Zero-Shot Transfer of Pretrained Speech Representations for Multilingual Emotion Recognition

Tomasz Kuczyński
Adam Mickiewicz University ul. Uniwersytetu Poznańskiego 4 61-614 Poznań, Poland
tomkuc2@st.amu.edu.pl

Abstract

Speech emotion recognition remains a challenging task, particularly in low-resource language settings. In this work, we explore the development of a system capable of identifying emotional states in Polish speech using training data exclusively from other languages. Our approach relies on a pretrained speech representation model and follows a strict zero-shot training paradigm, enabling cross-lingual knowledge transfer without access to any Polish data. The system was developed in the context of the Polish Speech Emotion Recognition Challenge (PolEval 2025), which required participants to train models solely on multilingual resources and evaluate them on Polish speech in a zero-shot setup. We present a complete solution encompassing model selection, audio preprocessing, and fine-tuning strategy, and discuss the potential of large-scale language models for cross-lingual emotion recognition.

1 Introduction

Speech Emotion Recognition (SER) is a critical subfield of affective computing, focusing on the automatic identification of human emotional states from vocal signals (Schuller et al., 2011). Speech is a rich communication channel that conveys not only linguistic but also paralinguistic information, making SER increasingly relevant in human-computer interaction, voice service analytics, assistive technologies, and remote psychological assessment (El Ayadi et al., 2011; George and Ilyas, 2024). Accurate detection of emotion in speech enables more natural and empathetic interactions in dialog systems and facilitates better understanding of users' intentions and mental states.

In recent years, advances in deep learning and self-supervised representation learning have significantly improved the performance of SER systems, especially for high-resource languages such as English, German, or Mandarin (Latif et al., 2021). Models such as wav2vec 2.0 (Baevski et al., 2020), HuBERT (Hsu et al., 2021), and WavLM (Chen et al., 2022) allow for the extraction of effective speech representations with minimal task-specific supervision. However, most of these developments have been concentrated on resource-rich languages, which limits their applicability in low-resource or underrepresented linguistic contexts.

The challenge of multilingual SER lies in the language dependence of emotional prosody. While certain acoustic correlates of emotion, such as pitch variation, speech rate, and energy, are considered relatively universal, the way emotions are expressed and perceived is strongly influenced by cultural, contextual, and phonological factors (Shochi et al., 2009; Pell et al., 2009). This makes cross-lingual transfer under zero-shot conditions particularly difficult. Models trained on emotions in one language may fail to recognize or may incorrectly classify equivalent expressions in another.

In recent years, numerous studies have attempted to address the challenges of cross-lingual knowledge transfer in speech emotion recognition, particularly in scenarios where no supervised data is available for the target language. Among the strategies explored to improve cross-lingual generalization are multilingual fine-tuning, techniques for aligning emotional representations across languages (for example, through contrastive learning or adversarial training), and the use of shared feature spaces derived from self-supervised speech models. Despite encouraging results in experimental settings, zero-shot performance, where models are trained without any access to data from the target language, remains substantially lower compared to supervised approaches based on direct fine-tuning.

The Polish Speech Emotion Recognition Challenge (PolEval 2025) directly addresses this problem by proposing a strict zero-shot setup in which participants are required to train their models ex-

Proceedings of the PolEval 2025 Workshop, pages 91–96
November 21, 2025 ©2025 Association for Computational Linguistics

clusively on multilingual data and evaluate them solely on Polish speech. This setup reflects real-world deployment conditions for SER systems in low-resource languages. While Polish does not entirely lack emotional speech corpora, high-quality, large-scale, publicly available datasets have only recently become accessible, such as the nEMO corpus (Christop, 2024). As a result, Polish remains underrepresented in the international SER research landscape, making it a compelling target language for empirical evaluation of zero-shot transfer.

In this work, we investigate whether pre-trained self-supervised speech models, specifically WavLM, can serve as a robust foundation for multilingual SER under a zero-shot paradigm. Our contributions are threefold:

- We design and implement a zero-shot SER system based on WavLM, fine-tuned exclusively on speech data from non-Polish corpora.
- We evaluate the system on the official PolEval 2025 benchmark, strictly adhering to the zero-shot evaluation constraints.
- We analyze the challenges of cross-lingual generalization in emotional representation and discuss future directions, including the use of larger models, multitask learning, and data augmentation techniques.

2 Task Description

The Polish Speech Emotion Recognition Challenge (PolEval 2025) is designed as a strict zero-shot classification task. The objective is to recognize one of six discrete emotional categories from spoken utterances: *anger*, *fear*, *happiness*, *neutral*, *sadness*, and *surprise*. No Polish data may be used during training, and all models are evaluated exclusively on Polish speech during testing.

2.1 Training Data

The training set consists of 29,714 utterances derived from 12 publicly available speech emotion recognition corpora. These corpora cover seven different languages and are unified within the multilingual CAMEO dataset. All samples are annotated using a consistent set of six emotion labels.

Table 1 summarizes the total number of training examples per emotion aggregated across all corpora. Although the dataset is relatively large, class imbalance remains a potential challenge.

Emotion	# Samples
Anger	5,212
Fear	4,241
Happiness	5,216
Neutral	7,161
Sadness	5,127
Surprise	2,757

Table 1: Total number of training samples per emotion (aggregated across all corpora).

2.2 Validation and Evaluation Data

The validation set used in this task is based on the existing nEMO corpus, which contains Polish emotional speech recordings labeled with six emotion categories: *anger*, *fear*, *happiness*, *neutral*, *sadness*, and *surprise*. For the purposes of the PolEval 2025 challenge, the entire corpus was adopted as the *dev* set and made available to participants for model tuning and selection. The class distribution within the validation set is shown in Table 2. While the dataset is relatively balanced, minor variations in class frequency are present.

In addition to the validation set, the organizers prepared two separate test sets for final evaluation:

- **test-A**, containing 1,583 recordings, was released during the competition to allow for intermediate evaluation under test-like conditions;
- **test-B**, containing 1,573 recordings, remained unreleased until the end of the challenge and was used to determine the final leaderboard.

For both test sets, emotion labels and class distributions were withheld. This ensured an unbiased zero-shot evaluation scenario, preventing any data leakage or task-specific tuning.

Emotion	# Samples
Anger	749
Fear	736
Happiness	749
Neutral	809
Sadness	769
Surprise	669

Table 2: Number of validation samples per emotion in the Polish nEMO corpus.

3 System Architecture

The architecture of the proposed system consists of three main components: an acoustic representation extraction layer, a classification layer, and a data preparation procedure. The system is based on the WavLM-Base+ model, which was further fine-tuned for the emotion classification task using a multilingual training set.

3.1 Representation Extraction

Raw audio signals were processed using the WavLM-Base+ model together with the feature extractor from the `transformers` library. All input samples were resampled to 16 kHz and standardized to a fixed length of 3.5 seconds, allowing for consistent batch processing. This duration corresponded to the typical utterance length in the test sets and ensured stable training in a resource-constrained environment.

Model parameters were fine-tuned for the classification task without modifying the internal architecture or layers. The weights were initialized using the official version released by the model's authors.

3.2 Classification Layer

The classification head was implemented as a single-layer neural classifier placed on top of the final hidden state of the base encoder. It consists of a dropout layer (with a dropout rate of 0.1) followed by a fully connected linear transformation projecting the encoder output into a 6-dimensional logits vector, corresponding to the six emotion categories defined in the dataset. The input to the classifier was the contextualized embedding of the [CLS] token, representing the aggregate sentence representation.

The output logits were passed to a softmax function during evaluation to produce class probabilities, while during training the raw logits were used directly in the cross-entropy loss function. No class weighting or label smoothing was applied.

This simple classification structure was chosen to avoid overfitting, preserve the generalization capacity of the base encoder, and ensure comparability with related works using similar low-parameter output layers. The design aligns with the standard setup in transformer-based text classification tasks.

3.3 Training Configuration

The model was trained using the AdamW optimizer with a learning rate of 2×10^{-5} and a weight decay of 0.01. A batch size of 16 was used, and gradient accumulation was set to 2 steps, effectively simulating a batch size of 32, which allowed efficient training on the available hardware.

We trained the model for 10 epochs. A linear warm-up was applied during the first 10% of total training steps, followed by a linear learning rate decay schedule. No early stopping or learning rate restarts were used. Dropout with a rate of 0.1 was applied before the final classification layer to reduce overfitting.

After each epoch, the model was evaluated on a held-out validation set, and the checkpoint achieving the highest macro-averaged F1 score was saved as the final model used for test evaluation.

Hyperparameter	Value
Optimizer	AdamW
Learning rate	3×10^{-5}
Batch size (per device)	4
Gradient accumulation	4 steps
Effective batch size	16
Number of epochs	5
Warm-up ratio	10% of total steps
Learning rate schedule	Linear decay
Dropout rate	0.1
Early stopping	No
Evaluation metric	Macro-averaged F1
Mixed precision (FP16)	Yes

Table 3: Summary of training hyperparameters used for fine-tuning the WavLM model.

3.4 Classification Layer

A single-layer classification head was appended to the model's output, consisting of a dropout layer followed by a fully connected linear layer. Its purpose was to assign each representation to one of six emotion classes. The model was trained using a standard cross-entropy loss without class weighting. Optimization was performed using the AdamW algorithm.

Training hyperparameters (number of epochs, learning rate, batch size, gradient accumulation) were selected empirically, considering available computational resources and training stability. The model was evaluated after each epoch on the validation set, and the checkpoint with the highest macro-averaged F1 score was selected as the final model.

3.5 Data Preparation

The training set was constructed based on metadata provided by the PolEval organizers, which included references to specific utterances from the CAMEO corpus. Only samples with clear assignments to one of the six target emotions were retained.

No data augmentation or additional language filtering was applied. The structure of the dataset was preserved in accordance with the competition rules and the zero-shot constraints, without using any Polish data during training.

4 Evaluation and Results

The system was evaluated on three datasets: a validation set (*dev*) and two independent test sets (*test-A* and *test-B*), all containing exclusively Polish speech. The final ranking in the PolEval 2025 competition was based on the macro-averaged F1 score obtained on the hidden *test-B* set, which was not accessible during model development.

Macro-averaged F1 was adopted as the primary evaluation metric, as it accounts for class imbalance and better reflects overall model performance across all emotion categories. For comparison, we also report accuracy, although its interpretative value is limited in the presence of imbalanced labels (Sokolova and Lapalme, 2009).

Dataset	F1-score	Accuracy
Dev	0.7045	0.7120
Test-A	0.5318	0.5338
Test-B	0.5233	0.5429

Table 4: Evaluation results on the validation and test sets.

As shown in Table 4, the system achieved an F1 score of 0.7045 on the validation set, indicating good generalization to previously unseen Polish data. Performance on the test sets was notably lower, highlighting the difficulty of the zero-shot emotion recognition task.

The gap in performance between the validation and test sets is a well-known phenomenon in cross-lingual transfer settings, where the model is evaluated on entirely new speakers, acoustic conditions, or lexical material. This illustrates the inherent challenge of transferring emotional representations across languages without any adaptation to the target language.

Despite these limitations, the system achieved competitive results under strict zero-shot constraints, confirming the potential of multilingual pretrained models for emotion recognition in under-resourced languages.

5 Discussion

The results obtained in this study confirm that multilingual pretrained models can serve as a reliable foundation for cross-lingual speech emotion recognition under zero-shot conditions. The use of the WavLM model enabled the extraction of speech representations that generalized successfully to Polish, despite the language being entirely absent from the training data. This suggests that the model was able to capture acoustic patterns relevant to emotion recognition that are not strictly language-dependent (Chen et al., 2022).

However, a detailed analysis of the results reveals certain limitations. The significant drop in performance between the validation set and the test sets indicates that full generalization to previously unseen data remains difficult. This is especially evident in the case of the test sets, which were not accessible during model selection and better reflect realistic deployment conditions. The observed discrepancy may have several causes.

First, emotional expression varies across languages due to phonological, cultural, and contextual factors (Pell et al., 2009; Shochi et al., 2009), which may have affected the model's ability to transfer knowledge from the training data to Polish. Second, the emotion categories used in the training set may not have fully matched the way those emotions are realized in Polish speech. While label definitions were harmonized, their acoustic realizations could still differ significantly across languages and corpora.

In addition, domain shift between the training and test data may have contributed to the performance degradation. Differences in speaker characteristics, recording quality, speaking style, or class distribution could have introduced inconsistencies that the model was unable to resolve. These factors are particularly important in speech emotion recognition, where prosodic cues are subtle and highly sensitive to contextual and acoustic variation.

It is also worth noting that the relatively higher score achieved on the validation set compared to the test sets may indicate a degree of implicit adaptation during model selection — even though no Polish data was used for training. The validation set was used as a criterion for selecting the best checkpoint, which, while in line with the competition rules, may have introduced a mild form of adaptation to that specific dataset.

Despite these limitations, the system achieved competitive results under strict zero-shot constraints. This confirms that models pretrained on large and diverse multilingual corpora can successfully transfer emotion-related features across languages. These findings support the validity of the zero-shot approach as a feasible strategy for emotion recognition in low-resource languages and highlight the need for further research into the mechanisms and limitations of cross-lingual transfer.

6 Conclusion

The primary objective of this study was to evaluate the feasibility of cross-lingual speech emotion recognition in a strict zero-shot scenario, where the target language (Polish) is completely absent from the training data. To this end, we developed a system based on the pretrained WavLM model, which was fine-tuned for emotion classification using the multilingual CAMEO corpus. A key constraint of the task was the strict separation between training and evaluation languages, reflecting realistic conditions for deploying systems in low-resource settings.

The results obtained on the official test sets of the PolEval 2025 challenge demonstrate that self-supervised models pretrained on large-scale, diverse audio data can effectively learn emotion-related representations that generalize across languages. Despite a noticeable performance drop compared to the validation set, the system maintained stable performance on unseen Polish data, suggesting that certain prosodic features associated with emotion are sufficiently language-independent to support zero-shot transfer.

The proposed approach highlights the practical potential of leveraging multilingual pretrained models for building SER systems in languages lacking annotated emotional speech corpora. This opens up opportunities for applications in voice-based interaction systems, affective computing in social media analytics, and mental health monitoring in conversational technologies.

Future work could explore several strategies to further improve cross-lingual generalization. These include domain adaptation methods, contrastive learning for more discriminative emotion representations, multitask learning that combines emotion recognition with auxiliary tasks such as speaker or language identification, and the use of larger or domain-specific pretrained models. It would also be valuable to investigate the impact of class imbalance, stylistic variation, and sociolinguistic factors on the quality of zero-shot transfer.

The findings presented in this paper represent a step toward building more robust and language-independent emotion recognition systems capable of operating in truly low-resource conditions.

Resources

To facilitate reproducibility and further research, we provide public access to the source code implementing our system at github.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 12449–12460.

Sanyuan Chen, Chengyi Wang, Yu Wu, Shujie Wu, Yanmin Qian, and Dong Yu. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

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.

M. El Ayadi, M. S. Kamel, and F. Karray. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587.

Swapna Mol George and P. Muhamed Ilyas. 2024. A review on speech emotion recognition: A survey, recent advances, challenges, and the influence of noise. *Neurocomputing*, 568:127015.

Wei-Ning Hsu, Bastian Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. In *Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL)*, pages 1474–1484.

Siddique Latif, Junaid Qadir, Adnan Qayyum, Muhammad Usama, and Shahzad Younis. 2021. Speech technology for healthcare: Opportunities, challenges, and state of the art. *IEEE Reviews in Biomedical Engineering*, 14:342–356.

Marc D. Pell, Laura Monetta, Silke Paulmann, and Sonja A. Kotz. 2009. Recognizing emotions in a foreign language. *Journal of Nonverbal Behavior*, 33(2):107–120.

B. Schuller, S. Steidl, and A. Batliner. 2011. Recognizing realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, 53(9–10):1062–1087.

T. Shochi, V. Aubergé, and A. Rilliard. 2009. Cross-cultural perception of prosodic functions in expressive speech: some japanese/french/chinese examples. *Intercultural Pragmatics*, 6(2):237–266.

M. Sokolova and G. Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437.