LanSER: Language-Model Supported Speech Emotion Recognition

Taesik Gong1, Josh Belanich2∗, Krishna Somandepalli2∗, Arsha Nagrani2, Brian Eoff2, Brendan Jou2

1KAIST, Republic of Korea
2Google Research
taesik.gong@kaist.ac.kr, {joshbelanich,ksoman,anagrani,beoff,bjou}@google.com

Abstract

Speech emotion recognition (SER) models typically rely on costly human-labeled data for training, making scaling methods to large speech datasets and nuanced emotion taxonomies difficult. We present LanSER, a method that enables the use of unlabeled data by inferring weak emotion labels via pre-trained large language models through weakly-supervised learning. For inferring weak labels constrained to a taxonomy, we use a textual entailment approach that selects an emotion label with the highest entailment score for a speech transcript extracted via automatic speech recognition. Our experimental results show that models pre-trained on large datasets with this weak supervision outperform other baseline models on standard SER datasets when fine-tuned, and show improved label efficiency. Despite being pre-trained on labels derived only from text, we show that the resulting representations appear to model the prosodic content of speech.

Index Terms: speech emotion recognition, large language models, weakly-supervised learning

1. Introduction

In conversations, humans rely on both what is said (i.e., lexical content), and how it is said (i.e., prosody), to infer the emotion expressed by a speaker. State-of-the-art methods in speech emotion recognition (SER) leverage the interplay of these two components for modeling emotional expression in speech. However, such methods still show limitations on in-the-wild scenarios due to the variability in natural speech, and the reliance on human ratings using limited emotion taxonomies. Extending model training to large, natural speech datasets labeled by humans for nuanced emotion taxonomies is expensive and is further complicated by the subjective nature of emotion perception.

Despite both lexical content and prosody being complementary for emotion perception, the two components are correlated, and in many cases the content is predictive of the prosody. For example, when someone says, “I won the lottery” – an upbeat and lively prosody would sound congruent, and one might perceive the emotional expression as elation or triumphant. In this work, we investigate how we might leverage the emotions congruent with lexical content in large unlabeled speech datasets to serve as weak supervision for developing SER models.

We turn to Large Language Models (LLMs) to infer expressed emotion categories in textual content. Due to the knowledge they embed from pre-training on large text corpora [1, 2], LLMs have demonstrated capabilities in numerous downstream tasks [3], including a few subjective tasks such as social and emotion reasoning [4]. In domains such as computer vision, LLMs were explored to reduce the need for labeled data, e.g., for visual question answering [5]. However, to our knowledge, they have not been studied for emotion recognition tasks, particularly from natural speech.

We propose LanSER, that uses LLMs to infer emotion categories from speech content i.e., transcribed text, which serve as weak labels for SER (Figure 1). Overall, LanSER enables pre-training a SER model on large speech datasets without human labels by (1) extracting text transcripts from utterances using automatic speech recognition (ASR), (2) using pre-trained LLMs to infer weak emotion labels with an engineered prompt and predetermined taxonomy, and (3) pre-training the SER model with the weak labels. We demonstrate that LanSER improves SER performance and label efficiency by fine-tuning on benchmark datasets. Moreover, we show that despite the emotion labels being derived from speech content only, LanSER captures speech prosody information that is relevant to SER.

2. Related Work

SER with LLMs: Recently, LLMs were used to generate pseudo-labels for semi-supervised learning for speech sentiment analysis [6]. Here, LLMs were fine-tuned on a labeled sentiment dataset to explore narrow sentiment classes of negative, positive, and neutral. In contrast, our work avoids fine-tuning LLMs on task-specific datasets by inferring weak labels via textual entailment, enabling exploration with wider emotion taxonomies. In the context of multi-modal emotion recognition, MEmoBERT [7] used audio, visual, and text information with prompt learning for unsupervised emotion recognition. Herein, the visual model is pre-trained on a large labeled emotion dataset. In contrast, in our work, pre-training on large human-annotated emotion datasets is not necessary.

Self-supervised learning: Self-supervised learning has become a popular method using large amounts of unlabeled speech data
for pre-training [8, 9]. Recent studies found that large pre-trained models via self-supervised learning show effectiveness in various downstream speech tasks [10], including many paralinguistic tasks [9]. We view self-supervised learning and our weak supervision from LLMs as complementary, since the two methodologies can be combined for training SER models.

3. Methodology

An overview of the training and inference process of LanSER in shown in Figure 1. During pre-training, we use ASR to generate transcripts from speech utterances, which are fed into a LLM with appropriate prompt to extract weak emotion labels in predetermined taxonomy via textual entailment. These labels are used to pre-train a SER model via weakly-supervised learning. The pre-trained SER model can then either be used directly to output emotion predictions according to the emotion taxonomy used to extract weak labels, or can be adapted for a different taxonomy or dataset by fine-tuning.

We note that the emotions inferred using LLMs from speech content are proxies for the emotion being expressed, and may not capture the larger context or intent of the speaker. Thus, we treat them as “weak” emotion labels in our work.

3.1. Weak label generation via textual entailment

There are multiple ways to use LLMs for extracting weak emotion labels. Two dominant approaches in the literature are (i) text generation [2] and (ii) filling mask [11, 1, 7]. Figure 2 demonstrates the behaviors of text generation and filling mask for weak emotion label prediction. We used representative LLMs for each approach: GPT-2 for text generation and BERT in [11] for filling mask. While these approaches show some success, the common limitation in a zero-shot setting is that they often output undesirable “noise”, like irrelevant words (text generation), or non-emotional responses (e.g., “himsself”) in filling mask in the Fig. 2).

Thus, we want to constrain the LLM model to output only words relevant to emotion perception. To this end, we use textual entailment [12] to generate weak labels that also allows us to constrain the emotion taxonomy apriori. Figure 2 illustrates the entailment-based weak emotion label generation; at a high-level, this method calculates the entailment scores between an input transcript (called hypothesis) and prompts with candidate labels from the taxonomy (called premise), and then selects the item with the highest score as the weak label. Formally, let \( x \in X \) denote ASR transcripts from speech and \( y \in Y \) denote a candidate label in taxonomy \( Y \). A prompting function \( g(\cdot) \) prepends a predefined prompt to the given input. \( f(x, g(y)) \) denotes the entailment score between a hypothesis \( x \) and a prompted label \( g(y) \). The resulting weak emotion label \( \hat{y} \) for a given transcript \( x \) is calculated as:

\[
\hat{y} := \arg \max_{\hat{y} \in Y} f(x, g(y)).
\]  

3.2. Prompt engineering

Prompt engineering is a task-specific description embedded in inputs to LLMs (e.g., a question format) [15]. It is a critical component affecting zero-shot performance of LLMs on various downstream tasks [1, 16, 17]. In Section 4.2 we explore various prompts in order to understand the impact of prompt engineering for the entailment task. Ultimately, we found that the prompt “The emotion of the conversation is {}” performed best, and we use this prompt throughout our experiments.

3.3. Taxonomy

The choice of emotion taxonomy is critical in developing SER models as emotion perception and expression is nuanced. Common SER benchmarks typically use 4–6 emotion categories [18, 19], which do not capture the variability in emotion perception [20]. Thus we experiment with BRAVE-43, a fine-grained taxonomy [21]. We adopted and modified the BRAVE taxonomy which originally contains 42 self-reported emotions labels. We converted several two-word emotions to one-word emotions for simplicity and added “shock” to capture a negative version of “surprise”, resulting in a total of 43 categories. Note this taxonomy is not speech-specific. We investigate the impact of taxonomy selection in Section 4.5. We expect fine-grained taxonomies to help learn effective representations by using the high degree of the expressiveness of LLMs.

4. Experiments

Our overarching hypothesis is that, given a sufficiently large amount of data, pre-training speech-only models on weak emotion labels derived from text improves performance on SER tasks. As such, throughout this paper, we focus on speech-only emotion recognition models. Additionally, our goal is not to obtain state-of-the-art results on downstream tasks but to assess, given a fixed model capacity, whether models pre-trained via LanSER achieve improved performance.

4.1. Data preparation

Pre-training data: We investigate two large-scale speech datasets for LanSER pre-training: People’s Speech [22] and Condensed Movies [23]. People’s Speech is currently the largest English speech recognition corpus, containing 30K hours of general speech. Condensed Movies is comprised of

<table>
<thead>
<tr>
<th>Text generation</th>
<th>Transcript</th>
<th>Prompt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input: “Stop. I hate you. Do you hear me?”</td>
<td>This person is expressing</td>
<td></td>
</tr>
<tr>
<td>Outputs:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. “disgust with me? I’m not that dumb, I hate you because of me...”</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. “her hatred, what for? All I did was say, ‘you’re not just some angry...”</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. “personal anger against you. I don’t know what to think about that’s you...”</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Filling mask

Input: “Stop. I hate you. Do you hear me?” | This person is expressing (MASK). |
Outputs: | |
1. “hate” |
2. “pain” |
3. “himsself” |

Textual entailment

Hypothesis: “Stop. I hate you. Do you hear me?” | |
Premise 1: “This person is expressing {MASK}.” | |
Premise 2: “This person is expressing {Hate}.” | |
Output: “anger” | |

Figure 2: Comparison of three weak label generation approaches: text generation, filling mask, and textual entailment.
of 1,000 hours of video clips from 3,000 movies, where we use only the audio. We explore these two large-scale speech datasets to understand the impact of the amount of data and their distributions; while People’s Speech has more samples from less emotional data sources (e.g., government, interview, health, etc.), Condensed Movies has fewer samples from a more emotional data source (movies). We use Whisper ASR [24] (“small” variant) to segment and generate transcripts for People’s Speech and Condensed Movies datasets, resulting in 4,321,002 and 1,030,711 utterances, respectively.

**Downstream tasks:** We use two common SER benchmarks for downstream tasks: IEMOCAP [18] and CREMA-D [19]. IEMOCAP is an acted, multi-speaker database containing 5,531 audio clips from 12 hours of speech. We follow the commonly used four-class (anger, happiness, sadness, and neutral) setup [7, 25, 10, 9] and use speaker-independent train/val/test splits. CREMA-D has 7,441 audio clips collected from 91 actors. An important characteristic of CREMA-D is that it is linguistically constrained, having only 12 sentences each presented using six different emotions (anger, disgust, fear, happy, neutral, and sad). We use CREMA-D to validate that our models indeed learn prosodic representations, and do not just learn to use language to predict the emotional expression.

4.1.1. Baselines

We compare LanSER models fine-tuned on downstream datasets with the following four baselines:

- **Majority:** Output the most prevalent class in the dataset [12].
- **GT Transcript + Word2Vec** [26]: Each word in a ground-truth transcript is converted to a Word2Vec embedding. We compute the cosine similarity between the averaged transcript embedding and each class label, outputting the class with the highest similarity.
- **GT Transcript + LLM + Entailment** [12]: Using the same methodology for producing weak labels, we process the ground-truth transcript with an LLM and entailment to output a classification according to the dataset’s taxonomy.
- **Supervised:** Supervised learning using the same model architecture as LanSER but without pre-training.

We include two language-based methods (Word2Vec and Entailment) to better understand how LanSER compares with baselines. Notably, LanSER pre-trained with the Condensed Movies is comprised of movies having more expressive speech; from the emotion recognition perspective, People’s Speech might introduce more noise than Condensed Movies. While People’s Speech is comprised of fairly neutral speech data (e.g., government, interviews, etc.), Condensed Movies is comprised of movies having more expressive speech; from the emotion recognition perspective, People’s Speech might introduce more noise than Condensed Movies.

4.2. Prompt engineering

We investigated the impact of various prompts to infer weak emotion labels using IEMOCAP. We chose IEMOCAP because it has transcripts and human-rated labels with majority agreement referred here as “ground-truth”. To evaluate the prompts, we compute accuracy by comparing the weak labels with the ground-truth. We also examined prompts used in previous emotion recognition studies [16, 7] and modified a few vision-specific prompts [16] for our study by replacing words such as “photo” or “image” with “speech”.

Table 1 shows the accuracy for each prompt. The prompt (“I am {}.”) used in the related sentiment work [7] was not as effective at capturing emotional signals. Similarly, adapting vision-specific prompts [16] was ineffective. This suggests that it is worthwhile to tailor the prompt to the SER task. Among the prompts we explored, “The emotion of the conversation is {}.” had the highest accuracy. We adopt this prompt to infer weak labels in all our experiments. We leave additional prompt tuning [30] as future work.

4.3. Fine-tuning

We fine-tune all models on the downstream tasks to evaluate their label efficiency and performance. To measure label efficiency, we varied the percentage of seen training data from 10% to 100% for each dataset. Table 2 shows the result. “LanSER (People’s Speech)” means pre-training with Peoples Speech, while “LanSER (Condensed Movies)” refers to pre-training with Condensed Movies. In all cases, we used the BRAVE taxonomy (see Sec. 3.3) as the label space.

First, NLP baselines (Word2Vec and Entailment) fail on CREMA-D, as they only use lexical speech content. Interestingly, LanSER’s results on CREMA-D suggest that the model can learn prosodic features even with labels from text only given a sufficiently large amount of data.

Overall, LanSER outperforms the NLP and majority class baselines. Notably, LanSER pre-trained with the Condensed Movies showed improved accuracy than with the People’s Speech. While People’s Speech is comprised of fairly neutral speech data (e.g., government, interviews, etc.), Condensed Movies is comprised of movies having more expressive speech; from the emotion recognition perspective, People’s Speech might introduce more noise than Condensed Movies.

To assess that performance improvements are being driven by the emotion labels inferred using LLMs, and not just the scale of the pre-training data, we compare the fine-tuning performance of LanSER to a model pre-trained on Condensed Movies.
Table 2: Unweighted accuracy (%) of fine-tuning for downstream tasks with varying the percentage of fine-tuning data (10%, 30%, 50%, 70%, and 100%). Bold fonts indicate the highest accuracy.

<table>
<thead>
<tr>
<th>Downstream task</th>
<th>Method</th>
<th>Fine-tuning data:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>10%</td>
</tr>
<tr>
<td>IEMOCAP [18]</td>
<td>Majority</td>
<td>30.9%</td>
</tr>
<tr>
<td></td>
<td>GT Transcript + Word2Vec [26]</td>
<td>34.9%</td>
</tr>
<tr>
<td></td>
<td>GT Transcript + Entailment [12]</td>
<td>47.5%</td>
</tr>
<tr>
<td></td>
<td>Supervised</td>
<td>38.5%</td>
</tr>
<tr>
<td>LanSER (People’s Speech)</td>
<td></td>
<td>42.0%</td>
</tr>
<tr>
<td>LanSER (Condensed Movies)</td>
<td></td>
<td>50.0%</td>
</tr>
<tr>
<td>CREMA-D [19]</td>
<td>Majority</td>
<td>17.1%</td>
</tr>
<tr>
<td></td>
<td>GT Transcript + Word2Vec [26]</td>
<td>19.1%</td>
</tr>
<tr>
<td></td>
<td>GT Transcript + Entailment [12]</td>
<td>16.1%</td>
</tr>
<tr>
<td></td>
<td>Supervised</td>
<td>37.8%</td>
</tr>
<tr>
<td>LanSER (People’s Speech)</td>
<td></td>
<td>35.5%</td>
</tr>
<tr>
<td>LanSER (Condensed Movies)</td>
<td></td>
<td>43.7%</td>
</tr>
</tbody>
</table>

Table 3: Unweighted accuracy (%) for fine-tuning on downstream tasks. LanSER (random labels) is pre-trained on Condensed Movies with BRAVE taxonomy labels assigned randomly.

<table>
<thead>
<tr>
<th>Downstream task</th>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>IEMOCAP [18]</td>
<td>LanSER (random labels)</td>
<td>47.6%</td>
</tr>
<tr>
<td></td>
<td>LanSER (weak labels)</td>
<td>54.5%</td>
</tr>
<tr>
<td>CREMA-D [19]</td>
<td>LanSER (random labels)</td>
<td>50.0%</td>
</tr>
<tr>
<td></td>
<td>LanSER (weak labels)</td>
<td>58.7%</td>
</tr>
</tbody>
</table>

Table 4: Zero-shot unweighted accuracy (%) of SER models.

<table>
<thead>
<tr>
<th>Downstream task</th>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>IEMOCAP [18]</td>
<td>Scratch</td>
<td>22.9%</td>
</tr>
<tr>
<td></td>
<td>LanSER (People’s Speech)</td>
<td>30.9%</td>
</tr>
<tr>
<td></td>
<td>LanSER (Condensed Movies)</td>
<td>34.3%</td>
</tr>
<tr>
<td>CREMA-D [19]</td>
<td>Scratch</td>
<td>16.3%</td>
</tr>
<tr>
<td></td>
<td>LanSER (People’s Speech)</td>
<td>15.9%</td>
</tr>
<tr>
<td></td>
<td>LanSER (Condensed Movies)</td>
<td>23.5%</td>
</tr>
</tbody>
</table>

4.4. Zero-shot classification accuracy

A unique advantage of LanSER over self-supervised learning [8, 9] is that it enables SER models to support zero-shot classification. Table 4 shows the zero-shot classification accuracy: for LanSER, SER models were pre-trained with the taxonomy of the downstream dataset instead of BRAVE and evaluated in a zero-shot setting. We use models with randomly initialized weights and no training as a lower-bound of performance, referred to as “Scratch”. Overall, LanSER shows higher accuracy than the baseline, although not as good as fine-tuning. These results suggest the potential of training large SER models that can perform well on various downstream tasks, without further fine-tuning. Improving zero-shot performance further using our proposed framework is part of our future work.

4.5. Impact of taxonomy

Figure 3 shows the impact of taxonomy selection. We compared the BRAVE taxonomy with downstream task’s taxonomies. “PS” and “CM” refers to People’s Speech and Condensed Movies, respectively. “IEMOCAP”, “CREMA-D”, and “BRAVE” means taxonomy used to generate weak labels. As shown, pre-training with the finer taxonomy (BRAVE) shows generally better accuracy when fine-tuned, with 4.2% accuracy improvement on average. This indicates that a fine-grained taxonomy is beneficial to learn effective representations by leveraging the high degree of the expressiveness of LLMs.

5. Caveats

Developing machine perception models of apparent emotional expression remains an open area of investigation. The models in this work do not aim to infer the internal emotional state of individuals, but rather model proxies from speech utterances. This is especially true when training on the output of LLMs, since LLMs may not take into account prosody, cultural background, situational or social context, personal history, and other cues relevant to human emotion perception. ASR transcription errors add another layer of noise.

The benchmark datasets we use in this work are relatively small and are labeled with limited emotion taxonomies. CREMA-D, while useful for its fixed lexical content, is an acted dataset where the emotional expression of its utterances may not well-represent natural speech.

6. Conclusion and Future Work

In this work, we proposed LanSER, a novel language-model supported speech emotion recognition method that leverages large unlabeled speech datasets by generating weak labels via textual entailment using LLMs. Our experimental results showed that LanSER can learn effective emotional representations including prosodic features.

We note several possible areas of future work. It may be possible to reduce the weak label noise via filtering mechanisms, or by modifying prompts to include more conversational context, like the previous and next utterances, or scene descriptions. Additionally, using LLMs to generate weak labels in an open-set taxonomy may better leverage their expressiveness. Finally, while in this work we used ResNet-50 as our backbone model, higher capacity models like Conformers [9] might better capture the complex relationship between speech and emotion on the pre-training datasets we explored. We believe that the initial investigation and findings of this work provide valuable insights for future SER research on large-scale unlabeled data.
7. References


