SPEECH-BASED EMOTION RECOGNITION WITH SELF-SUPERVISED MODELS USING ATTENTIVE CHANNEL-WISE CORRELATIONS AND LABEL SMOOTHING

Sofoklis Kakouros
University of Helsinki, Finland
Themos Stafylakis
Omilia - Conversational Intelligence, Athens, Greece
Ladislav Mošner
Brno University of Technology, Faculty of Information Technology, Speech@FIT, Czechia
Lukáš Burget
Brno University of Technology, Faculty of Information Technology, Speech@FIT, Czechia

INTRODUCTION

- Emotional expressions are a fundamental component of spoken interaction.
- However, recognizing emotions in speech remains a challenging problem.
- *What kind of acoustic representations are best for speech emotion recognition?*
- *How can we best model the long temporal context over which emotions take place?*
- *How can we best tackle the problem of ambiguous labels for the emotions?*

STANDARD POOLING METHODS

- Emotion recognition is an utterance-level task.
- Frame-level pooling:
  1. Mean
  2. Standard deviation
  3. Mean + standard deviation

Can we do things better and capture more informative representations from the successive frames?

CORRELATION POOLING

- Modelling correlations between channels.
- We first reduce the number of channels from 1024 to 256 using a learnable linear layer.
- Then, average pooling of the frame-wise outer products ⇒ correlation matrix:

$$\mathbf{C} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{o}_t \mathbf{o}_t'$$

But emotion information does not appear uniformly across our signals. What can we do?

ATTENTIVE CORRELATION POOLING

- We introduce a new flavor of attention by inserting weights in the estimates of the statistics:

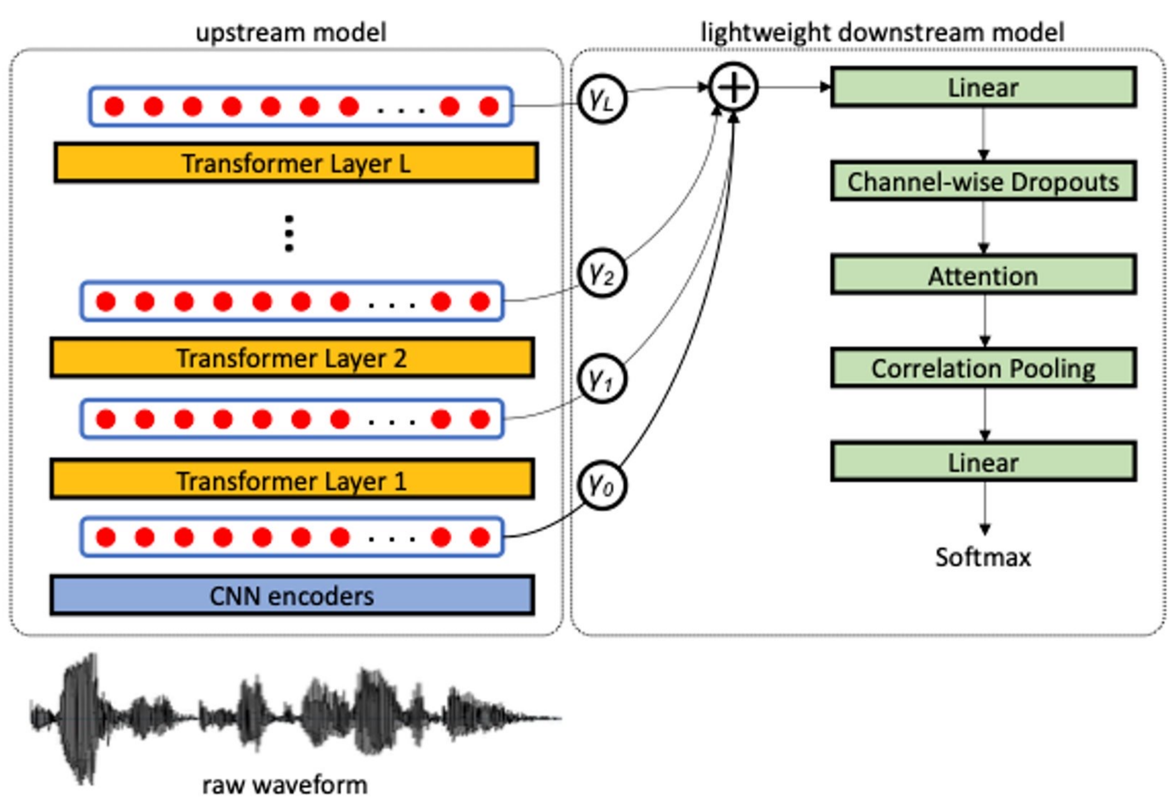
$$\boldsymbol{\mu} = \sum_{t=1}^{T} w_t \mathbf{v}_t, \;\; \boldsymbol{\Sigma} = \sum_{t=1}^{T} w_t (\mathbf{v}_t - \boldsymbol{\mu})(\mathbf{v}_t - \boldsymbol{\mu})'$$

- The proposed attention enables us:
  - to keep the multi-modality of multi-head attention since a single head is too weak to capture the phonetic, speaker, emotion and channel variability,
  - to robustly estimate the attention by aggregating the matrix similarities prior to the Softmax function.
- *However, our labels contain a certain degree of ambiguity that we need to address.*

LABEL SMOOTHING

- With label smoothing we soften the hard (one-hot) targets vectors.
- The aim of label smoothing is to reduce the confidence of the classifier on the target labels.
- Label smoothing replaces the one-hot encoded labels with a mixture of the one-hot encoded labels and the uniform distribution.
  - One-hot encoded labels maximize logit gaps that are fed into the Softmax function.
  - On the other hand, smoothed labels, encourage smaller logit gaps, thus reducing the confidence for the targets.

ARCHITECTURE



- Our setup is based on **SUPERB– Speech processing Universal PERformance Benchmark**
- SUPERB is a collection of benchmarking resources to evaluate the capability of a universal shared representation for speech processing
- We extract embeddings from all transformer layers:
  - HuBERT
  - WavLM
  - Wav2Vec 2.0
- Layer pooling
- Weigh embeddings
- Channel-wise dropouts
- Apply attention
- Correlation pooling

RESULTS

- Experiments run on **IEMOCAP**
- 5-fold cross-validation
- Our method yields results that surpass those on the benchmark setup of SUPERB
- SUPERB reports an accuracy of 70.62% with WavLM and 67.62% with HuBERT
- With our proposed approach we obtain 75.60% and 73.86% respectively

Table 1. Unweighted accuracy (% mean and std) between test sets for SER in IEMOCAP using HuBERT large, Wav2vec 2.0, and WavLM large self-supervised representations.

Pooling method	**HuBERT**	**Wav2vec 2.0**	**WavLM**
mean	65.73 (2.73)	66.86 (1.76)	69.44 (1.53)
mean-std	69.15 (1.61)	69.92 (1.17)	72.56 (1.67)
corr ($p_d = 0$)	69.82 (1.35)	68.44 (1.85)	72.34 (1.54)
corr ($p_d = 0.25$)	69.72 (1.19)	67.85 (1.84)	72.27 (1.45)
corr attentive	**73.86 (2.10)**	**70.01 (2.20)**	**75.60 (2.33)**

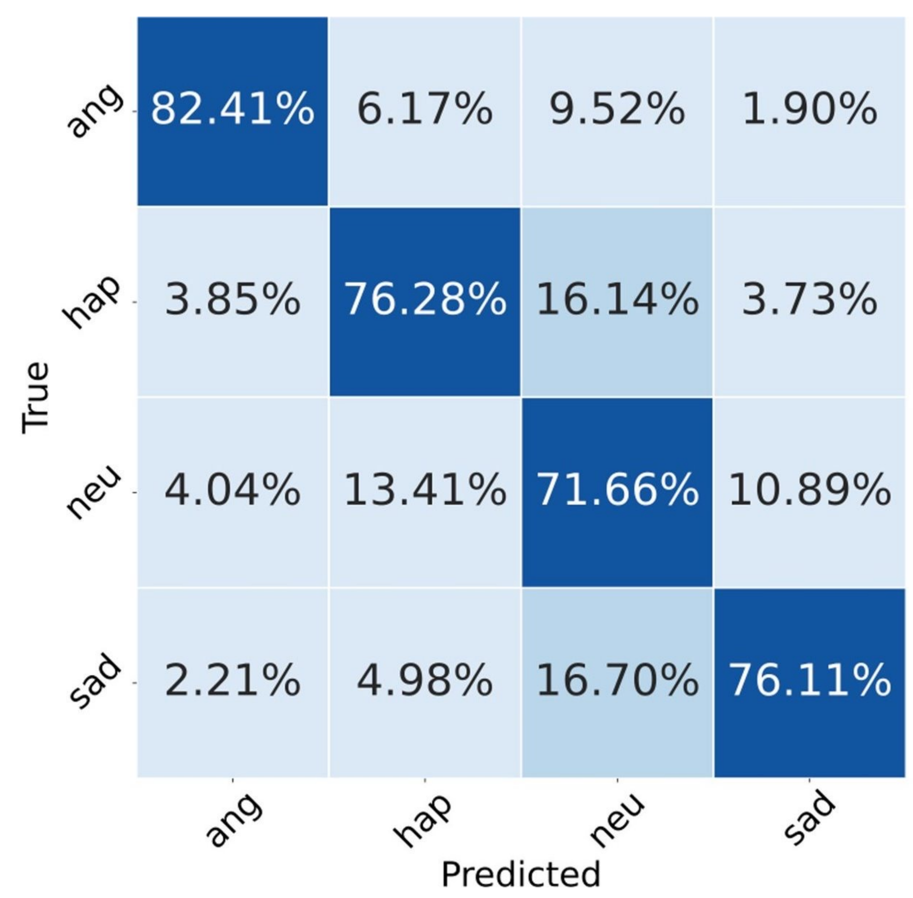


Figure: Attentive correlation pooling with WavLM

CONCLUSIONS

- SER framework that uses self-supervised representations and is based on label smoothing and a novel approach to attention: attentive correlation pooling.
- Our method does not require fine-tuning of the pre-trained SSL models but rather uses a light-weight classification head.
- Our method reaches high performance in all pre-trained models tested surpassing that of the literature in similar tasks.
- **Next steps**: extend the evaluation setup and validate the performance of our method on more datasets.

UNIVERSITY OF HELSINKI
FACULTY OF ARTS

Omilia
Conversational Intelligence

BRNO UNIVERSITY OF TECHNOLOGY

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI
HUMANISTINEN TIEDEKUNTA
HUMANISTISKA FAKULTETEN
FACULTY OF ARTS