SENTIMENT-AWARE AUTOMATIC SPEECH RECOGNITION PRE-TRAINING FOR ENHANCED SPEECH EMOTION RECOGNITION

Ayoub Ghriss[*]*, Bo Yang*[+]*, Viktor Rozgic*[+]*, Elizabeth Shriberg*[†]*, Chao Wang*[+]
[*] University of Colorado Boulder, ayoub.ghriss@colorado.edu, [†] Ellipsis Health,
elizabeth.shriberg@gmail.com, [+] Amazon Alexa, {amzbyang, rozgicv, wngcha}@amazon.com

ABSTRACT

We propose a novel multi-task pre-training method for Speech Emotion Recognition (SER). We pre-train SER model simultaneously on Automatic Speech Recognition (ASR) and sentiment classification tasks to make the acoustic ASR model more "emotion aware". We generate targets for the sentiment classification using text-to-sentiment model trained on publicly available data. Finally, we fine-tune the acoustic ASR on emotion annotated speech data. We evaluated the proposed approach on MSP-Podcast dataset, where we achieved the best reported concordance correlation coefficient (CCC) of 0.41 for valence prediction.

***Index Terms*—** Speech emotion recognition, automatic speech recognition, sentiment analysis, pre-training

1. INTRODUCTION

The origins of Speech Emotion Recognition (SER) date back to Blanton's work [1] when he wrote that "the language of the tones is the oldest and most universal of all our means of communication.", and culminated in studies on word-free voice samples with the goal of isolating qualitative features of the voice from those related to the articulated sounds patterns [2]. Emotion researchers define emotion either as a discrete construct or using emotion dimensions. The discrete construction theory credited to Ekman[3] isolates six basic emotion categories: anger, disgust, fear, happiness, sadness, surprise, but more comprehensive emotion category sets have been proposed [4]. The dimensional perspective, on the other hand, defines emotion as a point in a space defined by emotion dimensions, most commonly Activation, Valence, and Dominance (AVD).

A century later, the challenge of SER remains but new computational tools have emerged permitting more complex modelling. The motivations have also changed, it is no longer destined to understanding the emotion from a psychological perspective alone. It is further fueled by the ubiquitous speech-based interactions. Moreover, building smart systems capable of detecting the user emotional state has the potential of enhancing the interactive experience with different devices.

(Deep) neural network-based models have been a popular choice for SER in recent years [5, 6, 7]. However, training a Neural Network for SER requires a large training corpus, and non-neutral emotions are rare in speech. Recent work has addressed this limitation by pre-training on speech tasks, such as ASR [5, 8] or Speaker Identification [9, 10]. In this work, we propose a novel pre-training approach for SER. The proposed pre-training consists of building a "Sentiment-aware ASR" (SA2SR). The training of SA2SR objective is a combination of ASR and text-based sentiment classification, where the text sentiment labels are generated from a trained text sentiment model. This approach allows us to amass a large amount of text sentiment labels for speech data, and the effectiveness of these labels on improving SER is validated in our experiments.

2. RELATED WORK

A suitable choice of input representation is crucial for Speech Emotion Recognition (SER). Traditional approaches used classical signal processing methods (pitch, filter banks) or statistical measures (mean, variance, quantiles...) of the acoustic signal to train on classification/regression (Ekman categories vs AVD). Recent work on SER attempted to infer the emotion based on acoustic and textual cues, either simultaneously or separately.

In End-to-End SER with ASR [11], an acoustic-to-word ASR model is first trained then fine-tuned on a multi-task learning to jointly optimize ASR and SER objective. The emotion prediction block has two input fields: acoustic features similar to those used in ASR and the states of the ASR decoder. The authors also show that using this combination (raw acoustic inputs & ASR decoder features) outperforms an SER based on any single element of this combination.

Combining ASR and SER also outperforms a variant in which the SER block model takes the transcript of the ASR (word embeddings) instead of the decoder states. This result is expected since using the ASR transcript propagates the transcription inaccuracies to the SER block. The same reasoning can applied to text-based SER to point out their limitations, such as the one used in Sentiment Analysis based on speaker data [12]. In this speech-based Sentiment Analysis a pre-trained ASR is used for transcription of the utterance. The text is then fed to a feature extractor in parallel to a *Speaker*

[*] Work was done when Ayoub was working as an intern at Amazon
[†] Work was done while at Amazon

Identification feature block to provide an input to the emotion predictor.

A different approach that decouples text and acoustic features was introduced in Multi-modal SER [13], where the SER model encodes information from audio and text using two encoders. For the text encoder, the text input is the transcription from an ASR. This approach leverages the ability of text encoders to capture long term semantics better than the acoustic ones. However, this multi-modal approach assumes that the transcript is provided (via ASR), which limits its applicability when only the utterance (audio) is accessible.

To the extent of our knowledge, the only previous work that leveraged text-based sentiment labels was published recently[14] with three major differences: 1) we start with analyzing the correlation between text sentiment and speech emotion, thereby establishing a strong motivation for the proposed method, and an explanation for the observed performance boost, 2) the focus of this work is on SER with the widely used dimensional emotion representation (activation, valence and dominance), while that of [14] is on three-way speech sentiment classification, and 3) The approach in [14] uses out of the box encodings (ASR followed by Bert) and is oblivious to the feedback from sentiment labels. Our approach, on the other hand, leverages the proxy sentiment labels to induce the ASR embedding to incorporate emotional knowledge and yields better performance.

3. PROPOSED METHOD

We propose building a SER model that is pre-trained on a sentiment-aware ASR task. The sentiment-awareness is implemented by transferring sentiment knowledge from a text domain to the acoustic domain.

3.1. Correlation between text sentiment and speech emotion

We conjecture that text sentiment correlates with the valence dimension of speech emotion. Indeed, when a human listener tries to determine the emotion from a speech segment, the words in the speech also plays a role – the obvious examples are the speech segments that contain cursing words and phrases, or strong praising adjectives (e.g. excellent, beautiful, etc).

We test this correspondence between text-based sentiment and valence on the IEMOCAP dataset [15] and use a pre-trained text sentiment analysis (Roberta [16]) to get the sentiment labels: *negative, positive, neutral*. Table 1 shows the confusion matrix between text sentiment classes and speech emotion, with the dominant speech emotion highlighted in red in each sentiment class and the second dominant ones highlighted in blue.

As can be seen from Table 1, the negative text-sentiment utterances are mostly associated with negative speech emotion labels (Sad, Anger and Frustrated); while the positive text-sentiment utterances correspond to positive speech emotion labels (Happy). More interestingly, by investigating elements of the cell $\{Neutral, frustrated\}$ we can find transcripts such as : *"Nothing", "A vacation.", "I'm just saying."*}, while the inferred sentiment is neutral – this means the emotion in this case is likely conveyed through speech style and tone. Furthermore, by grouping $\{Sad, Frustrated, Anger\}$ into one class, we get a Spearman correlation of 0.22 between text sentiment and speech emotion. These observations motivate us to employ readily available text sentiment model to generate sentiment labels, which will serve as weak signal to train speech emotion models.

		Text sentiment		
		Negative	Neutral	Positive
Speech emotion	Sad	339	604	137
	Anger	**490**	518	94
	Frustrated	**658**	**1049**	141
	Neutral	253	**1251**	**204**
	Happy	252	848	**533**

Table 1. Confusion matrix between text-based sentiment [16] and speech emotion (IEMOCAP)

3.2. Sentiment-aware Automatic Speech Recognition (SA2SR)

In addition to the ASR model, a sentiment classifier (proxy classifier) is trained jointly on the acoustic encoder states. The architecture logic is similar to the one in combined ASR-SER[11]. The model takes as input the log filterbank energy (LFBE) features and contains two classifiers (Figure 1) that take the encoded acoustic sequence as input:

- Sequence to sequence classifier: A softmax layer for token classification
- Sentiment classifier: A sequence summarizer (recurrent neural network) followed by a softmax layer over the sentiment classes (*negative, neutral, positive*).

The proposed architecture is trained using a loss that is a linear combination (Equation 1) of (a) Connectionist Temporal Classification (CTC) loss [17] between the target sequence and the sequence of output probabilities over the token set, and (b) cross entropy loss between the predicted and proxy sentiment targets, i.e., sentiment obtained from the pre-trained text-to-sentiment model. The global loss is defined as:

$$L_{global} = L_{ASR} + \lambda L_{sentiment} \quad (1)$$

where $\lambda \geq 0$ is a hyper-parameter reflecting how the importance of sentiment classification vis-à-vis the ASR task.

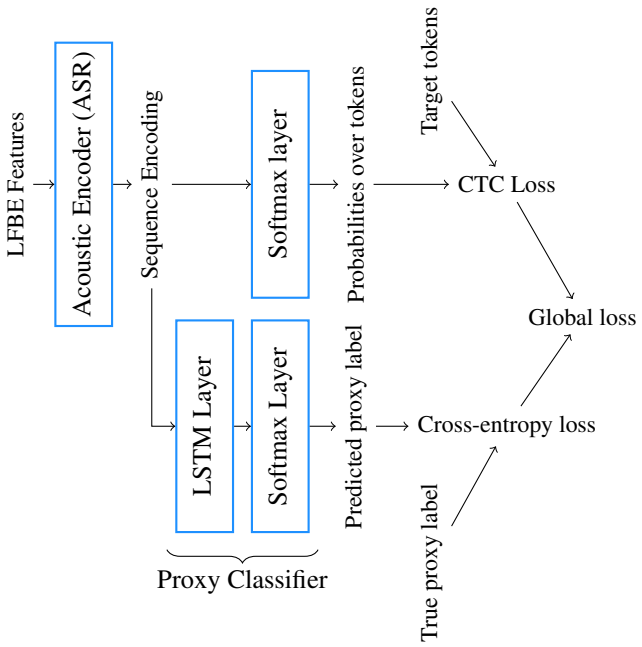

Fig. 1. Architecture of SA2SR pre-training network

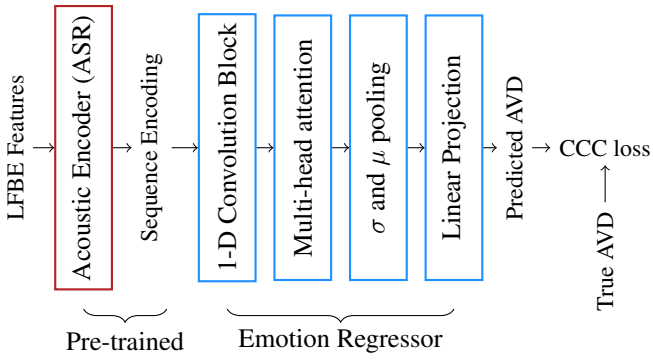

Fig. 2. Architecture of the fine-tuned SA2SR

3.3. Fine-tuning

During fine-tuning, we use the pre-trained acoustic encoder (Figure 2) and add an emotion regression transformer block that takes the encoding sequence as input and outputs the Activation, Valence, Dominance (AVD) values. The model is then trained to maximize the Concordance Correlation Coefficient (CCC) between prediction $\hat{y}$ and target values y.

$$CCC(y, \hat{y}) = \frac{2Cov(y, \hat{y})}{\sigma_y^2 + \sigma_{\hat{y}}^2 + (\mu_y - \mu_{\hat{y}})^2} \tag{2}$$

where Cov denotes the co-variance, σ and μ denote the sample variance and mean, respectively. During model training, these statistics are computed on mini-batches of data.

To enable easy comparison to previous works, the CCC objectives for each of the activation, valence and dominance are simply combined as:

$$\mathcal{L}_{CCC} = -\frac{1}{3}(CCC_A + CCC_V + CCC_D) \tag{3}$$

4. EXPERIMENT RESULTS

Datasets: We use the full Librispeech dataset [18] for pretraining a character based SA2SR. The dataset contains audios and transcripts for around 960 hours of audio, and we generate the proxy sentiment labels using text-to-sentiment RoBerta model trained on Twitter sentiment data [19]. We extract LFBE features with 40 frequencies using $25ms$ window and $10ms$ steps. Hence, for SA2SR pre-training, an input tuple consists of *(LFBE Features, transcript, proxy label)*.

Model fine-tuning and evaluation are conducted on MSP-Podcast [20] dataset. The MSP-Podcast contains 60K speech segments from podcast recordings which are perceptually annotated using crowd-sourcing. The ground truth emotion dimension scores are obtained averaging scores selected by individual annotators on seven-point *Likert* scale. The dataset was split following the same partitions provided in the official MSP corpus.

4.1. Networks Architecture

The Acoustic Encoder consists of 5 Bidirectional LSTM (Bi-LSTM) layers. All LSTM layers have 192 units with $tanh$ activation and $sigmoid$ recurrent activation.

The 1-D convolutional block in Emotion Regressor (Figure 2) is a stack of 2 one-dimensional *masked convolutions* with filter sizes (6,3) and strides (3,2) respectively. The convolutions accept input masks and process them accordingly to match the output of the convolution. The convolutions are followed by sample level normalization and a LeakyRELU$_{\alpha=0,3}$ is applied to the convolutional block output.

The multi-head attention uses 4 attention heads and 64 dimensions for all encoding spaces and feed-forward outputs. The output of the attention mechanism is pooled by computing the variance and the mean of the sequences yielding 128 features that are linearly projected into the AVD space.

4.2. Pre-training

For the global loss weight, we chose $\lambda = 200$. To optimize it, we use Adam optimizer with the parameters: $lr = 5 \times 10^{-5}$, $\beta_1 = 0.9, \beta_2 = 0.999$. The λ was chosen such that the ASR and sentiment classification losses achieve similar value on the validation set.

The LFBE features are normalized on a sample level. The training set is then augmented using speed augmentation of factors 0.9 and 1.1. This augmentation step triples the training set size. The features are then masked on time and frequency

dimensions with probability $p = 0.5$, as in the SpecAugment method [21]. We finally stack every adjacent three frames and skip every other two time steps, which leads to shorter sequences and improved model training time. The ASR training uses a token set of 29 characters.

After each epoch, we use the validation set to compute the Character Error Rate (CER) from the ASR task and Area Under the Curve (AUC) of the sentiment classifier. The pre-training terminates when the metric $M = CER - AUC$ does not improve for 25 epochs and the model that has the lowest M is evaluated. For the baseline model, we pre-train on the ASR task only and use the model from the epoch with the best CER on the validation set.

4.2.1. Effectiveness of ASR features for SER

To start, we test the effectiveness of ASR trained features for SER. Some previous work, for example [5, 8], reported limited transferability between ASR and SER. In our experiments, however, we found ASR features to be quite effective. It is possible that the differences are due to the large pre-training dataset (full Librispeech) and modern model architectures such as bidirectional LSTM and transformers.

Model	Activation	Valence	Dominance
No Pre-training	0.603	0.304	0.511
ASR features	0.649	**0.393**	0.544

Table 2. The baseline ASR pre-training achieves much improved CCC compared to no pre-training

We report these results in Table 2. The "ASR features" are pre-trained on Librispeech and fine-tuned on MSP-Podcast for SER. The "No pre-training" baseline is directly trained with the MSP-Podcast. We observe large performance boost when ASR pre-training is employed.

4.2.2. Effectiveness of SA2SR pre-training

In this experiment, we examine whether the additional proxy sentiment labels enhance SER performance. In particular, we expect that valence recognition to be improved, as we have seen correlation between text sentiment and valence in Section 3.1. The results are reported in Table 3. As a comparison, we also included the results from [7] that uses self-supervised pre-training as a strong contender.

Model	Activation	Valence	Dominance
ASR features	0.649	0.393	0.544
CPC-based pre-traing [7]	**0.706**	0.377	**0.639**
SA2SR features	0.679	**0.412**	0.564

Table 3. SA2SR produces good features for enhancing the CCC metric

We can see from Table 3 that recognition of valence, arguably the most important dimension of emotion, is further improved compared to the strong "ASR features" baseline. However, in this equal-weight multi-task emotion training setting, we see that activation and dominance dimension performs relatively weak compare to that of [7]. We view this as an encouraging result as in many applications, valence (positive v.s. negative) is of most interest.

Additionally, during the pre-training of the SA2SR model, we observed that the sentiment classification part achieves weighted-average-recall of 0.71 and 0.81 AUC. This indicates that indeed the model is trained to recognize these proxy sentiment labels. Therefore, we expect the learned representation to be suitable for the final SER task.

4.2.3. Importance of fine-tuning

Lastly, we are interested in examining whether we should freeze the learned representations during SER training. For both ASR and SA2SR features, we train models with and without freezing the acoustic encoder. The results are reported in Table 4.

Pretraining	Activation	Valence	Dominance
ASR features (frozen)	0.503	0.378	0.438
ASR features	0.649	0.393	0.544
SA2SR features (frozen)	0.508	0.403	0.483
SA2SR features	0.679	0.412	0.564

Table 4. Impact of finetuning the pre-trained encoders on CCC

As can be seen from Table 4, fine-tuning the encoder gives better SER performance compared to the methods with frozen encoders. This result proves that using out of the box ASR models for transcription without any fine-tuning would be outperformed when the gradient propagates to the ASR network weights. It also proves the point we made earlier about the importance of sentiment awareness in the SA2SR. Even without fine-tuning, SA2SR clearly outperforms the ASR embedding on 2 out of the 3 emotion AVD dimensions.

5. CONCLUSION

We proposed a novel pre-training method that utilizes proxy sentiment labels to aid ASR pre-training for SER. As text-sentiment and speech-emotion are correlated, this way we train speech representations capturing both phonetic and emotion-relevant info. We evaluated the proposed method on the MSP-Podcast dataset achieving state of the art performance on the challenging valence dimension.

Albeit we focused on ASR-based pre-training, the proxy sentiment classification task can be combined with other pre-training techniques, such as APC [22], CPC [23], which we will address in the future work.

6. REFERENCES

[1] Smiley Blanton, "The voice and the emotions," *Quarterly Journal of Speech*, vol. 1, no. 2, pp. 154–172, 1915.

[2] William F. Soskin and Paul E. Kauffman, "Judgment of Emotion in Word-Free Voice Samples," *Journal of Communication*, vol. 11, no. 2, pp. 73–80, 02 2006.

[3] Paul Ekman, "An argument for basic emotions," *Cognition and Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[4] Alan S. Cowen and Dacher Keltner, "Self-report captures 27 distinct categories of emotion bridged by continuous gradients," *PNAS, published online September 5, 2017*.

[5] Haytham M Fayek, Margaret Lech, and Lawrence Cavedon, "On the correlation and transferability of features between automatic speech recognition and speech emotion recognition.," in *Interspeech*, 2016, pp. 3618–3622.

[6] Vasudha Kowtha, Vikramjit Mitra, Chris Bartels, Erik Marchi, Sue Booker, William Caruso, Sachin Kajarekar, and Devang Naik, "Detecting emotion primitives from speech and their use in discerning categorical emotions," in *ICASSP 2020*. IEEE, 2020, pp. 7164–7168.

[7] 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*. IEEE, 2021, pp. 6329–6333.

[8] Egor Lakomkin, Cornelius Weber, Sven Magg, and Stefan Wermter, "Reusing neural speech representations for auditory emotion recognition," *arXiv preprint arXiv:1803.11508*, 2018.

[9] Michelle Bancroft, Reza Lotfian, John Hansen, and Carlos Busso, "Exploring the intersection between speaker verification and emotion recognition," in *ACIIW 2019*. IEEE, 2019, pp. 337–342.

[10] 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*. IEEE, 2020, pp. 7169–7173.

[11] Han Feng, Sei Ueno, and Tatsuya Kawahara, "End-to-End Speech Emotion Recognition Combined with Acoustic-to-Word ASR Model," in *Proc. Interspeech 2020*, 2020, pp. 501–505.

[12] S Maghilnan and M Rajesh Kumar, "Sentiment analysis on speaker specific speech data," in *2017 International Conference on Intelligent Computing and Control (I2C2)*, 2017, pp. 1–5.

[13] Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung, "Multimodal speech emotion recognition using audio and text," *CoRR*, vol. abs/1810.04635, 2018.

[14] Suwon Shon, Pablo Brusco, Jing Pan, Kyu J. Han, and Shinji Watanabe, "Leveraging pre-trained language model for speech sentiment analysis," *CoRR*, vol. abs/2106.06598, 2021.

[15] 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.

[16] Francesco Barbieri, José Camacho-Collados, Leonardo Neves, and Luis Espinosa Anke, "Tweeteval: Unified benchmark and comparative evaluation for tweet classification," *CoRR*, vol. abs/2010.12421, 2020.

[17] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd International Conference on Machine Learning*, 2006, pp. 369–376.

[18] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," *ICASSP 2015*, pp. 5206–5210, 2015.

[19] Mark Heitmann, Christian Siebert, Jochen Hartmann, and Christina Schamp, "More than a feeling: Benchmarks for sentiment analysis accuracy," *Communication & Computational Methods eJournal*, 2020.

[20] R. Lotfian and C. 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, October-December 2019.

[21] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *Interspeech 2019*, Sep 2019.

[22] Yu-An Chung and James Glass, "Generative pre-training for speech with autoregressive predictive coding," in *ICASSP 2020*. IEEE, 2020, pp. 3497–3501.

[23] Aaron van den Oord, Yazhe Li, and Oriol Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.