspresji, czyli sterowanie robotem

Automatyka i Robotyka, studia II stopnia

dr inż. Mateusz Pomianek • Politechnika Rzeszowska

Po tym wykładzie będziesz umiał/a:

- Zaprojektować mapowanie wektora (valence, arousal) na parametry ruchu robota
- Zastosować podstawy Laban Movement Analysis do projektowania gestów
- Sparametryzować prozodię głosu (pitch, tempo, energia) w syntezatorze mowy
- Wybrać między lokalnym syntezatorem mowy a API chmurowym pod kątem budżetu latencji z W1
- Zapewnić spójność multimodalną i bezpieczeństwo kinematyczne ekspresji



SZYBKIE PRZYWOŁANIE Z WYKŁADÓW 2, 3 I 4

Trzy wykłady percepcji dały jedno: wektor (valence, arousal, confidence). Co dalej?



Dziś etapy 4 i 5: decyzja i aktuacja. Robot wreszcie coś robi z tym, co wykrył.

Rozpoznać to nie to samo, co zareagować

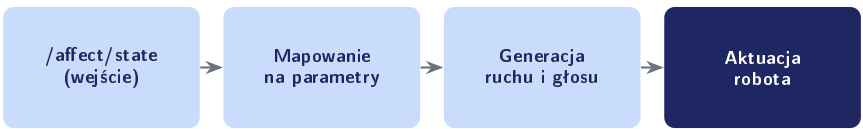
Robot poprawnie wykrył, że operator jest sfrustrowany: niska valence, wysoki arousal.

Nadal jednak wykonuje ruch tą samą, szybką i gwałtowną trajektorią co zawsze.

Doskonała percepcja bez odpowiedzi behawioralnej jest bezużyteczna, a czasem gorsza niż brak percepcji w ogóle.



Pipeline ekspresji, czyli etap 5 z W1



- Wejście: ten sam wektor (valence, arousal, confidence) niezależnie od źródła (wykłady 2 do 4)
- Mapowanie: funkcje przeliczające wektor na parametry ruchu i głosu
- Generacja: konkretna trajektoria ruchu oraz zsyntetyzowana mowa
- Aktuacja: wykonanie przez sterowniki robota (MoveIt2 Servo, silnik syntezy mowy)

Mapowanie afektu na parametry ruchu

- Arousal wpływa na prędkość i amplitudę: wysokie pobudzenie oznacza szybsze i większe ruchy
- Valence wpływa na jakość ruchu: wartość dodatnia daje ruch płynny i otwartą postawę, wartość ujemna daje ruch szarpany i postawę zamkniętą
- Confidence moduluje siłę efektu: niska pewność oznacza ekspresję stonowaną, bliską neutralnej
- To odwrócenie FACS z W2: tam obraz prowadził do emocji, tutaj emocja prowadzi do ruchu

Matematyka: skalowanie prędkości i amplitudy

$$v = \mathrm{clip}\big(v_0 \cdot (1 + k_a \cdot \mathrm{arousal}),\ v_{\min},\ v_{\max}\big)$$

v_0 to prędkość neutralna, k_a to czułość na arousal, a clip to ograniczenie do bezpiecznego zakresu (patrz slajd 20).

Przykład liczbowy: $v_0 = 0{,}3$ m/s, $k_a = 0{,}8$, arousal $= 0{,}7$, więc $v = 0{,}3 \cdot 1{,}56 \approx 0{,}47$ m/s.

Amplituda gestu skalowana analogicznie:

$$A = \mathrm{clip}\big(A_0 \cdot (1 + k_A \cdot |\mathrm{arousal}|),\ A_{\min},\ A_{\max}\big)$$

Funkcja clip to nie tylko wygoda numeryczna, to bramka bezpieczeństwa, której nigdy nie omijamy, niezależnie od intensywności wykrytej emocji.

Ekspresja przez ciało a ekspresja przez mimikę

Platforma	Kanał ekspresji	Ograniczenie
Unitree G1 EDU	Poza ciała, gest, chód, mowa syntezowana	Brak twarzy, cała ekspresja przez ruch i głos
Pepper	Ekran, oczy LED, gest ramion, mowa syntezowana	Ograniczona mimika, brak precyzyjnej manipulacji
Ameca	Silikonowa twarz, mikroekspresje	Statyczny tors, brak lokomocji i chwytu

Dla G1 EDU, który nie ma twarzy, cały ładunek ekspresji niesie poza ciała i głos. Stąd nacisk tego wykładu na kinematykę i prozodię, a nie na animację twarzy.

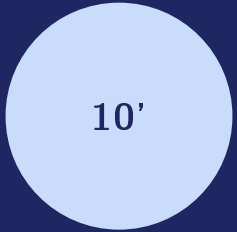
Laban Movement Analysis, czyli jakości ruchu

- **Przestrzeń**: trajektoria bezpośrednia (direct) albo elastyczna (indirect)
- **Czas**: ruch nagły (sudden) albo rozciągnięty w czasie (sustained)
- **Waga**: charakter silny i zdecydowany (strong) albo lekki (light)
- **Przepływ**: przebieg swobodny (free) albo kontrolowany (bound)

SZYBKIE PRZYWOŁANIE

Jakimi trzema jakościami LMA opisałbyś ruch robota wyrażającego radość?

30 sekund indywidualnie, potem porównaj odpowiedź z sąsiadem



PRZERWA

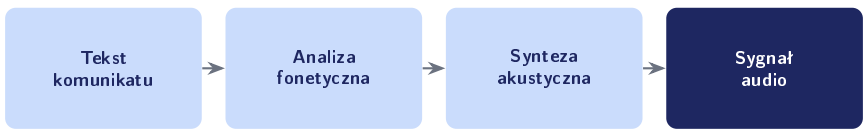
10 minut, wracamy o godzinie [uzupełnić]

Szybkie przypomnienie

- Arousal steruje prędkością i amplitudą, valence steruje jakością ruchu (płynność, otwartość)
- Laban Movement Analysis daje słownik do projektowania ruchu bez twarzy, jak u G1 EDU

Teraz zajmiemy się tym, jak ten sam wektor emocji ma brzmieć: synteza mowy afektywnej.

Synteza mowy: pipeline



Parametry prozodii, omówione na kolejnych dwóch slajdach, wstrzykuje się na etapie syntezy akustycznej. Ten sam tekst może wtedy zabrzmieć spokojnie albo w napięciu.

Prozodia: pitch, tempo, energia

- **Pitch (wysokość głosu)**: podniesiony przy pobudzeniu lub radości, spłaszczony przy smutku, mierzony w semitonach względem bazy
- **Tempo (prędkość mowy)**: przyspieszone przy wysokim arousal, zwolnione przy spokoju lub smutku
- **Energia (głośność)**: wyższa przy złości lub entuzjazmie, niższa przy zrezygnowaniu

Matematyka: od valence i arousal do prozodii

$$\text{pitch_shift} = p_0 + k_p \cdot \text{arousal} \quad \text{[semitony]}$$

$$\text{rate} = r_0 \cdot (1 + k_r \cdot \text{arousal}) \quad \text{[wielokrotność tempa bazowego]}$$

$$\text{energy}_{dB} = e_0 + k_e \cdot \text{arousal} + k_v \cdot \text{valence} \quad \text{[dB]}$$

p_0, r_0, e_0 to wartości neutralne, czyli baza głosu, a współczynniki k to czułości dobierane empirycznie i ograniczane funkcją clip jak na slajdzie 7.

Valence wpływa głównie na energię i barwę, natomiast arousal wpływa na pitch i tempo. To uproszczony, liniowy model. Rzeczywiste zależności są nieliniowe, ale liniowy punkt startowy dobrze sprawdza się w prototypie.

Synteza lokalna a synteza w chmurze

Silnik lokalny

Otwarte oprogramowanie, działa offline na sprzęcie edge, na przykład na module z W1.

Pełna kontrola parametrów prozodii i niska, przewidywalna latencja, mieszcząca się w budżecie z W1.

API chmurowe

Bardzo wysoka jakość i naturalność głosu, gotowe głosy emocjonalne.

Wymaga sieci: opóźnienie rzędu 100 do 300 milisekund lub więcej, co często przekracza

Kod: węzeł affect_expression sterujący ruchem

```
V_BASE, K_A = 0.3, 0.8
V_MIN, V_MAX = 0.05, 0.6

def on_affect(self, msg: AffectState):
    v = V_BASE * (1 + K_A * msg.arousal)
    v = max(V_MIN, min(v, V_MAX)) # clip, patrz slajd 7
    fluid = msg.valence >= 0
    self.set_arm_velocity(v, smooth=fluid)
    self.set_gesture_amplitude(
        base=0.4, k=0.6, arousal=msg.arousal)
```

Kod: synteza mowy z parametryzacją prozodii

```
from TTS.api import TTS
tts = TTS('tts_models/pl/mai/vits')

def speak_with_affect(self, text, msg: AffectState):
    pitch = 1.0 + 0.4 * msg.arousal
    rate = 1.0 + 0.3 * msg.arousal
    energy = 1.0 + 0.2*msg.arousal + 0.15*msg.valence
    wav = tts.tts(text, pitch=pitch,
                  speed=rate, energy=energy)
    self.audio_out.play(wav)
```

Spójność multimodalna: ruch i głos muszą się zgadzać

- Ruch mówiący o spokoju połączony z głosem mówiącym o podekscytowaniu daje sygnał niespójny
- Niespójność multimodalna budzi nieufność i dezorientuje operatora, czyli działa odwrotnie do celu z W1
- Rozwiązanie to jedno źródło prawdy: obie ścieżki, ruch i głos, czytają ten sam temat /affect/state
- To nie jest dodatkowa reguła, lecz naturalna konsekwencja architektury z W1, czyli wspólnego interfejsu

Bezpieczeństwo: bezpieczny zakres pracy robota

- Parametry ekspresji, czyli prędkość, amplituda, pitch i tempo, zawsze przechodzą przez funkcję clip
- Limity kinematyczne robota, takie jak maksymalna prędkość stawu czy zasięg ramienia, są nadrzędne wobec emocji
- Błędna klasyfikacja emocji, na przykład fałszywie wysoki arousal, nie może wywołać niebezpiecznego ruchu
- Ekspresja to warstwa modulująca zachowanie w bezpiecznych granicach, a nie warstwa je zastępująca

Kluczowe pojęcia dzisiejszego wykładu

Mapowanie afektu na parametry ruchu	**Laban Movement Analysis**	Prozodia: pitch, tempo, energia
Silnik lokalny a API chmurowe	**Spójność multimodalna**	Bezpieczny zakres pracy robota

Quiz kontrolny

1. Robot wykrył wysoką radość: arousal 0,9, valence 0,8. Jak zmienią się prędkość i pitch?

2. Dlaczego API chmurowe do syntezy mowy bywa ryzykowne w systemie z budżetem latencji 300 milisekund?

3. Co powinna zrobić funkcja clip, gdy klasyfikator błędnie zwróci arousal równy 5,0?

Wykład 6: architektury systemów afektywnych i etyka

- Cykl percepcja, ocena i reakcja jako całościowa architektura
- Integracja z ROS2: RoboMVP jako przykład systemu produkcyjnego
- Etyka: manipulacja emocjonalna, prywatność danych afektywnych, zaufanie

Praca własna przed W6:

Zastanów się nad sytuacją, w której robot świadomy Twoich emocji mógłby budzić Twój niepokój.

Dziękuję za uwagę

Pytania?