A Segment Level Approach to Speech Emotion Recognition using Transfer Learning: Supplementary Material

Sourav Sahoo[1][0000−0002−1956−2834], Puneet Kumar[2][0000−0002−4318−1353], Balasubramanian Raman[2][0000−0001−6277−6267] and Partha Pratim Roy[2][0000−0002−5735−5254]

[1] Dept. of Electrical Engineering, Indian Institute of Technology Madras, Chennai - 600036, India
ee17b040@smail.iitm.ac.in

[2] Dept. of Computer Science and Engineering, Indian Institute of Technology Roorkee, Roorkee - 247667, India
pkumar99@cs.iitr.ac.in,{balarfcs,proy.fcs}@iitr.ac.in

1 Example

In Table 1, we present the details of three clips that we obtained during the training process. All the arrays discussed in this material are 0-indexed.

Table 1. Test results for three randomly sampled audio clips

Clip	# Seg.	TC	Logit_0	Logit_1	Logit_2	Logit_3	Argmax of Logits
Clip # 1	4	0	0.50603974	-1.2699697	-4.3870134	-2.116255	0
			-1.16288	-0.24785714	-3.5117776	-2.2011812	1
			-1.6501933	-0.39175096	-1.8593655	-1.8029484	1
			-2.6610787	0.51635414	-3.0119264	-1.2123884	1
Clip # 2	3	3	-4.5419283	-0.60939735	-3.8518085	0.9972298	3
			-3.9391785	-0.46409646	-4.1760426	0.41488993	3
			-2.4461453	-0.6418519	-3.5422876	0.2814951	3
Clip # 3	3	3	-2.681143	-0.85285175	-4.1699333	0.95221466	3
			-1.5041095	0.5352828	-3.0267582	-1.6476822	1
			-0.43402833	-1.061139	-3.498886	-0.6317687	0

where # Seg.= Number of segments in the clip and TC = True Class of the clip.

2 Calculations

This section demonstrates the methods used to calculate the various accuracies defined in Section 3.3. The entire details are described in Table 3,4 of the paper.

2.1 Segment Accuracy

$$\text{total number of segments} = 4 + 3 + 3 = 10$$

$$\text{total number of segments predicted correctly} = 5$$

$$\text{Segment Accuracy} = 5/10 = 0.500$$

2.2 Absolute Clip Accuracy

Out of the three clips, only in the second case the model is able to predict the correct classes correctly for *each and every* segment of the clip. Hence,

$$\text{Absolute Clip Accuracy} = 1/3 = 0.333$$

2.3 Standard Clip Accuracy

In first clip, the mode of prediction is 1 but the correct class is 0. For the second clip, the mode as well as the correct class is 3. In the third case, there is no single mode as all the values appear maximum number of times (1 each). So, we state that the model has classified the third clip *incorrectly*. Hence,

$$\text{Standard Clip Accuracy} = 1/3 = 0.333$$

2.4 Average Logits Clip Accuracy

For the first clip, the average logits array:

$$\mathcal{A} = [-1.24202806, -0.34830592, -3.19252072, -1.83319325]$$

$$\underset{x}{\operatorname{argmin}}\, \mathcal{A} = 1$$

For the second clip, the average logits array:

$$\mathcal{B} = [-3.64241737, -0.5717819, -3.8567129, 0.56453828]$$

$$\underset{x}{\operatorname{argmin}}\, \mathcal{B} = 3$$

For the third clip, the average logits array:

$$\mathcal{C} = [-1.53976028, -0.45956932, -3.5651925, -0.44241208]$$

$$\underset{x}{\operatorname{argmin}}\, \mathcal{C} = 3$$

Hence,

$$\text{Average Logits Clip Accuracy} = 2/3 = 0.667$$

2.5 Best Clip Accuracy

In the first clip, 1 is predicted for maximum number of times but the correct class is 0. For the second clip, 3 is predicted for maximum number of times as well as the correct class is 3. In the third case, 0,1 and 3 are predicted maximum number of times (1 each). As the correct class $3 \in \{0, 1, 3\}$, we state that the model has classified the third clip *correctly*. Hence,

$$\text{Best Clip Accuracy} = 2/3 = 0.667$$