Appendix: Formal Proof of the theoretical guarantee of the empirical method implemented in Voice-Response-Based Emotion Intensity Classification in Assistive Robots

Hoashalarajh Rajendran, H. M. Ravindu T. Bandara, A. G. B. P. Jayasekara, D. P. Chandima

Aug 2023

Goal: To prove that the probability of selecting an incorrect label decreases exponentially as the number of aggregated audio segments increases.

Setup and Feature Extraction

Let T be the total duration of a given audio response. The audio is partitioned into N non-overlapping segments of length $w = 1.2$ seconds, such that $N = \lfloor T/w \rfloor$.

Let each 1.2 s segment i (where $i = 1, \ldots, N$) be composed of $m = 40$ low-level acoustic frames, each of duration 30 ms. We denote the raw acoustic feature matrix for segment i as $X_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,m}]$.

For each segment i, we extract a statistical feature vector V_i consisting of the temporal mean and standard deviation across the m frames:

$$V_i = [\mu(X_i), \sigma(X_i)]$$

where $\mu(X_i) = \frac{1}{m}\sum_{k=1}^{m} x_{i,k}$ and $\sigma(X_i) = \sqrt{\frac{1}{m}\sum_{k=1}^{m}(x_{i,k} - \mu(X_i))^2}$.

Let $h : \mathcal{V} \to \mathcal{L}$ represent the trained Machine Learning classifier, which maps the statistical feature space to the discrete label space $\mathcal{L} = \{l_1, l_2, l_3, l_4\}$ of size $C = 4$. For each segment i, the classifier outputs a predicted label:

$$Y_i = h(V_i)$$

Let the true label of the entire audio response be $l^* = l_1$.

Assumptions

1. **Independence:** The random variables $Y_1, \ldots, Y_N$ are independent and identically distributed (i.i.d.).

2. **Per-Segment Accuracy:** We assume that $m = 40$ frames provide a sufficiently stable empirical estimate of the acoustic distribution such that the statistical feature vector V_i allows the classifier to predict the true label with probability p:

   $$Pr(Y_i = l^*) = p$$

   For any incorrect class $j \neq 1$, let $q_j = Pr(Y_i = l_j)$. We define the classification margin as $\mu = \min_{j \neq 1}(p - q_j)$, and assume $\mu > 0$.

Plurality Decision

Define the empirical counts of the predicted labels across all segments:

$$N_j = \sum_{i=1}^{N} \mathbb{I}\{Y_i = l_j\}, \quad j = 1, \ldots, C$$

The overall label for the audio is determined by the plurality vote: $\hat{Y} = \arg\max_{1 \le j \le C} N_j$.

Theoretical Guarantee

Lemma 1 (Pairwise Margin)**.** *For each $j \neq 1$, define the pairwise difference random variable:*

$$Z_i^{(j)} = \mathbb{I}\{Y_i = l^*\} - \mathbb{I}\{Y_i = l_j\}$$

Then $Z_1^{(j)}, \ldots, Z_N^{(j)}$ are i.i.d., take values in $\{-1, 0, 1\}$, and their expected value is $\mathbb{E}[Z_i^{(j)}] = p - q_j \geq \mu$.

Theorem 1 (Error Bound of Plurality Aggregation)**.** *Given a classification margin $\mu > 0$, the probability that the plurality vote fails to select the true emotion-intensity label l^* is exponentially bounded by:*

$$Pr(\hat{Y} \neq l^*) \leq (C-1) \exp\left(-\frac{N\mu^2}{2}\right)$$

Proof. The plurality decision fails if at least one competing class $j \neq 1$ receives as many or more votes than the true class 1. That is, the failure event is $\exists j \neq 1 : N_1 \leq N_j$.

Let $S_N^{(j)} = \sum_{i=1}^{N} Z_i^{(j)} = N_1 - N_j$. The condition $N_1 \leq N_j$ is equivalent to $S_N^{(j)} \leq 0$.

Since the variables $Z_i^{(j)}$ are independent, bounded in $[-1, 1]$, and have an expectation of at least μ, we can apply Hoeffding's Inequality to bound the probability for a specific incorrect class j:

$$Pr(S_N^{(j)} \leq 0) = Pr(S_N^{(j)} - N(p-q_j) \leq -N(p-q_j)) \leq \exp\left(-\frac{2(N(p-q_j))^2}{N(1-(-1))^2}\right) = \exp\left(-\frac{N(p-q_j)^2}{2}\right)$$

Since $p - q_j \geq \mu$, we have:

$$Pr(S_N^{(j)} \leq 0) \leq \exp\left(-\frac{N\mu^2}{2}\right)$$

To find the total probability of failure across all $C-1$ competing classes, we apply the Union Bound:

$$Pr(\hat{Y} \neq l^*) = Pr(\exists j \neq 1 : S_N^{(j)} \leq 0) \leq \sum_{j \neq 1} Pr(S_N^{(j)} \leq 0) \leq (C-1) \exp\left(-\frac{N\mu^2}{2}\right)$$

This completes the proof. □

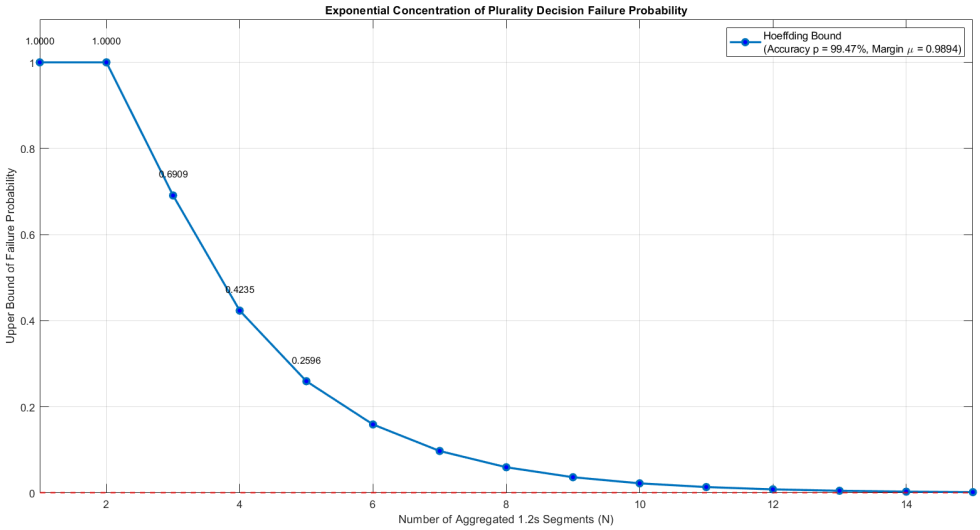

Figure 1: The curve shows the theoretical upper bound derived from Hoeffding's Inequality. The probability of an incorrect plurality vote decays exponentially as the number of aggregated 1.2 s segments (N) increases. The theoretical guarantee ensures the failure probability drops below 0.5 at $N = 4$ which corresponds to 4.8 s of audio. Near-absolute theoretical certainty is guaranteed after aggregating 8 to 10 segments which corresponds to 9.6 to 12 s of audio.