MULTI-VIEW LEARNING FOR SPEECH EMOTION RECOGNITION WITH CATEGORICAL EMOTION, CATEGORICAL SENTIMENT, AND DIMENSIONAL SCORES

Daniel Tompkins, Dimitra Emmanouilidou, Soham Deshmukh, Benjamin Elizalde

Microsoft
{daniel.tompkins, diemmano, sdeshmukh, benjaminm}@microsoft.com

ABSTRACT

Psychological research has postulated that emotions and sentiment are correlated to dimensional scores of valence, arousal, and dominance. However, the literature of Speech Emotion Recognition focuses on independently predicting the three of them for a given speech audio. In this paper, we evaluate and quantify the predictive power of the dimensional scores towards categorical emotions and sentiment for two publicly available speech emotion datasets. We utilize the three emotional views in a joined multi-view training framework. The views comprise the dimensional scores, emotions categories, and sentiment categories. We present a comparison for each emotional view or combination of, utilizing two general-purpose models for speech-related applications: CNN14 and Wav2Vec2. To our knowledge this is the first time such a joint framework is explored. We found that a joined multi-view training framework can produce results as strong or stronger than models trained independently for each view.

***Index Terms*—** Speech emotion recognition, sentiment analysis, affective computing, valence, arousal, dominance

1. INTRODUCTION

Emotion and sentiment recognition has been studied for decades in many domains including vision, language, and speech. Emotions commonly include categorical labels such as *Happy*, *Sad*, *Angry*, etc. while sentiment map emotions onto three categories: *Positive*, *Negative*, *Neutral*. In addition to these categorical labels, dimensional affect labels provide continuous values: valence describes the level of pleasantness, arousal describes the amount of excitement or energy in actions such as speaking, and dominance relates to the amount of control [1], with the latter often omitted due to its high correlation with arousal. Decades of psychological research has postulated that categorical emotions can be mapped to dimensional regions on a valence and arousal space—notably in the works related to Russell's circumplex model of affect [2].

Mainstream Machine Learing (ML) models break Speech Emotion Recognition (SER) into different tasks of detecting emotions, sentiment and emotional dimensions of valence-arousal-dominance. The SER model consists of an pretrained audio encoder trained on large scale speech data, which is then finetuned on the target speech emotion dataset. In [3], Papparagi et al. found that deep models pre-trained in speaker recognition and fine-tuned for SER performs better than a random network initialization, especially in the case of smaller or restricted datasets such as IEMOCAP [4]. Prior work in supplementary techniques such as data augmentation for Speech Emotion Recognition (SER) has showed no or little benefit, especially for larger or more diverse datasets [3, 5, 6]. With recent advances in transformers, the SER models have improved performance on predicting dimensional labels [7].

While SER research is dominated by the separate tasks of detecting emotions, sentiment, and emotional dimensions, there has not been much work in combining the continuous dimensional space with the discrete labels of emotions and/or sentiment. The psychological research evidently shows this relationship between the continuous space of emotional dimensions and the discrete emotional labels. In this paper, we have three main contributions:

- We evaluate and quantify the predictive power of the dimensional scores towards categorical emotions and categorical sentiment for two publicly available speech emotion datasets. To our knowledge this is the first time this information is presented.
- We utilize three emotional views of the available speech emotion corpora towards a joined multi-view training framework. The views comprise the dimensional scores, emotions categories, and sentiment categories. To our knowledge this is the first time such a joint framework is explored.
- We present a comparison for each emotional view or combination of, utilizing two general-purpose audio models, CNN14 and Wav2Vec2.

2. DATASETS AND METRICS

2.1. Datasets

MSP-Podcast v1.10: This large speech emotion dataset of $\sim$ 166 hours was collected from podcast recordings of more than 600 speakers [8]. In this work we considered the following 7 classes: $\{$*Happy*, *Neutral*, *Sad*, *Angry*, *Disgust Fear*, *Contempt*$\}$; these categorical labels were converted to sentiment labels as $\{$*Pos*, *Neu*, *Neg*, *Neg*, *Neg*, *Neg*, *Neg*$\}$, respectively. Segments without judge consensus or labeled *Other* were excluded, as was class *Surprise*, since it could represent either positive or negative valence. We used the standard split ($77,783$ files total), with *Test1* as the test set ($12,731$ files), and *Development* for validation ($8,283$ files).

IEMOCAP: The Interactive Emotional Dyadic Motion Capture dataset is an acted dataset of scripted and improvised dialogues by 10 speakers, with a duration of $\sim$ 12 hours [4]. We used data from 8 emotional classes (excluding *Surprise*) $\{$*Happy*, *Neutral*, *Sad*, *Angry*, *Disgust Fear*, *Excited Frustrated*$\}$. sentiment labels were created by converting class labels to $\{$*Pos*, *Neu*, *Neg*$\}$ in a similar manner as above. In total, there were $7,451$ files remaining after excluding labels *Other* or of no agreement.

The three types of emotional views used in this work are: i) dimensional scores from valence, arousal, and dominance; ii) categorical labels of emotions enumerated above, and iii) sentiment labels created by casting the emotion classes to labels of *Pos*, *Neu*, *Neg*. Depending on the modeling paradigm, data from all classes per dataset are used (*allC*), or from only the first 4 classes (*4C*), or the first 5 (*5C*) classes.

2.2. Performance metrics

We use the following performance metrics for assessing the categorical ($categ$) and sentiment ($senti$) models: sample-weighted accuracy, $\mathbf{wACC}$, which corresponds to the percentage of correctly classified samples over all samples; unweighted (or balanced) accuracy, $\mathbf{uACC}$, which is the average of the individual class accuracies and is not affected by imbalanced classes; and weighted F1 score, $\mathbf{F1w}$. For predicting continuous values of the dimensional emotions arousal, valence, and dominance, the Concordance Correlation Coefficient (CCC) [9] is used; CCC provides a measure of agreement between two sets of data, similarly to a correlation coefficient, but is more conservative as it takes into account the bias of the true values.

3. PREDICTING EMOTION AND SENTIMENT FROM DIMENSIONAL SCORES

Many psychological works have plotted emotions on the dimensional valence/arousal graph from Posner et al [10]. To compare the theoretical emotional mappings, we used the labeled data from the MSP-Podcastdataset to plot the valence and arousal scores with the categorical emotions, as shown in Fig. 1. The figure overlays the mean and standard deviation of valence/arousal scores of each emotion with a common theoretical circumplex layout. There is a clear delineation between the mean coordinates of most emotions, especially *Angry* and *Sad* but there is significant overlap between many of the emotions in the standard deviation lines. While the separation of emotions in valence/arousal space is visible, it is much less distinct than in the theoretical circumplex models.

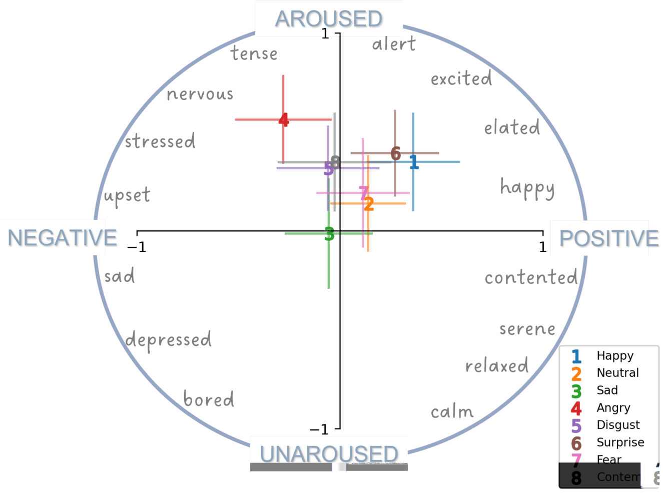

Fig. 1. The outer circle and emotions are the circumplex model of affect, adapted from [10]. The x-axis represents valence while the y-axis represents arousal. We plotted inside the circle the mean valence and arousal values and their standard deviations of 8 emotions from the MSP-Podcast dataset.

3.1. Experimental setup

In this section, we look at how predictive dimensional scores of arousal, valence and dominanceare of i) the emotional categories and ii) the sentiment categories.

We used a weighted linear Support Vector Machines (SVM) classifier with a one-vs-one multi-class strategy, with fixed γ=$1/(feat\_dim)$ and parameter C=$\arg\max_C(uACC$+$wACC)$ over the dev split set, with logarithmic search for $C \in [0.0001, 20]$. Valence (V), Arousal (A) and Dominance (D) values were used as input features, and categorical ($categ$) or sentiment ($senti$) values as output. Results are presented in Table 1, where column @$\mathbf{randm}$ shows performance at chance level. The definitions below help the reader navigate Table 3.2:

$\circ$ Subscripts $\{$V, A, D$\}$: input to the SVM model - either valence only (V) or both valence and arousal (V, A) or all three valence, arousal and dominance (V, A, D).

$\circ$ Superscripts $4C$, $5C$, and $allC$: the number of input data classes (see Section 2.1).

3.2. Results

For case $allC$, the categorical models ($categ$) have a 14.29% (12.50%) chance at random for MSP-Podcast (IEMOCAP). Sentiment models ($senti$) use the converted $\{$*Pos*, *Neu*, *Neg*$\}$

labels and have a 33.33% chance at random. Sentiment models ($senti$) were not expected to benefit from adding arousal to the feature set - a small $< 2\%$ increase in performance was observed. For categorical classification ($categ$), adding dominance showed no particular benefit ($< 3\%$ performance increase), which is in agreement with some prior behavioral and psychological studies [11, 1]. Comparing the last to the one-but-last row, we see a $\mathbf{uACC}$ increase from 34.00 to 53.29 for the {V,A,D} case compared to {V,A}, which is a misleading result of having two classes in the IEMOCAP test set with $\leq$ #2 samples (for *Disgust* and *Fear*). In all cases, the use of RBF - instead of a linear - kernel showed $< 2\%$ performance increase (not showed here).

Table 1. Predicting emotion ($categ$) and sentiment ($senti$) labels from the ground truth labels of arousal (A), valence (V), dominance (D); for 4 emotional classes ($4C$) or 5 ($5C$) or for all classes ($allC$).

	$\mathbf{uACC}$	$\mathbf{wACC}$	$\mathbf{F1w}$	$\mathbf{@randm}$
MSP-Podcast - v1.10 (%)				
SVM$^{4C}_{V}$ $senti$	72.62	71.2	71.32	33.33
SVM$^{5C}_{V}$ $senti$	71.66	70.75	70.7	33.33
SVM$^{allC}_{V}$ $senti$	68.86	67.91	67.78	33.33
SVM$^{4C}_{\{V,A\}}$ $categ$	71.81	65.24	66.56	25.00
SVM$^{4C}_{\{V,A,D\}}$ $categ$	72.13	65.42	66.82	25.00
SVM$^{5C}_{\{V,A,D\}}$ $categ$	60.63	59.17	61.18	20.00
SVM$^{allC}_{\{V,A\}}$ $categ$	43.18	49.80	51.31	14.29
SVM$^{allC}_{\{V,A,D\}}$ $categ$	43.48	45.62	47.30	14.29
IEMOCAP - Fold #1 (%)				
SVM$^{4C}_{V}$ $senti$	74.99	75.85	75.19	33.33
SVM$^{allC}_{V}$ $senti$	74.09	78.71	78.57	33.33
SVM$^{4C}_{\{V,A\}}$ $categ$	68.23	67.23	67.16	25.00
SVM$^{4C}_{\{V,A,D\}}$ $categ$	70.77	68.84	68.60	25.00
SVM$^{allC}_{\{V,A\}}$ $categ$	34.00*	42.91	45.55	12.50
SVM$^{allC}_{\{V,A,D\}}$ $categ$	53.29*	44.15	45.16	12.50

*misleading $\mathbf{uACC}$ increase caused by classes of $\leq$ #2 data samples in the test set.

Overall findings suggest that valence, arousal and dominance have ∼65-71% $\mathbf{wACC}$ predictive power for sentiment or categorical (4-class) labels, and ∼49% prediction rate for finer-grain labels (7 classes - *categ*), for MSP-Podcast. This may indicate an upper bound for sentiment or categorical SER models architectured to learn from the ground truth labels of arousal and valence.

4. MULTI-VIEW LEARNING FOR PREDICTING EMOTION, SENTIMENT AND DIMENSIONAL

4.1. Architecture

The architecture used for multi-view learning experiments are shown in Figure 2. Let the training data be $D = \{(a_i, y_i)\}_{i=1}^{i=N}$ where a_i and y_i represents the raw audio and labels respectively. Let $f(a)$ be the audio encoder. The audio encoder $f(a)$ first produces a log Mel Spectrogram from raw audio followed by a learnable embedding function. For a batch size of b, this results in:

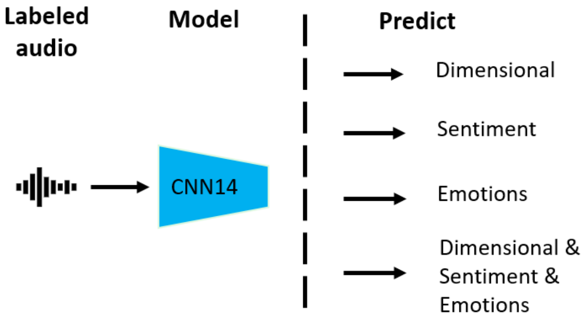

Fig. 2. Multi-view framework for prediction of Dimensional scores, Sentiment Categories, and Emotion Categories. The output layer varies per model: three regression outputs for three Dimensional classes (valence, arousal and dominance); three classes for sentiment (*Pos*, *Neu*, *Neg*); five classes for Emotions; and the combinations (Dimensional+Sentiment, Dimensional+Emotion, or all 3).

$$x_a = \{f(a_i)\}_{i=1}^{i=b} \tag{1}$$

where $x_a \in \mathbb{R}^{b \times v}$ are the audio representations of dimension v. The audio representation is then passed through a projection layer $l_a(x_a)$ as shown in Figure 2. The projection layer consists of a liner layer with ReLU activation.

$$\hat{x}_a = l_a(x_a) \tag{2}$$

where $\hat{x}_a \in \mathbb{R}^{b \times d}$ is the final representation consumed by different tasks. Let the tasks be represented by $\{t_i\}_{i=1}^{i=5}$. Each task has its independent loss $\{\ell_i\}_{i=1}^{i=5}$. We use linear layer for each task. For Emotion and sentiment tasks, the predictions are passed through Softmax activation and the choice of loss is Binary Cross Entropy. For Valence, Arousal and Dominance prediction we use CCC as loss. The loss for final model is average of individual task losses:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{i=5} \ell_i \tag{3}$$

4.2. Experimental Setup

We use sampling rate of 16 kHz for all the models. We use a Convolutional Neural Network architecture CNN14 [14] and a transformer-based model Wav2Vec2 [15] as audio encoders because these are both common model architectures used in the field. The CNN14 encoder is pretrained on AudioSet [16], while Wav2Vec2 is pretrained on 960h of speech data. The Wav2Vec2 audio encoder directly works with raw

Table 2. Rows *a* to *l* show the performance for Emotion Classification (Emo), Sentiment Classification (Senti), Dimensional regression (V/A/D), and the proposed multi-view framework (Dimensional+Sentiment, Dimensional+Emotion, or all 3), for the 5 classes of MSP-Podcast corpus ($5C$). Column 'Train/Predict' shows the input (and output) labels for each model. *indicates sentiment score derived indirectly from emotion outputs. The top section of the table references prior work.

Model	Train/Predict	Dimensional			Sentiment (%)			Emotions (%)		
		$\mathrm{CCC_V}$	$\mathrm{CCC_A}$	$\mathrm{CCC_D}$	wACC	uACC	F1w	wACC	uACC	F1w
				Prior work						
preCPC [12]	V/A/D	.377	.706	.639	-	-	-	-	-	-
ResNet-sc [6]	Emo	-	-	-	-	-	-	-	-	50.71
ResNet-SID [6]	Emo	-	-	-	-	-	-	-	-	58.62
CNN14-sc [7]	V/A/D	.248	.658	.564	-	-	-	-	-	-
w2v2-b [13]	V/A/D	.363	.728	.636	-	-	-	-	-	-
				Proposed work						
a. CNN14	V/A/D	.345	.618	.514	-	-	-	-	-	-
b. CNN14	Senti	-	-	-	57.8	48.5	55.1	-	-	-
c. CNN14	Emo	-	-	-	58.5*	46.8*	53.5*	57.9	29.7	52.3
d. CNN14	V/A/D, Senti	.340	.623	.518	59.0	48.3	55.3	-	-	-
e. CNN14	V/A/D, Emo	.344	.610	.514	59.2*	46.4*	52.9*	58.1	29.3	52.9
f. CNN14	V/A/D, Senti, Emo	.351	.617	.516	58.8	48.3	55.1	58.5	30.2	53.6
g. w2v2-b	V/A/D	.515	.641	.542	-	-	-	-	-	-
h. w2v2-b	Senti	-	-	-	61.9	51.4	58.2	-	-	-
i. w2v2-b	Emo	-	-	-	47.5*	33.3*	30.6*	47.5	20.0	30.6
j. w2v2-b	V/A/D, Senti	.520	.637	.536	63.4	53.1	60.1	-	-	-
k. w2v2-b	V/A/D, Emo	.478	.605	.499	62.1*	51.3*	58.2*	61.2	32.9	56.3
l. w2v2-b	V/A/D, Senti, Emo	.500	.632	.525	63.5	53.6	60.5	62.2	34.1	57.3

audio waveforms. For Cnn14 audio encoder, the audio is represented by 64-bin Log Mel Spectrogram, hop size of 320, window size 1024 and frequency range between 50 to 14000 Hz. We finetune both CNN14 and Wav2Vec2 encoders along with the task linear layers. Similar to [13], the initial CNN layers in Wav2Vec2 remain frozen. All models are trained with PyTorch, batch size of 128 and learning rate of 10^{-4}.

4.3. Results

The results of our experiments are shown in Table 2. We report the average of two runs per model, except row *l*, reporting a single run. We found that multi-view models trained to predict both categorical emotions, categorical sentiment, and dimensional scores of valence, arousal and dominance produce results comparable and sometimes better than single-view models with the same architecture that were trained on each separately. For example, the multi-view CNN14 model results from row *f*, which was trained to predict 11 classes (dimensional, emotions, sentiment), exceeds all of the scores of the CNN14 of row *c* (5 emotion classes), and exceeds almost all scores of row *a* (3 dimensional values), while all rows share the same model architecture. This shows that the information needed to predict dimensional scores and categorical emotions is complementary enough that a joined single model can be trained to effectively predict all types.

These findings agree with those in Figure 1 and Table 1 which indicate that valence, arousal and dominance correlate with categorical emotions, but are not very strong predictors due to overlap in the space. If the correlation were stronger, the performance difference between single-view and multi-view might have been smaller.

Models from previous works trained on MSP-Podcast can be seen at the top of Table 2, although it should be noted that our version of MSP-Podcast is newer (v1.10 which includes an updated test set) and differs somewhat than the cited publications, so they may not be directly comparable. Also, authors in [7] trained the CNN-14 from scratch rather than using the pretrained model from [14]. Nevertheless, we show similar trends: valence scores lower than other dimensional labels, and Wav2Vec2.0 scores higher than CNN14—although the gap is much smaller in our experiments.

5. CONCLUSION

We analyzed the relationship between dimensional scores of valence, arousal, and dominance, categorical emotions and categorical sentiment and compared our results with theoretical psychological works of the circumplex of affect. We found that a joined multi-view training framework that simultaneously predict dimensional scores, emotions categories, and sentiment categories can produce results as strong or stronger than models trained independently for each view. Our analysis of the relationship between the three views in human-assigned labels showed overlap of several emotions in dimensional space rather than the theoretical distinct spaces in psychological research. Further exploration of the relationship between dimensional and categorical emotional data can offer additional insights into the core technology of emotion recognition and understanding.

6. REFERENCES

[1] Iris Bakker, Theo Van Der Voordt, Peter Vink, and Jan De Boon, "Pleasure, arousal, dominance: Mehrabian and russell revisited," *Current Psychology*, vol. 33, no. 3, pp. 405–421, 2014.

[2] James A Russell, "A circumplex model of affect.," *Journal of personality and social psychology*, vol. 39, no. 6, pp. 1161, 1980.

[3] Raghavendra Pappagari, Tianzi Wang, Jesus Villalba, Nanxin Chen, and Najim Dehak, "X-vectors meet emotions: A study on dependencies between emotion and speaker recognition," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7169–7173.

[4] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[5] Mohammed Abdelwahab and Carlos Busso, "Study of dense network approaches for speech emotion recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5084–5088.

[6] Raghavendra Pappagari, Jesús Villalba, Piotr Żelasko, Laureano Moro-Velazquez, and Najim Dehak, "Copy-paste: An augmentation method for speech emotion recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6324–6328.

[7] Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Florian Eyben, and Björn W Schuller, "Dawn of the transformer era in speech emotion recognition: closing the valence gap," *arXiv preprint arXiv:2203.07378*, 2022.

[8] Reza Lotfian and Carlos Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, 2019.

[9] Lawrence I-Kuei Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, vol. 45, no. 1, pp. 255–268, 1989.

[10] Jonathan Posner, James A Russell, and Bradley S Peterson, "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology," *Development and psychopathology*, vol. 17, no. 3, pp. 715–734, 2005.

[11] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert, "Norms of valence, arousal, and dominance for 13,915 english lemmas," *Behavior Research Methods*, vol. 45, no. 4, pp. 1191–1207, 2013.

[12] Mao Li, Bo Yang, Joshua Levy, Andreas Stolcke, Viktor Rozgic, Spyros Matsoukas, Constantinos Papayiannis, Daniel Bone, and Chao Wang, "Contrastive unsupervised learning for speech emotion recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6329–6333.

[13] Sundararajan Srinivasan, Zhaocheng Huang, and Katrin Kirchhoff, "Representation learning through cross-modal conditional teacher-student training for speech emotion recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6442–6446.

[14] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[15] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12449–12460, 2020.

[16] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.