PolEval 2025 Task 4: Polish Speech Emotion Recognition Challenge

Iwona Christop **Maciej Czajka**
Adam Mickiewicz University
ul. Uniwersytetu Poznańskiego 4
61-614 Poznań, Poland
{iwona.christop, maciej.czajka}@amu.edu.pl

Abstract

This paper introduces the Polish Speech Emotion Recognition Challenge, a shared task aimed at advancing research on cross-lingual emotion recognition in low-resource languages. The challenge's objective was to develop systems that could recognize emotional states in Polish speech using only multilingual training data, with no access to Polish training examples. The final test set consisted of newly recorded Polish speech samples created specifically for the challenge, ensuring a fully blind evaluation. Participants submitted emotion predictions for six target classes. System performance was assessed using the macro-averaged F1 score as the primary metric.

1 Introduction

Speech emotion recognition (SER) is a growing area of research focused on identification of human emotional states from speech signals. It relies not only on the lexical content of utterances, but also on prosodic and acoustic cues, such as intonation, pitch, and rhythm.

This capability plays an essential role in a wide range of real-world applications, including human-computer interaction, conversational agents, and accessibility technologies. Despite its importance, SER remains challenging due to cross-speaker variability, culturally dependent emotional expression, and the lack of large, balanced datasets, particularly for low-resource languages.

The main objective of the Polish SER Challenge was to promote research on cross-lingual emotion recognition by developing systems able to recognize emotions in Polish speech without using any Polish training data. Participants were asked to classify each audio recording into one of six predefined emotional states. The focus on Polish was motivated by the scarcity of labeled emotional speech resources in this language and the practical need for robust multilingual SER systems.

The corpus prepared for this task was based on the large multilingual CAMEO dataset (Christop and Czajka, 2025), which aggregates multiple established emotional speech corpora and is publicly available through the Hugging Face platform. The training set consisted exclusively of emotional speech recordings in seven non-Polish languages. Polish speech samples used for validation came from the nEMO dataset (Christop, 2024), which was integrated into CAMEO. Importantly, the test data consisted of unseen audio samples recorded specifically for this challenge. These recordings followed the same labeling scheme and comparable recording conditions but had not been released publicly. The test set labels were withheld throughout the competition, ensuring that the final evaluation was conducted on completely unseen material and represented as a genuine cross-lingual assessment scenario.

2 Task Definition

The Polish SER Challenge required participants to build systems capable of predicting emotion labels for Polish speech samples from the hidden test set. The training was restricted to the provided multilingual training data. No Polish speech samples could be used for training or data augmentation, including the validation split, which was provided strictly for evaluation purposes. Manual annotations of the test set, semi-annotated labeling, crowdsourcing, or any indirect form of labeling was strictly forbidden. It was also prohibited to use external datasets or resources not included in the official training data. Pretrained models and transfer learning approaches were allowed only if they had not been trained or fine-tuned on Polish data or specifically on the nEMO dataset. These constraints ensured a fair comparison across systems and a true simulation of a low-resource zero-shot learning scenario.

Proceedings of the PolEval 2025 Workshop, pages 72–81
November 21, 2025 ©2025 Association for Computational Linguistics

Split	**anger**	**fear**	**happiness**	**neutral**	**sadness**	**surprise**	**Total**
`train`	5 212	4 241	5 216	7 161	5 127	2 757	29 714
`dev`	749	736	749	809	769	669	4 481
`test-A`	276	267	246	271	268	255	1 583
`test-B`	258	240	271	271	267	266	1 573
Total	6 495	5 484	6 482	8 512	6 431	3 947	37 351

Table 1: Distribution of samples per emotional state across train, validation and test splits.

Dataset	**Language**	**anger**	**fear**	**happiness**	**neutral**	**sadness**	**surprise**	**Total**
CaFE	French	144	144	144	72	144	144	792
CREMA-D	English	1 271	1 271	1 271	1 087	1 271	–	6 171
EMNS	English	133	–	158	149	150	153	743
Emozionalmente	Italian	986	986	986	986	986	986	5 916
eNTERFACE	English	210	210	207	–	210	210	1 047
JL-Corpus	English	240	–	240	240	240	–	960
MESD	Spanish	143	144	144	143	144	–	718
Oréau	French	73	71	72	71	72	72	431
PAVOQUE	German	601	–	584	3 126	556	–	4 867
RAVDESS	English	192	192	192	96	192	192	1 056
RESD	Russian	219	223	218	191	162	–	1 013
SUBESCO	Bengali	1 000	1 000	1 000	1 000	1 000	1 000	6 000

Table 2: Distribution of samples per emotional state in train set.

3 Dataset

The dataset for the Polish SER Challenge was split into three parts – training, validation, and test. Table 1 shows the distribution of samples per emotional state across the splits.

3.1 Train Set

The training split consisted of 29 714 speech samples from 12 multilingual datasets within CAMEO – CaFE (Gournay et al., 2018), CREMA-D (Cao et al., 2014), EMNS (Noriy et al., 2023), Emozionalmente (Catania et al., 2025), eNTERFACE (Martin et al., 2006), JL-Corpus (James et al., 2018), MESD (Duville et al., 2021a,b), Oréau (Kerkeni et al., 2020), PAVOQUE (Steiner et al., 2013), RAVDESS (Livingstone and Russo, 2018), RESD (Amentes et al., 2022), and SUBESCO (Sultana et al., 2021). Together, these covered seven non-Polish languages. All training samples were annotated using the same six emotional categories – anger, fear, happiness, sadness, surprise, and a neutral state. Table 2 shows the distribution of samples by emotional state in the training set across all datasets.

The audio recordings and metadata for train split were accessible as a part of the CAMEO collection through the Hugging Face platform[1]. Table 3 shows an overview of available metadata fields.

Field	**Description**
file_id	Unique identifier of the audio sample.
audio	File path, raw waveform, and sampling rate.
emotion	Expressed emotional state.
transcription	Orthographic transcription of the utterance.
speaker_id	Unique speaker identifier.
gender	Gender of the speaker.
age	Age of the speaker.
dataset	Dataset of origin.
language	Primary language of the sample.
license	Original dataset license.

Table 3: Overview of the metadata fields available in the CAMEO collection.

Additionally, the input files, `in.tsv`, were available in the challenge repository on GitHub platform[2]. Each line specified the CAMEO split name

[1] `https://huggingface.co/datasets/amu-cai/CAMEO`

[2] `https://github.com/poleval/2025-speech-emotion`

Speaker	Gender	Age	anger	fear	happiness	neutral	sadness	surprise	Total
SB0	M	24	68	69	65	65	69	67	403
JS0	M	25	68	59	55	67	64	55	368
MC0	M	25	70	68	70	70	70	69	417
IC0	F	30	70	70	70	70	70	70	420
KJ0	M	60	54	66	57	68	70	69	384
MC1	M	22	70	69	67	70	70	69	415
SO0	F	21	69	62	70	69	70	63	403
JW0	F	24	65	44	63	63	52	59	346
Total	3F / 5M		534	507	517	542	535	521	3 156

Table 4: Distribution of samples per emotional state across all speakers in the test set.

and the file identifier, ensuring precise mapping between the provided lists and the datasets hosted on Hugging Face, as shown in Figure 1.

```
cafe    e9d4b7b83bd1f6825dabca3fc51.flac
ravdess 4dd5629990a4931959da7735b28.flac
resd    ebcea26cf1ffffdb66eed7d7468.flac
```

Figure 1: Example of the `in.tsv` file available for train and validation splits.

3.2 Validation Set

The validation split consisted of 4 481 recordings from the Polish nEMO dataset. These samples were fully labeled but restricted to evaluation only. The validation set allowed participants to estimate cross-lingual generalization performance on Polish speech prior to submitting systems for the hidden test evaluation.

As with the training set, the audio recordings and metadata for the validation set were available on the Hugging Face platform. The input file was provided in the GitHub repository. Both the metadata and the input file had the same structure as the training set.

3.3 Test Set

The test set contained unseen Polish speech recordings that were obtained specifically for this challenge. These samples were not part of any public dataset and their labels were hidden throughout the competition. This dataset provided a controlled and unbiased evaluation of system performance on Polish emotional speech from previously unheard speakers and newly recorded material.

The test set consisted of recordings from eight speakers, including five men and three women, ranging in age from 21 to 60 years old. The test set contains a total of 3 156 utterances, which are distributed relatively evenly across the six emotion categories. Each speaker contributed a comparable number of recordings to ensure that no single speaker dominated the test data. This composition allows for a fair and robust evaluation of emotion recognition performance across speakers and emotional categories. Table 4 shows the distribution of samples by speaker across all emotional states.

A total of 70 distinct utterances were recorded to create the test set. The complete list of these sentences is provided in the Appendix A. The intended number of samples per emotional state was equal to the number of utterances (70). However, recordings that did not sufficiently represent the targeted emotional state were manually rejected and excluded from the final test set. This resulted in minor deviations from the ideal count.

```
{
    "file_id": "bb7ee27f3e.flac",
    "transcription": "Ochronię cię.",
    "speaker_id":"SB0",
    "gender":"male",
    "age":"24",
    "dataset": "test",
    "language": "Polish",
    "license": "CC BY-NC-SA 4.0"
}
```

Figure 2: Example record from the JSONL metadata files provided for test splits.

The test set was divided into two splits, `test-A` and `test-B`, comprising 1 583 and 1 573 samples, respectively. Both splits were released to participants, but the labels were hidden. `test-B` was used exclusively for the final evaluation and leaderboard ranking. The input files for both `test-A` and `test-B` contained the split name and file identifier for each sample. Additionally, JSONL metadata files were provided to ensure the same set of metadata fields was available to participants across all splits. An example metadata record is shown in Figure 2. The corresponding audio recordings for both test splits were distributed as a compressed

TAR archive via the challenge GitHub repository.

4 Evaluation

4.1 Submission Format

The evaluation was conducted based on participant submissions in the form of a TSV file. This file was supposed to contain exactly one emotion label per line, corresponding to each sample listed in the input file. Figure 3 shows an example of the output file.

```
anger
neutral
happiness
```

Figure 3: Example of an output file.

4.2 Metrics

System performance was measured using standard classification metrics – macro-averaged F1 score as the primary metric and accuracy as a supplementary metric. The macro F1 score was computed by calculating the F1 score separately for each emotion class and averaging these values without weighting, according to the Formula 1. This way, the performance across all emotional states was emphasized equally regardless of dataset imbalance.

$$\mathrm{F1}_{\mathrm{macro}} = \frac{1}{K} \sum_{i=1}^{K} \mathrm{F1}_i \qquad (1)$$

where K – number of classes, $\mathrm{F1}_i$ – F1-score for class i.

The accuracy was calculated according to Formula 2, as the fraction of correctly predicted samples across the entire test set.

$$\mathrm{Accuracy} = \frac{\sum_{i=1}^{K} \mathrm{TP}_i}{N} \qquad (2)$$

where TP_i – number of correctly predicted samples for class i, N – total number of samples.

For both metrics, values ranged from 0 to 1, with 1 being the highest possible score.

4.3 Post-processing Strategy

To assist participants in creating valid submissions, an implementation of the post-processing strategy introduced by Christop and Czajka (2025) was provided. This tool was designed primarily to normalize outputs generated by large language models or other systems that might produce descriptive responses or use different part of speech than expected.

The post-processing strategy involved tokenization the generated response and calculating the Levenshtein ratio between each target label and each word in the prediction. Similarity scores below the predefined threshold of 0.57 were filtered out for each label, and the remaining values were summed to yield an aggregated score for that label. The class with the highest aggregated similarity score was then selected as the best match. This strategy was only used if the generated response was not an exact match for any of the labels.

The usage of the post-processing strategy was optional.

4.4 Baseline

Several open source baseline systems were evaluated on the validation (`dev`) set, as well as on the two test splits (`test-A` and `test-B`), to provide reference points for this challenge. The macro-averaged F1 scores obtained by these models are shown in Table 5. In addition to a variety of audio languange models, a cascaded system comprising Whisper-Large-v3 (Radford et al., 2022) and Llama-3.3-70B-Instruct (Grattafiori et al., 2024) was evaluated. In this system, Whisper first generates a transcription of each utterance. Then, Llama 3.3, a text-only large language model, analyzes the transcription to predict the emotional label.

Model	dev	test-A	test-B
Audio Flamingo 3	0.0829	0.0768	0.0875
GAMA	0.0433	0.0486	0.0528
Qwen2-Audio	**0.1977**	**0.1500**	0.1492
Qwen-Audio-Chat	0.1444	0.1363	0.1236
Ultravox v0.6	0.1334	0.1418	**0.1576**
Whisper + Llama 3.3	0.1173	0.1283	0.1147

Table 5: F1-macro results obtained by selected open source systems on validation and test sets.

Overall, baseline performance was relatively low, reflecting the task's difficulty and the strict zero-shot, cross-lingual setting. ultravox-v0_6-llama-3_3-70b (fixie-ai, 2025) achieved the strongest performance among the evaluated systems on the final hidden test set, obtaining a macro F1 score of 0.1576. This system also exhibited consistent behavior across splits, producing comparable score on the validation and `test-A` sets.

Qwen2-Audio-7B-Instruct (Chu et al., 2023) obtained the highest score on the validation set (0.1977), but it exhibited a noticeable drop in performance on both test splits. This suggests that it has limited generalization to newly recorded Polish speech data. Similarly, Qwen-Audio-Chat (Chu et al., 2023) and Whisper + Llama 3.3 demonstrated moderate performance on the validation set, yet neither outperformed Ultravox v0.6 on test-B.

The remaining baselines, Audio Flamingo 3 (Goel et al., 2025) and GAMA (Ghosh et al., 2024), achieved substantially lower macro F1 scores across all splits, indicating their limited effectiveness in cross-lingual speech emotion recognition in this setup. The gap observed between validation and test performance across several models further highlights the challenges posed by domain mismatch, unseen speakers, and newly recorded test material.

To facilitate a deeper analysis of baseline behavior, the confusion matrices for all baseline systems evaluated using the validation and test splits are provided in Appendix B.

An analysis of the confusion matrices for all the evaluated baseline systems further illustrates the task's challenges. Predictions were heavily biased toward a small subset of emotions across models, most notably *neutral*. For all emotional states, substantial confusion between emotionally proximate classes remained evident. These patterns suggest that the baseline systems had difficulty capturing subtle emotional distinctions in Polish speech in a zero-shot, cross-lingual setting, often defaulting to more common or less distinctive emotional states.

The confusion matrices also reveal that no baseline system achieved balanced performance across all six emotional states. This directly explains the low macro-averaged F1 scores observed in Table 5. Even the strongest baseline, Ultravox v0.6, exhibited notable confusion between *happiness* and other affective states. This suggests limited sensitivity to emotional prosody in the unseen language. The weakest baselines exhibited near-random behavior for several classes, rarely predicting certain emotions.

The confusion patterns observed for the cascaded system of Whisper + Llama 3.3, suggest that relying solely on lexical information, without direct access to acoustic cues, is insufficient for robust speech emotion recognition.

Overall, these baseline results established a challenging lower bound for the task and underscored the need for more specialized modeling approaches tailored to cross-lingual and low-resource speech emotion recognition.

5 Results

A total of six participants submitted solutions that were evaluated on at least one test set. However, only five participants provided predictions for the final evaluation set, test-B, and were therefore included in the official leaderboard. The F1-macro results obtained by all participants are shown in Table 6.

Rank	User	test-A	test-B
1	maciejlachut	0.5161	**0.5412**
2	tomasz	0.5318	0.5247
3	tomek	**0.5319**	0.5129
4	kondziu98	0.3915	0.3833
5	cyrta	–	0.0966
–	pawlew	0.1273	–

Table 6: F1-macro results obtained by participants on test sets.

The best performing system, submitted by maciejlachut, reached a macro F1 score of 0.5412, significantly outperforming the baseline models. tomasz followed closely behind with a score of 0.5247. The third-place finished, tomek, scored 0.5129 on test-B, while kondziu98 placed fourth with a score of 0.3833. The fifth-place participant, cyrta, achieved substantially lower score of 0.0966 on the hidden test set. pawlew did not submit results for test-B and therefore did not receive a ranking on the leaderboard.

Comparing these results to the baseline models (Table 5) shows that the baseline systems performed notably lower. The best-performing baseline model, Ultravox v0.6, achieved an F1-macro score of 0.1576 on test-B, significantly lower than the top participants' scores.

Overall, almost all of the participants' systems clearly outperformed the baseline models. maciejlachut achieved an F1-macro score lead of over 0.38 on hidden test set compared to the best baseline. This suggests that the participants' models leveraged the provided data effectively through fine-tuning and strategies tailored to the unique challenge. These results underscore the difficulty of the task and the effectiveness of the competing systems in addressing cross-lingual speech emotion recognition.

Acknowledgements

The authors would like to thank the PolEval challenge organizers and all the participants who submitted solutions. They are especially grateful to the volunteers who generously allowed their voices to be recorded, which made it possible to create the newly collected Polish test sets used for the final evaluation. The authors also acknowledge the creators of the original datasets included in the CAMEO collection. Their commitment to open science made this challenge possible. Finally, the authors encourage all users of the CAMEO collection to properly cite the original sources and authors of the contributing corpora in order to recognize their essential contributions to this research.

References

Artem Amentes, Nikita Davidchuk, and Ilya Lubenets. 2022. Russian Emotional Speech Dialogs with annotated text. https://huggingface.co/datasets/Aniemore/resd_annotated.

Houwei Cao, David Cooper, Michael Keutmann, Ruben Gur, Ani Nenkova, and Ragini Verma. 2014. CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5:377–390.

Fabio Catania, Jordan Wilke, and Franca Garzotto. 2025. Emozionalmente: A Crowdsourced Corpus of Simulated Emotional Speech in Italian. *IEEE Transactions on Audio, Speech and Language Processing*, PP:1–14.

Iwona Christop. 2024. nEMO: Dataset of emotional speech in Polish. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12111–12116, Torino, Italia. ELRA and ICCL.

Iwona Christop and Maciej Czajka. 2025. CAMEO: Collection of Multilingual Emotional Speech Corpora. *Preprint*, arXiv:2505.11051.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-Audio: Advancing Universal Audio Understanding via Unified Large-Scale Audio-Language Models. *Preprint*, arXiv:2311.07919.

Mathilde Marie Duville, Luz Alonso-Valerdi, and David I. Ibarra-Zarate. 2021a. Mexican emotional speech database based on semantic, frequency, familiarity, concreteness, and cultural shaping of affective prosody. *Data*, 6.

Mathilde Marie Duville, Luz Alonso-Valerdi, and David I. Ibarra-Zarate. 2021b. The mexican emotional speech database (mesd): elaboration and assessment based on machine learning. volume 2021.

fixie-ai. 2025. Ultravox v0.6 (LLaMA-3.3-70B) Audio-Text-to-Text Model. https://huggingface.co/fixie-ai/ultravox-v0_6-llama-3_3-70b.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2024. GAMA: A Large Audio-Language Model with Advanced Audio Understanding and Complex Reasoning Abilities. *Preprint*, arXiv:2406.11768.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. 2025. Audio Flamingo 3: Advancing Audio Intelligence with Fully Open Large Audio Language Models. *Preprint*, arXiv:2507.08128.

Philippe Gournay, Olivier Lahaie, and R. Lefebvre. 2018. A canadian french emotional speech dataset. pages 399–402.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Jesin James, Li Tian, and Catherine Inez Watson. 2018. An open source emotional speech corpus for human robot interaction applications. In *Interspeech 2018*, pages 2768–2772.

Leila Kerkeni, Catherine Cleder, Youssef Serrestou, and Kosai Raoof. 2020. French emotional speech database - Oréau.

Steven R. Livingstone and Frank A. Russo. 2018. 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*, 13(5):1–35.

O. Martin, I. Kotsia, B. Macq, and I. Pitas. 2006. The eNTERFACE'05 Audio-Visual Emotion Database. In *Proceedings of the 22nd International Conference on Data Engineering Workshops*, ICDEW '06, page 8, USA. IEEE Computer Society.

Kari Ali Noriy, Xiaosong Yang, and Jian Jun Zhang. 2023. EMNS /Imz/ Corpus: An emotive single-speaker dataset for narrative storytelling in games, television and graphic novels. *Preprint*, arXiv:2305.13137.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. *Preprint*, arXiv:2212.04356.

Ingmar Steiner, Marc Schröder, and Annette Klepp. 2013. The PAVOQUE corpus as a resource for analysis and synthesis of expressive speech. In *Phonetik Phonologie 9. Phonetik Phonologie (PP-9), October 11-12, Zurich, Switzerland*, pages 83–84. UZH, Peter Lang.

Sadia Sultana, M. Shahidur Rahman, M. Reza Selim, and M. Zafar Iqbal. 2021. SUST Bangla Emotional Speech Corpus (SUBESCO): An audio-only emotional speech corpus for Bangla. *PLOS ONE*, 16(4):1–27.

A List of Utterances Used for Test Set

This appendix presents the complete list of the 70 distinct utterances used to record the Polish speech samples that make up the test set for the Polish Speech Emotion Recognition Challenge. Each sentence was recorded by each speaker for each targeted emotional state, providing a consistent, controlled basis for eliciting emotional speech. Using a fixed set of utterances across speakers and emotions ensured the comparability of the recordings, allowing emotional variation to be expressed through prosody and vocal delivery rather than lexical content.

1. Biała parasolka jest tej pani.
2. Oni zaczęli.
3. Czy posiada pani kartę kredytową?
4. O co tyle hałasu?
5. Czy lubisz rap?
6. Woli pan herbatę?
7. Zawsze piję rano dwie filiżanki kawy.
8. Nie zapomnij swoich rzeczy.
9. Poproszę filiżankę kawy.
10. Jaki kolor ma twoja sukienka?
11. Wybałuszyła oczy.
12. Niech nie wie lewica, co robi prawica.
13. W restauracji zamówiłem zestaw z chrzanem.
14. Otworzyłem pudło, ale ono było puste.
15. Damy radę.
16. Uczniowie nie posłuchali swojego nauczyciela.
17. Uratujemy ich.
18. To moja płyta, nie?
19. Który rekomendujesz?
20. Ona żyje!
21. Chyba go polubisz.
22. Idę do kina.
23. Dopóki nie była całkiem syta, jadła jednego cukierka za drugim.
24. Zwolnią go.
25. Wciąż o pani myślę.
26. Nosi okulary.
27. Pocimy się w tym upale.
28. Ochronię cię.
29. Maria była na Węgrzech.
30. Iloma językami dobrze się posługujesz?
31. Pokazał język swojemu nauczycielowi.
32. Maria potrafi grać na pianinie.
33. Słyszę muzykę.
34. Idź drogą w prawo.
35. Ciężko oddychali.
36. Lubię koreańskie jedzenie.
37. Miałem operację.
38. Jej uczucia są trochę urażone.
39. To płonie.
40. Języki programowania to jego hobby.
41. Wczoraj wieczorem ukradziono mi rower.
42. Nie oczekuję, że pan odpowie.
43. Ona pije kawę.
44. Mianowali Janka kierownikiem.
45. Uczę się chińskiego w Pekinie.
46. Są zajęci.
47. Zaufała mi.
48. Próbowałem pisać moją lewą ręką.
49. On nie jada surowych ryb.
50. Ohydne to mleko.

51. Czy zarezerwowałeś już pokój w hotelu?

52. Ona uprawia wiele gatunków kwiatów.

53. Niech pan mnie ochrania!

54. Oto Japonia.

55. Na Węgrzech każdy mówi po węgiersku.

56. Boli cię głowa?

57. Uważaj na lewe taryfy.

58. Uczymy się języka hiszpańskiego.

59. Para dobrych okularów pomoże ci czytać.

60. Oliwa sprawiedliwa zawsze na wierzch wypływa.

61. Do stomatologa obowiązują zapisy.

62. Mówisz moim językiem.

63. Pojechałem do Paryża.

64. Jego rada niewiele pomogła.

65. Świeże owoce i warzywa są dobre dla twojego zdrowia.

66. Tomek nigdy nie był w Bostonie.

67. Moim hobby jest łowienie ryb.

68. Prawa ręka nie wie co czyni lewa.

69. Lubię happy endy.

70. Kiedy wychodzimy?

B Confusion Matrices for Baseline Systems

Figures 4– 9 show confusion matrices for all baseline models evaluated on `dev`, `test-A`, and `test-B` splits. Rows correspond to true labels and columns to predicted labels, following the fixed label order: anger (**A**), fear (**F**), happiness (**H**), neutral (**N**), sadness (**Sa**), and surprise (**Su**).

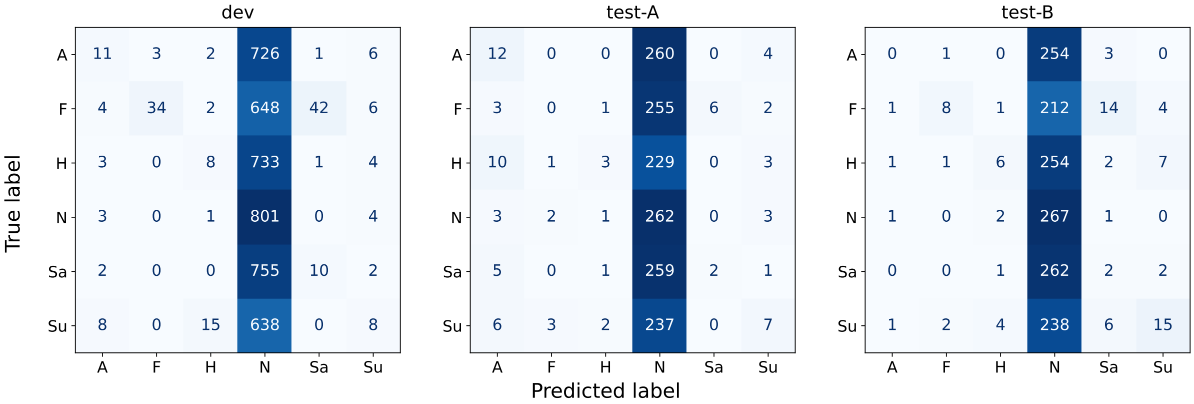

Figure 4: Confusion matrices for Audio Flamingo 3 model evaluated on validation and test splits.

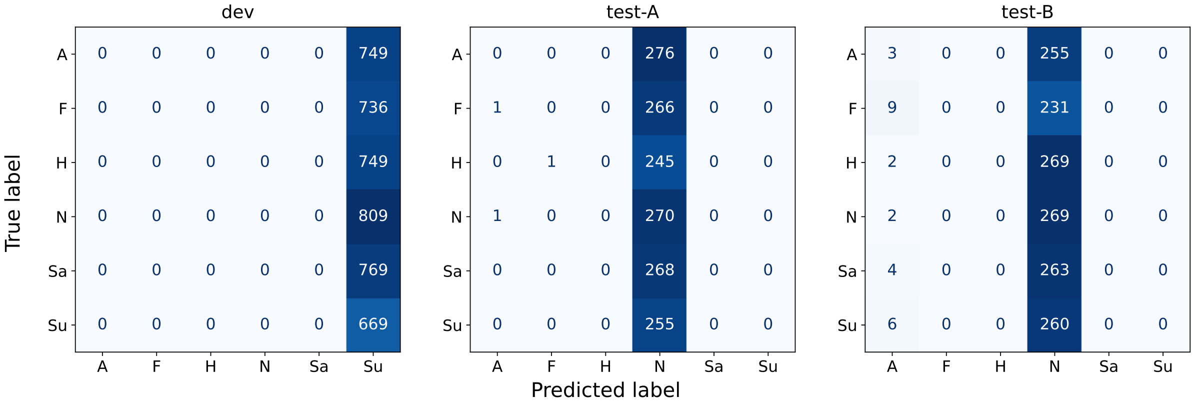

Figure 5: Confusion matrices for GAMA model evaluated on validation and test splits.

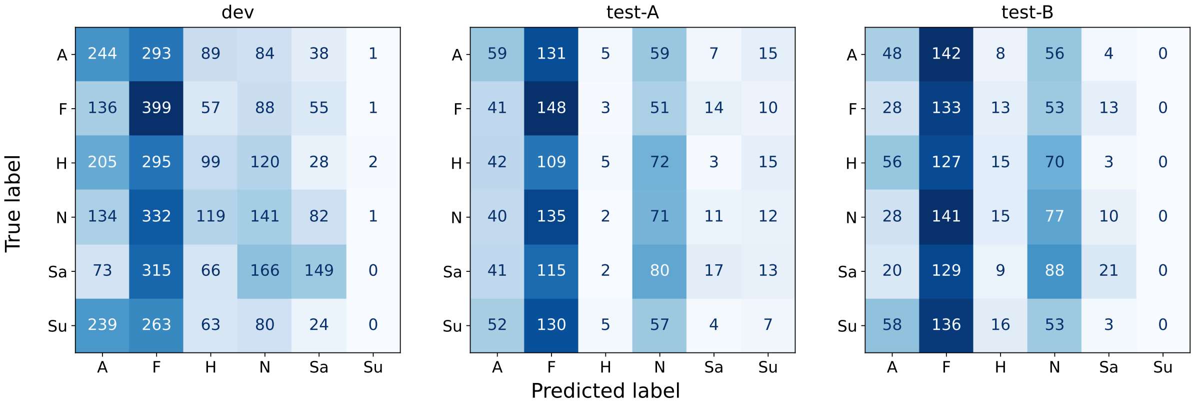

Figure 6: Confusion matrices for Qwen2-Audio model evaluated on validation and test splits.

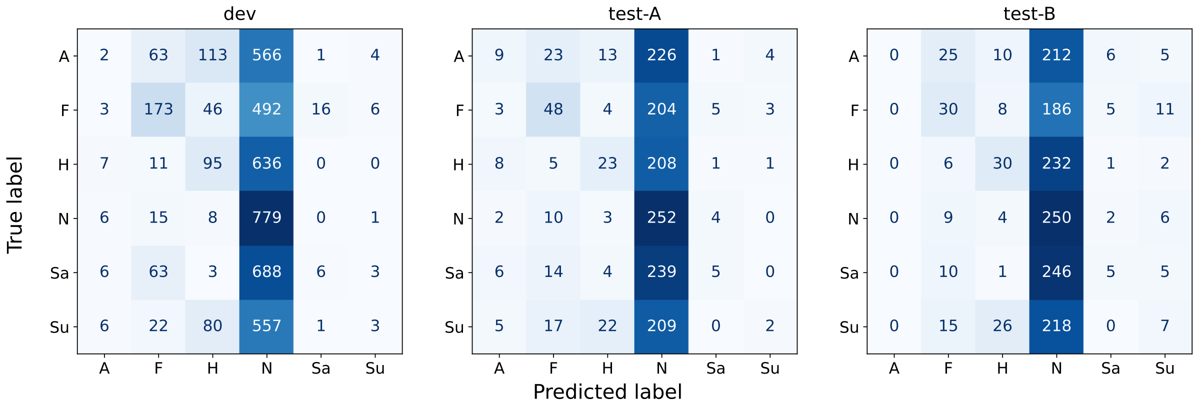

Figure 7: Confusion matrices for Qwen-Audio-Chat model evaluated on validation and test splits.

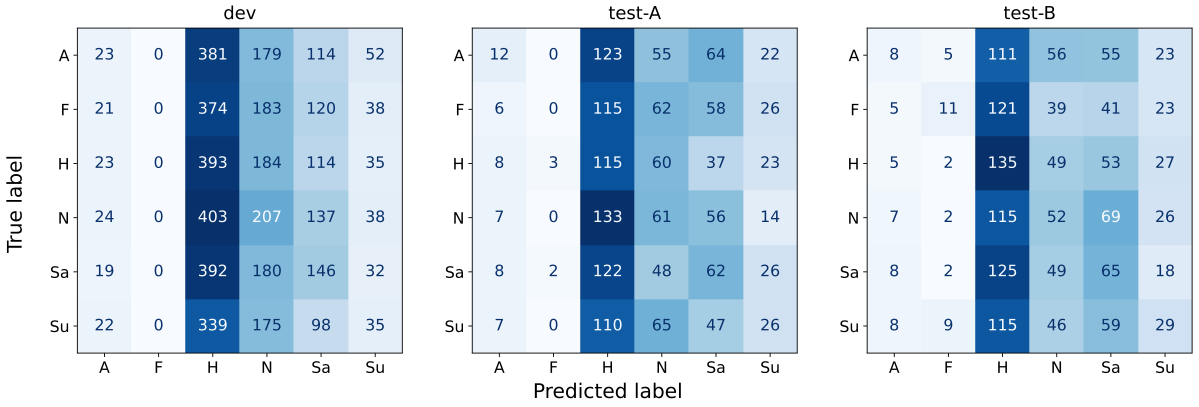

Figure 8: Confusion matrices for Ultravox v0.6 model evaluated on validation and test splits.

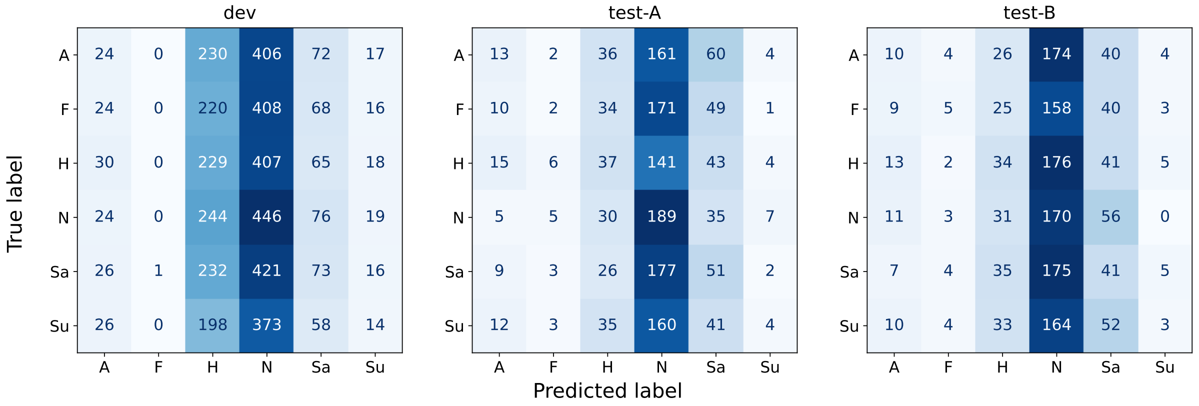

Figure 9: Confusion matrices for Whisper + Llama 3.3 cascaded system evaluated on validation and test splits.