Inference-Only Speaker Adaptation Improves Cross-Lingual Speech Emotion Recognition

Maciej Łachut
Poznan University of Technology
Poland
maciej.lachut@student.put.poznan.pl

Abstract

Cross-lingual Speech Emotion Recognition (SER) is frequently hindered by speaker-specific prosodic variations that obscure universal emotional cues. Standard models often fail to generalize across languages due to the domain shift caused by differing acoustic standards. To address this, we present a novel SER approach that integrates unsupervised speaker adaptation directly at inference time. Our architecture utilizes a frozen, pretrained HuBERT encoder and introduces a Greedy Cluster Assignment Algorithm. This method groups a speaker's utterances to form emotion-dependent centroids, enforcing speaker-consistent labeling without the computational cost of retraining. We evaluated this approach in a cross-lingual setting using the Polish nEMO dataset, which was excluded from training. Our method achieved the best performance in the POL-EVAL 2025 Task 4, improving the Macro F1 score from 0.619 to 0.753 on validation data and securing 1st place on the official leaderboard. Results demonstrate that inference-only clustering effectively disentangles ambiguous high-arousal categories, such as Fear and Surprise, by calibrating to the individual speaker's vocal range.

1 Introduction

Cross-lingual Speech Emotion Recognition (SER) is frequently hindered by speaker-specific prosodic variations that obscure universal emotional cues. Standard models often fail to generalize across languages due to the domain shift caused by differing acoustic standards. Recent findings have highlighted that integrating speaker-specific vocal characteristics through adaptation is crucial for improving SER accuracy in these challenging scenarios (Ihori et al., 2025; Shi et al., 2025). While supervised adaptation typically requires computationally expensive retraining, inference-time strategies offer a more efficient alternative.

To address this, we present a novel SER approach that integrates speaker-specific vocal characteristics through an efficient inference-only adaptation procedure. Our architecture is built upon a pretrained HuBERT encoder (Hsu et al., 2021) fine-tuned on the Dusha dataset (Kondratenko et al., 2022). We further introduce a *Greedy Cluster Assignment Algorithm*, which groups speaker embeddings during inference to enforce speaker-consistent labeling and capture emotion-dependent clusters without the computational cost of retraining.

We evaluated this method in a cross-lingual setting using the Polish nEMO dataset (Christop, 2024), which was excluded from the multilingual CAMEO training set. This approach achieved the best performance in the POL-EVAL 2025 Task 4: Polish Speech Emotion Recognition Challenge. Experimental results demonstrate that the proposed clustering strategy significantly outperforms a direct inference model, improving the Macro F1 score from 0.619 to 0.753.

Finally, while our primary contribution is to SER, this work has significant implications for generative tasks. Controlling emotional expressivity remains a persistent challenge in Text-to-Speech (TTS) (Li et al., 2023). By effectively disentangling speaker identity from emotional state, our proposed speaker-adaptation procedure provides the granular control necessary to support high-fidelity, emotion-aware synthesis.

2 Dataset

For model training, we employ the CAMEO dataset (Christop and Czajka, 2025; Gournay et al., 2018; Cao et al., 2014; Noriy et al., 2023; Catania et al., 2025; Martin et al., 2006; James et al., 2018; Duville et al., 2021b,a; Christop, 2024; Kerkeni et al., 2020; Steiner et al., 2013; Livingstone and Russo, 2018; Amentes et al., 2022; Sultana

Proceedings of the PolEval 2025 Workshop, pages 82–90

et al., 2021), which provides multilingual emotional speech samples annotated with categorical emotion labels. The Polish subset (Christop, 2024) is excluded from training and reserved solely for cross-lingual evaluation. To ensure label consistency across languages, we restrict the training data to the six emotion classes represented in the nEMO subset: anger, fear, happiness, neutral, sadness, and surprise. We discard samples from CAMEO sub-datasets that contain additional categories.

2.1 Training Set

The training set consists of 29,714 audio recordings aggregated from 12 different sub-datasets: 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., 2021b,a), Oreau (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). The distribution of samples per language and emotion category within the training set is detailed in Table 4.

2.2 Validation Set

The validation set consists solely of the nEMO split (Christop, 2024) of the CAMEO dataset. This set comprises 4,481 audio recordings in the Polish language. The distribution of samples per emotion is presented in Table 5.

2.3 Data Augmentation

To improve the model's robustness to channel variations and prevent overfitting to speaker-specific traits, we applied a comprehensive on-the-fly data augmentation pipeline during training. The augmentation strategy was designed to simulate diverse recording conditions and speaker variations without altering the underlying emotional semantics. We utilized the `torch-audiomentations` library alongside custom PyTorch implementations for time-domain transformations.

The pipeline applies the following transformations probabilistically:

- **Additive Noise:** We injected white noise with a signal-to-noise ratio (SNR) sampled uniformly between 15 and 40 dB ($p = 0.3$). Additionally, background environmental noise was added with an SNR between 10 and 30 dB ($p = 0.2$).
- **Signal Degradation & Filtering:** To simulate varying microphone qualities, we applied random low-pass (2–7 kHz), high-pass (100–2000 Hz), band-pass, and band-stop filters, each with a probability of $p = 0.15$. We also introduced algorithmic reverberation ($p = 0.25$) to mimic room acoustics.
- **Temporal & Pitch Perturbation:** We employed two distinct strategies to disentangle pitch and tempo. *Speed perturbation* was applied via resampling factors in $[0.9, 1.1]$ ($p = 0.25$), affecting both pitch and duration. Separately, *time stretching* was performed using a phase vocoder with rates in $[0.85, 1.20]$ ($p = 0.25$) to alter speed while preserving pitch.
- **Pitch Shifting:** We shifted the pitch by ± 3 semitones ($p = 0.25$) to encourage invariance to speaker fundamental frequency (F_0).
- **SpecAugment-style Masking (Park et al., 2019):** We applied random time masking, zeroing out segments between 0.05 and 0.5 seconds ($p = 0.25$) to force the model to rely on contextual cues.

All augmentations were applied to the raw waveform prior to feature extraction. The final audio was clamped to the range $[-1, 1]$.

3 Model Architecture

The proposed model architecture is based on the pretrained HuBERT transformer encoder (hubert-large-ls960-ft) (Hsu et al., 2021), which is subsequently fine-tuned on the large-scale Dusha speech emotion recognition (SER) dataset (Kondratenko et al., 2022). The pretrained HuBERT encoder serves as the backbone of the system, upon which we introduce an attention-based pooling mechanism that aggregates frame-level representations into a fixed-dimensional utterance-level embedding. This embedding is passed to a fully connected classification head that outputs probabilities over six predefined emotion categories.

Model parameters are optimized using the AdamW optimizer. The learning rate is set to 1×10^{-5} for the HuBERT backbone and 5×10^{-5} for both the attention-pooling module and the classification head. A cosine learning-rate schedule with a linear warm-up phase comprising 10% of the total training steps is employed. Weight decay is set to 0.01, and training is performed with a batch size of 8 for a total of four epochs over the CAMEO dataset (excluding nEMO).

The full code is publicly available[1].

3.1 Greedy Cluster Assignment for Speaker-Adaptive Inference

To incorporate speaker-specific structure during inference, we remove the emotion-classification head from the model and use the encoder with attention pooling to generate fixed-dimensional embeddings for each utterance. For a given speaker, these embeddings are expected to form emotion-dependent clusters (e.g., a cluster corresponding to *sad* utterances and another to *happy* ones).

Our inference procedure consists of the following steps. First, for each speaker, we group all of that speaker's utterances and generate their embeddings. If only a single utterance is available, we directly apply the emotion-classification head and assign the predicted label.

For speakers with multiple utterances, we perform K-means clustering over their embeddings, using $k = \min(n, 6)$, where n is the number of utterances. Prior to clustering, we apply standardization to ensure that all embedding dimensions contribute equally. For each cluster, we compute its centroid in the original embedding space.

We then estimate emotion probabilities for each centroid using the pretrained classification head. Each centroid is forwarded through the head (GELU $\rightarrow$ Dropout $\rightarrow$ Linear), producing a probability distribution over the emotion classes.

To assign emotions to clusters, we use a greedy matching strategy. We construct a list of all $(\text{probability}, \text{cluster}, \text{emotion})$ triples and sort them in descending order by probability. Iterating through this list, we assign an emotion to a cluster if: (i) the cluster has not yet been assigned an emotion, and (ii) the emotion has not yet been used. This ensures a one-to-one mapping between clusters and emotion labels whenever possible. If any cluster remains unassigned after the greedy pass, we assign it the emotion with the highest centroid-level probability, even if that emotion has already been used.

Finally, all utterances inherit the emotion label assigned to the cluster to which they belong. This produces a speaker-consistent labeling that prevents conflicting assignments within the same speaker while allowing emotion distributions to vary between speakers.

Code for this algorithm is presented in Algorithm 1.

4 Results

We demonstrate that the proposed method performs particularly well when a sufficiently large number of utterances is available for a given speaker. It achieves substantially better results than the direct baseline model and exhibits robust generalization to previously unseen languages. Moreover, the approach provides a simple and effective means of improving the performance of existing models. This method was also successfully applied in the POL-EVAL Task 4 competition, where it competed alongside alternative solutions.

To further investigate the source of these improvements, we report a comparative per-emotion analysis in Table 3.

We can expect a performance improvement of several percentage points when applying this method to an already pretrained model. A key observation from our experiments is that the Greedy Cluster Assignment algorithm achieves top-tier performance (Table 2) despite relying on the HuBERT (Hsu et al., 2021) encoder, which predates current state-of-the-art foundation models.

Results suggest that WavLM-Large (Chen et al., 2022) baseline (without clustering) attains competitive results primarily due to its substantially larger pre-training corpus (94k hours vs. 60k hours) and the inclusion of an explicit denoising objective. This indicates that the performance of our system is currently bottlenecked by the quality of the underlying embeddings, rather than by the clustering strategy itself.

We further argue that our inference-only adaptation procedure is model-agnostic. Replacing the backbone with a more powerful direct model would likely yield a more pronounced separation between emotion clusters in the latent space. As a result, the clustering algorithm would encounter fewer ambiguous centroids, which could plausibly raise the Macro F1 score well above the current benchmark of 0.753. Therefore, our method should be regarded as a performance multiplier whose effectiveness scales with the representational strength of the underlying encoder.

Figure 1 presents the confusion matrices for both

[1] https://github.com/NeonFeline/SER-Inference-Only-Speaker-Adaptation

Table 1: Ablation Study: Impact of Inference-Time Clustering. The proposed clustering mechanism consistently outperforms the direct model. On the challenging hidden test set, removing the clustering logic results in a sharp performance drop (0.4822 F1), confirming that the +5.9% gain is a robust property of the adaptation method, mirroring the trend seen in validation.

Dataset	Direct Model (F1)	With Clustering (F1)	Absolute Gain
Validation (nEMO)	0.6190	**0.7530**	+13.4%
Hidden Test (Official)	0.4822	**0.5412**	+5.9%

Table 2: Official Top 3 Final Standings for POL-EVAL 2025 Task 4. Our submission (*maciejlachut*) secured 1st place using the older HuBERT backbone enhanced with inference-time clustering, demonstrating that adaptive methods can yield SOTA performance without requiring the newest foundation models.

Rank	Participant	Score (F1)
1	**maciejlachut (Ours)**	**0.5412**
2	tomasz	0.5247
3	tomek	0.5129

the direct model and the clustering method. It is evident that the direct model struggles to differentiate between *Fear* and *Sadness*, as well as between *Happiness* and *Surprise*. These errors are notably less pronounced in the clustering approach. Furthermore, Figure 2 illustrates a PCA analysis of the backbone's final embeddings. The visualization reveals a clear separation of emotions into distinct clusters, providing empirical grounds for the effectiveness of our clustering method.

5 Limitations

While our Greedy Cluster Assignment algorithm significantly improves performance, it relies on two key assumptions. First, the method assumes the availability of accurate speaker diarization, as it requires grouping utterances by speaker ID prior to inference. In real-world in-the-wild scenarios, diarization errors (e.g., merging two speakers) could degrade the purity of the clusters and degrade assignment accuracy. Second, the greedy matching strategy enforces a one-to-one mapping between clusters and emotion labels. This assumes a speaker expresses a specific emotion (e.g., "Anger") in a unimodal way. In cases where a speaker exhibits multimodal expressions of a single emotion (e.g., "cold anger" vs. "hot anger"), the algorithm may force one of these clusters into an incorrect category to satisfy the unique-label constraint. Finally, because the method requires aggregating a speaker's utterances to form clusters, it functions as an offline or buffered batch-processing approach rather than a low-latency streaming solution.

6 Future work

While our current inference-only adaptation confirms the efficacy of speaker-based clustering for cross-lingual SER, future research will focus on integrating this adaptation directly into an end-to-end training pipeline. Building upon the few-shot personalization frameworks proposed by Ihori et al. (2025), we propose a context-aware architecture where the neural network dynamically aggregates speaker information. Specifically, we intend to employ a Transformer encoder that attends to a buffer of past utterances to predict the emotional state of the current target, effectively learning to perform speaker adaptation on the fly. We anticipate that this dynamic, in-context learning approach will further improve cross-lingual robustness and offer significant benefits for controlling expressivity in downstream Text-to-Speech applications.

Table 3: Comparative Performance: Direct vs. Clustering (Macro Averages)

Emotion	Support	Direct Method			Clustering Method		
		Prec	Rec	F1	Prec	Rec	F1
Anger	749	0.883	0.648	0.747	**0.908**	**0.816**	**0.859**
Fear	736	0.735	0.387	0.507	**0.741**	**0.693**	**0.716**
Happiness	749	0.609	0.758	0.676	**0.707**	**0.793**	**0.748**
Neutral	809	0.595	0.805	0.684	**0.835**	**0.815**	**0.825**
Sadness	769	0.537	**0.886**	0.669	**0.747**	0.817	**0.780**
Surprise	669	**0.792**	0.296	0.431	0.597	**0.580**	**0.588**
Mean (Macro)	*4481*	0.692	0.630	0.619	**0.756**	**0.752**	**0.753**

Table 4: Distribution of samples per emotion in the Training Set. Dash (-) indicates the emotion is missing from that subset.

Dataset	Lang.	Total	Ang.	Fear	Hap.	Neu.	Sad.	Sur.
CaFE	French	792	144	144	144	72	144	144
CREMA-D	English	6,171	1,271	1,271	1,271	1,087	1,271	-
EMNS	English	743	133	-	158	149	150	153
Emozionalmente	Italian	5,916	986	986	986	986	986	986
eNTERFACE	English	1,047	210	210	207	-	210	210
JL-Corpus	English	960	240	-	240	240	240	-
MESD	Spanish	718	143	144	144	143	144	-
Oreau	French	431	73	71	72	71	72	72
PAVOQUE	German	4,867	601	-	584	3,126	556	-
RAVDESS	English	1,056	192	192	192	96	192	192
RESD	Russian	1,013	219	223	218	191	162	-
SUBESCO	Bengali	6,000	1,000	1,000	1,000	1,000	1,000	1,000
Total	**-**	**29,714**	**5,212**	**4,241**	**5,216**	**7,161**	**5,127**	**2,757**

Table 5: Distribution of samples per emotion in the Validation Set (nEMO).

Emotion	# Samples
Anger	749
Fear	736
Happiness	749
Neutral	809
Sadness	769
Surprise	669
Total	**4,481**

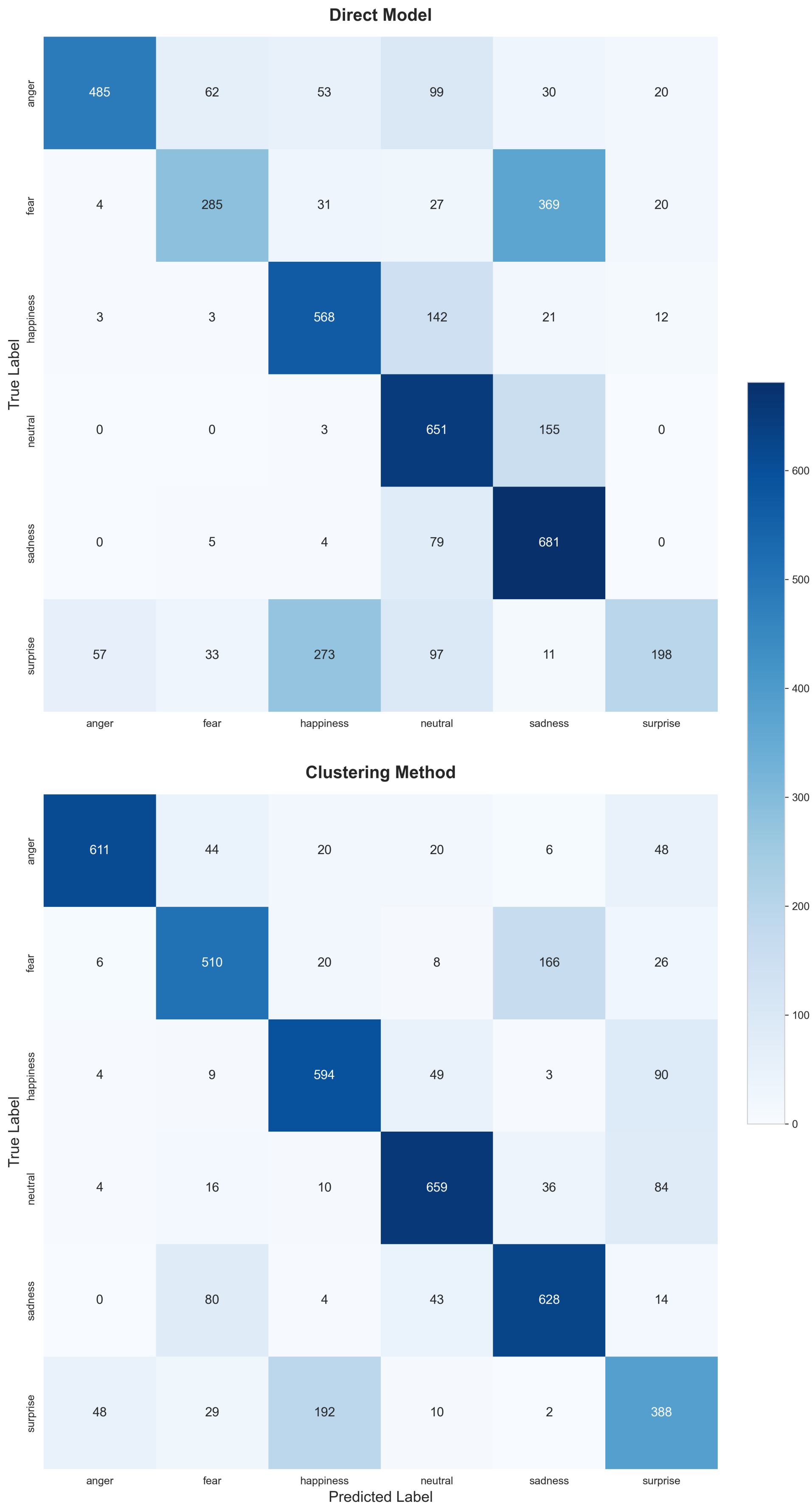

Figure 1: Confusion Matrix Comparison: Direct Model (Top) vs. Clustering Method (Bottom). The clustering method significantly reduces confusion between ’Fear’ and ’Surprise’.

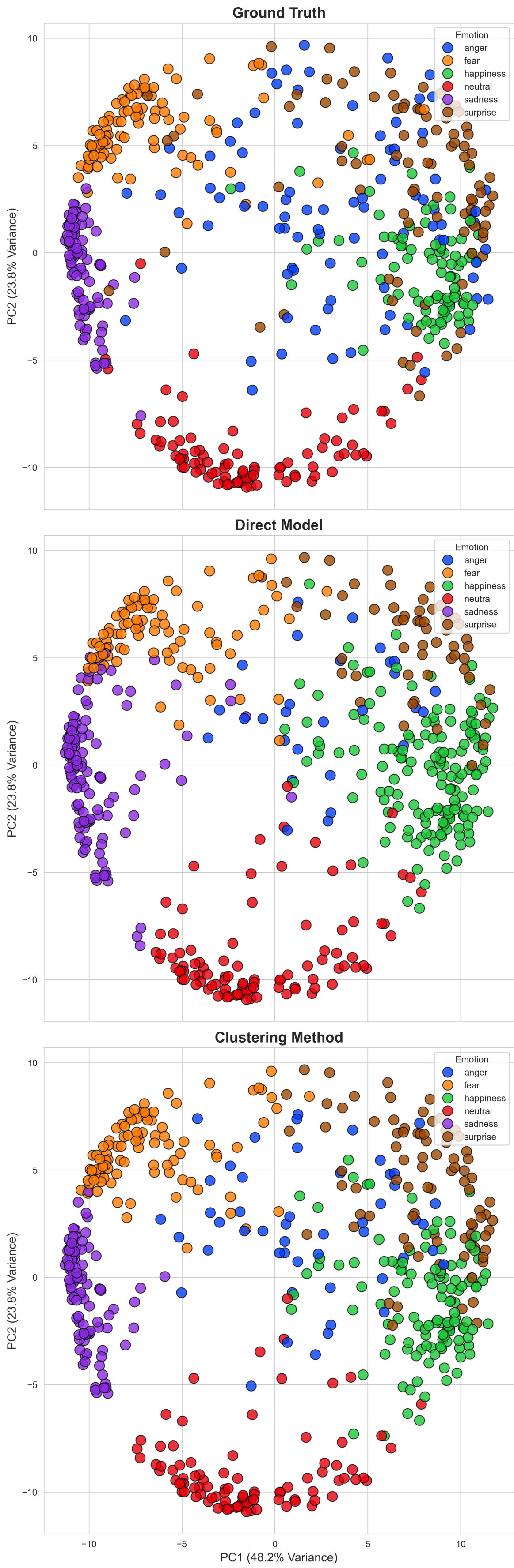

Figure 2: PCA Visualization of embeddings for one person: The clustering method (Bottom) shows better separation of emotion classes compared to the direct model (Center) and approaches Ground Truth (Top).

Algorithm 1: Greedy Cluster Assignment for Speaker-Adaptive Inference

Input : Embeddings $E = \{e_1, \ldots, e_N\}$, speaker IDs $S = \{s_1, \ldots, s_N\}$, pretrained emotion head H, maximum clusters $K = 6$

Output : Predicted emotion labels $L = \{l_1, \ldots, l_N\}$

Initialize $L \leftarrow$ array of default label (e.g., `neutral`);
Group indices by speaker: $G \leftarrow$ map from speaker s to indices i where $s_i = s$;
foreach *speaker* s *in* G **do**
    $I \leftarrow G[s]$; `// Global indices for speaker` s
    $E_s \leftarrow \{e_i \mid i \in I\}$; `// Speaker embeddings`
    $n \leftarrow |I|$;
    if $n = 1$ **then**
        Compute $p = H(E_s)$;
        $L[I[0]] \leftarrow \arg\max_e p$;
        continue ;
    $k \leftarrow \min(n, K)$;
    Standardize E_s ;
    Perform K-means clustering on E_s with k clusters $\rightarrow$ cluster labels $C \in \{0, \ldots, k-1\}^n$ and centroids M;
    Compute emotion probabilities for centroids: $P \leftarrow H(M)$;
    Initialize $cluster\_to\_emotion \leftarrow \emptyset$, $used\_emotions \leftarrow \emptyset$;
    Create list of tuples $(P[c, e], c, e)$ for all clusters c and emotions e;
    Sort list in descending order by probability;
    foreach *tuple* $(prob, c, e)$ *in sorted list* **do**
        if $c \notin cluster\_to\_emotion$ ***and*** $e \notin used\_emotions$ **then**
            $cluster\_to\_emotion[c] \leftarrow e$;
            Add e to $used\_emotions$;
            if $|cluster\_to\_emotion| = k$ **then**
                break
    `// Fallback: Assign remaining clusters to their highest probability emotion`
    for $c \leftarrow 0$ **to** $k-1$ **do**
        if $c \notin cluster\_to\_emotion$ **then**
            $cluster\_to\_emotion[c] \leftarrow \arg\max_e P[c, e]$;
    `// Map local cluster labels back to global utterance indices`
    for $j \leftarrow 0$ **to** $n-1$ **do**
        $i \leftarrow I[j]$; `// Get global index`
        $c_{\text{label}} \leftarrow C[j]$; `// Get local cluster label`
        $L[i] \leftarrow cluster\_to\_emotion[c_{\text{label}}]$;
return L;

References

Artem Amentes, Nikita Davidchuk, and Ilya Lubenets. 2022. Russian Emotional Speech Dialogs with annotated text.

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.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*.

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.

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.

Philippe Gournay, Olivier Lahaie, and Roch Lefebvre. 2018. A Canadian French Emotional Speech Dataset. In *Proceedings of the 9th ACM Multimedia Systems Conference*, MMSys '18, pages 399–402, New York, NY, USA. Association for Computing Machinery.

Wei-Ning Hsu, Benjamin 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. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Mana Ihori, Taiga Yamane, Naotaka Kawata, Naoki Makishima, Tomohiro Tanaka, Satoshi Suzuki, Shota Orihashi, and Ryo Masumura. 2025. Few-shot personalization via in-context learning for speech emotion recognition based on speech-language model. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. Accepted.

Jesin James, Li Tian, and Catherine Watson. 2018. An Open Source Emotional Speech Corpus for Human Robot Interaction Applications. pages 2768–2772.

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

Vladimir Kondratenko, Artem Sokolov, Nikolay Karpov, Oleg Kutuzov, Nikita Savushkin, and Fyodor Minkin. 2022. Large raw emotional dataset with aggregation mechanism. *arXiv preprint arXiv:2212.12266*.

Yinghao Aaron Li, Cong Han, Vinay S. Raghavan, Gavin Mischler, and Nima Mesgarani. 2023. StyleTTS 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. *arXiv preprint arXiv:2306.07691*.

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 *22nd International Conference on Data Engineering Workshops (ICDEW'06)*, pages 8–8.

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.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. SpecAugment: A simple data augmentation method for automatic speech recognition. In *Proceedings of Interspeech*, pages 2613–2617.

Jiacheng Shi, Hongfei Du, Y. Alicia Hong, and Ye Gao. 2025. EMO-TTA: Improving test-time adaptation of audio-language models for speech emotion recognition. *arXiv preprint arXiv:2509.25495*.

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 (P&P-9)*, pages 83–84, Zurich, Switzerland. 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.