MODELING SPEECH EMOTION WITH LABEL VARIANCE AND ANALYZING PERFORMANCE ACROSS SPEAKERS AND UNSEEN ACOUSTIC CONDITIONS

Vikramjit Mitra, Amrit Romana, Dung T. Tran & Erdrin Azemi
Apple
Cupertino, CA 95014, USA
{vmitra,aromana,dung_tran,eazemi}@apple.com

ABSTRACT

Spontaneous speech emotion data usually contain perceptual grades where graders assign emotion score after listening to the speech files. Such perceptual grades introduce uncertainty in labels due to grader opinion variation. Grader variation is addressed by using consensus grades as groundtruth, where the emotion with the highest vote is selected. Consensus grades fail to consider ambiguous instances where a speech sample may contain multiple emotions, as captured through grader opinion uncertainty. We demonstrate that using the probability density function of the emotion grades as targets instead of the commonly used consensus grades, provide better performance on benchmark evaluation sets compared to results reported in the literature. We show that a saliency driven foundation model (FM) representation selection helps to train a state-of-the-art speech emotion model for both dimensional and categorical emotion recognition. Comparing representations obtained from different FMs, we observed that focusing on overall test-set performance can be deceiving, as it fails to reveal the models generalization capacity across speakers and gender. We demonstrate that performance evaluation across multiple test-sets and performance analysis across gender and speakers are useful in assessing usefulness of emotion models. Finally, we demonstrate that label uncertainty and data-skew pose a challenge to model evaluation, where instead of using the best hypothesis, it is useful to consider the 2- or 3-best hypotheses.

1 INTRODUCTION

Speech-based emotion models aim to estimate the emotional state of a speaker from their speech utterances. Real-time speech-emotion models can help to improve human-computer interaction Mitra et al. (2019); Kowtha et al. (2020) and facilitate health applications Stasak et al. (2016); Niu et al. (2023); Provost et al. (2024). Speech emotion research has pursued two distinct definitions of emotion: (1) *categorical emotions*: for example, fear, anger, joy, sadness, disgust, and surprise Ekman (1992), and (2) *dimensional emotions*: that represent emotion using a 3-dimensional model of *Valence*, *Activation* and *Dominance* Posner et al. (2005). Early studies on speech emotion detection focused on acted or elicited emotions Busso et al. (2008), however, models trained with acted emotions often fail to generalize for spontaneous emotions Douglas-Cowie et al. (2005). Recently, attention has been given to datasets with spontaneous emotions Mariooryad et al. (2014) where graders listen to each audio file and assign emotion labels. Such perceptual grading is difficult due to utterances containing mixed, shifting, subtle, or ambiguous emotions. To account for this, Mariooryad et al. have multiple graders review and grade each audio file. Traditionally, researchers addressed label variance by taking the grader consensus Chou et al. (2024). However, modeling such variance Prabhu et al. (2022); Chou et al. (2024); Tavernor et al. (2024) can be useful to account for audio samples that were perceptually difficult to annotate. In this work, we investigate training models with distributions of grader decisions for categorical emotions, instead of consensus grades, as the target. We hypothesize that modeling label uncertainty can help to improve the model's robustness because consensus grades fail to account for mixed, shifting, subtle, or ambiguous emotions.

Recent studies have shown that pre-trained foundation model (FM) representations are useful for emotion recognition from speech Srinivasan et al. (2022); Mitra et al. (2022; 2023). Given that

the FMs may not have been trained with emotion labels, the final layer representations may not be optimal for emotion recognition. Earlier studies have investigated intermediate FM representations for various speech tasks Alain & Bengio (2016); Mitra & Franco (2020); Mitra et al. (2024a); Yang et al. (2024). In this work, we investigate saliency based FM layer selection for the downstream emotion modeling task. To summarize, in this work, we:

1. Account label uncertainty through the use of categorical emotion pdf as targets.
2. Explore saliency-driven intermediate FM layer representations for emotion recognition.
3. Evaluate performance across speakers, gender and unseen acoustic conditions.

We observed that models that provide state-of-the-art (SOTA) results, may not generalize well across speakers and varying acoustic conditions. We found that having a diverse evaluation set along with a diverse evaluation metric is useful for model selection. We found that the traditional 1-best hypothesis used in emotion literature may get biased by the training data-skew, in which case 2- or 3-best hypotheses may be useful to account for speech samples containing multiple emotions.

2 DATA

We have used the MSP-Podcast dataset (ver. 1.11) Mariooryad et al. (2014); Lotfian & Busso (2017) that contains ≈ 238 hours of speech data spoken by English speakers ($N > 1800$), consisting of $\approx 152K$ speaking turns. The speech segments contain single-speaker utterances with a duration of 3 to 11 seconds. The data contain manually assigned valence, activation and dominance scores and categorical emotions (9 categories) from multiple graders. Grader decisions for categorical emotions were converted to a pdf (reflecting the probability of each of the 9 emotions), which was used as the target for our model training. The data split is shown in Table 4 in Appendix A.2. To make our results comparable to Ghriss et al. (2022); Srinivasan et al. (2022), we report results on Eval1.6 and Eval1.11 (see Table 4). For evaluating model robustness, we have added noise to the MSP test-set at SNR levels 15 dB and 5dB (see $Eval_{15dB}$ and $Eval_{5dB}$ in Table 4, Appendix A.2). We report categorical emotion recognition performance on six emotions: neutral, happy, angry, sad, contempt and surprise. We have used CMU-Mosei, Zadeh et al. (2018) and a 5 hour in-house conversational speech data from 85 speakers for cross-corpus speech emotion recognition analysis.

3 REPRESENTATIONS

We explore speech embeddings as features to a TC-GRU model (see Figure 1). We use the following pre-trained models to generate those embeddings: (i) **HuBERT** large Hsu et al. (2021), a transformer based acoustic model, pre-trained on 60K hours of Libri-light speech data, generating 1024-dimensional embedding. (ii) **WavLM** large Chen et al. (2022), a transformer based acoustic model, generating 1024 dimensional embedding. WavLM has been pre-trained on 60K hours of Libri-light, 19K hours of GigaSpeech and 25K hours of VoxPopuli. (iii) **Whisper** medium Radford et al. (2023) acoustic model that generates 1024 dimensional embeddings from 24 transformer encoder layers. Whisper is trained with 680K hours of noisy and diverse speech data from the web.

Motivated by Mitra et al. (2024b;a) we explore obtaining layer-saliency to obtain the optimal FM layer representation for emotion modeling. Let the N dimensional representation from the k^{th} layer of a FM for an utterance y be represented by a vector $H_k^y(t) = [X_{1,k}, \ldots, X_{t,k}, \ldots, X_{M,k}]$, where M denotes the sequence length. For a regression task, let the sequence targets be L^y, where $L^y \in \mathbb{R}^D$, where the D dimensional vector L denotes the output targets, for each utterance. $\overline{H}_k^y$ in eq. 1 is obtained from H_k^y by taking the mean across all the frames for utterance y. The cross-correlation based saliency (CCS) of i^{th} dimension of the k^{th} layer is given by:

$$S_{CCS,i,k} = \left| \frac{Cov(\overline{H}_{k,i}^y, L^y)}{\sigma_{H_{k,i}^y} \sigma_{L^y}} \right| + \gamma_i, \quad where, \;\; \overline{H}_k^y = \frac{1}{M} \sum_{t=1}^{M} H_k^y(t) \tag{1}$$

$$\gamma_i = \frac{1}{N-1} \sum_{j=1, j \neq i}^{N} w_j \left\| \frac{Cov(\overline{H}_{i,k}^y, \overline{H}_{j,k}^y)}{\sigma_{\overline{H}_i^y} \sigma_{\overline{H}_j^y}} \right\|, \quad where, \;\; w_j = \left\| \frac{Cov(\overline{H}_j^y, L^y)}{\sigma_{\overline{H}_j^y} \sigma_{L^y}} \right\| \tag{2}$$

$$\mu_{CCS,k} = \frac{1}{D} \sum_{l=1}^{D} S_{CCS_{k,l}}. \tag{3}$$

γ_i is the sum of the weighted cross-correlation between the i^{th} dimension and all other dimensions, as shown in eq. 2. In our experiments we have used $\mu_{CCS,k}$ given in eq. 3 to select salient layers of a pre-trained FM, which is obtained from a randomly sampled 30K utterances in the Train1.11.

3.1 Model Training

We have trained a multi-task (dimensional and categorical) emotion recognition model. It consists of temporal convolution (kernel size of 3), followed by a 2-layered gated recurrent unit (TC-GRU) network, consisting of 256 neurons in each layer and an embedding layer of 256 neurons. The model architecture is illustrated in Fig. 1 and the model parameters are described in Appendix A.8. The model was trained with Train1.11 data (see Table 4), where the performance on Valid1.11 set was used for model selection and early stopping. Concordance correlation coefficient (CCC) Lawrence & Lin (1989) is used as the loss function, see Appendix A.1. Models were trained with a mini-batch of 32 and a learning rate of 0.0005.

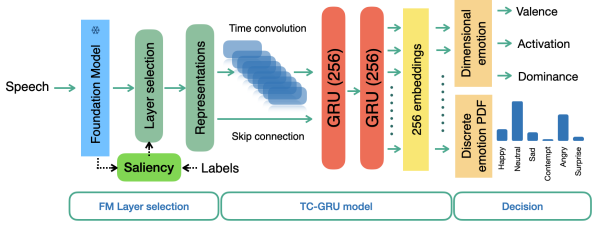

Figure 1: Multi-task emotion recognition model

4 Results

We trained multi-task emotion recognition models with embeddings from HuBERT, WavLM, and Whisper FMs. In addition, we trained a baseline model with mel-filterbank and pitch (MFBF0) feature. In Table 1, we report dimensional emotion estimation performance obtained from the trained systems and compared them with the state-of-the-art results reported in the literature (see Table 1). Note that in Srinivasan et al. (2022) ASR generated transcripts were used, which was not used for the other systems in Table 1. Finally, we compared categorical emotion recognition performance obtained from the TC-GRU models with respect to results reported in the literature (see Table 2).

Table 1: Dimensional emotion estimation performance ($CCC \uparrow$) and comparison with SOTA

Systems	**Eval1.6**			**Eval1.11**			**Eval$_{15dB}$**			**Eval$_{5dB}$**		
	Act.	Val.	Dom.	Act.	Val.	Dom.	Act.	Val.	Dom.	Act.	Val.	Dom.
MFBF0 TC-GRU	0.73	0.34	0.66	0.62	0.39	0.56	0.69	0.26	0.61	0.53	0.14	0.48
HuBERT TC-GRU	**0.77**	0.65	**0.70**	**0.66**	0.59	**0.59**	0.74	0.62	0.64	0.61	0.54	0.49
WavLM TC-GRU	**0.77**	0.70	**0.70**	**0.66**	0.63	0.58	0.73	**0.71**	**0.66**	0.62	0.64	0.53
Whisper TC-GRU	0.75	**0.71**	0.69	0.65	**0.64**	0.58	0.73	**0.71**	**0.66**	**0.66**	**0.69**	**0.60**
Mitra et al. (2024b)	0.75	0.66	0.67	-	-	-	-	-	-	-	-	-
Srinivasan et al. (2022)	**0.77**	0.69	0.68	-	-	-	-	-	-	-	-	-

Table 2: Categorical emotion recognition performance and comparison with SOTA models

Systems	**Eval1.6**		**Eval1.11**		**Eval$_{15dB}$**		**Eval$_{5dB}$**		*Mosei*		*Inhouse*	
	$F1_m$	UAR	$F1_m$	UAR	$F1_m$	UAR	$F1_m$	UAR	$F1_m$	UAR	$F1_m$	UAR
MFBF0 TC-GRU	0.34	0.45	0.44	0.58	0.46	0.53	0.46	0.50	0.40	0.56	0.19	0.34
HuBERT TC-GRU	0.49	0.67	0.48	0.67	0.49	0.65	0.50	0.64	**0.48**	0.66	0.58	0.64
WavLM TC-GRU	0.50	**0.70**	0.48	**0.69**	0.50	**0.69**	0.50	**0.68**	0.47	**0.69**	**0.61**	**0.67**
Whisper TC-GRU	**0.52**	0.69	**0.50**	0.68	**0.52**	0.68	**0.52**	0.67	**0.48**	0.68	0.59	0.65
Das et al. (2024)	-	0.67	-	-	-	-	-	-	-	-	-	-
Feng & Narayanan (2023)	-	0.67	-	-	-	-	-	-	-	-	-	-
Wu et al. (2024)	0.35	-	-	-	-	-	-	-	-	-	-	-

Next we investigated how these models perform across speakers, where we accumulated model decisions by speaker, and computed the UAR for the categorical emotion predictions. We have used Eval1.11 and Inhouse sets to compare the performance of the models. For performance evaluation across speakers, we introduced a metric: paUAR-X, which measures the percentage of speakers

who are above a UAR of X%, where we have used two thresholds: X: 75% and 50%, respectively. Table 3 shows paUAR-75 and paUAR-50 for categorical emotion, obtained across speakers. Note that Tables 1 and 2 show that overall WavLM TC-GRU model performed better than the HuBERT TC-GRU, however table 3 shows that a better system may not necessarily generalize across speakers.

Table 3: Emotion recognition performance across speakers
where paUAR-X is the percentage of speakers who are above a UAR of X%.

Eval Sets		**MFBF0 TC – GRU**		**HuBERT TC – GRU**		**WavLM TC – GRU**		**Whisper TC – GRU**	
	hyps.	paUAR-75	paUAR-50	paUAR-75	paUAR-50	paUAR-75	paUAR-50	paUAR-75	paUAR-50
`Eval1.11`	1-best	2.3	21.1	3.1	34.4	3.1	40.6	4.7	39.1
`Inhouse`		2.3	14.0	10.0	30.5	9.3	35.0	11.6	39.5
`Eval1.11`	2-best	11.7	46.9	21.1	61.7	28.1	68.8	22.7	68.0
`Inhouse`		23.0	100.0	93.0	100.0	100.0	100.0	100.0	100.0
`Eval1.11`	3-best	20.3	68.0	36.0	86.7	46.1	85.9	36.7	89.8
`Inhouse`		53.1	100.0	95.4	100.0	100.0	100.0	100.0	100.0

In terms of the 1-best hypothesis paUAR-75 and paUAR-50, Whisper TC-GRU model performed better than the others, likely due the fact it was pre-trained with a noisy, more diverse, and larger set of speech. However, even with this best performing model, only 5% and 12% of speakers had UAR above 0.75 for Eval1.11 and Inhouse sets, respectively. In Appendices A.5 and A.6, we explore potential explanations for the speaker-level performance differences including whether gender or emotion label distributions play a role. We find that gender has a significant impact on results, where 7% of female speakers had UAR above 0.75 compared to $\approx 14\%$ of male speakers for the Inhouse evaluation set. This gap illustrates the importance of evaluating model performance at speaker and group levels. Interestingly, even if Tables 1 and 2 show that WavLM TC-GRU model overall performed much better than MFBF0 TC-GRU, their paUAR-75 were comparable for Eval1.11, indicating that the usage of overall metrics while assessing the usefulness of a model can be deceiving. Also note that the speaker level performance obtained from Eval1.11 and the inhouse set was quite different for each of the models investigated, where the performance for Eval1.11 was found to be lower, as it is a harder and larger containing more speakers than the inhouse set (see table 4 in A.2). Note that for Eval1.11, the best model demonstrated an UAR above 0.75 for only $\approx 5\%$ of the speakers. The poor performance across speakers can be attributed to the uncertainty in the labels and the overall skew toward "neutral" emotion. For example, in many instances different graders assigned different emotions to the same speech file, which reveals that a speech file can contain a mix-of-emotions due to mixed, shifting, subtle, or ambiguous emotions. Additionally, data skew due to one emotion category being present overwhelmingly in the training set (e.g., "neutral") can lead the model to over-estimate that emotion, in which case a 1-best hypothesis may lead to pessimistic results. Appendix A.7 illustrates the relationship between 1-best and 2-best hypothesis, and how by studying both we can obtain better clarity regarding the models generalization capacity. Table 3, we explored paUAR-X if the target emotion exists within the 2-best or 3-best hypotheses. We find a paUAR-75 of more than 28% can be obtained by considering the 2-best hypothesis and as high as 46% can be obtained with a 3-best hypothesis. These findings indicate that (1) in case of data with uncertain labels and distribution skew, it is helpful to consider multiple model hypothesis and (2) label distribution skew impacts model's generalization capacity across speakers.

5 Conclusions

In this work, we demonstrated SOTA results for both dimensional and categorical emotion recognition. The models were found to perform well for unseen datasets (Mosei and Inhouse) and demonstrated reasonable noise robustness. Interestingly, the models failed to generalize across speakers, where we observed that the model performed with an overall UAR of above 0.75 for less than 10% of the speakers. The model offered UAR above 0.5 for $\approx 60\%$ of the speakers. This indicated that using metrics that reflect the overall performance on an eval set may not be prudent, speaker-level and gender-level performance are crucial to assess how well the model will perform across users. We also observed that instead of using the 1-best hypothesis from the model, it is useful to consider 2-best or 3-best hypothesis, as certain utterances may contain multiple emotions, in which case the model may provide more than one likely emotion categories. With 2-best and 3-best hypothesis, we observed that UAR above 0.75 was obtained for $> 60\%$ and $> 85\%$ of the speakers, respectively. The findings from this study opens the question regarding performance metrics, which can account for co-occurrences of semantically closer emotions, such as "angry", "contempt", "disgust", which may have a higher chance of confusion with each other.

REFERENCES

G. Alain and Y. Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, J.N. Kim, S.and Chang, S. Lee, and S.S. Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359, 2008.

S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

Huang-Cheng Chou, Lucas Goncalves, Seong-Gyun Leem, Ali N Salman, Chi-Chun Lee, and Carlos Busso. Minority views matter: Evaluating speech emotion classifiers with human subjective annotations by an all-inclusive aggregation rule. *IEEE Transactions on Affective Computing*, 2024.

N. Das, S. Dingliwal, S. Ronanki, R. Paturi, D. Huang, P. Mathur, J. Yuan, D. Bekal, X. Niu, S.M. Jayanthi, et al. Speechverse: A large-scale generalizable audio language model. *arXiv preprint arXiv:2405.08295*, 2024.

E. Douglas-Cowie, L. Devillers, J.C. Martin, R. Cowie, S. Savvidou, S. Abrilian, and C. Cox. Multimodal databases of everyday emotion: Facing up to complexity. In *Proc. of Interspeech*, 2005.

P. Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

T. Feng and S. Narayanan. Peft-ser: On the use of parameter efficient transfer learning approaches for speech emotion recognition using pre-trained speech models. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 1–8. IEEE, 2023.

A. Ghriss, B. Yang, V. Rozgic, E. Shriberg, and C. Wang. Sentiment-aware automatic speech recognition pre-training for enhanced speech emotion recognition. *Proc. of ICASSP*, pp. 7347–7351, 2022.

W.N. Hsu, B. Bolte, Y.H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

V. Kowtha, V. Mitra, C. Bartels, E. Marchi, S. Booker, W. Caruso, S. Kajarekar, and D. Naik. Detecting emotion primitives from speech and their use in discerning categorical emotions. In *Proc. of ICASSP*, pp. 7164–7168. IEEE, 2020.

I. Lawrence and K. Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, pp. 255–268, 1989.

R. Lotfian and C. Busso. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Trans. on Affective Computing*, 10(4): 471–483, 2017.

S. Mariooryad, R. Lotfian, and C. Busso. Building a naturalistic emotional speech corpus by retrieving expressive behaviors from existing speech corpora. In *Proc. of Interspeech*, 2014.

V. Mitra and H. Franco. Investigation and analysis of hyper and hypo neuron pruning to selectively update neurons during unsupervised adaptation. *Digital Signal Processing*, 99:102655, 2020.

V. Mitra, W. Wang, C. Bartels, H. Franco, and D. Vergyri. Articulatory information and multiview features for large vocabulary continuous speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5634–5638. IEEE, 2018.

V. Mitra, S. Booker, E. Marchi, D.S. Farrar, U.D. Peitz, B. Cheng, E. Teves, A. Mehta, and D. Naik. Leveraging acoustic cues and paralinguistic embeddings to detect expression from voice. *Proc. Interspeech*, pp. 1651–1655, 2019.

V. Mitra, H.Y.S. Chien, V. Kowtha, J.Y. Cheng, and E. Azemi. Speech emotion: Investigating model representations, multi-task learning and knowledge distillation. *Proc. of Interspeech*, 2022.

V. Mitra, V. Kowtha, H.Y.S. Chien, E. Azemi, and C. Avendano. Pre-trained model representations and their robustness against noise for speech emotion analysis. In *Proc. of ICASSP*, pp. 1–5. IEEE, 2023.

V. Mitra, A. Chatterjee, K. Zhai, H. Weng, A. Hill, N. Hay, et al. Pre-trained foundation model representations to uncover breathing patterns in speech. *arXiv preprint arXiv:2407.13035*, 2024a.

V. Mitra, J. Nie, and E. Azemi. Investigating salient representations and label variance in dimensional speech emotion analysis. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 11111–11115. IEEE, 2024b.

Minxue Niu, Amrit Romana, Mimansa Jaiswal, Melvin McInnis, and Emily Mower_Provost. Capturing mismatch between textual and acoustic emotion expressions for mood identification in bipolar disorder. In *Interspeech*. Interspeech, 2023.

J. Posner, J.A. Russell, and B.S. Peterson. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, 17(3):715–734, 2005.

N.R. Prabhu, N. Lehmann-Willenbrock, and T. Gerkmann. Label uncertainty modeling and prediction for speech emotion recognition using t-distributions. In *Proc. of ACII*, pp. 1–8. IEEE, 2022.

Emily Mower Provost, Sarah H Sperry, James Tavernor, Steve Anderau, Anastasia Yocum, and Melvin G McInnis. Emotion recognition in the real-world: Passively collecting and estimating emotions from natural speech data of individuals with bipolar disorder. *IEEE Transactions on Affective Computing*, 2024.

A. Radford, J.W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.

S. Srinivasan, Z. Huang, and K. Kirchhoff. Representation learning through cross-modal conditional teacher-student training for speech emotion recognition. *Proc. of ICASSP*, pp. 6442–6446, 2022.

B. Stasak, J. Epps, N. Cummins, and R. Goecke. An investigation of emotional speech in depression classification. In *Proc. of Interspeech*, pp. 485–489, 2016.

James Tavernor, Yara El-Tawil, and Emily Mower Provost. The whole is bigger than the sum of its parts: Modeling individual annotators to capture emotional variability. *arXiv preprint arXiv:2408.11956*, 2024.

H. Wu, H-C. Chou, K-W. Chang, L. Goncalves, J. Du, J-S.R. Jang, C-C. Lee, and H-Y. Lee. Emosuperb: An in-depth look at speech emotion recognition. *arXiv preprint arXiv:2402.13018*, 2024.

S-W. Yang, H-J. Chang, Z. Huang, A.T. Liu, et al. A large-scale evaluation of speech foundation models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

A.B. Zadeh, P.P. Liang, S. Poria, E. Cambria, and L.-P. Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018.

A Appendix

A.1 Concordance Correlation Coefficient

Concordance correlation coefficient based loss (L_{ccc}) is defined by:

$$L_{ccc} = -\frac{1}{N}\sum_{i=1}^{N} CCC_i \tag{4}$$

where L_{ccc} is the mean of CCC's obtained from each of the N output targets. CCC is defined by:

$$CCC = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}. \tag{5}$$

where μ_x and μ_y are the means, σ_x^2 and σ_y^2 are the corresponding variances for the estimated and groundtruth variables, and ρ is the correlation coefficient between them.

A.2 DATA SPLIT

Table 4: MSP-podcast data split, noise-degraded test sets and out-of-domain MOSEI and Inhouse evaluation set

Split	**Hours**	**Speakers**	**Description**
MSP Train1.11	135.4	1411	Training set
MSP Valid1.11	31.7	456	Validation set
MSP Eval1.6	16.6	51	Podcast1.6 evaluation set
MSP Eval1.11	48.9	244	Podcast1.11 Eval set 1
MSP $Eval_{15dB}$	28.4	51	Eval + noise within 10-20 dB
MSP $Eval_{5dB}$	28.4	51	Eval + noise within 0-10 dB
CMU-Mosei	70.6	-	Mosei segments
Inhouse data	5.0	85	Conversational speech segments

A.3 LAYER SALIENCY MEASURE

Neural saliency was used in Mitra et al. (2024b) to reduce the number of representations for the downstream task with a goal of model size reduction. "Saliency" in this work focuses on layer-saliency as outlined in section 3, where the saliency measure was modified to provide a layer-wise collective measure, that informs which transformer layer in the foundation model is more relevant. This measure is particularly important, as given the large number of transformer layers in an FM, it may not be possible to perform layer-wise experimentation of which layer offers the best representation. Layer-wise saliency measure offers a data-driven solution to figure out which layers in the transformer network are better suited for the downstream task, without the need to train downstream models for representations from each individual layer.

We observed that valence is more sensitive to transformer layer representation, compared to activation and dominance (see Figure 2). Earlier studies Chen et al. (2022) have found that for WavLM intermediate layers (specifically layers 19 and 20) are better for intent classification. Valence plays an important role in emotion discrimination, such as Happy versus Angry or Sad versus Calm. In Figure 3 we show how saliency based on individual valence, happy and angry scores vary by WavLM transformer layer representation. Figures 2 and 3 show that intermediate transformer layers of WavLM offer better representations (paralinguistic cues) for downstream emotion detection compared to the final layer. We observed that the intermediate layers correlated strongly with articulatory features (extracted using the model in Mitra et al. (2018)), speech rate, pitch and voicing information, compared to the final layer.

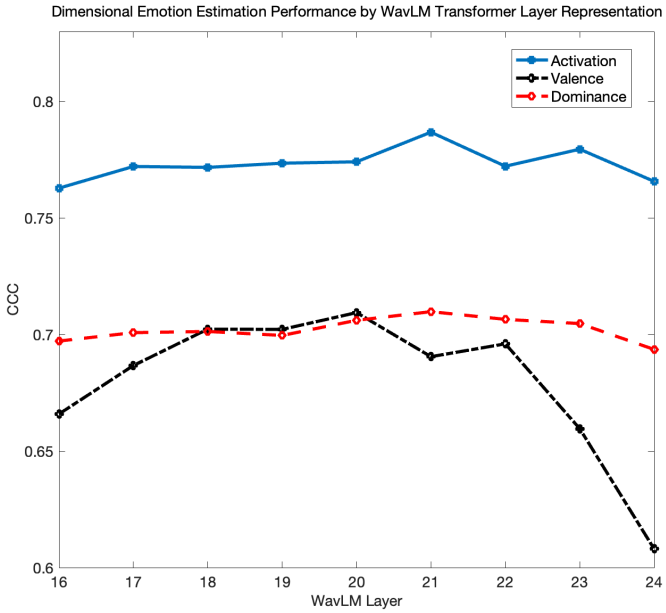

Figure 2: Dimensional emotion estimation for different transformer layers in WavLM

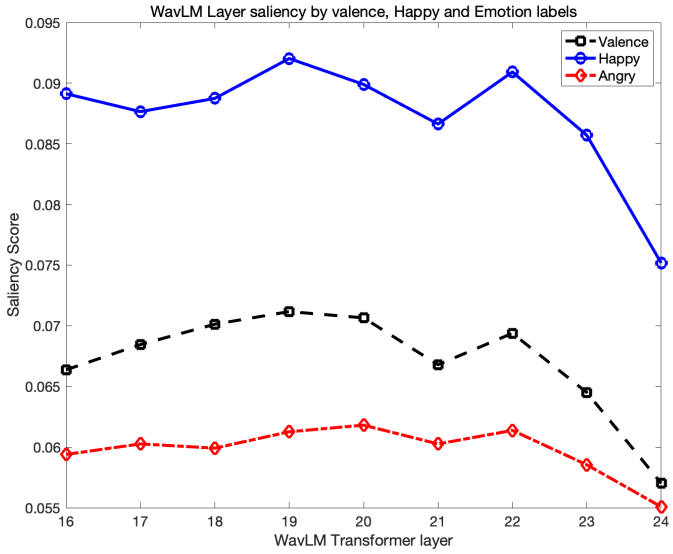

Figure 3: WavLM layer saliency by valence, happy and angry emotion

A.4 Emotion Model Details

Table 5 shows that representations from emotion-salient layer as compared to the final FM layer, resulted in improvement in emotion recognition performance. It is also interesting to note that relative improvement in valence was higher ($> 8\%$ relative) compared to the other dimensional

emotions. For unseen-noise sets ($Eval_{15dB}$ and $Eval_{5dB}$), the relative improvement was higher (16.5% for dimensional and 10% for categorical emotion) than other evaluation sets. [1]

Table 5: Dimensional and categorical emotion estimation using (1) MFBF0 feature, (2) FM representations from final layer and (3) FM representations from the salient layer

Test set	Reps.	Layer	Dim. Emo. CCC ↑			Cat. Emo. UAR ↑
			Act.	Val.	Dom.	
Eval1.3	MFBF0	–	0.73	0.33	0.67	0.55
	HuBERT	Final	0.75	0.60	0.69	0.65
		Salient	**0.78**	0.65	**0.71**	0.66
	WavLM	Final	0.77	0.61	0.70	0.66
		Salient	0.77	0.70	**0.71**	**0.71**
	Whisper	-	0.76	**0.71**	0.69	**0.71**
Eval1.6	MFBF0	–	0.73	0.34	0.66	0.54
	HuBERT	Final	0.75	0.60	0.68	0.64
		Salient	**0.77**	0.65	**0.70**	0.66
	WavLM	Final	0.76	0.61	0.69	0.65
		Salient	**0.77**	0.70	**0.70**	**0.70**
	Whisper	-	0.75	**0.71**	0.69	0.61
Eval1.11	MFBF0	–	0.62	0.39	0.56	0.58
	HuBERT	Final	0.64	0.55	0.57	0.65
		Salient	**0.66**	0.59	**0.59**	0.67
	WavLM	Final	0.65	0.57	0.58	0.65
		Salient	**0.66**	0.63	0.58	**0.69**
	Whisper	-	0.65	**0.64**	0.58	0.68
Eval$_{15dB}$	MFBF0	–	0.69	0.26	0.61	0.50
	HuBERT	Final	0.70	0.57	0.60	0.62
		Salient	**0.74**	0.62	0.64	0.64
	WavLM	Final	0.72	0.58	0.62	0.63
		Salient	0.73	**0.71**	**0.66**	**0.68**
	Whisper	-	0.73	**0.71**	**0.66**	0.67
Eval$_{5dB}$	MFBF0	–	0.53	0.14	0.48	0.44
	HuBERT	Final	0.56	0.49	0.43	0.56
		Salient	0.61	0.54	0.49	0.60
	WavLM	Final	0.61	0.50	0.48	0.60
		Salient	0.62	0.64	0.53	**0.66**
	Whisper	-	**0.66**	**0.69**	**0.60**	0.65
Mosei	MFBF0	–	-	-	-	0.56
	HuBERT	Final	-	-	-	0.64
		Salient	-	-	-	0.66
	WavLM	Final	-	-	-	0.66
		Salient	-	-	-	**0.69**
	Whisper	-	-	-	-	0.68

A.5 Performance by Gender

Table 6 shows performance variance across male and female speakers for Eval1.11 and Inhouse test sets. We find performance is considerably lower for female speakers across both datasets, and the gap between performance for male and female speakers increases with the paUAR threshold. The training set is skewed toward male speakers, which likely contributes to the observation in Table 6.

[1] performance gains from the salient FM-layer representations were statistically significant ($p < 0.05$) compared to the results reported in the literature

Table 6: Emotion recognition performance (paUAR-75 and paUAR-50) by gender for Whisper TC-GRU model

Eval Sets	paUAR-75		paUAR-50	
	Female	Male	Female	Male
`Eval1.11`	2.7	7.1	45.8	50.4
`Inhouse`	7.1	13.8	28.6	44.8

A.6 Performance by Speaker's Emotion Distributions

Figure 4 shows performance plotted against emotion distributions for each speaker in Eval1.6. Because UAR is the unweighted average across recall on all emotions, we do not find a strong relationship between UAR and emotion distribution. This suggests UAR is robust to these speaker-level changes and can capture other important factors in speaker-level performance.

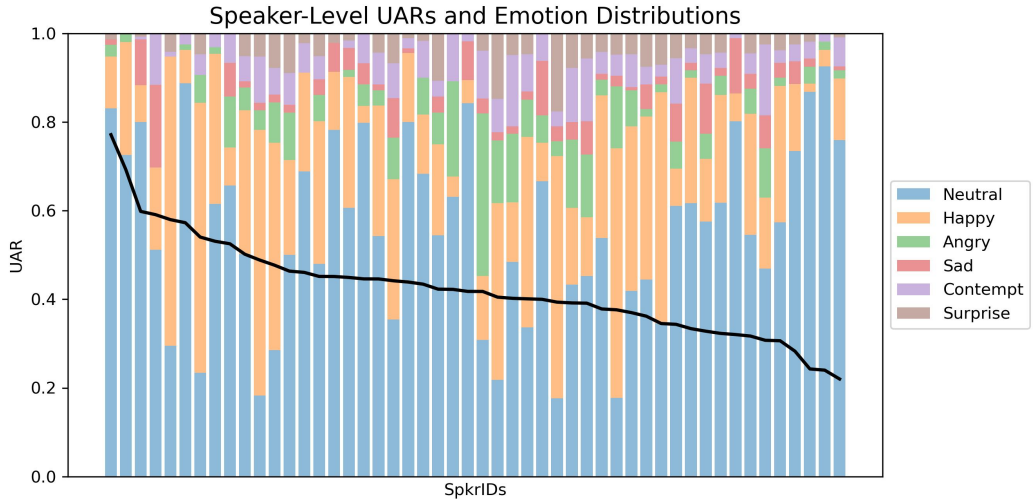

Figure 4: Speaker-level performance (UAR from Whisper TC-GRU) plotted against emotion distributions, for speakers in Eval1.6.

A.7 Relationship Between 1st and 2nd Best Model Hypotheses

We find that the model's first and second hypotheses show a clear relationship, and that the first hypothesis alone may not fully reflect the model's understanding. Figure 5 illustrates these details, with the samples accurately labeled by the first hypothesis outlined by the horizontal gray bars, and the samples accurately labeled by the second hypothesis outlined by the vertical gray bars. The first hypotheses are highly accurate for happiness and anger, indicated by the white squares within the horizontal gray outlines. However, for most sadness samples, the model identifies neutral as the most likely emotion and sadness as the second most likely emotion, indicated by the white square within the vertical gray outline. Similarly, for surprise samples, the model identifies happiness as the most likely emotion and surprise as the second most likely emotion, where this hierarchy likely results from the closer relationship between happiness and surprise with the former class having more representation in the training data. We also see considerable confusion between contempt, anger, and neutral. When we explore the models second best hypotheses, we find the model correctly detects the overall sentiment but does not distinguish correctly between them. This finding supports our analysis into considering the model's second best hypotheses when determining model predictions.

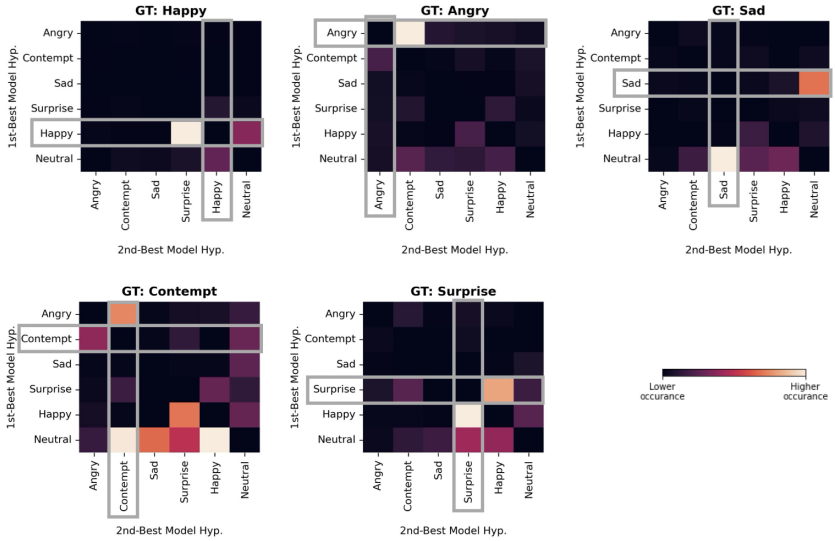

Figure 5: Confusion matrices showing the relationship between 1st and 2nd best model hypotheses from Whisper TC-GRU and the Eval1.6 test set.

A.8 Model Parameters

The TC-GRU models had 1.6M parameters (2.1MB), whereas the MFBF0 was 700KB in size, for saliency based layer selection, we were able to reduce the computation needed by WavLM (16%) and by HuBERT (8%) by reducing the number of transformer layers needed to generate the representations, see Table 7. Note that layers were all frozen for feature extraction, i.e., none of the FM transformer layers were fine-tuned for the given task as shown in Figure 1. Earlier work Mitra et al. (2024b) has shown that saliency-based representation selection can help to reduce the downstream model size, however that was not the focus of this work. The goal of this work is to investigate layers that are relevant for downstream emotion task, where joint modeling of categorical and dimensional emotion would result in better performance, as compared to using the final layers. Note that most studies have used FM final layer representations to train teacher models to distill information into simpler downstream models, in this work we show that better teacher models can be obtained by proper selection of representation layers.

Table 7: Model Parameters

Model	**Full FM Params** (Not loaded)	**Salient Layer FM Params** (Loaded, frozen)	**Saliency Size Reduction**	**Trainable Params** (TC-GRU)	**Total Loaded Model Size**
MFBF0 TC-GRU	-	-	-	0.5M	0.7MB
WavLM TC-GRU	315.5M	265.1M	16%	1.6M	2.2MB
HuBERT TC-GRU	315.4M	290.2M	8%	1.6M	2.2MB
Whisper TC-GRU	315.7M	315.7M	0%	1.6M	2.2MB