# Investigating Transformer Encoders and Fusion Strategies for Speech Emotion Recognition in Emergency Call Center Conversations.

Theo Deschamps-Berger, Lori Lamel, Laurence Devillers

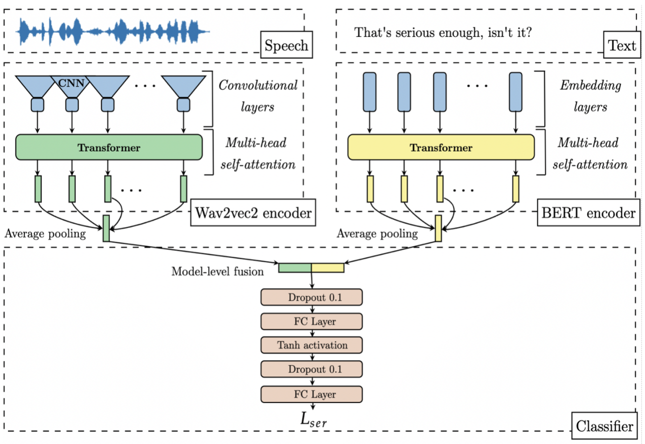
# Investigating Transformer Encoders and Fusion Strategies for Speech Emotion Recognition in Emergency Call Center Conversations.

THEO DESCHAMPS-BERGER, LISN - CNRS, Paris-Saclay University, France

LORI LAMEL, LISN - CNRS, France

LAURENCE DEVILLERS, LISN - CNRS and Sorbonne University, France

**Model**




**Corpus**

| CEMO$_s$ | ANGER | FEAR | NEUTRAL | POSITIVE | Total |
|---|---|---|---|---|---|
| #Speech segments | 1056 | 1056 | 1056 | 1056 | 4224 |
| #Speakers | 149 | 537 | 450 | 544 | 812 |
| #Dialogues | 280 | 504 | 425 | 516 | 735 |
| Durations (min, s) | 39 | 52 | 49 | 20 | 160 |
| Duration mean (s) | 2.2 | 2.9 | 2.8 | 1.1 | 2.3 |
| Vocabulary size (#words) | 1146 | 1500 | 1150 | 505 | 2600 |
| Average word count | 9.3 | 11.9 | 7.9 | 3.8 | 8.2 |

**Results**

| Model name | Modality | UA |
|---|---|---|
| Wav2vec2-Fr-3K$_{LARGE}$ | Speech | 73.1 |
| FlauBERT$_{LARGE}$ | Text | 67.0 |
| Model-level fusion: | | |
| Wav2vec2-Fr-3K$_{LARGE}$ + FlauBERT$_{LARGE}$ | Speech + Text | 76.8 |
| Late fusion: | | |
| Wav2vec2-Fr-3K$_{LARGE}$ + FlauBERT$_{LARGE}$ | Speech + Text | 77.1 |

Fig. 1. Model wav2vec2 - BERT (model-level fusion), corpus (CEMO) and main results in this work

There has been growing interest in using deep learning techniques to recognize emotions from speech. However, real-life emotion datasets collected in call centers are relatively rare and small, making the use of deep learning techniques quite challenging. This research focuses on the study of Transformer-based models to improve the speech emotion recognition of patients' speech in French emergency call center dialogues. The experiments were conducted on a corpus called CEMO, which was collected in a French emergency call center. It includes telephone conversations with more than 800 callers and 6 agents. Four emotion classes were selected for these experiments: Anger, Fear, Positive and Neutral state. We compare different Transformer encoders based on the wav2vec2 and BERT models, and explore their fine-tuning as well as fusion of the encoders for emotion recognition from speech. Our objective is to explore how to use these pre-trained models to improve model robustness in the context of a real-life application. We show that the use of specific pre-trained Transformer encoders improves the model performance for emotion recognition in the CEMO corpus. The Unweighted Accuracy (UA) of the french pre-trained wav2vec2 adapted to our task is 73.1%, whereas the UA of our baseline model (Temporal CNN-LSTM without pre-training) is 55.8%. We also tested BERT encoders models: in particular FlauBERT obtained good performance for both manual 67.1% and automatic 67.9% transcripts. The late and model-level fusion of the speech and text models also improve performance (77.1% (late) - 76.9% (model-level)) compared to our best speech pre-trained model, 73.1% UA. In order to



Manuscript submitted to ACM

place our work in the scientific community, we also report results on the widely used IEMOCAP corpus with our best fusion strategy, 70.8% UA. Our results are promising for constructing more robust speech emotion recognition system for real-world applications.

CCS Concepts: • **Computing methodologies** → **Supervised learning**; **Speech recognition**; *Discourse, dialogue and pragmatics*; Cross-validation.

Additional Key Words and Phrases: real-life emotional corpus, emergency call center, speech emotion recognition, Transformer-based models, late fusion, models-level fusion

**ACM Reference Format:**

Theo Deschamps-Berger, Lori Lamel, and Laurence Devillers. 2022. Investigating Transformer Encoders and Fusion Strategies for Speech Emotion Recognition in Emergency Call Center Conversations.. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '22 Companion), November 7–11, 2022, Bengaluru, India.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3536220.3558038

## 1 INTRODUCTION

Research on multimodal analysis of users's behavior, such as speech and text analysis, has demonstrated the potential for estimating a user's emotion from of these modalities [40]. Such studies suggest that similar analyses can serve in the context of emergency call centers where human agents must listen and understand callers quickly, taking into account their emotional state.

Calls to emergency services can be made by the patient or by a third party (family, friend, colleague, neighbor) or home/health assistant. Such call center data constitute a particular form of natural conversational data that is collected in a real-life context from a large number of speakers. The call centers are staffed 24 hours a day, 7 days a week and are intended to be used as a resource by individuals in times of crises. The operators are trained to quickly assess callers' states of mind, the crisis level, and the urgency of the situation and to decide which services are needed: ambulance assistance, psychiatric care, or advice to call their doctor.

Real-life emotional data collected in call centers are relatively rare and are difficult to share due to privacy constraints. Although there is a vast quantity of raw data, annotated corpora are generally small due to the high annotation costs including data anonymization. These small data sets make the popular deep learning techniques challenging to us. Recent advances have shown that Transformer-based methods, such as wav2vec2 and BERT pre-trained models, outperformed previous approaches for emotion recognition tasks[40]. The research objective of this work is to explore the use of these speech and text pre-trained models applied to speaker's emotion recognition in order to increase the model robustness for recordings in an emergency context.

More specifically, this paper reports experiments using Transformer-based encoder pre-trained models on speech data and their transcripts for Emotion Recognition from speech. We used the 4 major emotions (Anger, Fear, Positive and Neutral) occurring in a real corpus of agent-patient conversations called CEMO [14]. In our baseline study, we explored the acoustic modality of speech emotion recognition with spectrogram-like representations and CNN Bi-LSTM models [12]. In this study, we investigate the effects of different wav2vec2 and fine-tuning variations [3, 33]. An off-the-shelf system was used to generate automatic transcripts, and a performance comparison was conducted on manual and automatic transcripts for the text modality using French pre-trained Transformers such as FlauBERT [25] and CamemBERT ([29]). Finally, we explored the complementarity of the wav2vec2 and BERT pre-trained models and different fusions (late and model) of the speech and text modes.

The remainder of this paper is as follows. The next section overviews the related work, followed by the description of the emergency call center corpus in Section 3. Sections 4 and 5 describe the experimental setup and results: the architectures and the datasets tested (CEMO end IEMOCAP), and contrastive experiments with CEMO, presenting and analyzing the results. Section 6 summarises our best results on CEMO and provides results obtained on IEMOCAP. Finally, Section 7 discusses ethical aspects and reproducibility and Section 8 highlights some discussion points and conclusions.

## 2 RELATED WORK FOR SPEECH EMOTION RECOGNITION

### 2.1 Classical approaches

Over the years, the research community has explored many approaches to address emotion recognition tasks. On the one hand, researchers tried to find reliable and relevant speech representations, investigating a wide range of features ranging from low-level descriptors like the GeMAPS configuration [18] to audio transformations like Mel-spectrograms [35]. On other hand, the research community also explored different speech emotion recognition systems, such as Convolutional Neural Networks to extract speech features, or Recurrent Neural Networks to detect near and longer dependencies in audio features or linguistic information [16].

### 2.2 Pre-trained Transformer encoders

A few years ago the concept of intra- or self-attention started attracting growing interest by the research community [8]. Moreover, the progress made on attention [38] for RNNS led to the introduction of new models called Transformers. These new models have been used primarily in the fields of natural language processing and computer vision. Transformer-based encoders have been growing in popularity and have led to language representation models such as BERT [15] which was designed to serve as a pre-trained core model. It is trained in a self-supervised mode on huge amounts of data in order to serve for a wide range of applications, with the idea of being able to refine it to tackle low-resource domains [1]. The BERT model was applied to the French language with CamemBERT [29] and FlauBERT [25]. In other research work, the Transformer encoder has been extended to speech segments to address speech recognition in low-resourced situations, with wav2vec [36] and wav2vec2 [3]. The wav2vec2 model is comparable to the masked BERT language model, but it introduces a contrastive task in the pre-training that aims at discretizing audio frames [2]. It was also applied to French by LeBenchmark [17]. Transformer-based encoder models have shown good capabilities for solving downstream tasks, including speech emotion recognition [27, 33]. According to [23, 40], they appear to be invariant to domain, speaker, and gender characteristics. In this study, we investigated the performance of these pre-trained models based on the speech or/and text modalities to detect real-life emotions in an emergency call center context.

### 2.3 Fusion approaches

Given the importance of both speech and text for speech emotion recognition, multimodal fusion approaches have been addressed in a number of research works with classical approaches and recently with pre-trained models as in [40] and have yielded significant improvement of results for speech emotion recognition [6, 28]. Previous research has explored various strategies for combining the two modalities, which can be classified into three groups: early fusion; model-level fusion; and late fusion [44]. Early fusion consists of concatenating the features of the different modalities and using these as the input to the model [6, 42]. Model-level fusion involves combining the hidden representation of modalities during training [19, 21], which can be achieved with attention mechanisms [7, 11, 20, 26, 37]. Finally late fusion is the aggregation of scores according to specific criteria [6, 13, 30]. In this work, we explored the relevance of model-level and late fusions in the context of a real application to assess their help in disambiguating real-life emotions.

## 3 REAL-LIFE EMOTIONAL DATA IN EMERGENCY CALL CENTER

The CEMO corpus used in this work was collected in an Emergency call center in France. The caller can be either the patient or a third party (family, friend, colleague, neighbor, home assistant). The agents are trained to quickly assess callers' states of mind, level of crisis, and urgency of the situation.

### 3.1 Emotional annotations

Table 1. Top 10 most represented emotions and emotion mixtures for patients and agent. (FEA)R, (NEU)TRAL, (POS)ITIVE, (ANG)ER, (SAD)NESS, (HUR)T, (SUR)PRISE, OTHER: total of all remaining classes

| Patient | #Segments | #Speakers | Agent | #Segments | #Speakers |
|---|---|---|---|---|---|
| Total | 17679 | 870 | Total | 16523 | 7 |
| FEA | 7397 | 825 | NEU | 10059 | 7 |
| NEU | 7329 | 822 | POS | 4310 | 7 |
| POS | 1187 | 566 | ANG | 1213 | 6 |
| ANG | 417 | 146 | FEA | 437 | 7 |
| HUR | 261 | 67 | FEA/POS | 122 | 4 |
| SUR | 144 | 118 | ANG/POS | 65 | 4 |
| FEA/POS | 130 | 103 | ANG/FEA | 57 | 3 |
| FEA/SAD | 128 | 71 | POS/SUR | 24 | 4 |
| FEA/HUR | 116 | 55 | FEA/SUR | 16 | 4 |
| OTHER | 294 | 171 | OTHER | 52 | 3 |

The complete CEMO corpus and the emotional data annotation scheme are described in [14, 39]. During the annotation process, two coders were given the opportunity to choose one major and one minor emotion for each extracted segment and the 21 fine-grained labels were grouped into seven macro-classes: Fear (Anxiety, Stress, Fear, Panic, Dismay, Embarrassment), Anger (Impatience, Annoyance, Cold Anger, Hot Anger), Sadness (Resignation, Disappointment, Sadness, Despair), Pain, Surprise, Positive (Relief, Interest, Compassion, Amusement), and Neutral. Table 1 gives the details of the 34202 annotated segments, from 756 real emergency call center conversations between 870 callers and 7 agents. In previous work [14], reports the inter-annotator agreement on the major macro-classes annotations between the two coders with a Kappa value of 0.54 for the callers and 0.35 for the agents. Since the agent's Kappa value is significantly lower, it was hypothesized that this could be due to the fact that agents need to remain calm and lucid in emergency situations and thus control their emotions, resulting in more complex emotional segments for the coders to annotate. In addition, as shown in Table 1, the agents' major emotions differ from those of the patients: in his or her work, the agent must be able to support the patients both morally and via medical assistance, thus there are fewer emotional segments for the agents.

### 3.2 Transcriptions

The manual transcripts were performed by 2 coders, using transcription guidelines similar to those used for spoken dialogues in Amities project [22]. Some additional markers were added to denote named-entities, breath, silence, unintelligible speech, throat clearing and other noises. The manual transcripts contain 2393 speech markers included 1472 silences, 751 mouth noise (i.e. breath) and 170 non-intelligible speech. The vocabulary size of the manual transcripts is 2.6k, with a mean and median of about 10 words per segment (minimum 1 word, maximum 47 words). An off-the-shelf Automatic Speech Recognition system (Factored TDNN, 70k-word vocabulary) for conversational telephone speech from VOCAPIA Research, a partner of the CNRS-LISN, was used to generate automatic transcripts for the calls. In addition to the word level transcription, the system also hypothesizes long silences and filler words. In our segments, there were 251 filler words located.

## 4 EXPERIMENTAL SETUP

In this section, we describe the data sets used for training and evaluation, as well as the models and strategies used in our research study. Appropriate data is critical for training deep learning models, especially for real-world applications.

### 4.1 Datasets: CEMO and IEMOCAP

- CEMO: In application specific data, emotions are scarce accounting from 10% (banking context) [22] to 30% (emergency call context) of speech turns/segments [14]. Data preparation is a key step to achieving good performance and robustness. Here we describe the selection of a balanced subset of the CEMO data used for model training and validation.

Table 2. Details of the CEMO subset based on speech signals and manual transcripts.

| **CEMO$_s$** | **ANGER** | **FEAR** | **NEUTRAL** | **POSITIVE** | **Total** |
|---|---|---|---|---|---|
| #Speech segments | 1056 | 1056 | 1056 | 1056 | 4224 |
| #Speakers | 149 | 537 | 450 | 544 | 812 |
| #Dialogues | 280 | 504 | 425 | 516 | 735 |
| Durations (min, s) | 39 | 52 | 49 | 20 | 160 |
| Duration mean (s) | 2.2 | 2.9 | 2.8 | 1.1 | 2.3 |
| Vocabulary size (#words) | 1146 | 1500 | 1150 | 505 | 2600 |
| Average word count | 9.3 | 11.9 | 7.9 | 3.8 | 8.2 |

We mainly focused on negative emotions such as stressed (Fear) or impatient (Anger) which could lead to bad decision making in an emergency call center. The positive emotions such as relief or interest were merged in a unique class Positive. In order to build a balanced database, first segments with matching annotations for the four major macroclasses: Anger, Fear, Positive and Neutral were selected, after which we excluded segments outside of 0.4 and 7.5 seconds to avoid high computational costs. Note that the data distribution reported above in Table 1 is prior to this trimming. After this first filtering, there are only 386 samples of Anger, 1056 Positive samples and around 6000 Fear and Neutral segments. The Anger class was then completed with segments from the agents[1] to obtain 1056 samples, balancing across the agents. The Fear and Neutral classes were subsampled, maintaining 1056 samples for each class, prioritizing a broad representation of speaker diversity and segments for which the annotators were in agreement. This results in a smaller number of speakers for Anger compared to the other classes as can be seen in Table 2, as there are only 7 agents in the CEMO corpus. Compared to the subset of the CEMO corpus that was used in a prior study [12], we selected a more balanced and richer subset for this paper which we refer to as $\text{CEMO}_s$. $\text{CEMO}_s$ is comprised of 4224 segments (2h40) equally distributed over the 4 main emotion classes. As can be noted in Table 2, the Positive class has the largest number of speakers and dialogues, potentially being richer and more heterogeneous than the other classes. However, at the same time, the total duration of the Positive segments and the average number of words, is less than that of the other classes. Consequently, the vocabulary size is also largely reduced.

- IEMOCAP: So as to place our studies within the scientific community, we also report results on the widely used IEMOCAP [5] data in Section 6. This database was recorded from ten actors in dyadic sessions during hypothetical oral communication scenarios for the purpose of eliciting emotions. The Table 3 summarizes the characteristics of the IEMOCAP dataset.

Table 3. Details and distribution of the IEMOCAP corpus.

| IEMOCAP | ANGER | SADNESS | NEUTRAL | HAPPY | Total |
|---|---|---|---|---|---|
| #speech segments | 1103 | 1084 | 1708 | 595 | 4490 |
| #Speakers | 10 | 10 | 10 | 10 | 10 |
| #Dialogues | 84 | 70 | 135 | 76 | 151 |
| Duration (min, s) | 83 | 99 | 111 | 43 | 336 |
| Duration mean (s) | 4.5 | 5.5 | 3.9 | 4.3 | 4.5 |

### 4.2 Architectures

This section describes the pre-trained models which are publicly available off-the-shelf systems at "https://huggingface.co/" that were explored in our experiments on speech emotion detection with the French corpus CEMO. We tested wav2vec 2.0 models, BERT models and fusions.

- Wav2vec2 models: Wav2vec2 is a self-supervised pre-trained model that learns to predict a masked part of the signal provided as input. Our aim is to assess the impact of a pre-trained model to provide powerful representations that can be adapted to the CEMO task. The databases used in pre-trained models are detailed in [9, 36]. They differ in the choice of the languages (mono or multilingual), the types and amount of data composing the training database, the styles of speech (read, spontaneous/acted, conversational), the number of speakers (accents, gender, age), the emotional content and the quality of the recordings (noise, distance of the microphone). We selected, among the available pre-trained wav2vec2 encoders, 2 encoders which included in their training database at least one of the following criteria: French data, spontaneous dialogs, telephone-recorded data and emotional content, as shown in the Table 4.

[1]In Table 1 it can be seen that there are more segments annotated with anger for the agents than for the patients.

Table 4. Speech corpora used to train the publicly available encoder models

| Encoder model | Pre-training | Total hours | French | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Total | Read | Broadcast | Spontaneous | Acted telephone | Acted emotional |
| Wav2vec2-xlsr-53$_{LARGE}$ | 53 languages | 56 K | 1500 | 1500 | - | - | - | - |
| Wav2vec2-FR-3K$_{LARGE}$ | French | 2.9 K | 2900 | 1100 | 1600 | 123 | 38 | 29 |

Table 5. Statistics for the text corpora used to train the encoder models

| Encoder model | Tokenizer | Masking strategy | Parameters | French pre-training data | Number of tokens | Size |
|---|---|---|---|---|---|---|
| CamemBERT$_{BASE}$<br>CamemBERT$_{LARGE}$ | SentencePiece 32K | Whole-word | 110M<br>335M | CCNet<br>(135 GB of text) | 32.7B | 59.4M<br>documents |
| FlauBERT$_{BASE}$<br>FlauBERT$_{LARGE}$ | BPE 50K | Sub-word | 138M<br>373M | 24 French subcorpora<br>(71 Gb of text) | 12.79 B | 488.78M<br>sentences |

- BERT models: We selected, among the available pre-trained models based on the BERT architecture, two widely used French language models: CamemBERT [29] and FlauBERT [25]. More specifically we used BERT$_{BASE}$ (12 layers, 768 hidden dimensions, 12 attention heads) and BERT$_{LARGE}$ (24 layers, 1024 hidden dimensions, 16 attention heads). CamemBERT$_{BASE}$ and CamemBERT$_{LARGE}$ both of which were trained on the French dataset CCNet [24]) with different filtering processes of the CommonCrawl database[2]. According to the authors, CCNet was constructed with a language model trained on Wikipedia, making it able to filter out noise (tables, code, etc). FlauBERT$_{BASE}$ and $_{LARGE}$ are also trained on a filtered part of the CommonCrawl database in addition to some sources from Wikipedia, books, news and subtitles [25]. The tokenizer and masking strategy also differ in both models see Table 5. According to Wang et al. [41] and Fan et al. [32], the training of large Transformers is known to be sensitive to instability and normalization techniques are typically used to train these models [25].
- Fine-tuning: In order to adapt the wav2vec2 and BERT models, we added a classifier on top of them, adapted to our speech emotion recognition task, as detailed in Figure 2. To reduce the computational overhead of an experiment we chose to use a pre-trained encoder as a feature extractor, and subsequently simply train a classifier on the generated features. We tested different variations of fine-tuning with the wav2vec2 and BERT encoders: No encoder fine-tuning, Convolutional layers (wav2vec)/ embedding layers (BERT) frozen and Full fine-tuning, as shown in Figure 2.
- Fusion strategies: We explored two fusion strategies [20]: late fusion (or decision-level fusion) which consists of combining predictions (in this paper we simply average the emotional class scores of the model outputs) and the model-level fusion which concatenates the intermediate representation of each model to learn potential hidden correlations between features as shown in Figure 3.

[2]https://commoncrawl.org/

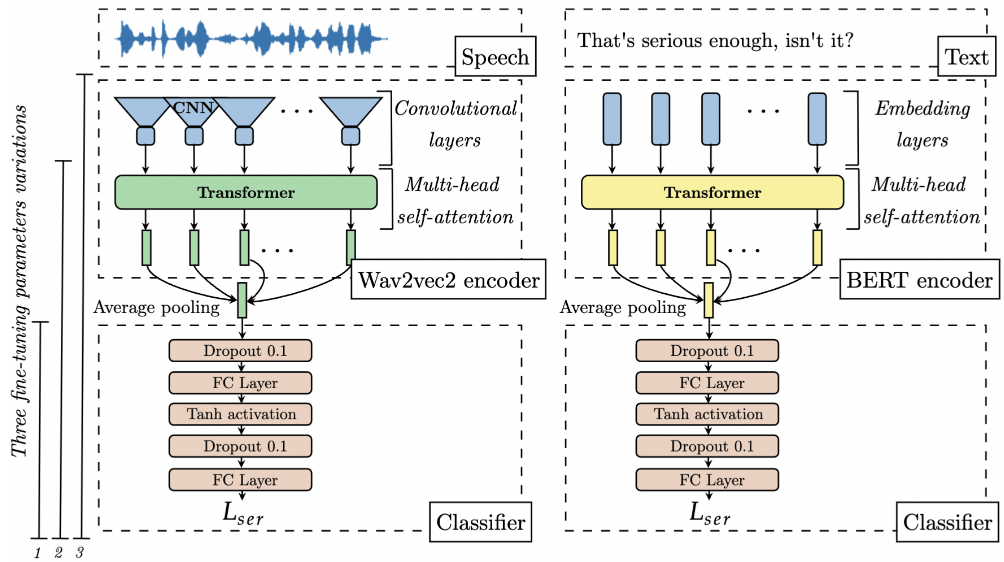



Fig. 2. Three variations of fine-tuning of the learning parameters:
(1) No encoder fine-tuning, (2) Convolutional layers (wav2vec) / Embedding layers (BERT) frozen, (3) Full fine-tuning.

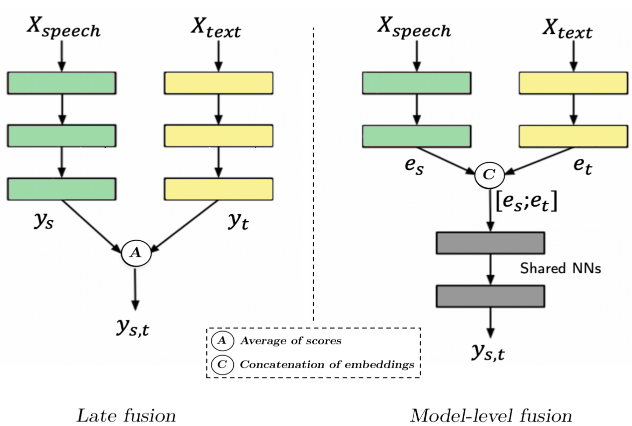



Fig. 3. Two fusions strategies: Late fusion and Model-level fusion,
$y_s$: Speech outputs, $y_t$: Text outputs, $e_s$: Speech embeddings, $e_t$: Text embeddings, $y_{s,t}$: Speech and Text outputs

## 5 EXPERIMENTAL RESULTS ON CEMO

In this section we present results obtained with several pre-trained models (wav2vec2-xlsr-53, wav2vec2-Fr-3K, CamemBERT and FlauBERT) and with a baseline model (Temporal CNN-BiLSTM) similar to that used in previous work [12]. For each pre-trained model we added a classification layer including the averaging of the outputs of our pre-trained layers followed by twice a sequence of a dropout of 10% layer and a dense linear (adapted to the size of the pre-trained model), with a tanh activation layer in between as shown in Figure 2. All evaluations are performed on 5 folds with a classical cross-speaker folding strategy that is speaker independent between training, validation and test sets. During each fold, system training is optimized on the best Unweighted Accuracy (UA) of the validation set. The outputs of each fold are combined for the final results.

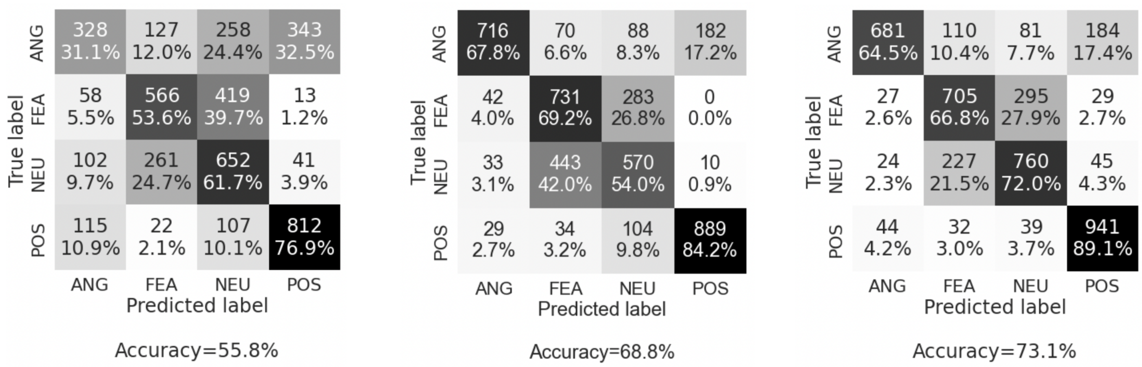
### 5.1 Speech modality

The results in this subsection compare the performance of the pre-trained models without and with adaptation using the CEMO data. Table 4 gives the emotion recognition results with three models using the speech modality: the baseline CNN-BiLSTM model similar to [12], the wav2vec2-xlsr-53 model [10] and the wav2vec2-French model [17]. To evaluate the contribution of the pre-trained models, we randomly initialized their parameters and compared them to runs with the pre-trained parameters. We also tested three versions of fine-tuning following the procedure of the fine-tuning section in 4.2 as shown in Figure 2.

Table 6. Comparative experiments for the speech modality.
(Train)able (p)arameters, Unweighted Accuracy (UA: %)

| **Encoder model** | **Model name** | **Pre-trained language** | **Fine-tuning** | **Train. p.** | **ANG** | **FEA** | **NEU** | **POS** | **UA** |
|---|---|---|---|---|---|---|---|---|---|
| Temporal CNN-BiLSTM | - | - | - | 8M | 31.1 | 53.6 | 61.7 | 76.9 | **55.8** |
| Wav2vec 2.0$_{LARGE}$ | - | Random initialization | No encoder fine-tuning | 1.1 M | 34.5 | 9.2 | 44.9 | 31.6 | 30.0 |
| | | | Convolutional layers frozen | 312 M | 7.1 | 22.6 | 12.7 | 74.1 | 29.1 |
| | | | Full fine-tuning | 316 M | 39.9 | 8.7 | 30.6 | 16.1 | 23.8 |
| | Wav2vec2-xlsr-53$_{LARGE}$ | 53 languages including French - 1.5K hours | No encoder fine-tuning | 1.1 M | 7.0 | 17.3 | 63.1 | 47.1 | 33.6 |
| | | | Convolutional layers frozen | 312 M | 54.1 | 49.1 | 63.7 | 80.8 | 61.9 |
| | | | Full fine-tuning | 316 M | 67.8 | 69.2 | 54.0 | 84.2 | **68.8** |
| | Wav2vec2-FR-3K$_{LARGE}$ | French - 3K hours | No encoder fine-tuning | 1.1 M | 26.6 | 16.7 | 35.9 | 31.8 | 27.7 |
| | | | Convolutional layers frozen | 312 M | 64.5 | 66.8 | 72.0 | 89.1 | **73.1** |
| | | | Full fine-tuning | 316 M | 60.9 | 60.0 | 76.5 | 83.9 | 70.3 |

The Positive and Neutral classes seem to be the easiest to model in our context. The two wav2vec2 encoders in their best configurations outperform the Temporal CNN-LSTM baseline model (68.8% and 73.1% versus 55.8%, c.f. Table 6). Pre-training the convolutional layers appears to provide a better parameter initialization compared to random initialization (entry Convolutional layers frozen: 73.1% and 61.9% vs. 29.1%). The results with fine-tuning exhibit different trends for the wav2vec2-xlsr-53$_{LARGE}$ and the wav2vec2-FR-3K$_{LARGE}$ models. For the multilingual model, full fine-tuning gives the best result, where as for the French model, the best result is obtained by freezing the convolutional layers. Effectively, transfer learning shows its potential for speech emotion recognition when combined with an appropriate version of fine-tuning (either Convolution layers frozen or Full fine-tuning) (68.8% and 73.1%, see Table 6). The CEMO balanced subset contains only 2.6 hours 2, so the inclusion of additional knowledge about spoken French (1500 hours for Wav2vec2-xlsr-53$_{LARGE}$ and 2900 hours for Wav2vec2-FR-3K$_{LARGE}$, see Table 4) is seen to improve the results. Moreover, the similarity of the datasets used for pre-training with the CEMO corpus might have led to better performance. It appears that the difference in performance of the two pre-trained models (8.3%) can be attributed to differences in the pre-trained datasets. Indeed, the presence of monolingual (French) versus multilingual samples or the presence of spontaneous speech (123 hours), acted telephone (38 hours) and acted emotional (29 hours) samples, as detailed in Table 4 which are similar to the CEMO context, could explain the better performance of the Wav2vec2-FR-3K$_{LARGE}$ model.




(a) Temporal CNN-BiLSTM model

(b) Wav2vec2-xlsr-53$_{LARGE}$ (Full fine-tuning)

(c) Wav2vec2-FR-3K$_{LARGE}$ (Convolutional layers frozen)

Fig. 4. Confusion matrices of three speech emotion recognition models with CEMO

Comparing the confusion matrix of our baseline model in Figure 4a, a significant improvement in the separation of our classes and particularly for our pair Anger-Positive can be seen in Figure 4c. If we compare the best configuration of each wav2Vec2 model in the Table 6 and plot their confusion matrices in Figures 4b and 4c, we notice that the wav2vec2-xlsr-53$_{LARGE}$ still has a difficulty discriminating the Fear-Neutral pair (as is the case for the CNN-BiLSTM model), while the wav2vec2-FR-3K$_{LARGE}$ does a better job at disambiguating this pair. Indeed, the overall improvement in Unweighted Accuracy seems to come from the model pre-training on appropriate datas.

### 5.2 Textual modality

Four models were explored with the textual modality: CamemBERT$_{BASE}$, CamemBERT$_{LARGE}$, FlauBERT$_{BASE}$ and FlauBERT$_{LARGE}$. First, two sizes for each model (base and large) were tested with No encoder fine-tuning on our manual transcripts to select the best off-the-shelf system on our task. We followed the same fine-tuning nomenclature as in Section 4.2. The results of the base models, CamemBERT$_{BASE}$, and FlauBERT$_{BASE}$ were less good so we decided to use the larger models in the remaining experiments.

Table 7. Comparative experiments with text-based models using manual transcripts with the No encoder fine-tuning configuration (Train)able (p)arameters, UA: %

| **Model name** | **Train. p.** | **ANG** | **FEA** | **NEU** | **POS** | **Total** |
|---|---|---|---|---|---|---|
| CamemBERT$_{LARGE}$ | 1.1 M | 53.6 | 60.8 | 58.7 | 67.6 | 60.2 |
| FlauBERT$_{LARGE}$ | 1.1 M | 56.4 | 60.2 | 70.8 | 81.0 | **67.1** |

FlauBERT$_{LARGE}$ produced the best overall score for emotion detection. A possible explanation for the better performance of FlauBERT$_{LARGE}$ compared to CamemBERT$_{LARGE}$ could be the larger vocabulary size (50K vs 32K tokens, see Table 5) of FlauBERT$_{LARGE}$. Due to its bigger vocabulary, FlauBERT$_{LARGE}$ might be able to better process French real-life conversational transcripts and provide better contextual representation. We used FlauBERT in its *LARGE* configuration (1024 dimensional output vectors) for the following experiments, as it yielded the best results.

Table 8. Comparative fine-tuning experiments on manual transcripts.
(Train)able (p)arameters, UA: %

| Models | Config. | Train. p. | ANG | FEA | NEU | POS | Total |
|---|---|---|---|---|---|---|---|
| $\text{FlauBERT}_{LARGE}$ | No encoder fine-tuning | 1.1 M | 56.4 | 60.2 | 70.8 | 81.0 | 67.1 |
| | Embedding layers frozen | 303 M | 57.0 | 63.4 | 67.4 | 83.3 | **67.8** |
| | Full fine-tuning | 374 M | 62.3 | 61.3 | 63.1 | 78.8 | 66.4 |

We fine-tuned the FlauBERT large model at three levels as described the section 4.2. As illustrated in Table 8, the fine-tuning on our task is not very useful for the classification. The fine-tuning version (Embedding layers frozen) is slightly better than the No encoder fine-tuning, but there is a huge gap in the number of trainable parameters, the FlauBERT (Embedding layers frozen) configuration model is 300 times bigger than the FlauBERT (No encoder fine-tuning). We kept the $\text{FlauBERT}_{LARGE}$ (No encoder fine-tuning) for our next experiment, where we use automatic transcriptions rather than the manual ones for emotion classification.

It can be seen in Table 9 that compares the emotion recognition results with the automatic transcripts are quite close to those obtained with the manual ones. This is a very interesting result as it means that we can use a semi-automatic protocol for transcribing and annotating more data in future experiments.

Table 9. Comparative experiments on transcript configurations.
(Tr)anscriptions (type): Manual/Auto., UA: %

| Models | Tr. type | ANG | FEA | NEU | POS | Total |
|---|---|---|---|---|---|---|
| $\text{FlauBERT}_{LARGE}$ | Manual | 56.4 | 60.2 | 70.8 | 81.0 | **67.1** |
| | Auto. | 57.6 | 66.2 | 65.9 | 81.9 | **67.9** |

### 5.3 Fusion strategies

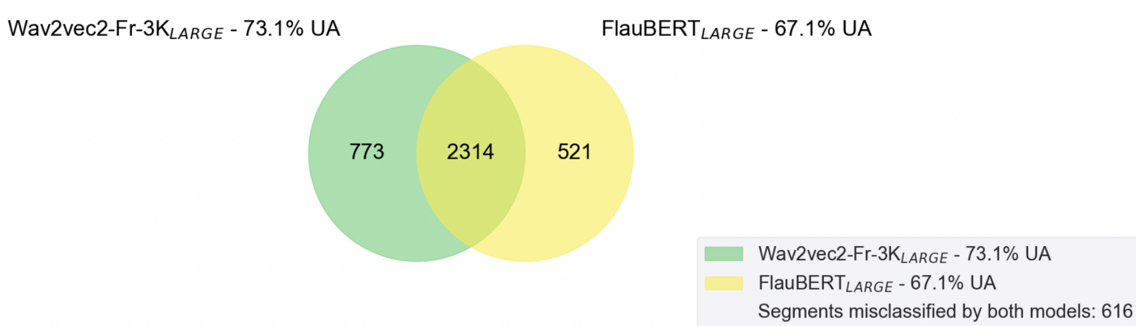



Fig. 5. Correct classifications for the 2 models on our dataset (4224 samples), i.e. $\text{Wav2vec2-FR-3K}_{LARGE}$ (Convolutional layers frozen) correctly classified $(773 + 2314)/4224 \equiv 73,1\%$ UA.

To assess the potential fusion of the two modalities (speech and text), a comparison was made of the segments correctly classified by the best pre-trained models, respectively, $\text{wav2vec2-FR-3K}_{LARGE}$ (frozen convolutional layers)

Table 10. Fusion strategy experiments, with speech and manual transcripts, UA: %

| Model name | Modality | UA |
|---|---|---|
| Wav2vec2-Fr-3K$_{LARGE}$ | Speech | 73.1 |
| FlauBERT$_{LARGE}$ | Text | 67.1 |
| Model-level fusion: | | |
| Wav2vec2-Fr-3K$_{LARGE}$ + FlauBERT$_{LARGE}$ | Speech + Text | 76.8 |
| Late fusion: | | |
| Wav2vec2-Fr-3K$_{LARGE}$ + FlauBERT$_{LARGE}$ | Speech + Text | 77.1 |

and FlauBERT$_{LARGE}$ (No encoder fine-tuning). The Venn diagram in Figure illustrates their complementarity, explicitly showingthe number of segments correctly classified by each model individually and the number of segments classified by both. The two models share 2314 segments correctly classified (54.8% UA), and there is still over 1200 segments correctly classified by only one of the models that could be exploited with a multimodal system.

We explicitly show here two transcribed segments that were classified differently by the speech encoder (wav2vec2-FR-3K$_{LARGE}$ (frozen convolutional layers)) and the text encoder (FlauBERT$_{LARGE}$ (no encoder fine-tuning)):

*Example 1:* **Manual transcript**: "et j' ai appelé le le médecin SOS" - *"and I called the the SOS doctor"*
*Example 2:* **Manual transcript**: "Je je pas à côté mais juste en face" - *"I I not next door but right across the street"*

Example 1 was correctly classified as Fear by FlauBERT$_{LARGE}$ but classified as Neutral by wav2vec2-FR-3K$_{LARGE}$. Indeed, this segment sounds completely neutral when we listen to it but the content clearly demonstrates an urgency and was annotated as Fear (The two annotators detected Anxiety at a fined-grained emotion level). The second example was correctly classified as Fear by wav2vec2-FR-3K$_{LARGE}$, but classified as Neutral by FlauBERT$_{LARGE}$ and had been annotated as Fear (One coder indicated Stress and Anxiety at a micro emotion level). Even in an emergency context the agent needs to get precise information about the location of the callers, such a description is often provided by the patients and could be detected as Neutral segments by the linguistic model. This shows the importance of using both modalities to detect emotions in real-life spontaneous conversational speech segments.

As mentioned in Section 4.2, we studied two fusion strategies, model-level fusion and late fusion. As can be seen in 10, both fusions improve our system with the addition of explicit linguistic information to original the speech information, obtaining a UA of about 77% for both fusion strategies.

## 6 EXPERIMENTAL RESULTS ON CEMO AND IEMOCAP

In order to situate our work for the scientific community, we report results on IEMOCAP with our best configurations. All models use a five-fold cross-validation strategy, independent of the speaker. The scientific contributions have no fixed pattern to determine the validation set [16], so for example in IEMOCAP (10 speakers) we dedicate in each fold, 4 sessions for training (8 speakers) and divide the last session for validation (1 speaker) and testing (1 speaker). IEMOCAP is an english corpus, so we used English pre-trained models, the wav2vec2$_{BASE}$ and RoBERTa$_{BASE}$ which shared the same neural network architecture with wav2vec2-Fr-3K$_{LARGE}$ and FlauBERT$_{LARGE}$ in a different size.

Table 11. Experimental results on CEMO and related work with IEMOCAP. The results in the top part of the table are ours, with some of the closest reported work on IEMOCAP in the bottom, UA: %

| Model | Modality | IEMOCAP | CEMO |
|---|---|---|---|
| Wav2vec2$_{BASE}$ / Wav2vec2-Fr-3K$_{LARGE}$ | Speech | 65.4 | 73.1 |
| RoBERTa$_{BASE}$ / FlauBERT$_{LARGE}$ | Text | 56.2 | 67.1 |
| <u>Late fusion:</u> | | | |
| Wav2vec2$_{BASE}$ + RoBERTa$_{BASE}$ / Wav2vec2-Fr-3K$_{LARGE}$ + FlauBERT$_{LARGE}$ | Speech + Text | 70.6 | 77.1 |
| <u>Model-level fusion:</u> | | | |
| Wav2vec2$_{BASE}$ + RoBERTa$_{BASE}$ / Wav2vec2-Fr-3K$_{LARGE}$ + FlauBERT$_{LARGE}$ | Speech + Text | **70.8** | 76.8 |
| LSTM w. attention [31] | Speech | 58.8 | - |
| CNN-LSTM [35] | Speech | 59.4 | - |
| TDNN-LSTM w. attention [34] | Speech | 60.7 | - |
| Wav2vec [4] | Speech | 64.3 | - |
| BiLSTM w. GloVe embedding [43] | Text | 57.8 | |
| LSTM model-level fusion [43] | Speech + Text | 67.7 | - |
| LSTM attention fusion [43] | Speech + Text | 70.9 | - |

We validated these fusion strategies against similar work with IEMOCAP in the same configuration, as shown in the table 11. With CEMO, we explored the use of pre-trained models and classical fusion strategies to improve the performances of our models and showed better results with late fusion of the speech and text models (77.1 %).

## 7 ETHICS AND REPRODUCIBILITY

The use of the CEMO database or any subsets of it, carefully respected ethical conventions and agreements ensuring the anonymity of the callers. All the experiments were carried out using Pytorch on two GPUs (GeForce GTX 1080 Ti with 11 Gbytes of RAM). We used Adam Optimizer with a learning rate of $2x10^{-5}$ (IEMOCAP) and $10^{-4}$ (CEMO) per step. To ensure the reproducibility of the runs, we set a random seed to 0 and prevent our system from using non-deterministic algorithms.

To be comparable to related work with IEMOCAP (10 speakers) cited in this work, we dedicate in each fold, 4 sessions for training (8 speakers) and divide the last session for validation (1 speaker) and testing (1 speaker).

## 8 DISCUSSION AND CONCLUSION

In these studies, we explored and adapted several pre-trained models for speech emotion recognition in a real-world context. The use and adaptation of self-supervised representations, such as Transformer encoders previously trained on large and varied of corpora, provided reasonable generalizations of performance on unseen data. In particular, we compared several pre-trained models and our hypothesis is that the good performance of the French wav2vec2-Fr-3K$_{LARGE}$ model on the CEMO corpus can be attributed to the amount of French data used during pre-training and its similarity with the CEMO corpus. These characteristics may in part explain the better performance of this model over its multilingual version wav2vec2-xlsr-53$_{LARGE}$, (73.1% versus 68.8% UA). Nevertheless both models showed an overall gain in performance compared to the baseline Temporal CNN-LSTM model (55.8%). In addition, fine-tuning of the pre-trained models for Speech Emotion Recognition with CEMO$_s$ was essential for both wav2vec2 models' performance. Indeed wav2vec2-xlsr-$53_{LARGE}$ increased from 33.6% (without fine-tuning) to 68.8% (with fine-tuning) and the wav2vec2-Fr-$3K_{LARGE}$ increased from 27.7% (without fine-tuning) to 73.1% (with fine-tuning). For the textual modality, fine-tuning the core layers of the BERT-based models was not useful, undoubtedly due to the limited vocabulary variety and training set size of CEMO, compared to the amount of data used to pre-train the BERT-based models. The good predictions on automatic transcripts bodes a strong future for speech emotion recognition in real-world applications. We also studied the complementarity of speech and text modalities with the manual transcripts, combining them with two variants of fusion mechanisms (model-level and late fusion) obtaining respectively 76.8% and 77.1% UA. This combined approach helped predicting more complex segments of the CEMO corpus, when speech characteristics deviate from the surface meaning of the transcripts, which occurs when expressing irony or when people attempt to control or exaggerate their emotions.

Future research will now focus on finding creative joint-encoding methods across modalities and using semi-automated methods to annotate a larger corpus of call center conversations.

## 9 ACKNOWLEDGMENTS

### ACKNOWLEDGMENTS

The PHD thesis of Theo Deschamps-Berger is supported by the ANR AI Chair HUMAAINE at LISN-CNRS, led by Laurence Devillers and reuniting researchers in computer science, linguists and behavioral economists from the Paris-Saclay University. The data annotation work was partially financed by several EC projects: FP6-CHIL and NoE HUMAINE. The authors would like to thank, M. Lesprit and J. Martel for their help with data annotation. The work is conducted in the framework of a convention between the APHP France and the LIMSI-CNRS.

## 10 CITATIONS AND BIBLIOGRAPHIES

### REFERENCES

[1] Francisca Acheampong, Henry Nunoo-Mensah, and Wenyu Chen. 2021. *Transformer Models for Text-based Emotion Detection: A Review of BERT-based Approaches.*

[2] Alexei Baevski, Steffen Schneider, and Michael Auli. 2020. Vq-Wav2vec: Self-Supervised Learning of Discrete Speech Representations. *arXiv:1910.05453 [cs]* (Feb. 2020). arXiv:1910.05453 [cs]

[3] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *arXiv:2006.11477 [cs, eess]* (Oct. 2020). arXiv:2006.11477 [cs, eess]

[4] Jonathan Boigne, Biman Liyanage, and Ted Östrem. 2020. Recognizing More Emotions with Less Data Using Self-supervised Transfer Learning. https://doi.org/10.48550/arXiv.2011.05585 arXiv:2011.05585 [cs, eess]

[5] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: Interactive Emotional Dyadic Motion Capture Database. *Language Resources and Evaluation* 42, 4 (Nov. 2008), 335. https://doi.org/10.1007/s10579-008-9076-6

[6] Carlos Busso, Serdar Yildirim, Murtaza Bulut, Chul Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, and Shrikanth Narayanan. 2004. *Analysis of Emotion Recognition Using Facial Expressions, Speech and Multimodal Information.* 211 pages. https://doi.org/10.1145/1027933.1027968

[7] Ming Chen and Xudong Zhao. 2020. A Multi-Scale Fusion Framework for Bimodal Speech Emotion Recognition. In *Interspeech 2020*. ISCA, 374–378. https://doi.org/10.21437/Interspeech.2020-3156

[8] Jianpeng Cheng, Li Dong, and Mirella Lapata. 2016. Long Short-Term Memory-Networks for Machine Reading. *arXiv:1601.06733 [cs]* (Sept. 2016). arXiv:1601.06733 [cs]

[9] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised Cross-lingual Representation Learning for Speech Recognition. https://doi.org/10.48550/arXiv.2006.13979 arXiv:2006.13979 [cs, eess]

[10] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised Cross-lingual Representation Learning for Speech Recognition. *arXiv:2006.13979 [cs, eess]* (Dec. 2020). arXiv:2006.13979 [cs, eess]

[11] Jean-Benoit Delbrouck, Noé Tits, and Stéphane Dupont. 2020. Modulated Fusion Using Transformer for Linguistic-Acoustic Emotion Recognition. *arXiv:2010.02057 [cs]* (Oct. 2020). arXiv:2010.02057 [cs]

[12] Theo Deschamps-Berger, Lori Lamel, and Laurence Devillers. 2021. *End-to-End Speech Emotion Recognition: Challenges of Real-Life Emergency Call Centers Data Recordings*. 8 pages. https://doi.org/10.1109/ACII52823.2021.9597419

[13] Laurence Devillers and Laurence Vidrascu. 2006. *Real-Life Emotions Detection with Lexical and Paralinguistic Cues on Human-Human Call Center Dialogs*.

[14] Laurence Devillers, Laurence Vidrascu, and Lori Lamel. 2005. Challenges in Real-Life Emotion Annotation and Machine Learning Based Detection. *Neural Networks* 18, 4 (May 2005), 407–422. https://doi.org/10.1016/j.neunet.2005.03.007

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]* (May 2019). arXiv:1810.04805 [cs]

[16] Caroline Etienne, Guillaume Fidanza, Andrei Petrovskii, Laurence Devillers, and Benoit Schmauch. 2018. CNN+LSTM Architecture for Speech Emotion Recognition with Data Augmentation. 21–25. https://doi.org/10.21437/SMM.2018-5

[17] Solene Evain, Ha Nguyen, Hang Le, Marcely Zanon Boito, Salima Mdhaffar, Sina Alisamir, Ziyi Tong, Natalia Tomashenko, Marco Dinarelli, Titouan Parcollet, Alexandre Allauzen, Yannick Esteve, Benjamin Lecouteux, Francois Portet, Solange Rossato, Fabien Ringeval, Didier Schwab, and Laurent Besacier. 2021. LeBenchmark: A Reproducible Framework for Assessing Self-Supervised Representation Learning from Speech. *arXiv:2104.11462 [cs, eess]* (June 2021). arXiv:2104.11462 [cs, eess]

[18] Florian Eyben, Klaus R. Scherer, Bjorn W. Schuller, Johan Sundberg, Elisabeth Andre, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong. 2016. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing* 7, 2 (April 2016), 190–202. https://doi.org/10.1109/TAFFC.2015.2457417

[19] Jing Han, Zixing Zhang, Nicholas Cummins, Fabien Ringeval, and Björn Schuller. 2017. Strength Modelling for Real-World Automatic Continuous Affect Recognition from Audiovisual Signals. *Image and Vision Computing* 65 (Sept. 2017), 76–86. https://doi.org/10.1016/j.imavis.2016.11.020

[20] Jing Han, Zixing Zhang, Zhao Ren, and Björn Schuller. 2021. EmoBed: Strengthening Monomodal Emotion Recognition via Training with Crossmodal Emotion Embeddings. *IEEE Transactions on Affective Computing* 12, 3 (July 2021), 553–564. https://doi.org/10.1109/TAFFC.2019.2928297 arXiv:1907.10428 [cs, eess]

[21] Jing Han, Zixing Zhang, Fabien Ringeval, and Björn Schuller. 2017. Prediction-Based Learning for Continuous Emotion Recognition in Speech. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 5005–5009. https://doi.org/10.1109/ICASSP.2017.7953109

[22] H. Hardy, K. Baker, L. Devillers, L. Lamel, S. Rosset, T. Strzalkowski, Cristian Ursu, and N. Webb. 2003. Multi-Layer Dialogue Annotation for Automated Multilingual Customer Service. *undefined* (2003).

[23] Wei-Ning Hsu, Anuroop Sriram, Alexei Baevski, Tatiana Likhomanenko, Qiantong Xu, Vineel Pratap, Jacob Kahn, Ann Lee, Ronan Collobert, Gabriel Synnaeve, and Michael Auli. 2021. Robust Wav2vec 2.0: Analyzing Domain Shift in Self-Supervised Pre-Training. *arXiv:2104.01027 [cs, eess]* (Sept. 2021). arXiv:2104.01027 [cs, eess]

[24] Zilong Huang, Xinggang Wang, Yunchao Wei, Lichao Huang, Humphrey Shi, Wenyu Liu, and Thomas S. Huang. 2020. CCNet: Criss-Cross Attention for Semantic Segmentation. https://doi.org/10.48550/arXiv.1811.11721 arXiv:1811.11721 [cs]

[25] Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised Language Model Pre-training for French. *arXiv:1912.05372 [cs]* (March 2020). arXiv:1912.05372 [cs]

[26] Pengfei Liu, Kun Li, and Helen Meng. 2022. Group Gated Fusion on Attention-based Bidirectional Alignment for Multimodal Emotion Recognition. https://doi.org/10.48550/arXiv.2201.06309 arXiv:2201.06309 [cs, eess]

[27] Manon Macary, Marie Tahon, Yannick Estève, and Anthony Rousseau. 2021. On the Use of Self-supervised Pre-trained Acoustic and Linguistic Features for Continuous Speech Emotion Recognition. In *IEEE Spoken Language Technology Workshop*. Virtual, China.

[28] Mariana Rodrigues Makiuchi, Kuniaki Uto, and Koichi Shinoda. 2021. Multimodal Emotion Recognition with High-Level Speech and Text Features. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. 350–357. https://doi.org/10.1109/ASRU51503.2021.9688036

[29] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: A Tasty French Language Model. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020), 7203–7219. https://doi.org/10.18653/v1/2020.acl-main.645 arXiv:1911.03894

[30] Angeliki Metallinou, Sungbok Lee, and Shrikanth Narayanan. 2010. Decision Level Combination of Multiple Modalities for Recognition and Analysis of Emotional Expression. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2462–2465. https://doi.org/10.1109/ICASSP.2010.5494890

[31] Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang. 2017. *Automatic Speech Emotion Recognition Using Recurrent Neural Networks with Local Attention*. https://doi.org/10.1109/ICASSP.2017.7952552

[32] Toan Q. Nguyen and Julian Salazar. 2019. Transformers without Tears: Improving the Normalization of Self-Attention. In *Proceedings of the 16th International Conference on Spoken Language Translation*. Association for Computational Linguistics, Hong Kong.

[33] Leonardo Pepino, Pablo Riera, and Luciana Ferrer. 2021. Emotion Recognition from Speech Using Wav2vec 2.0 Embeddings. *arXiv:2104.03502 [cs, eess]* (April 2021). arXiv:2104.03502 [cs, eess]

[34] Mousmita Sarma, Pegah Ghahremani, Daniel Povey, N. Goel, K. K. Sarma, and N. Dehak. 2018. Emotion Identification from Raw Speech Signals Using DNNs. In *INTERSPEECH*. https://doi.org/10.21437/Interspeech.2018-1353

[35] Aharon Satt, Shai Rozenberg, and Ron Hoory. 2017. Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms. In *Interspeech 2017*. ISCA, 1089–1093. https://doi.org/10.21437/Interspeech.2017-200

[36] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. Wav2vec: Unsupervised Pre-training for Speech Recognition. *arXiv:1904.05862 [cs]* (Sept. 2019). arXiv:1904.05862 [cs]

[37] Panagiotis Tzirakis, Anh Nguyen, Stefanos Zafeiriou, and Björn W. Schuller. 2021. Speech Emotion Recognition Using Semantic Information. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 6279–6283. https://doi.org/10.1109/ICASSP39728.2021.9414866

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *arXiv:1706.03762 [cs]* (Dec. 2017). arXiv:1706.03762 [cs]

[39] Laurence Vidrascu and Laurence Devillers. 2005. Detection of Real-Life Emotions in Call Centers. 1841–1844.

[40] Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Florian Eyben, and Björn Schuller. 2022. *Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap.*

[41] Haoyu Wang, Ming Tan, Mo Yu, Shiyu Chang, Dakuo Wang, Kun Xu, Xiaoxiao Guo, and Saloni Potdar. 2019. *Extracting Multiple-Relations in One-Pass with Pre-Trained Transformers.* 1377 pages. https://doi.org/10.18653/v1/P19-1132

[42] Matthias Wimmer, Björn Schuller, Dejan Arsic, Gerhard Rigoll, and Bernd Radig. 2008. *Low-Level Fusion of Audio, Video Feature for Multi-Modal Emotion Recognition.* 151 pages.

[43] Haiyang Xu, Hui Zhang, Kun Han, Yun Wang, Yiping Peng, and Xiangang Li. 2020. Learning Alignment for Multimodal Emotion Recognition from Speech. *arXiv:1909.05645 [cs, eess]* (April 2020). arXiv:1909.05645 [cs, eess]

[44] Zhihong Zeng, Maja Pantic, and Glenn Roisman. 2009. A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE transactions on pattern analysis and machine intelligence* 31 (Feb. 2009), 39–58. https://doi.org/10.1109/TPAMI.2008.52