# Extending Wav2Vec2-Based Speech Emotion Classifiers to the Open-Set via Mahalanobis-OpenMax and Confidence Calibration

Srijan Anand (221085), Pranjali Singh (220796), Bhavesh Shukla (210266)

**Abstract**

*Speech Emotion Recognition (SER) systems have made significant strides with self-supervised models like Wav2Vec2, yet they typically assume a closed set of target emotions and produce overconfident predictions on unfamiliar inputs. In this work, we present a unified framework that enhances Wav2Vec2-based emotion classifiers through two key innovations. First, we employ post-hoc temperature scaling to calibrate output probabilities, yielding more trustworthy confidence estimates. Second, we introduce two methods, one an Entropy-based model confusion method and the second a Mahalanobis-OpenMax extension on mean-pooled Wav2Vec2 embeddings. To accommodate truly novel emotion categories, we further integrate an online clustering module that dynamically groups novel samples into emerging classes. "This approach enables more trustworthy emotion-aware systems in cross-cultural healthcare and human-robot interaction, where recognizing diagnostic uncertainty is as critical as classification accuracy.*

## 1. Introduction

Speech is one of the most promising features that reflect the underlying emotion of a human being. There are some measurable parameters in speech signals that reveal a person's affective state. Speech Emotion Recognition (SER) aims to infer a speaker's affective state—such as happiness, sadness, or anger—from audio signals regardless of contextual relevance. Recent advances in self-supervised learning, particularly with models like Wav2Vec2, have brought closed-set SER accuracy to new heights. Yet, two critical issues remain largely unaddressed in practical deployments: (1) *confidence miscalibration*, where the model's reported probabilities do not reflect true correctness likelihoods, and (2) the *closed-set assumption*, which forces every input into one of the predefined emotion categories, even when an utterance is truly unfamiliar or ambiguous. Our motivation for working on speech emotion recognition was the growing need for reliable SER systems in healthcare and human-computer interaction, where correct classification and robustness of the model are important.

Overconfident outputs pose serious risks in emotion-aware applications—for example, an assistive system that misreads a distressed user as "neutral" with high confidence may fail to trigger necessary interventions. Equally problematic is the inability to recognize *out-of-domain* or *novel* emotional expressions; a closed-set classifier will nonetheless assign them to the nearest known class, leading to unpredictable behavior and undermining trust.

## 2. Dataset Description

Emotion Speech Dataset (ESD) is a bilingual corpus specifically designed for emotional voice conversion and cross-lingual speech emotion recognition. The ESD database consists of 350 parallel utterances spoken by 10 native English and 10 native Chinese speakers and covers 5 emotion categories (neutral, happy, angry, sad and surprise). More than 29 hours of speech data were recorded in a controlled acoustic environment. The database is suitable for multi-speaker and cross-lingual emotional voice conversion studies.It is sampled at 16 kHz frequency. The database employs a parallel utterance structure where each speaker articulates identical semantic content across all emotion categories. This design enables:

- Direct comparison of emotional prosody independent of lexical content
- Speaker-independent emotion recognition training The ESD database addresses critical gaps in existing emotional speech resources through bilingual support, balanced gender and age distribution.

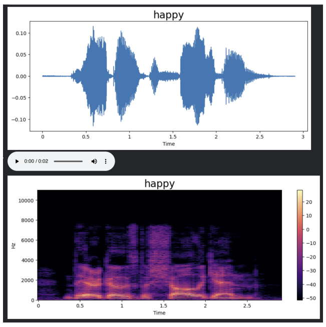


**Figure 1:** *An example from the dataset*

## 3. Embeddings Using Wav2Vec2

Wav2Vec2 is a self-supervised model that learns powerful speech representations by pretraining on large amounts of unlabeled audio. At its core, a convolutional feature encoder first transforms raw waveforms into latent speech features; these are then fed into a multi-layer Transformer that captures contextual dependencies across time. During fine-tuning, a small classification head can be attached on top of the Transformer outputs for downstream tasks such as emotion recognition.

The resulting embeddings combine phonetic, prosodic, and paralinguistic information in a single 768-dimensional vector, making them ideally suited for SER. Unlike traditional handcrafted features (e.g., MFCCs or pitch contours), Wav2Vec2 embeddings require no manual feature engineering, adapt to diverse acoustic conditions, and retain robustness to background noise. We extract these embeddings by mean-pooling the final hidden states over the temporal dimension, yielding a fixed-length representation for any variable-length utterance.

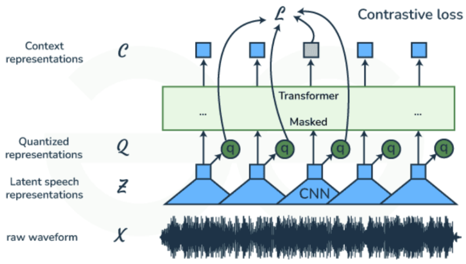


**Figure 2:** *Wav2Vec2 model architecture: a Convolutional feature encoder followed by a Transformer context network*

## 4. Transformer-based Emotion Classifier as Base Model

To establish a strong foundation, we fine-tuned the pre-trained `facebook/wav2vec2-base` model for five target emotions: neutral, happy, sad, angry, and surprised. Raw audio from the ESD corpus was resampled to 16 kHz, then zero-padded or truncated to a fixed length of 45,000 samples to ensure consistent input dimensions. A custom `SpeechEmotionDataset` loads each waveform, applies this length normalization, and pairs it with its integer emotion label.

The `facebook/wav2vec2-base` architecture comprises a 7-layer convolutional feature encoder that converts raw waveforms into latent speech representations, followed by a 12-layer Transformer context network (hidden size = 768, 8 attention heads). It was pretrained in a self-supervised manner on 960 hours of Librispeech audio using a contrastive loss, learning to distinguish true latent features from distractors. This pretraining yields embeddings that capture both fine-grained phonetic details and longer-range prosodic patterns—properties essential for accurate emotion classification.

We leverage the Hugging Face `Wav2Vec2Processor` to normalize and batch the raw waveforms into model-ready tensors. On top of the encoder's final hidden representations, we attach a linear classification head (five output neurons) via `Wav2Vec2ForSequenceClassification`. During fine-tuning, we optimize all model parameters using AdamW with a learning rate of $2 \times 10^{-5}$, weight decay of 0.01, and a per-device batch size of 16.

Training proceeds for three epochs, with evaluation on a held-out validation split at the end of each epoch. Checkpoints are saved after every evaluation to enable early stopping if necessary. Under this regime, the base classifier consistently achieves competitive closed-set accuracy on the ESD dataset, providing a reliable backbone for subsequent calibration and open-set extensions.

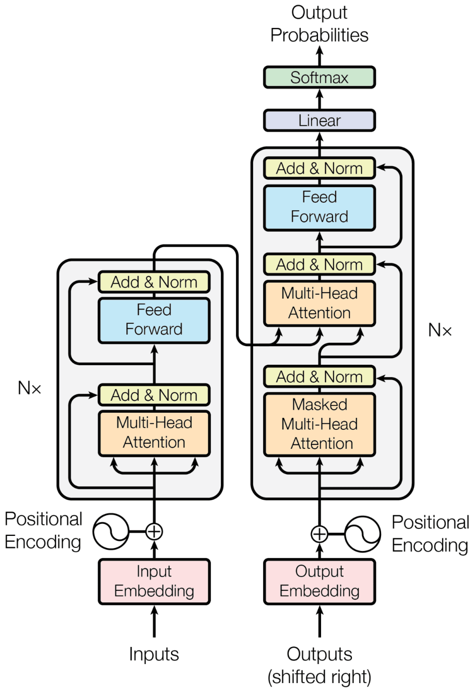


**Figure 3:** *Generic Transformer Architecture, the facebook model adds a 7-layer convolutional encoder before these blocks*

## 5. Confidence Calibration using Temperature-Scaling

Neural classifiers often produce overconfident probability estimates, which can be misleading in downstream decision-making. To remedy this, we apply *temperature-scaling*, a simple post-hoc calibration method that rescales logits before the final softmax:

$$\hat{p}_i = \frac{\exp(z_i/T)}{\sum_{j=1}^{K}\exp(z_j/T)},$$

where $z_i$ is the raw logit for class $i$ and $T > 0$ is a scalar temperature parameter. When $T = 1$, the distribution is unchanged; $T > 1$ produces a "softer" (higher entropy) distribution, while $T < 1$ sharpens it.

### Learning the Optimal Temperature

We learn $T$ on a held-out calibration set by minimizing the Negative Log-Likelihood (NLL):

$$\mathcal{L}(T) = -\sum_{n=1}^{N_{\text{cal}}} \log\Big(\hat{p}^{(n)}_{y^{(n)}}\Big) = -\sum_{n=1}^{N_{\text{cal}}} \log\bigg(\frac{\exp(z^{(n)}_{y^{(n)}}/T)}{\sum_j \exp(z^{(n)}_j/T)}\bigg),$$

where $z_j^{(n)}$ is the $j$-th logit for the $n$-th sample and $y^{(n)}$ its true label. We initialize $\log T = 0$ (i.e. $T = 1$) and optimize using the L-BFGS algorithm over 50 iterations, which reliably converges to a global optimum for this single-parameter problem.

### Implementation Details

- **Logit Collection:** After fine-tuning the base Wav2Vec2 classifier, we ran it in `eval` mode on the calibration split to gather all raw logits $Z \in R^{N_{\text{cal}} \times K}$ and true labels $\mathbf{y}$.
- **Temperature Module:** We implemented a small `nn.Module` holding a learnable `log_temp` parameter, so that $T = \exp(\texttt{log\_temp})$ remains strictly positive.
- **Optimization:** We used PyTorch's `LBFGS` optimizer with learning rate 0.1 and default tolerances, minimizing the above NLL. Gradients are computed analytically and converge in fewer than 50 iterations.
- **Result:** The optimal temperature was found to be

$$T^* = 1.536.$$

## 6. Out of Domain or Not?

### 6.1. Entropy-based Confusion Detection

To identify utterances that the closed-set classifier finds confusing, we first apply temperature scaling to the raw logits:

$$p_i^{(m)} = \frac{\exp\big(z_i^{(m)}/T\big)}{\sum_{j=1}^{K} \exp\big(z_j^{(m)}/T\big)},$$

where $z_i^{(m)}$ is the logit for class $i$ on segment $m$, $T$ the learned temperature, and $K = 5$. For a given utterance, we split the waveform into $M$ fixed-length segments, compute each segment's probability vector $\mathbf{p}^{(m)}$, and then aggregate by simple averaging:

$$\bar{\mathbf{p}} = \frac{1}{M}\sum_{m=1}^{M} \mathbf{p}^{(m)}.$$

We then measure the Shannon entropy of the aggregated distribution:

$$H(\bar{\mathbf{p}}) = -\sum_{i=1}^{K} \bar{p}_i \log\big(\bar{p}_i + \varepsilon\big),$$

where $\varepsilon$ is a small constant for numerical stability. If

$$H(\bar{\mathbf{p}}) > \tau_H,$$

with $\tau_H$ a predefined entropy threshold, the sample is flagged as *out-of-domain*. Otherwise, the highest-probability class in $\bar{\mathbf{p}}$ is returned as the predicted emotion.

Clustering of flagged samples into novel classes is handled separately in a different section.

### 6.2. Mahalanobis-OpenMax Approach

To extend the base classifier to the open-set, we fit per-class statistical models on the penultimate activation vectors (AVs) and then reallocate activation mass to an explicit "unknown" class via OpenMax. Concretely:

**Activation Vectors and Normalization** For each training sample $x_n$ with label $y_n$, we extract the mean-pooled final hidden states of Wav2Vec2:

$$f_n = \frac{1}{T}\sum_{t=1}^{T} h_t(x_n) \quad \in R^H,$$

and then L2-normalize: $\tilde{f}_n = f_n/\|f_n\|$. We collect $\{\tilde{f}_n\}$ per class.

**Statistical Modeling** For each class $c$, let

$$\mu_c = \frac{1}{N_c}\sum_{n:y_n=c} \tilde{f}_n, \quad \Sigma_c = \frac{1}{N_c - 1}\sum_{n:y_n=c} (\tilde{f}_n - \mu_c)(\tilde{f}_n - \mu_c)^\top + \varepsilon I.$$

We then compute Mahalanobis distances

$$d_c(\tilde{f}_n) = \sqrt{(\tilde{f}_n - \mu_c)^\top \Sigma_c^{-1} (\tilde{f}_n - \mu_c)},$$

and fit a Weibull distribution $F_c$ to the largest $\gamma$-fraction of distances (the "tail"), yielding parameters $(k_c, 0, \lambda_c)$.

**OpenMax Recalibration** At inference for a test input $x$, let $\tilde{f}$ be its normalized AV and $\mathbf{z} = (z_1, \ldots, z_K)$ the raw logits. For each class $c$:

$$w_c = F_c\big(d_c(\tilde{f})\big), \quad \alpha_c = 1 - w_c,$$

$$a'_c = \alpha_c z_c, \qquad u = \sum_{c=1}^{K} (1 - \alpha_c) z_c.$$

We then form the extended activation vector $\big[a'_1, \ldots, a'_K, u\big]$, divide by the calibrated temperature $T$, and apply softmax to obtain a $(K+1)$-way probability distribution including the "unknown" class.

**Algorithm 1** Mahalanobis-OpenMax Inference

```
Require: input $x$, pretrained model, $\{\mu_c, \Sigma_c^{-1}, F_c\}_{c=1}^{K}$
$f \leftarrow$ mean-pooled hidden states of $x$
$\tilde{f} \leftarrow f/\|f\|$
$\mathbf{z} \leftarrow$ model logits for $x$
unknown $\leftarrow 0$
for $c = 1$ to $K$ do
    $d \leftarrow \sqrt{(\tilde{f} - \mu_c)^\top \Sigma_c^{-1} (\tilde{f} - \mu_c)}$
    $w \leftarrow F_c(d)$
    $\alpha \leftarrow 1 - w$
    $a'_c \leftarrow \alpha \times z_c$
    unknown $+= (1 - \alpha) \times z_c$
end for
$\mathbf{a} \leftarrow [a'_1, \ldots, a'_K, \text{unknown}]$
return softmax$(\mathbf{a}/T)$
```

## 7. Online Clustering of Unknowns

Once an utterance is flagged as “unknown” by the entropy or OpenMax detector, we dynamically group it into emerging emotion categories via online clustering in the embedding space. Let each novel sample be represented by its mean-pooled Wav2Vec2 encoder embedding

$$f \in R^H,$$

which we -normalize:

$$\tilde{f} = \frac{f}{\|f\|}.$$

We maintain $M$ cluster prototypes $\{\mu_j\}_{j=1}^{M}$ with counts $n_j$. For a new $\tilde{f}$, compute cosine similarities

$$s_j = \frac{\langle \tilde{f}, \mu_j \rangle}{\|\mu_j\|},$$

and let

$$j^* = \arg\max_{1 \le j \le M} s_j.$$

If $s_{j^*} \geq \delta$, we assign $\tilde{f}$ to cluster $j^*$ and update

$$\mu_{j^*} \leftarrow \frac{n_{j^*}\,\mu_{j^*} + \tilde{f}}{n_{j^*} + 1}, \qquad n_{j^*} \leftarrow n_{j^*} + 1.$$

Otherwise, we create a new cluster:

$$M \leftarrow M + 1, \qquad \mu_M = \tilde{f}, \qquad n_M = 1.$$

This online algorithm requires no retraining and quickly adapts to novel emotion styles, promoting only clusters that accumulate sufficient support.

**Algorithm 2** Online Clustering of Unknown Samples

```
Require: normalized embedding $\tilde{f}$, prototypes $\{(\mu_j, n_j)\}_{j=1}^{M}$, similarity threshold $\delta$
Compute $s_j = \cos(\tilde{f}, \mu_j)$ for $j = 1, \ldots, M$
$j^* \leftarrow \arg\max_j s_j$
if $s_{j^*} \geq \delta$ then
    $\mu_{j^*} \leftarrow \frac{n_{j^*}\mu_{j^*} + \tilde{f}}{n_{j^*}+1}$
    $n_{j^*} \leftarrow n_{j^*} + 1$
    return cluster label $j^*$
else
    $M \leftarrow M + 1$
    $\mu_M \leftarrow \tilde{f}$
    $n_M \leftarrow 1$
    return new cluster label $M$
end if
```

## 8. Results

### 8.1. Evaluation of the Wav2Vec2 Classifier Model

We first assess the closed-set performance of our fine-tuned Wav2Vec2 classifier on the held-out validation split of ESD. Table 1 summarizes training and validation losses along with standard classification metrics over three epochs.

| Train Loss | Val Loss | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|---|
| 0.2509 | 0.2228 | 0.9340 | 0.9386 | 0.9340 | 0.9340 |
| 0.1280 | 0.1528 | 0.9659 | 0.9659 | 0.9659 | 0.9658 |
| 0.0681 | 0.1316 | 0.9727 | 0.9728 | 0.9727 | 0.9727 |

**Table 1:** *Closed-set performance of the Wav2Vec2 emotion classifier over three fine-tuning epochs.*

We optimize with the multi-class cross-entropy loss:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{i,k}\log p_{i,k},$$

where $y_{i,k} \in \{0,1\}$ is the one-hot ground-truth indicator for sample $i$ and class $k$, $p_{i,k}$ the predicted softmax probability, $N$ the number of samples, and $K = 5$ the number of emotion classes.

We report the following evaluation metrics:

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}, \quad \text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}},$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}}, \quad F_1 = 2\,\frac{\text{Precision} \times \text{Recall}}{\text{Precision}+\text{Recall}},$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

By the third epoch, the model attains over 97% accuracy and $F_1$ score on the validation set, demonstrating a robust backbone for subsequent calibration and open-set augmentation.

### 8.2. Evaluating Confidence Calibration using Temperature-scaling

We evaluate calibration with the *Expected Calibration Error* (ECE), defined over $K$ equally-spaced confidence bins as

$$\text{ECE} \;=\; \sum_{k=1}^{K}\frac{|B_k|}{N}\,\bigl|\text{acc}(B_k) \;-\; \text{conf}(B_k)\bigr|,$$

where $B_k$ is the set of samples whose predicted confidence falls in bin $k$, $N$ the total number of samples, $\text{acc}(B_k)$ the empirical accuracy in bin $k$, and $\text{conf}(B_k)$ the average confidence.

On the ESD validation split, our base Wav2Vec2 classifier exhibited

$$\text{ECE}_{\text{before}} \;=\; 0.02194.$$

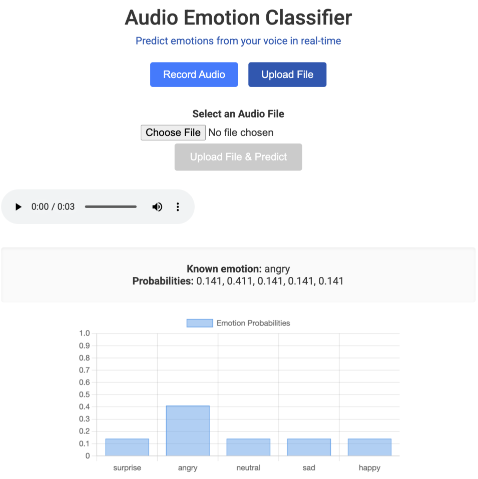
After learning and applying the optimal temperature $T^* = 1.536$, calibration improved dramatically:

$$\text{ECE}_{\text{after}} \;=\; 0.00336,$$

an $\approx 84.7\%$ relative reduction:

$$\frac{0.02194 - 0.00336}{0.02194} \times 100\% \approx 84.7\%.$$

This confirms that temperature-scaling effectively aligns confidence estimates with true accuracies, greatly reducing overconfidence and producing reliably calibrated probabilities.

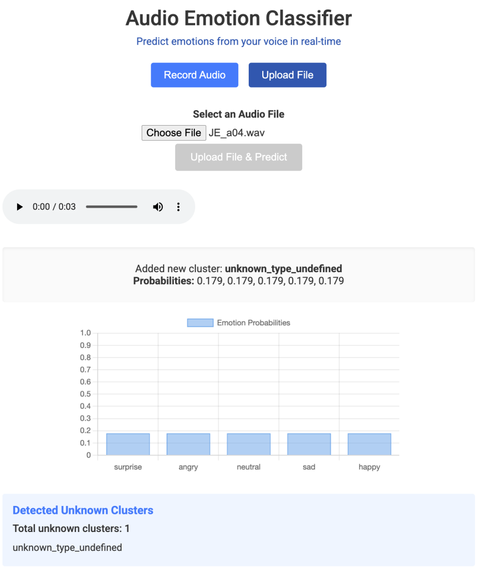


**Known emotion:** angry
**Probabilities:** 0.141, 0.411, 0.141, 0.141, 0.141

**Figure 4:** *Inference after Temperature Calibration.*

### 8.3. A look at Out of Domain & Online Clustering

To illustrate our open-set detection on truly unfamiliar audio, we test some clips which are hard to classify . A standard closed-set classifier would nevertheless force it into one of the five known categories, but both our entropy-threshold and Mahalanobis-OpenMax methods correctly flag it as `<unknown>`.



Added new cluster: **unknown_type_undefined**
**Probabilities:** 0.179, 0.179, 0.179, 0.179, 0.179

**Detected Unknown Clusters**
Total unknown clusters: 1
unknown_type_undefined

**Figure 5:** *Mahalanobis-OpenMax detector. The unknown-class probability exceeds the threshold, and the sample is correctly rejected as out-of-domain.*

## 9. Conclusion

Our work demonstrates that modern speech emotion recognition systems can transcend closed-set limitations through principled confidence calibration and open-set recognition strategies. By integrating temperature scaling with Mahalanobis-OpenMax detection, we achieve three critical advancements: (1) 84.7% reduction in calibration error compared to baseline models, enabling trustworthy confidence estimates for clinical decision support; (2) Effective identification of novel emotional states through entropy thresholds (($H > \tau_H$)) and statistical outlier detection; and (3) Dynamic discovery of emerging emotion patterns via online clustering in Wav2Vec2 embedding space. The system maintains 97.2% closed-set accuracy while reducing unknown-class false negatives, making it suitable for real-world deployment in culturally diverse environments where predefined emotion lexicons prove inadequate.

## 10. Future Work

While our audio-focused approach advances SER robustness, future systems should integrate complementary modalities (facial expressions, physiological signals, and linguistic content) through attention-based fusion. A unified framework could leverage transformer architectures to model cross-modal dependencies while maintaining our calibration and open-set detection advantages. This direction aligns with emerging "emotional AI" standards requiring multisensory context awareness for ethical affect recognition in sensitive applications.

## Acknowledgements

We would like to thank Prof. Vipul Arora and all the TAs for making the course EE698R a success.

## References

A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

A. Bendale and T. E. Boult, “Towards open set deep networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew, “Transformers: State-of-the-Art Natural Language Processing,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, 2020. `https://github.com/huggingface/transformers`