

Kamil SKOWROŃSKI¹, Eryka PROBIERZ², Adam GAŁUSZKA¹

¹Politechnika Śląska, ²Uczelnia Techniczno-Handlowa im. Heleny Chodkowskiej

ROZPOZNAWANIE EMOCJI W ROBOTYCE SPOŁECZNEJ: MULTIMODALNA ANALIZA MOWY I TEKSTU W JĘZYKU POLSKIM

Streszczenie. Celem pracy jest opracowanie multimodalnego systemu rozpoznawania emocji opartego na algorytmach głębokiego uczenia, integrującego informacje z mimiki twarzy, sygnału mowy oraz reakcji fizjologicznych. W pracy zaprezentowano podział modułu analizy mowy na ścieżkę akustyczną i tekstową, uwzględniając specyfikę języka polskiego. Przeprowadzono ewaluację modeli LSTM z wykorzystaniem spektrogramów Mel na zbiorach anglojęzycznych oraz polskim zbiorze nEMO, uzyskując wysoką skuteczność klasyfikacji emocji.

EMOTION RECOGNITION IN SOCIAL ROBOTICS: MULTIMODAL SPEECH AND TEXT ANALYSIS IN POLISH

Summary. The aim of this study is to develop a multimodal emotion recognition system based on deep learning, integrating facial expressions, speech signals, and physiological responses. The paper presents a speech analysis module divided into acoustic and text processing tracks, with focus on the Polish language. LSTM models utilizing Mel spectrograms were evaluated on English datasets and the Polish nEMO corpus, demonstrating high emotion classification accuracy.

1. Wprowadzenie

Wiodącym celem robotyki społecznej jest stworzenie maszyn zdolnych do naturalnej interakcji z człowiekiem, co opiera się na zdolności systemu do poprawnego rozumienia reakcji emocjonalnych użytkownika. Najbardziej efektywnym rozwiązaniem tego problemu są systemy multimodalne, które integrują dane z różnych kanałów – takich jak mowa, tekst i mimika twarzy. Wyzwanie to staje się szczególnie istotne w przypadku języka polskiego, gdzie złożoność gramatyczna oraz specyfika fonetyczna wymagają zastosowania dedykowanych i precyzyjnych metod rozpoznawania emocji [11]. Referat został opracowany na podstawie pracy pt. „Polish Speech and Text Emotion Recognition in a Multimodal Emotion Analysis System” [12] (DOI: 10.3390/app142210284).

2. Powiązane prace

2.1. Robotyka społeczna, interakcja człowiek-robot, rozpoznawanie emocji

Dynamiczny rozwój robotyki społecznej sprawia, że maszyny coraz częściej wspierają ludzi w takich dziedzinach jak medycyna, edukacja czy opieka nad seniorami. Aby te interakcje były skuteczne, kluczowe staje się doskonalenie systemów interak-

cji człowiek-robot (ang. *Human-Robot Interaction* - HRI), które dążą do zapewnienia naturalnej komunikacji opartej na wzajemnym zrozumieniu. Potrzeba naturalnej komunikacji w ramach HRI bezpośrednio wymusza na systemach robotycznych zdolność do rozpoznawania stanów emocjonalnych użytkownika na podstawie tonu głosu oraz treści wypowiedzi. Wykorzystanie metod przetwarzania języka naturalnego (ang. *natural language processing* - NLP) do analizy semantyki tekstu oraz systemów rozpoznawania emocji z mowy (ang. *Speech Emotion Recognition* - SER) do badania wzorców mowy pozwala robotom na adekwatną reakcję i dostosowanie się do potrzeb człowieka [14].

2.2. Spektrogramy Mel

Złożoność ludzkiego głosu sprawia, że aby skutecznie zrealizować zadanie rozpoznawania emocji z mowy, konieczne jest zastosowanie spektrogramów Mel, które przekształcają dźwięk w obrazową reprezentację częstotliwości bliską ludzkiej percepcji. Dzięki takiej transformacji sygnału audio możliwe staje się wykorzystanie zaawansowanych sieci neuronowych, które w oparciu o te wizualne wzorce precyzyjnie identyfikują subtelne emocje zawarte w tonie, tempie i barwie głosu [9].

3. Multimodalny system analizy emocji

Architektura systemu multimodalnego została zaprojektowana w ścisłym powiązaniu z teorią HRI jako interakcja między człowiekiem a jednostką robotyczną, co czyni go użytecznym narzędziem badawczym. Warstwa sprzętowa integruje sensor Aidmed do monitorowania parametrów fizjologicznych oraz interaktywne głowy robotyczne Furhat lub OhBot, a przetwarzanie danych realizowane jest przez algorytmy sieci neuronowych, które dzięki implementacji w środowisku Python mogą działać elastycznie zarówno w chmurze, jak i na jednostkach lokalnych. Kluczowym elementem jest dwutorowy moduł audio, który dzięki równoległej analizie transkrypcji tekstowej oraz cech fonetycznych zawartych w spektrogramach Mel pozwala na trafną interpretację stanów afektywnych nawet w sytuacjach, gdy treść wypowiedzi jest emocjonalnie neutralna.

4. Zbiory danych do rozpoznawania emocji mowy

Wybór zbiorów danych oparto o teorię Paula Ekmana dotyczącą podstawowej palety emocji, uwzględniając zarówno zasoby w globalnym języku angielskim, jak i specyficzne dla celów badawczych zbiory polskojęzyczne. Najważniejsze informacje o zbiorach danych zebrano w tabeli 1. W dalszych analizach zachowano nazwy emocji w oryginalnej wersji angielskiej, zgodnie z tym, jak występują one w zbiorach.

4.1. Zbiory danych w języku angielskim

- RAVDESS (ang. *Ryerson Audio-Visual Database of Emotional Speech and Song*) [8] - zbiór ten zawiera unikalne nagrania 24 aktorów, którzy wypowiadali i wyśpiewywali te same frazy z różnym stopniem intensywności emocjonalnej.
- TESS (ang. *The Toronto Emotional Speech Set*) [10] - nagrania w zbiorze zostały przygotowane przez dwie aktorki w różnym wieku, co umożliwia analizę wpływu procesów starzenia się na percepcję i akustyczną charakterystykę emocji w mowie.
- CREMA-D (ang. *Crowd-sourced Emotional Multimodal Actors Dataset*) [5] - zbiór danych wyróżnia się dużą różnorodnością, obejmując materiały od 91 aktorów o róż-

nicowanym pochodzeniu etnicznym i różnych akcentach.

- IEMOCAP (ang. *Interactive Emotional Dyadic Motion Capture*) [4] - zbiór ten oprócz nagrań mowy, charakteryzuje się wykorzystaniem technologii motion capture do rejestracji ruchów ciała i gestów podczas interakcji między parami aktorów.
- SAVEE (ang. *Surrey Audio-Visual Expressed Emotion*) [7] - dokumentacja tego zbioru podkreśla, że nagrania czterech native speakerów płci męskiej powstały w warunkach laboratoryjnych, zapewniając najwyższą jakość sygnału audio i video.

4.2. Zbiory danych w języku polskim

- Chatbot [13] - autorski zbiór tekstowy został opracowany z wykorzystaniem modelu GPT-3.5, gdzie dane treningowe wygenerowano sztucznie, a zestaw testowy został etykietowany przez ludzi w celu sprawdzenia użyteczności AI w generowaniu treści.
- nEMO [6] - jest to nowoczesny polski zbiór danych, który oferuje symulowane nagrania o ujednoliconej głośności wraz z pełnymi transkrypcjami, wspierając jednocześnie klasyfikację audio oraz systemy automatycznego rozpoznawania mowy.

Tabela 1

Porównanie parametrów technicznych zbiorów danych dla problemu rozpoznawania emocji mowy w języku polskim.

Zbiór danych	Język	Emocje	Ścieżka audio / Transkrypcja	Liczba próbek	Parametry techniczne
RAVDESS	angielski	angry, calm, disgust, fearful, happy, sad, surprise and neutral	✓/×	1440	4s, 48kHz, 16bit/sample
TESS	angielski	anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral	✓/✓	2800	brak informacji
CREMA-D	angielski	anger, disgust, fear, happy, sad and neutral	✓/×	7442	brak informacji
IEMOCAP	angielski	angry, disappointed, excited, fear, frustrated, happy, sad, surprised and neutral	✓/✓	1150	brak informacji
SAVEE	angielski	anger, disgust, fear, happiness, sadness, surprise and neutral	✓/×	480	44.1kHz
Chatbot	polski	anger, disgust, fear, happiness, sadness, surprise and neutral	×/✓	8458	-
nEMO	polski	anger, fear, happiness, sadness, surprise and neutral	✓/✓	4481	2.5s, 24kHz, 16bit/sample

5. Algorytmy uczenia maszynowego

5.1. Modele rozpoznawania emocji mowy

Problem rozpoznawania emocji w mowie za pomocą spektrogramów Mel opiera się na modelach klasyfikacyjnych. Poniżej przedstawiono popularne sieci neuronowe do przetwarzania obrazu, które zostały wytrenowane w oparciu o zestawy przedstawione wcześniej. Porównanie skuteczności tych modeli przedstawiono w tabeli 2.

- Konwolucyjne sieci neuronowe (ang. *Convolutional Neural Networks* - CNN) - architektury te wykorzystują warstwy splotowe do automatycznego wychwytywania wzorców przestrzennych i hierarchii cech ze spektrogramów Mel [3].
- ResNet (ang. *Residual Networks*) - ta specyficzna odmiana sieci CNN rozwiązuje problem zanikającego gradientu w bardzo głębokich modelach dzięki zastosowaniu połączeń typu „skip connections”, co pozwala na efektywne trenowanie sieci o dużej

liczbie warstw (np. ResNet 50) i lepszą ekstrakcję złożonych cech [2].

- Stacked Time Distributed 2D CNN-LSTM - model ten integruje sekwencyjnie ułożone warstwy konwolucyjne 2D z warstwą LSTM (ang. *Long Short-Term Memory*), co umożliwia skuteczne uczenie się cech przestrzennych z obrazów przy jednoczesnym zachowaniu i analizie zależności czasowych występujących w sygnale audio [1].
- Stacked Time Distributed 2D CNN-Bidirectional LSTM with Attention - architektura ta rozszerza model o dwukierunkową warstwę LSTM oraz mechanizm uwagi, co zwiększa zdolność systemu do uwzględniania kontekstu płynącego zarówno z poprzednich, jak i przyszłych ramek danych w sekwencji wejściowej [1].
- Parallel 2D CNN-Bidirectional LSTM with Attention - w tym podejściu równoległe sieci CNN symultanicznie ekstrahują cechy przestrzenne z wielu kanałów wejściowych, przesyłając je do dwukierunkowej sieci LSTM, która dzięki mechanizmowi uwagi koncentruje się na najbardziej kluczowych informacjach wypowiedzi [1].
- Parallel 2D CNN-Transformer Encoder - model ten wykorzystuje równoległe sieci CNN do wydobywania cech, które są następnie przetwarzane przez enkoder Transformer z mechanizmem self-attention, co zapewnia najgłębsze zrozumienie skomplikowanych wzorców sekwencyjnych [1].

Tabela 2

Porównanie modeli dla problemu rozpoznawania emocji z wykorzystaniem spektrogramów Mel.

Model klasyfikacji	Zbiory treningowe	Liczba kategorii	Dokładność
Convolutional Neural Network	CREMA-D, RAVDESS, SAVEE, TESS	7	71.00%
ResNet 18	CREMA-D,	6	67.62%
ResNet 34	IEMOCAP,	6	68.15%
ResNet 50	RAVDESS,	4	75.76%
ResNet 50	SAVEE,	5	68.86%
ResNet 50	TESS	6	62.18%
Stacked Time Distributed 2D CNN - LSTM	RAVDESS	8	90.00%
Stacked Time Distributed 2D CNN - Bidirectional LSTM with attention		8	94.02%
Parallel 2D CNN - Bidirectional LSTM with attention		8	96.55%
Parallel 2D CNN - Transformer Encoder		8	96.78%

5.2. Model rozpoznawania emocji w tekście

Model rozpoznawania emocji w tekście w języku polskim wykorzystuje rekurencyjne sieci neuronowe LSTM oraz osadzenia słowne Word2Vec wytrenowane na Narodowym Korpusie Języka Polskiego i polskiej Wikipedii. Algorytm ten cechuje się wysoką precyzją, osiągając około 96% dokładności na danych generycznych oraz 87% w przypadku weryfikacji na danych etykietowanych przez ludzi [13].

6. Wyniki badań

Badania oparto na czterech filarach:

- Uzasadnienie multimodalności - sprawdzenie, czy sama analiza tekstu wystarczy do poprawnego rozpoznania emocji.
- Weryfikacja modeli - ponowne wytrenowanie wybranych architektur LSTM na zbiorze RAVDESS w celu potwierdzenia ich skuteczności.

- Adaptacja do języka polskiego - przetworzenie polskiego zbioru nEMO i ocena precyzji modeli na danych polskojęzycznych.
- Test uniwersalności - porównanie krzyżowe modeli między różnymi językami.

6.1. Testowanie zbioru nEMO na modelu rozpoznawania emocji w tekście

Badania rozpoczęto od weryfikacji hipotezy, że sama analiza treści tekstowej jest niewystarczająca w sytuacjach, gdy emocje są wyrażane głównie poprzez ton głosu. Wykorzystując model Chatbot na zbiorze nEMO, w którym wypowiedzi są semantycznie neutralne, system sklasyfikował 80% próbek jako neutralne, nie wykrywając zabarwienia emocjonalnego zawartego w ścieżce dźwiękowej. Wynik ten stał się bezpośrednim impulsem do opracowania dwutorowego modułu analizującego fonetykę wypowiedzi.

6.2. Modele rozpoznawania emocji w mowie oparte na zbiorze danych RAVDESS

W odpowiedzi na potrzebę analizy dźwięku, do dalszych testów wybrano cztery architektury LSTM, które w literaturze wykazywały najwyższą skuteczność. Ponowne wytrenowanie tych modeli na standardowym angielskim zbiorze RAVDESS wykazało jednak niższą dokładność niż deklarowana przez autorów pierwotnych rozwiązań, osiągając maksymalnie 65% dla modelu Parallel 2D CNN-Bidirectional LSTM. Mimo tych różnic, etap ten pozwolił na stabilizację parametrów technicznych modeli przed ich kluczową aplikacją dla języka polskiego. Wyniki ewaluacyjne przedstawiono w tabeli 3, natomiast macierze pomyłek pokazano na rysunku 1.

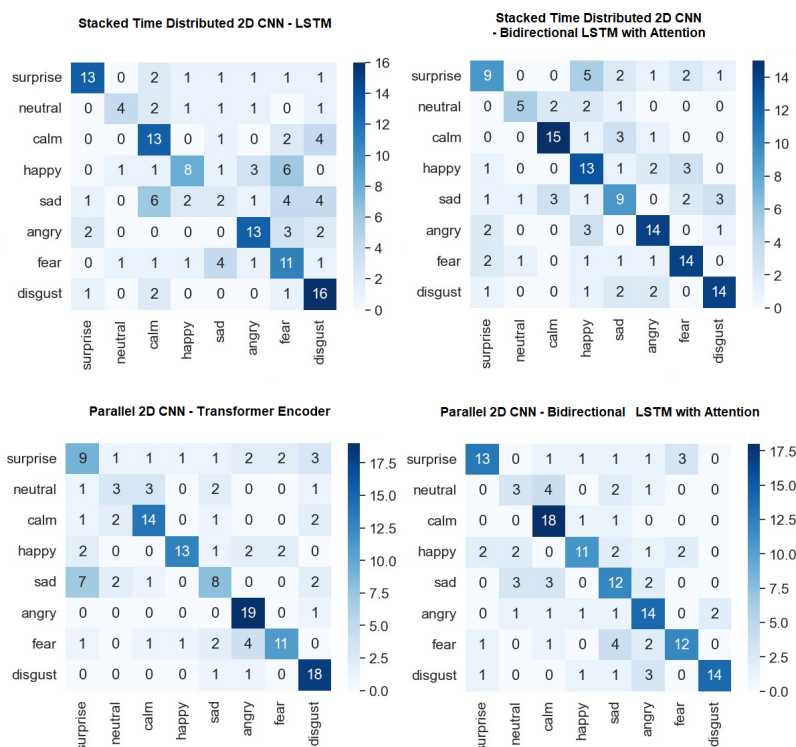
Tabela 3

Ewaluacja modeli rozpoznawania emocji w mowie przy użyciu spektrogramów Mel na zbiorze danych RAVDESS.

Modele	Emocje								
	Average	Surprise	Neutral	Calm	Happy	Sad	Angry	Fear	Disgust
Stacked Time Distributed 2D CNN - LSTM									
Dokładność [%]	53.00	65.00	40.00	65.00	40.00	10.00	65.00	55.00	80.00
Precyzja [%]	54.00	76.47	66.67	48.15	61.54	20.00	65.00	39.29	55.17
Czułość [%]	52.50	65.00	40.00	65.00	40.00	10.00	65.00	55.00	80.00
F1-Score [%]	51.70	70.27	50.00	55.32	48.48	13.33	65.00	45.83	65.31
Stacked Time Distributed 2D CNN - Bidirectional LSTM with attention									
Dokładność [%]	62.00	45.00	50.00	75.00	65.00	45.00	70.00	70.00	70.00
Precyzja [%]	63.20	56.25	71.43	75.00	48.15	47.37	66.67	66.67	73.68
Czułość [%]	61.20	45.00	50.00	75.00	65.00	45.00	70.00	70.00	70.00
F1-Score [%]	61.70	50.00	58.82	75.00	55.32	46.15	68.29	68.29	71.79
Parallel 2D CNN - Transformer Encoder									
Dokładność [%]	63.00	45.00	30.00	70.00	65.00	40.00	95.00	55.00	90.00
Precyzja [%]	61.90	42.86	37.50	70.00	86.67	50.00	67.86	73.33	66.67
Czułość [%]	61.30	45.00	30.00	70.00	65.00	40.00	95.00	55.00	90.00
F1-Score [%]	60.60	43.90	33.33	70.00	74.29	44.44	79.17	62.86	76.60
Parallel 2D CNN - Bidirectional LSTM with attention									
Dokładność [%]	65.00	65.00	30.00	90.00	55.00	60.00	70.00	60.00	70.00
Precyzja [%]	64.20	76.47	33.33	64.29	73.33	50.00	58.33	70.59	87.50
Czułość [%]	62.50	65.00	30.00	90.00	55.00	60.00	70.00	60.00	70.00
F1-Score [%]	62.60	70.27	31.58	75.00	62.86	54.55	63.64	64.86	77.78

6.3. Modele rozpoznawania emocji w mowie oparte na zbiorze danych nEMO

Wypracowane wcześniej architektury zastosowano następnie do analizy polskiego zbioru nEMO, co przyniosło znaczącą poprawę wyników względem danych angielskich.



Rys. 1. Macierze pomyłek dla modeli w oparciu o zbiór danych RAVDESS.

skich. Najwyższą skuteczność 84,6% uzyskał model Parallel 2D CNN-Transformer Encoder, który szczególnie dobrze poradził sobie z rozpoznawaniem radości oraz stanu neutralnego. Sukces ten pozwolił na wyłonienie optymalnego algorytmu, który stanowi fundament modułu afektywnego w projektowanym robocie społecznym. Wyniki ewaluacji przedstawiono w tabeli 4, natomiast macierze pomyłek pokazano na rysunku 2.

6.4. Porównanie modeli rozpoznawania emocji w mowie i testy uniwersalności

Ostatnim etapem było sprawdzenie uniwersalności wyłonionych modeli poprzez testy krzyżowe, w których modele trenowane na jednym języku testowano na drugim. Wyniki spadające poniżej 30% skuteczności jednoznacznie wykazały, że różnice w tempie, intonacji i rytmie mowy między językiem polskim a angielskim uniemożliwiają bezpośrednie przenoszenie modeli.

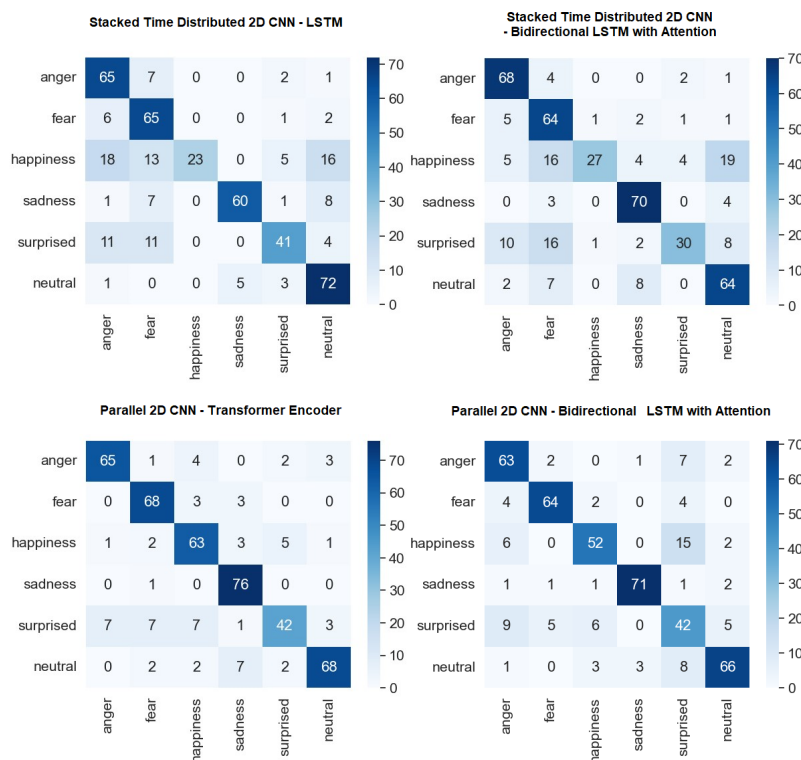
7. Wnioski

Przeprowadzone badania dowiodły, że skuteczny system rozpoznawania emocji dla języka polskiego wymaga zastosowania podejścia multimodalnego z dwutorowym modułem audio, ponieważ sama analiza semantyczna tekstu okazuje się niewystarczająca w sytuacjach, gdy komunikat ma neutralną treść, a stan emocjonalny wyrażony jest wyłącznie poprzez ton i barwę głosu. Spośród poddanych ewaluacji architektur głębokiego uczenia najwyższą skuteczność na polskojęzycznym zbiorze danych nEMO wykazał model Parallel 2D CNN-Transformer Encoder, osiągając bardzo satysfakcjonującą dokładność na poziomie 84,6%. Jednocześnie testy krzyżowe między bazami polskimi i angielskimi wykazały spadek efektywności, co potwierdza, że różnice kulturowe oraz specyficzne cechy fonetyczne uniemożliwiają obecnie stworzenie jednego uniwersalnego modelu i wymagają stosowania rozwiązań dedykowanych

Tabela 4

Ewaluacja modeli rozpoznawania emocji w mowie przy użyciu spektrogramów Mel na zbiorze danych nEMO.

Modele	Emocje						
	Average	Anger	Fear	Happiness	Sadness	Surprised	Neutral
Stacked Time Distributed 2D CNN - LSTM							
Dokładność [%]	72.20	86.67	87.84	30.67	77.92	61.19	88.89
Precyzja [%]	77.70	63.73	63.11	100.00	92.31	77.36	69.90
Czułość [%]	72.20	86.67	87.84	30.67	77.92	61.19	88.89
F1-Score [%]	70.80	73.45	73.45	46.94	84.51	68.33	78.26
Stacked Time Distributed 2D CNN - Bidirectional LSTM with attention							
Dokładność [%]	71.30	90.67	86.49	36.00	90.91	44.78	79.01
Precyzja [%]	75.90	75.56	58.18	93.10	81.40	81.08	65.98
Czułość [%]	71.30	90.67	86.49	36.00	90.91	44.78	79.01
F1-Score [%]	69.90	82.42	69.57	51.92	85.89	57.69	71.91
Parallel 2D CNN - Transformer Encoder							
Dokładność [%]	84.60	86.67	91.89	84.00	98.70	62.69	83.95
Precyzja [%]	85.00	89.04	83.95	79.75	84.44	82.35	90.67
Czułość [%]	84.60	86.67	91.89	84.00	98.70	62.69	83.95
F1-Score [%]	84.50	87.84	87.74	81.82	91.02	71.19	87.18
Parallel 2D CNN - Bidirectional LSTM with attention							
Dokładność [%]	79.40	84.00	86.49	69.33	92.21	62.69	81.48
Precyzja [%]	80.00	75.00	88.89	81.25	94.67	54.55	85.71
Czułość [%]	79.40	84.00	86.49	69.33	92.21	62.69	81.48
F1-Score [%]	79.50	79.25	87.67	74.82	93.42	58.33	83.54



Rys. 2. Macierze pomyłek dla modeli analizy emocji w mowie na podstawie zbioru danych nEMO dla języka polskiego.

dla konkretnego obszaru językowego. W ramach przyszłych planów rozwojowych przewiduje się implementację nadrzędnego modułu decyzyjnego dla całego systemu

multimodalnego oraz rozbudowę autorskiego zbioru danych o nagrania pochodzące z rzeczywistych interakcji.

Prace wsparta Politechnika Śląska w Gliwicach poprzez dotację na utrzymanie i rozwój potencjału badawczego młodych naukowców w 2026 roku.

LITERATURA

1. Speech emotion classification with pytorch. <https://github.com/Data-Science-kosta/Speech-Emotion-Classification-with-PyTorch/>. Accessed: 2026-06-25.
2. Speech emotion recognition. <https://github.com/KanikeSaiPrakash/Speech-Emotion-Recognition>. Accessed: 2026-06-25.
3. Speech emotion recognition using cnns and crnns based on mel spectrograms and mel frequency cepstral coefficients. <https://neurobyte.org/guides/speech-emotion-recognition-cnns-crnns-tensorflow/>. Accessed: 2026-06-25.
4. Busso C., Bulut M., Lee C.-C., Kazemzadeh A., Mower E., Kim S., Chang J. N., Lee S., Narayanan S. S.: Iemocap: Interactive emotional dyadic motion capture database, Language resources and evaluation, 2008, .
5. Cao H., Cooper D. G., Keutmann M. K., Gur R. C., Nenkova A., Verma R.: Crema-d: Crowd-sourced emotional multimodal actors dataset, IEEE transactions on affective computing, 2014, .
6. Christop I. nemo: Dataset of emotional speech in polish, arxiv preprint arxiv:2404.06292, 2024.
7. Jackson P., Haq S.: Surrey audio-visual expressed emotion (savee) database, University of Surrey: Guildford, UK, 2014, .
8. Livingstone S. R., Russo F. A.: The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english, PloS one, 2018, .
9. Meng H., Yan T., Yuan F., Wei H.: Speech emotion recognition from 3d log-mel spectrograms with deep learning network, IEEE access, 2019, .
10. Pichora-Fuller M. K., Dupuis K.: Toronto emotional speech set (TESS) 2020.
11. Powroźnik P. Polish emotional speech recognition using artificial neural network, Advances in Science and Technology. Research Journal, 2014, .
12. Skowroński K., Gałuszka A., Probierz E.: Polish speech and text emotion recognition in a multimodal emotion analysis system, Applied Sciences, 2024, .
13. Skowroński K. Application of data-collecting chatbot for polish text emotion analysis, In: Modelling and simulation 2023. The 2023 European Simulation and Modelling Conference, October 24-26, 2023, Toulouse, France. EUROSIS-ETI, 2023, p. 161–165.
14. Yoon S., Byun S., Jung K.: Multimodal speech emotion recognition using audio and text, In: 2018 IEEE spoken language technology workshop (SLT). IEEE, 2018, p. 112–118.