Mutual impact of acoustic and linguistic representations for continuous emotion recognition in call center conversations

Manon Macary, Marie Tahon, Yannick Estève, Daniel Luzzati

Abstract—The goal of our research is to automatically retrieve the satisfaction and the frustration in real-life call-center conversations. This study focuses an industrial application in which the customer satisfaction is continuously tracked down to improve customer services. To compensate the lack of large annotated emotional databases, we explore the use of pre-trained speech representations as a form of transfer learning towards AlloSat corpus. Moreover, several studies have pointed out that emotion can be detected not only in speech but also in facial trait, in biological response or in textual information. In the context of telephone conversations, we can break down the audio information into acoustic and linguistic by using the speech signal and its transcription. Our experiments confirms the large gain in performance obtained with the use of pre-trained features. Surprisingly, we found that the linguistic content is clearly the major contributor for the prediction of satisfaction and best generalizes to unseen data. Our experiments conclude to the definitive advantage of using CamemBERT representations, however the benefit of the fusion of acoustic and linguistic modalities is not as obvious. With models learnt on individual annotations, we found that fusion approaches are more robust to the subjectivity of the annotation task. This study also tackles the problem of performances variability and intends to estimate this variability from different views: weights initialization, confidence intervals and annotation subjectivity. A deep analysis on the linguistic content investigates interpretable factors able to explain the high contribution of the linguistic modality for this task.

Index Terms—Continuous Speech Emotion Recognition, Pre-trained Features, Multi-modalities, AlloSat

1 INTRODUCTION

NOWADAYS, relations between customers and companies are increasingly based on call centers [1]. Within these structures, massive speech data is collected and automatically processed everyday by companies, since such data contains crucial information for these companies to improve their commercial relations with customers. With the huge improvements in Automatic Speech Recognition and Spoken Language Understanding processing, it is now possible to extract automatically linguistic and semantic information for speech analytics. In addition, paralinguistic cues can be useful to evaluate the customer level of commitment or attention to the agent discourse. One of the main paralinguistic cue of interest in such speech data is the emotional state of the speaker. In particular, frustration and satisfaction hold key factors of the customer relationship, and more precisely their evolution according time during the conversation. In this paper, we focus on the automatic continuous extraction of such factors in the whole speech conversation.

Emotional states have been extensively studied and many theories exist [2], [3]. Among these, the continuous theory, also called dimensional theory, has been introduced by Wundt et al. [4] and Scholsberg [5], and consider that all affective states arise from independent fundamental neuro-physiological systems. According to this authors, these systems can be defined by three independent dimensions characterized by their extremum values: pleasant-unpleasant, tension-relaxation, and excitation-calm. These three dimensions were soon-to-be found overlapping. Russell [2] introduced the circumplex model in which emotion categories are arranged on a circle controlled by two dimensions:

valence (positive-negative) and arousal (weak-strong). Consequently, each emotion category can be understood as a linear combination of these two dimensions, or as varying degrees of both valence and arousal. While most emotional theories consider affective states from the point of view of psychology and psychiatry, machine learning systems usually takes one input among speech, vision, or physiological signals. More precisely, a Speech Emotion Recognition (SER) system consider that emotion in speech is conveyed by both linguistic and acoustic modalities. For example, Alva et al. [6] proved that arousal is better recognized from acoustic features and valence from linguistic features. Considering these facts, we investigated the fusion of both acoustic and linguistic modalities in our work as many studies have proven its utility in comprehension related domain [7], [8], [9], [10]. In a previous work [11], we defined a new axis within the circumplex model that goes from satisfaction to frustration through a neutral state in the middle. This axis has been proposed for the specific analysis of customer relationships in the context of call-center conversations.

SER systems are subject to different forms of variability which make commercial applications difficult to set up. The first variability lies in the references used to train the models: Emotion perception is highly subjective, and several manual annotations are required to reach a kind of “ground truth”. This variability is usually measured with annotator agreements (kappa values or correlation coefficients). Second, the reliability of the performances increases with the number of audio samples used to evaluate the models. In SER, this number is relatively small due to the high data collection and annotation cost, therefore, confidence inter-

vals for the performances of the systems are highly required. Finally, the third variability comes from the initialization of the parameters of the models, and possible shuffle of the data during the training stage. In this paper, we intend to bring some insights to these three forms of variability with the investigation of individual annotations, the systematic addition of confidence intervals to the regression performances, and the evaluation of initialization impact on the performances.

The different experiments detailed in this article conclude that the linguistic modality is the major contributor for satisfaction recognition in call-center conversations. A deep analysis on the linguistic content of some conversations is carried on to investigate interpretable factors able to explain the high contribution of the linguistic modality for this specific task.

The main contributions of our study are the followings:

- The use of pre-trained models for satisfaction recognition
- The fusion of acoustic and linguistic modalities
- The addition of protocols to evaluate performance variabilities (annotation, initialization, confidence intervals).
- The proposition of interpretable linguistic cues which explain the performances of our model

The paper is organized as follows: Section 2 presents the related works, followed by our motivations in Section 3 and the global overview of our experimental protocols in Section 4. Satisfaction recognition using either acoustic or linguistic modalities experiments and results detailed in Section 5, while the fusion experiments and results are described in Section 6. Section 7 presents a complete analysis regarding annotation and linguistic content. The conclusion is drawn in the last Section.

2 RELATED WORKS

2.1 Features for speech emotion recognition

Looking for speech cues that gives the best emotion recognition model has always been a “holy grail” [12], [13], [14]. However most studies agree that emotion mainly lies in prosody which is a combination of different factors such as intensity, intonation, rhythm and voice quality. These high level factors are usually estimated from low level descriptors: pitch, spectral features, MFCCs, energy, etc... Therefore, to analyze emotion in speech, researchers usually rely on various voice parameters set that are related to emotion [15], [16], [17] including fundamental frequency, speech rate, pauses, voice intensity, voice onset time, jitter (pitch perturbations), shimmer (loudness perturbations), voice breaks, pitch jumps, and measures of voice quality. Para-linguistic sets used in Speech Emotion Recognition (SER) such as ComParE [18], and GeMAPS [13] used in Interspeech Emotion Challenges [19], are designed to capture prosody. Other features like spectral ones can also be extracted: among them, mel frequency cepstral coefficients (MFCCs) are clearly the most often used as they are robust to noisy signals, even if they have not been designed to retrieve prosodic information nor emotion as concluded in Tahn et al. [14].

For a while, SER has been dominated by the acoustic modality. However, emotions are not only conveyed by prosody but also by words. While automatic speech recognition systems (ASR) are more and more efficient, linguistic features can be extracted with high reliability. In the field of Sentiment Analysis (SA) where the goal is to find emotion in written text, different features were proposed such as POS-tagged (Part-Of-Speech-tagged) words [20], [21], polarity dictionaries (SenticNet [22], FAN [23]) or features extracted with the GloVe representation [24] n-grams/bag-of-words [25]. It should be noticed however, that spoken language differs from written text in the grammatical correctness, disfluences and non-verbal vocalizations such as laughter, breathing, and so on [26].

2.2 Modality fusion

Due to the small amounts of training data, SER has late moved to the neural paradigm. First studies have used RNN (Recurrent Neural Networks), especially with LSTMs (Long Short Term memory) to retrieve emotional categories [27] or continuous dimensions [28], [29]. CNNs (Convolutional networks) have also been used to predict SEWA continuous dimensions [30] however LSTMs seems to better generalize on call-center data [31].

In order to take advantage of the linguistic content in SER, the fusion of both textual and audio information gains on popularity [32], [33], [34]. Three strategies are usually applied for multi-modal fusion: (a) at the feature level by concatenating the inputs of different modalities, (b) at the decision level with majority voting, or (c) at the model level by merging intermediate representations [7], [10], [35], [36]. More precisely, the fused model (c) is done by concatenated outputs of two distinct networks corresponding to each modality to feed next layers [37]. Many other modalities can be used in SER to better represent affective states. For example, audio representation, facial cues from video, textual information are used in the work of Chen et al. [38] and Poria et al. [39] while Wu et al. [40] focuses on semantics labels and audio features. Modality fusion always improve the performances obtained on speech only.

2.3 Pre-trained features for NLP

Expert features have the advantage to convey human understandable information but there are not the only way to represent data. From other research domains such as SA, we assist to the rising of pre-trained self-supervised feature to represent the data, especially with word embeddings such as GloVe [24] or Word2Vec [41]. As there are trained on a massive amount of data, they tend to be able to efficiently represent data, without the need of human annotation. Very recently, these pre-trained features spread in SER. Atmaja et al. [42] uses acoustic features consisting mostly in time and spectral domain features, and Word2Vec embedding for the linguistic part by performing a feature level fusion. While Yenigalla [43] et al. uses spectrogram and phoneme embeddings merged at the model level.

The self-supervised learning of speech or language representations has been proposed these last few years, for instance with the BERT system [44], used for textual representation. Such representations, computed by neural models

trained on huge amounts of unlabeled data, have shown their effectiveness on some tasks under certain conditions, for instance in ASR [45], [46], or speech translation [47]. Recently Wav2Vec [48], Mockingjay [46] and Audio ALBERT [49] were introduced in ASR and speaker identification as one of the first pre-trained approaches to extract context dependent features from raw signals for ASR tasks but they have not been used for SER yet. Very recently a BERT-like model for French has been developed [50]. To the best of the authors' knowledge, such pre-trained features have not been yet used for SER.

3 MOTIVATION

The goal of our research is to continuously recognize satisfaction and frustration in real-life call-center conversations. To do so, we are using AlloSat [11] French corpus to train speech emotion recognition network.

Moreover, several studies point out that emotion information can be detected not only in speech but also in facial traits, biological responses or linguistic and semantic information. Traditionally, emotion recognition models use only the acoustic modality [51], even if some works have shown that linguistic modality also convey important information [52]. In our work, we investigate the use of the acoustic signal and its linguistic transcription, separately or jointly. To compensate the lack of training data dedicated to the targeted task, we also explore the benefit of using models pre-trained on huge amount of data for both modalities such as Wav2Vec [48], Word2Vec [53] or BERT [54].

To design application for real industrial end users, one of the main concern is to be able to reproduce the results on multiple GPU clusters, thus to reduce all possible variabilities during the evaluation process. In the scope of neural networks paradigm, weights initialization has always been pointed out as crucial as it impacts both the training time and the phenomena of being stuck in a local minima [55].

Therefore we will estimate how much the weight initialization affects the performances of the satisfaction recognition. Because the Test set is, of course, not representative of all possible realizations, we decided to include an confidence interval to our scores. This aims at given an idea of how much the performances could vary when evaluating on different conversations, considering that all non-deterministic sources are fixed for that matter. In the field of continuous emotion recognition, the reference generally consists of the averaged value over all annotators. In our study, we tackle the problem of the subjectivity of the annotation by considering individual annotation instead of the averaged reference.

Our major conclusion is that models learnt on features extracted from the transcripts only are very accurate in the prediction of satisfaction and frustration. Therefore, this work analyses the linguistic content, and proposes relevant linguistic clues which are strongly related with the perception of the emotional state.

4 GLOBAL OVERVIEW

This section presents the speech material used to train and evaluate the models and the general architecture of the neural network used for SER.

4.1 Speech emotional data : AlloSat corpus

While past emotional speech corpora were annotated with discrete emotion categories [56], [57], the current trend is to move towards continuous annotations of affective dimensions. Among the most popular corpora annotated continuously, we can cite SEMAINE [58] composed of English interactions with virtual or human operators, or RECOLA [59] targeting French dyadic online conversations. Both corpora are annotated at least according to arousal and valence dimensions. The recent cross-cultural Emotion Database SEWA [60] was presented for the 2018 Audio/Visual Emotion Challenge [61] which aimed to retrieve arousal, valence and liking dimensions from semi-supervised dyadic conversations.

In order to fit with our target task, we choose to carry on our experiments on AlloSat corpus [11] composed of real-life call-center conversations, annotated along the satisfaction axis. AlloSat was precisely built to continuously predict the evolution of the customer satisfaction on call-centers audio recordings of French speaking adult callers (i.e. customers). Various information are asked by the callers: contract information, global details on the company, or complains.

All conversations were recorded at 8kHz between July 2017 and November 2018 in call-centers located in French-speaking countries. The agents are employees of various companies in different domains, mainly energy, travel agency, real estate agency and insurance. The two telephone channels were recorded separately. Due to commercial constraints, we discarded the part of the receiver (i.e agent). Consequently, there is no overlap in the conversations.

AlloSat contains 303 conversations for a total duration of 37h 23' as summarized in Table 1. There is generally one single speaker per conversation even if some conversations can involve multiple speakers, for instance when the caller gives the telephone to someone else. In order to preserve the speakers' privacy, all personal information were obfuscated with a jazzy sound letting the annotator knows that there was private information at this very moment. This anonymization process ensures to respect the General Data Protection Regulation (GDPR) recommendation. Because we removed the agent speech, there can be long moments of silence in the remaining caller speech. To minimize the annotator effort, we decided to replace these silences by 2 seconds of white noise, allowing the annotators to identify these moments of silence. In order to avoid collecting too many conversations with poor emotional content, we decided to apply three selection criterion based on prosodic and linguistic content.

- 1) Speech duration: conversations longer than 30 seconds containing more than three speech turns;
- 2) Intonation: standard deviation of the fundamental frequency (F_0) over 40 Hz. F_0 is extracted with YAPPT algorithm [62] which is adapted to telephone signals;
- 3) Linguistic valence: the valence score computed on the transcriptions is below 4.98 (negative) or above 5.02 (positive). Word scores are given by FAN French dictionary [23] and unknown words are at 5.00.

Statistics	Value
number of conversations	303
number of speakers	308
number of women	191
number of men	117
total duration	37h23m27s
min duration conversations	32s
max duration conversations	41m
mean duration conversations	7m24s

TABLE 1: Summary of AlloSat characteristics

Emotion annotation is known to be a highly subjective task. To compensate for the subjectivity of the annotation task, three annotators rated continuously the 303 conversations along the satisfaction axis. This axis range from frustration to satisfaction with a neutral state in the middle and is sampled every 0.25 seconds. Individual annotations were averaged to get a gold reference, used in the prediction task. For more details about the coherence of the annotations, please refer to our previous work [11]. An automatic transcription were provided by Allo-Media for each conversation.

The corpus has been divided into three subsets: The train set contains 201 conversations corresponding to about 25h of audio signal and 16h of speech; The development set is composed of 42 conversations; and the test set contains 60 conversations. Both Development and Test sets are composed of about 6h of audio signal and 3h of speech.

4.2 SER neural network model

4.2.1 Baseline architecture

We designed a regressive baseline neural network to continuously predict the satisfaction along the conversation. To do so, a recurrent network, inspired from [30], is used for the prediction task using bidirectional Long Short-Term Memory units (biLSTM).

The sizes of the different layers have been optimized in our previous work, and the final architecture is composed of 4 biLSTM layers of respectively 200, 64, 32, 32 units with a tanh activation as shown on Figure 1. A single output neuron is also used to predict the regression value each 250 ms at the emotional segment level. Neither dropout nor batch normalisation is used in this approach.

The baseline network is fed with expert acoustic, respectively linguistic, feature sets of low dimension (40, respectively 48) described in the next section 5. When moving to pre-train features, the input dimension explodes up to hundreds as they intend to represent huge amounts of speech data.

A mean and variance normalization of the input features is done over the training data for all experiments.

4.2.2 Loss and evaluation function

The concordance correlation coefficient (CCC) [63] goes from 0 (chance level) to 1 (perfect) and is calculated according to eq. 1, where x is the prediction and y the reference. μ_x and μ_y are the means for the two variables and σ_x and σ_y their corresponding variances. ρ is the correlation coefficient between the two variables x and y .

$$CCC = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \quad (1)$$

Fig. 1: Baseline network architecture. Number of neurons of each layer are written in red.

In previous experiments on the prediction of emotional dimensions [28], [30], the loss function to be minimized during the training phase is defined according to eq. 2, where the CCC is computed over all concatenated conversations within a batch.

$$\mathcal{L}_c = 1 - CCC \quad (2)$$

The CCC is also used as the evaluation metric on the Development and Test subsets. The score is computed at once on all the concatenated conversations of a given data subset, as described in AVEC challenges [61].

4.2.3 Confidence interval for CCC score

As mentioned previously, our work also intend to assess the robustness of the models from an industrial perspective. More precisely, as the number of samples used to evaluate the models is relatively small, we need to estimate how reliable is the final CCC score with a confidence interval [64], [65]. The definition of the confidence interval for CCC is given in Appendix A. On AlloSat evaluations, the confidence interval widths for the CCC are between 0.006 (lower CCC) and 0.002 (high CCC). In the following experiments, a difference in performance will be judged as consistent if the two confidence intervals do not overlap.

4.2.4 Hyper-parameters

All networks are implemented under Pytorch framework [66]. Preliminary experiments on the development set, helped to settle the baseline network architecture (number of biLSTM layers and number of neurons per layer) and the following hyper-parameters: training is done on batches from 8 to 20 conversations using the Adam optimiser, depending on the size of the input embedding and memory constraints. All the conversations are kept without any padding. The learning rate is optimized at 0.001 by empirical method, tested on a range from 0.001 to 0.02 by a 0.005 step. After preliminary experiments, we noticed that networks were not improving after the first 400 epochs, so the maximum number of epochs is set to 500. For each training process, the final model is the one extracted from

the epoch that gets the best score on the Development set. This final model is then evaluated on the Test set.

4.2.5 Initialization

the initialization of the model can have a huge impact on both the execution time and the accuracy of the resulting system. To handle with this hypothesis, 5 random initializations are tested on our best decision fusion system. In additional experiments¹, the final CCC score of one of the fusion approaches varies from .873 to .911 depending on the seed used for the initialization. It is a high variability which is considered to be relevant if we refer to the confidence interval, allowing us to conclude that the initialization is crucial. In such a situation, if a new model is trained with same data and same architecture, there is a significant uncertainty on the final performances. This will not be investigated in the rest of the article.

5 ACOUSTIC AND LINGUISTIC FEATURES

This section describes features used in input of the network. While acoustic features are extracted directly from the speech signal, linguistic features are obtained from the transcription. Baseline features consists of the traditional inputs used to represent the signal, *i.e.* Mel Frequency Cepstral Coefficients (MFCCs), or textual information, *i.e.* Word2Vec. Pre-trained features are indeed embeddings which are learnt on huge amount of data for an external task, here automatic speech recognition.

5.1 Acoustic modality

5.1.1 Baseline features: MFCCs

Two baseline acoustic sets are used as input of the network: MFCCs and the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS). While eGeMAPS intend to precisely capture and represent prosody in speech, MFCCs are known to be robust to low quality audio signals such as telephone. In previous experiments, we have shown that MFCCs better achieve to predict satisfaction than eGeMAPS features [67], therefore only MFCCs are considered in the remainings. In speech processing, the spectral content is considered as constant on small audio segments of around 30 ms. Our signal is sampled at 8 kHz, therefore MFCC 1-12 and their delta values are extracted on 30 ms frames each 10 ms with torchaudio toolkit².

Mean and standard deviation of each coefficient are computed over the emotional segment in order to get a 48 dimensional vector each 250 ms.

5.1.2 Pre-trained features : Wav2Vec

Self-supervised learning approaches have been designed in order to take benefit of huge amount of unlabelled data. Wav2Vec (1.0) [48] is a neural model trained through self-supervision to compute speech representations from raw audio. This model is composed of two distinct convolutional neural networks. A first encoder network converts the audio signal into a new representation that is given to the second

network, the “context network”, which takes care of the context by aggregating multiple time step representations into a contextualized tensor that matches to a receptive field of about 210 ms. Both are then used to minimize a contrastive loss function. The resulting embedding is a 512-dimensional feature vector. As the training of such model demands a lot of data and calculation power, we use the large pre-trained model provided by Schneider et al. in [48], trained on Librispeech corpus [68] consisting of 960 hours of English audio book samples at 16 kHz. Our features were extracted on an upsampled version of AlloSat³. In order to investigate the influence of the acoustic context on Wav2Vec representations, embeddings are extracted either on the current 250 ms emotional segment (without context) or on the whole conversation input (with context).

In the end, each emotional segment is represented by a 512-dimensional vector which consists of the averaged values of obtained embeddings over each segment of 250 ms.

5.2 Linguistic modality

5.2.1 Baseline features : Word2Vec

Word2Vec embeddings have been extensively used for sentiment analysis or opinion mining from text [41], [42], this motivated us to use such representation for the prediction of satisfaction. In the following experiments, a Word2Vec model has been trained with the toolkit GENSIM [69], using private data owned by Allo-Media composed of manual call transcriptions received by call centers, totaling over 500 hours of speech, with CBoW algorithm [53]. No stop list is used before extracting the embeddings. In a first step, the output size embedding is fixed to 40 in order to have similar dimension with baseline MFCC features (*i.e.* 48). It is also motivated with empirical results showing that in the range between 20 and 60, the dimension 40 gave the best results. We also did the experiment with a more standardized output size, fixed at 100.

5.2.2 Pre-trained features : CamemBERT

Inspired by RoBERTa [70] and BERT, CamemBERT [54] is a multi-layer bidirectional Transformer. CamemBERT is trained on the Masked Language Modeling (MLM) task which consists of replacing some tokens by either the token <MASK> or a random token and asking the model to correct the tokens. The network uses a cross-entropy loss. The input consists of a mix of whole words and sub-words in order to take advantage of the context.

We use the “camemBERT-base” pre-trained model delivered by the authors and trained on the French part of OSCAR corpus [71] consisting of a set of monolingual corpora extracted from Common Crawl snapshot and totaling 138GB of raw text and 32.7B tokens after sub-word tokenization. Text representations were extracted on Allosat by using this pre-trained model, and we summarized the results by averaging the continuous representations of sub-words occurring in the current emotional segment. In total, we use a 768-dimensional feature vector. In order to investigate the influence of the linguistic context on CamemBERT

1. The results are not presented here

2. <https://pytorch.org/audio/stable/index.html>

3. We used FFMpeg resampling function with *sinc* interpolation function

representations, embeddings are extracted either on the words pronounced during the current emotional segment (without context) or on the whole conversation input (with context).

5.3 Results

Modality	# size	Satisfaction		
		Dev	Test	Diff. (%)
AUDIO				
MFCC	48	.851 [0.0]	.651 [0.0]	-23.5
Wav2Vec woc	512	.844 [-0.8]	.806 [+23.8]	-4.5
Wav2Vec wc	512	.823 [-3.2]	.656 [+0.8]	-20.3
TEXT				
Word2Vec	40	.885 [0.0]	.861 [0.0]	-0.1
Word2Vec	100	.853 [-3.5]	.812 [-5.7]	-4.7
CamemBERT woc	768	.916 [+3.5]	.817 [-5.2]	-10.8
CamemBERT wc	768	.917 [+3.7]	.924 [+7.3]	-0.8

TABLE 2: Comparison of the audio and text modalities in terms of CCC computed on Development and Test sets on AlloSat. Shuffle is activated within batches. woc: without context; wc: with context. Relative difference between Dev and Test sets and relative improvement between baseline and pre-trained features, are given in %.

All results on acoustic and linguistic modalities are reported in Table 2. We confirm that pre-trained features are achieving awesome results in comparison to baseline features. Especially, the performance impressively increases on the Test set (+23.8%) when using Wav2Vec pre-trained features extracted without context (CCC=.806) instead of MFCCs features (CCC=.651). The relative improvement on Test set (+7.3%) obtained when using CamemBERT pre-trained features extracted with context instead of Word2Vec is not as spectacular as the one obtained on acoustics because Word2Vec (CCC=.861) features already reach good results in comparison to MFCCs (CCC=.651). However this modality seems more robust as it improves for both Dev and Test sets.

To confirm the reliability of our results, we can notice that the performance obtained by our models trained on acoustic features computed by the English Wav2Vec1.0 model is consistent, and even better, to the one obtained on the same data and presented in a recent study [72] that used a Wav2Vec2.0 model to extract acoustic features to feed smaller neural models.

Deeper experiments on the number of features used to train Word2Vec representations confirm that the best performance on Dev and Test set are obtained with a size of 40. Increasing the number of features to 100 degrades the score on Dev (-3.5%) and Test (-5.7%) sets. Regarding to confidence intervals detailed in appendix B, we confirm that all mentioned improvements are significant.

A lot of differences exist by nature between CamemBERT and Word2Vec: complexity of the neural architecture, context-dependent dynamic embeddings *vs.* static embeddings, sub-words *vs.* words, . . . The computation of CamemBERT needs a lot of GPUs, data and time. However, we do not have the means to train such a model on specialized data with call-center conversations. Fortunately with the help of the pre-trained model kindly distributed by the authors and it is possible to get very good results on the targeted SER task without owning such amount of resources.

As described in Table 2, the different acoustic and linguistic representations of the speech signal have different sizes which can impact the training of the network. We previously investigated the impact of this dimension gap on system performances [67], by comparing the network presented in Figure 1 with another one designed to reduce the dimension of input features. This reduction was done by adding an optional dense layer after the inputs and before the first biLSTM layer in order to reduce the input size to 40, resp. 48, for linguistic, resp. acoustic, modalities. We concluded that both architectures achieved comparable results and the addition of a dense layer was not necessary and that the input size does not significantly affect the results.

To conclude from Table 2, we confirm the relevance of using pre-trained features for satisfaction recognition. Surprisingly, we also found that linguistic embeddings, are able to capture a lot of emotional information directly from the transcribed speech as it performs slightly better than the acoustic one, especially when using CamemBERT features extracted with context. At this point, we should notice that pre-trained linguistic features are extracted from textual transcriptions, however Word2Vec and CamemBERT models are trained on speech signals. Therefore some acoustic information (mainly phonetics) is, in a sense, also included in these linguistic features. However, we do not know at this stage how prosodic and para-linguistic information is captured by pre-trained linguistic features.

6 MODALITY FUSION

As discussed in Section 2, many studies confirm that emotion is conveyed by many modalities, especially acoustic and linguistic modalities as presented in Section 2. However, there is no consensus on the independence of acoustic and linguistic modalities, or there synchronicity with time. To address this problem, we experiment three types of fusion : feature, model and decision fusion. In our case, the output value is return each 250 ms. Therefore acoustic and linguistic vectors must be aligned together with respect to time.

6.1 Feature fusion

Feature fusion methods enable a new representation of the speech signal which is the concatenation of individual modality features from the two modalities (Fig. 2a). A single model is then trained with a unique vector corresponding to a joint representation of the acoustic and linguistic features. The input size is therefore the sum of the two acoustic and linguistic feature sizes. Good fusion performances at the feature level would probably mean that acoustic and linguistic modalities are synchronously used to perceive the satisfaction.

6.2 Model fusion

We experiment two types of model fusion :

- Early fusion: Outputs of the first layers of acoustic and linguistic modalities are concatenated to feed the second layer (Fig. 2b).
- Late fusion: Outputs of the third layers of acoustic and linguistic modalities are concatenated to feed the last biLSTM layer (Fig. 2c).

(a) Feature fusion by concatenating input features.

(b) Model fusion by concatenating the first acoustic and linguistic layers.

(c) Model fusion by concatenating the last acoustic and linguistic layers.

(d) Decision fusion by averaging predictions from audio and text modalities.

Fig. 2: Description of the four used fusions.

6.3 Decision fusion

To perform a decision fusion, two models are trained independently on each modality and the predicted numerical values are averaged to compute new predictions (Fig. 2d). In this configuration, it can be relevant to compute the global prediction (CCC_G) as the weighted average (Eq. 3) of the individual acoustic CCC_a and the linguistic CCC_l scores, in order to give more importance to one of the two modalities.

$$CCC_G = w_a \cdot CCC_a + w_l \cdot CCC_l \quad (3)$$

We optimize the weights of each modality from 0.1 to 0.9 with a step of 0.01. The final configuration is the one which gives the better score on the development set. Good fusion performances at the decision levels would probably mean that synchronicity is useless for the perception of satisfaction on a 250 ms frame.

6.4 Results

The CCC scores obtained on Dev and Test sets with baseline (resp. pre-trained) features are summarized in Table 3 (resp. Table 4). The relative differences between Test and Dev results are given in the last column to estimate the generalization power of the model. Relative improvements are also included with the best single model as reference, *i.e.* Word2Vec or CamemBERT. Detailed scores with confidence interval can be found in appendix B.

Table 3 shows that whatever the fusion level, fusion performs better than Word2Vec only on the Dev set and

lower on the Test set. The poor performances on the Test set can be explained by the very small CCC obtained with MFCCs ($CCC=.651$). The best improvement on the Dev set is obtained when using the late model fusion (+3.9%), however this is the configuration that less generalizes on the Test set (−11.1%).

Table 4 shows that the addition of the acoustic modality to CamemBERT embeddings does not improve performances on Dev set with feature fusion but with model or decision fusion. We confirm the fact that acoustic modality alone does not generalize well on the Test (−4.5%) while CamemBERT does (+0.8%). The best improvement on Dev

Fusion level	Satisfaction		
	Dev	Test	Diff. (%)
SINGLE BEST			
MFCC	.851	.651	-23.5
Word2Vec	.883 [0.0]	.881 [0.0]	-0.1
FEATURE			
MFCC $\oplus$ Word2Vec	.895 [+1.4]	.833 [-5.6]	-6.9
MODEL			
Early	.904 [+2.4]	.807 [-8.5]	-10.7
Late	.917 [+3.9]	.815 [-7.6]	-11.1
DECISION			
.66 Word2Vec + .34 MFCC	.897 [+1.6]	.840 [-4.8]	-6.4

TABLE 3: Comparison of four fusion approaches. CCC results on Dev and Test sets. Shuffle is activated within batches. Relative differences between Dev and Test sets and relative improvements between baseline and pre-trained features, are given in %.

Fusion level	Satisfaction		
	Dev	Test	Diff. (%)
SINGLE BEST			
Wav2Vec woc	.844	.806	-4.5
CamemBERT wc	.917 [0.0]	.924 [0.0]	+0.8
FEATURE			
Wav2Vec $\oplus$ CamemBERT	.907 [-1.1]	.884 [-4.3]	-2.5
MODEL			
Early	.924 [+0.8]	.897 [-2.9]	-2.9
Late	.945 [+3.1]	.893 [-3.4]	-5.5
DECISION			
.72 CamemBERT + .28 Wav2Vec	.932 [+1.6]	.920 [-0.4]	-1.3

TABLE 4: Comparison of four fusion approaches. CCC results on Dev and Test sets. Shuffle is activated within batches. woc:w ithout context; wc: with context. Relative differences between Dev and Test sets and relative improvements between baseline and pre-trained features, are given in %.

is obtained with a late model fusion (+3.1%), however this is the configuration that less generalizes on the Test set (−5.5%). The decision fusion better generalizes on Test set (−1.3%) than other fusion approaches and have the advantage of slightly improving the Dev score (+1.6%) while not much degrading on the Test (−0.4%) in comparison to CamemBERT features only.

Unexpectedly, our results concludes that the linguistic modality (without the addition of acoustic features) best generalize to unseen data. They also confirms the relevance of pre-trained features such as CamemBERT, to a lesser extent Wav2Vec, for satisfaction recognition in call-center conversations. While the model late fusion does not reach the best results in Test set, it significantly outperforms single linguistic modality on Dev, confirming the multi-modal aspects of emotion. The advantage of this fusion method is that it requires less computing ressources to be trained. Therefore, acoustic information is still useful but is less robust to unseen data.

7 ANALYSIS AND DISCUSSION

This section deeper analysis our results in order to better understand the importance of the linguistic modality. We investigates two axes: Annotator subjectivity and linguistic content.

7.1 Influence of annotation subjectivity

Our first analysis interrogates the subjectivity of the annotation task regarding acoustic and linguistic modalities. To do so, we modify the reference: Instead of training a single model on the averaged value over the three annotators, we train three different models per annotator, in which each reference is the single values for this annotator. The predictions of these models are evaluated regarding individual annotations (top part of Table 5) or the ground truth defined as the average of the three individual annotations (bottom part of Table 5). The AVG column gives the average performance over the three individual models. The CV column gives the coefficient of variation (standard deviation over mean) over the three individual models. Diff1 is the relative difference between linguistic and acoustic taken independently and gives an idea of the gain per annotator.

Individual annotations: From the upper part of Table 5, we can notice that the coefficient of variation (CV) for single features, is higher with acoustic features than with linguistic features when the references are individual annotations, especially on the Test set. More precisely, regarding annotator a_3 , the performance of the acoustic modality severally drops on the Test set (CCC=.597). Our hypothesis is that the variability in the acoustic space is highly diverse, and the same acoustic realization might be perceived with different satisfaction levels by the same annotator, what produces bad performances on the acoustic modality. In the previous Section 6, we have shown that the fusion of the modalities improves performances on Dev but degrades on Test. This is not true when models are train and evaluated on individual annotations: fusion improves performances in most configurations and the best performance in average is reached with the model early fusion (CCC=.854 on Test set). The improvement on Test is highest with annotator a_2 (+3.7% with model early fusion). This can be explained by the very small difference between the performances obtained on independent modalities for this annotator (+6.2%), maybe indicating that both modalities carry different information for this specific annotator.

From these results, we hypothesize that, at the annotator level, acoustic and linguistic modalities convey complementary emotional information, however, while the linguistic part is well shared among annotators, the perception of the acoustic part seems quite individual. Of course additional experiments with cross-annotations are needed to confirm this hypothesis.

Averaged annotations: Regarding individual models evaluated with averaged annotations (bottom part of Table 5), we notice that annotator a_2 has the lowest performances when using only linguistic features. The model built upon this annotator reaches the lowest performances using any type of fusion on Dev and Test sets. Thus confirming the importance of high linguistic performances for the general evaluation. This result can be explained by the fact that among the three annotators, we have shown that a_2 had the lowest intra-annotator agreement (see [11]). We also confirm the fact that the fusion helps to improve the performances per annotator in all cases. The early model fusion has the advantage of having higher averaged performances than CamemBERT and of being the model less affected by individual annotations (CV =0.020 on Test set).

From these experiments, we conclude that while the fusion approaches degrade the global performances in comparison to CamemBERT only (see Table 4), it seems that they are more robust to the subjectivity of the annotation task. We found that the early model fusion was the best compromise between performance and robustness. Our insights also interrogates the evaluation process using the average values of the three annotators: averaged values have no perceptive reality, but individual values do.

7.2 Linguistic analysis

In the context of call-center conversations, the experiments described below conclude that the satisfaction-frustration axis is more supported by linguistic than acoustic content. Regarding to the circumflex model, this axis is very close

Reference	Annotator		a_1		a_2		a_3		AVG		CV	
	Fusion level		Dev	Test	Dev	Test	Dev	Test	Dev	Test	Dev	Test
Individual annotations	SINGLE	Wav2Vec	.834	.734	.731	.785	.841	.597	.802	.705	.077	.138
		CamemBERT	.898	.877	.833	.834	.900	.804	.877	.838	.043	.044
	Diff1 (%)		7.7	19.5	14.0	6.2	7.0	34.7	-	-	-	-
	FEATURE		.884	.870	.815	.753	.883	.834	.861	.819	.046	.073
	MODEL	Early	.883	.870	.855	.865	.888	.826	.875	.854	.020	.028
		Late	.911	.875	.814	.837	.921	.799	.882	.837	.067	.045
Averaged annotations	DECISION		.913	.882	.840	.849	.916	.793	.890	.841	.048	.053
	SINGLE	Wav2Vec	.862	.736	.774	.731	.779	.710	.805	.726	.061	.019
		CamemBERT	.916	.878	.755	.793	.851	.833	.841	.835	.096	.051
	Diff1 (%)		6.3	19.3	-2.5	8.5	9.2	17.3	-	-	-	-
	FEATURE		.896	.845	.741	.688	.868	.861	.835	.798	.099	.120
	MODEL	Early	.911	.833	.809	.824	.879	.856	.866	.838	.060	.020
		Late	.914	.899	.763	.784	.844	.841	.840	.841	.090	.068
	DECISION		.938	.882	.795	.778	.868	.874	.867	.845	.082	.069

TABLE 5: Fusion results for each annotator. Models are trained and evaluated on individual labels. CamemBERT are extracted with context while Wav2Vec are extracted without context. AVG: averaged over the three annotators. CV: coefficient of variation over the three annotators. Improvement corresponds to the absolute difference between CamemBERT and the decision fusion. Best fusion is chosen on Dev subset. Diff2: relative improvement between CamemBERT and best fusion.

to the valence axis, what could explain in some extend the importance of word for the detection of satisfaction. In this section, we intend to provide elements that could explain the importance of linguistics to retrieve the satisfaction. This analysis have been done on 13 conversations selected in order to cover different dynamics of the satisfaction dimension: Globally flat, occurrences of high frustration (ground truth < 4) and occurrences of strongly decreasing satisfaction (frustration drops). The analysis has been done using the automatic transcription, the reference satisfaction annotation and tags corresponding to *high frustration* and *frustration drop*.

Our hypothesis is that frustrated speech mainly correspond to the accentuation of the oral phenomena. Consequently, we specifically investigated the following orality clues:

- Amount of disfluencies,
- Hesitations, repairs, repetitions, babbling,
- Importance of self-breaks defined as “the points where the utterance flow is broken” [73],
- Usage of interrogations and negations,
- Semantic evidences of frustration or unhappiness,
- Amount of meaningful segments *vs.* semantically empty segments.

Based on these clues, the analysis concludes to different observations. There are semantic evidences of frustration in the conversations such as the usage of the negation (*ça ne m’amuse pas, c’est inadmissible*), strong markers (*c’est gonflé, putain de ...*) and weak markers (*quand même, franchement*). It seems also that the amount of meaningful segments, self-breaks and disfluencies, are generally correlated with high frustration or satisfaction drops. The syntactic structure of interrogative utterances seems also correlated with frustration.

In a second step, we intend to go further in this analysis with the automatic extraction of orality clues. Of course, moving from manual to automated extraction implies to do some choices in the definition of the clues. Trying to model the amount of meaningful segments, we extract POS tags using MACAON [74] directly from automatic transcriptions and compute the number of verbs and nouns with respect

to time. To capture the other orality clues, we decided to extract automatically the seven features mentioned in Table 7.

The idea is not to provide an exhaustive analysis on the whole dataset but to provide some explainable clues. We focus here on the deep analysis of a single conversation about a certified letter. All the occurrences of features summarized in Table 7 are synchronized in time together with the annotated satisfaction reference. The number of verbs and nouns does not give relevant information and is not represented here. The dynamic linguistic analysis of each conversation is shown on Fig. 3. This conversation has been annotated with a strong drop of satisfaction before 200 sec. The automatic transcription obtained just before this drop is given in Table 6. Just before the drop, the occurrences of single words repetition and *c’est* are important, whereas after the drop, the number of filled pauses and negation marker (*pas*) increases. We also notice that a strong marker (*réclamation*) happen just before the drop, probably meaning that this specific word induces the perception of noticeable frustration.

In the context of AlloSat speech data, emotional information seems to lies more in the words than in the prosodic and acoustic content. In such data, the expression of frustration is mainly related to the accentuation of the oral phenomena: semantic content and above all self-breaks, disfluencies, hesitations, repairs and repetitions.

8 CONCLUSION

This paper present the independent use of acoustic and linguistic pre-trained features and the fusion of these two modalities for the continuous recognition of satisfaction in call-center conversations. We also present a further analysis on the influence of annotation subjectivity on the performances. We also investigate possible linguistic clues able to explain the supremacy of linguistic features for this task.

Conducted on the AlloSat corpus, built for the recognition of satisfaction and frustration in real-life call-center conversations, we observe that Wav2Vec acoustic and CamemBERT linguistic pre-trained features, better represent satisfaction than baseline features such as MFCC and Word2Vec.

French	English translation
- <u>voilà</u> et <u>la deuxième lettre</u> // c'est pareil <i>mais</i> <u>bon</u> <u>cette lettre</u> // elle est où maintenant... pas comprendre pourquoi on n'a pas retiré <u>la lettre...</u> <u>la deuxième lettre</u> // c'est pareil <i>mais</i> <u>elle</u> venait d'où // <u>cette lettre...</u> c'était <u>qui</u> // <u>qui</u> a envoyé <u>cette lettre...</u> parce que c'est important // on est une société // nous... quand on sait pas qui c'est // ... comment on peut savoir <u>qui c'est</u> <i>ouais</i> <i>mais</i> ça va pas du tout <i>hein</i> ça va pas du tout // <u>ça</u>	- <u>there we are</u> and <u>the second letter</u> // it is the same <i>but</i> yes <u>this letter</u> // where is it now ... not understand why no one removed <u>this letter</u> ... <u>the second letter</u> // it is the same but where does <u>it</u> come from // <u>this letter</u> ... it is <u>who</u> // <u>who</u> sent <u>this letter</u> ... because it is important // we are a society // we ... when we don't know who it is // ... how can we know who it is <i>yeah</i> <i>but</i> it's not ok <i>eh</i> <i>it's not ok</i> // <u>it</u>

TABLE 6: Extract (137 - 166 sec.) from a conversation about a certified letter. Disfluencies: *italic*; Hesitations, repairs, babbling: underline; Semantic evidences of frustration: **bold**; self-breaks: //

Fig. 3: Dynamic analysis of frustration of a conversation about a certified letter. Number of occurrences of the seven linguistic features are plotted with respect to time. The gold satisfaction reference is represented with red dashed line.

Features	# occurrences
single word repetitions (deg1)	26
bi-grams repetitions (deg2)	4
filled pauses (<i>euh, bah, hein, eh, etc.</i>)	22
strong markers (<i>important, inquiet, scandaleux, etc.</i>)	14
weak markers (<i>quand même, franchement, etc.</i>)	3
negation marks (<i>pas, ne, n'</i>)	30
<i>c'est</i>	44
# words in the conversation	1050
# utterances in the conversation	152

TABLE 7: Seven features and their occurrences number used to model the orality clues supposed to be responsible for frustration in the conversations. The total number of utterances and words are included for the complete conversation about the certified letter.

On the Test set, the CCC score increases from 0.651 with 48 MFCC features to 0.806 with Wav2Vec; and from 0.861 with 40 Word2Vec to 0.904 with CamemBERT. In our experiments, we found that linguistic representations clearly outperform acoustic representations, thus questioning the need for acoustic in such task. However, linguistic pre-trained features are extracted on automatic transcriptions directly obtained from the acoustic signals. So we definitely need acoustic and we do not know at this stage how prosodic and para-linguistic information is captured by these pre-trained features.

Our results clearly affirm the advantage of using CamemBERT representations, however the benefit of the fusion of acoustic and linguistic modalities is not as obvious. With models learnt on individual annotations, we found that fusion approaches are more robust to the differences

in annotations. The early model fusion has the advantage of slightly degrading performances in comparison to model trained on CamemBERT features only, and being more robust to the subjectivity of the annotation task.

This article also investigates the robustness of the proposed approach towards industrial applications. We pointed out the fact that the initialization process induces a large variability in the performances of the network. Further investigations are needed to cope with this issue. We demonstrate that the use of fused models improves the robustness of the models regarding annotation subjectivity. Finally a deep linguistic analysis allows us to propose relevant linguistic clues (negation and semantic markers, repetitions, filled pauses, etc.) that somewhat explains why the linguistic content is so important for this task. We conclude that para-linguistic information is mainly included in words and their syntax.

In a future work, we intend to develop some approaches in order to cope with the initialization issue in order to provide reproducible experiments. Additional experiments with cross-annotations approaches are needed in order to investigate the differences in perception due to the linguistic and acoustic modalities. This work raises the question of the place of acoustic cues, especially prosody features. To pursue our investigation, we aim at applying the presented protocol on additional speech data, for instance broadcast news, political debates, etc.

Fusion level	Satisfaction			
	Dev		Test	
SINGLE AUDIO				
MFCC	.8507	[.8491; .8523]	.6506	[.6477; .6536]
Wav2Vec woc	.8437	[.8420; .8453]	.8055	[.8036; .8073]
Wav2Vec wc	.8234	[.8215; .8252]	.6559	[.6529; .6589]
SINGLE TEXT				
Word2Vec - 40	.8848	[.8836; .8860]	.8613	[.8598; .8627]
Word2Vec - 100	.8526	[.8510; .8541]	.8124	[.8105; .8143]
CamemBERT woc	.9159	[.9150; .9168]	.8166	[.8148; .8185]
CamemBERT wc	.9171	[.9162; .9180]	.9239	[.9231; .9248]
FEATURE FUSION				
MFCC $\oplus$ Word2Vec.- 40	.8952	[.8941; .8964]	.8331	[.8315; .8348]
Wav2Vec woc $\oplus$ CamemBERT wc	.9066	[.9056; .9076]	.8840	[.8828; .8851]
MODEL FUSION				
MFCC $\oplus$ Word2Vec - 40 (early)	.9039	[.9028; .9049]	.8067	[.8047; .8086]
MFCC $\oplus$ Word2Vec - 40 (late)	.9166	[.9157; .9175]	.8149	[.8131; .8168]
Wav2Vec woc $\oplus$ CamemBERT wc (early)	.8926	[.8914; .8937]	.8937	[.8926; .8947]
Wav2Vec woc $\oplus$ CamemBERT wc (late)	.9450	[.9444; .9456]	.8927	[.8915; .8938]
DECISION MODEL				
.66 Word2Vec-40 + .34 MFCC	.9149	[.9139; .9158]	.8351	[.8334; .8367]
.72 CamemBERT + .28 Wav2Vec	.9315	[.9307; .9322]	.9202	[.9194; .9210]

TABLE 8: Confidence intervals

APPENDIX A

CONFIDENCE INTERVAL FOR CCC

Concordance correlation coefficient between two distributions X and Y .

$$\hat{\rho}_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \quad (4)$$

Where:

- standard deviation: $\sigma_x = \frac{1}{N} \sum_i (x_i - \mu_x)^2$
- covariance $\sigma_{xy} = \frac{1}{N} \sum_i (x_i - \mu_x)(y_i - \mu_y)$
- correlation coefficient: $\rho = \frac{\sigma_{xy}}{\sigma_x\sigma_y}$

Applying the Fisher transformation is desirable to better meet the normal approximations. We call $\hat{Z}$ the estimator of the CCC [65]

$$\hat{Z} = \tanh^{-1}(\hat{\rho}_c) = \frac{1}{2} \ln \left(\frac{1 + \hat{\rho}_c}{1 - \hat{\rho}_c} \right) \quad (5)$$

And the standard deviation of this estimate is:

$$\sigma_{\hat{Z}}^2 = \frac{\frac{(1 - \rho^2)\hat{\rho}_c^2}{(1 - \hat{\rho}_c^2)^2} + \frac{2\hat{\rho}_c^3(1 - \hat{\rho}_c)u^2}{\rho(1 - \hat{\rho}_c^2)^2} - \frac{\hat{\rho}_c^4 u^4}{2\rho^2(1 - \hat{\rho}_c^2)^2}}{N - 2} \quad (6)$$

With the location shift relative to the scale parameter: $u = \frac{\mu_x - \mu_y}{\sigma_x\sigma_y}$.

Finally the confidence interval at 95% for the CCC is:

$$[\tanh(\hat{Z} - 1.64\sigma_{\hat{Z}}); \tanh(\hat{Z} + 1.64\sigma_{\hat{Z}})] \quad (7)$$

APPENDIX B

FULL SCORES WITH CONFIDENCE INTERVAL

Table 8 summarizes the complete fusion CCC results with their confidence intervals. woc: without context, wc: with context.

ACKNOWLEDGMENTS

The authors thanks AlloMedia company for its technical support in the collection of the AlloSat data, including its transcription. The authors would like to thank Nathan Guilhot, internship at LIA in 2021, who helped towards the reproducibility of the results.

REFERENCES

- [1] K. Cheong, J. Kim, and S. So, "A study of strategic call center management: relationship between key performance indicators and customer satisfaction," *European Journal of Social Sciences*, vol. 6, no. 2, pp. 268–276, 2008.
- [2] J. A. Russell, "A circumplex model of affect," *Journal of personality and social psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.
- [3] K. R. Scherer, "What are emotions? and how can they be measured?" *Social science information*, vol. 44, no. 4, pp. 695–729, 2005.
- [4] W. M. Wundt, *Outlines of psychology*. W. Engelmann, Leipzig, 1897.
- [5] H. Schlosberg, "Three dimensions of emotion," *Psychological review*, vol. 61, no. 2, pp. 81–88, 1954.
- [6] Y. Alva M, N. Muthuraman, and J. Paulose, "A comprehensive survey on features and methods for speech emotion detection," in *Proc. of IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, Coimbatore, Tamilnadu, India, 2015, pp. 1–6.
- [7] M. Wöllmer, M. Kaiser, E. F., B. Schuller, and G. Rigoll, "LSTM-modeling of continuous emotions in an audiovisual affect recognition framework," *Image and Vision Computing*, vol. 31, no. 2, pp. 153–163, 2013.
- [8] F. Alam and G. Riccardi, "Fusion of acoustic, linguistic and psycholinguistic features for speaker personality traits recognition," in *Proc. of ICASSP*, Florence, Italy, 2014, pp. 955–959.
- [9] P. Atrey, M. Hossain, A. El Saddik, and M. Kankanhalli, "Multimodal fusion for multimedia analysis: A survey," *Multimedia Syst.*, vol. 16, no. 1, pp. 345–379, 2010.
- [10] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proc. of ACL*, Melbourne, Australia, 2018, p. 2247–2256.
- [11] M. Macary, M. Tahon, Y. Estève, and A. Rousseau, "AlloSat: A new call center french corpus for satisfaction and frustration analysis," in *Proc. of Language Resources and Evaluation Conference (LREC)*, Virtual Conference, 2020, pp. 1590–1597.
- [12] A. Batliner, S. Steidl, B. Schuller, D. Seppi, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, V. Aharonson, L. Kessous, and N. Amir, "Whodunnit – searching for the most important feature types signalling emotion-related user states in speech," *Computer Speech & Language*, vol. 25, no. 1, pp. 4–28, 2011.

- [13] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. André *et al.*, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.
- [14] M. Tahon and L. Devillers, "Towards a small set of robust acoustic features for emotion recognition: Challenges," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 1, pp. 16–28, 2016.
- [15] D. Ververidis, C. Kotropoulos, and I. Pitas, "Automatic emotional speech classification," in *Proc. of ICASSP*, Montreal, Canada, 2004, pp. 593–596.
- [16] T. Vogt and E. Andre, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition," in *International Conference on Multimedia and Expo*, Amsterdam, Netherlands, 2005, pp. 474–477.
- [17] O. W. Kwon, K. Chan, J. Hao, and T. W. Lee, "Emotion recognition by speech signals," in *European Conference on Speech Communication and Technology*, Geneva, Switzerland, 2003, pp. 125–128.
- [18] B. Schuller, S. Steidl, A. Batliner, J. Krajewski, J. Epps, F. Eyben, F. Ringeval, and *al.*, "The INTERSPEECH 2014 computational paralinguistics challenge: Cognitive & physical load," in *Proc. of INTERSPEECH*, Singapore, 2014, pp. 427–431.
- [19] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer *et al.*, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proc. of INTERSPEECH*, Lyon, France, 2013, pp. 148–152.
- [20] S. Vanaja and M. Belwal, "Aspect-level sentiment analysis on e-commerce data," in *Proc. of International Conference on Inventive Research in Computing Applications (ICIRCA)*, Coimbatore, Tamil Nadu, India, 2018, pp. 1275–1279.
- [21] S. Dhar, S. Pednekar, K. Borad, and A. Save, "Sentiment analysis using neural networks: A new approach," in *Proc. of International Conference on Inventive Communication and Computational Technologies (ICICCT)*, New Delhi, India, 2018, pp. 1220–1224.
- [22] S. Poria, A. Gelbukh, A. Hussain, D. Das, and S. Bandyopadhyay, "Enhanced SenticNet with affective labels for concept-based opinion mining," *IEEE Intelligent Systems*, vol. 28, p. 31–38, 2013.
- [23] C. Monnier and A. Syssau, "Affective norms for French words (FAN)," *Behavior Research Methods*, vol. 46, no. 4, pp. 1128–1137, 2014.
- [24] H. Meisheri and L. Dey, "TCS research at SemEval-2018 task 1: Learning robust representations using multi-attention architecture," in *Proc. of The International Workshop on Semantic Evaluation*, New Orleans, Louisiana, USA, 2018, pp. 291–299.
- [25] S. Chaffar and D. Inkpen, "Using a heterogeneous dataset for emotion analysis in text," in *Proc of Advances in Artificial Intelligence*, St. John's, NF, Canada, 2011, pp. 62–67.
- [26] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communication of ACM*, vol. 61, no. 5, p. 90–99, 2018.
- [27] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Proc. of INTERSPEECH*, Dresden, Germany, 2015, pp. 1537–1540.
- [28] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, and *al.*, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. of ICASSP*, Shanghai, China, 2016, pp. 5200–5204.
- [29] M. Schmitt and B. Schuller, "Deep recurrent neural networks for emotion recognition in speech," in *DAGA*, Munich, Germany, 2018, pp. 1537–1540.
- [30] M. Schmitt, N. Cummins, and B. W. Schuller, "Continuous emotion recognition in speech - do we need recurrence?" in *Proc. of INTERSPEECH*, Graz, Austria, 2019, pp. 2808–2812.
- [31] M. Macary, M. Lebourdais, M. Tahon, Y. Estève, and A. Rousseau, "Multi-corpus experiment on continuous speech emotion recognition: convolution or recurrence?" in *Proc. of Conference on Speech and Computer (SPECOM)*, Virtual Conference, 2020.
- [32] B. T. Atmaja and M. Akagi, "Dimensional speech emotion recognition from speech features and word embeddings by using multitask learning," *APSIPA Transactions on Signal and Information Processing*, vol. 9, no. e17, pp. 1–12, 2020.
- [33] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," in *Proc. of Spoken Language Technologies Workshop (SLT)*, Athens, Greece, 2018, pp. 112–118.
- [34] S. Sahu, V. Mitra, N. Seneviratne, and C. Y. Espy-Wilson, "Multimodal learning for speech emotion recognition: An analysis and comparison of asr outputs with ground truth transcription," in *Proc. of INTERSPEECH*, Graz, Austria, 2019, pp. 3302–3306.
- [35] J. Sebastian and P. Pierucci, "Fusion Techniques for Utterance-Level Emotion Recognition Combining Speech and Transcripts," in *Proc. of INTERSPEECH*, Graz, Austria, 2019, pp. 51–55.
- [36] S. Planet and I. Iriondo, "Comparison between decision-level and feature-level fusion of acoustic and linguistic features for spontaneous emotion recognition," in *Proc. of the Iberian Conference on Information Systems and Technologies (CISTI)*, Madrid, Spain, 2012, pp. 1–6.
- [37] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O'Reilly, S. Han, P. Liu, M. Chen, and Y. Tong, "Feature-level and model-level audiovisual fusion for emotion recognition in the wild," in *Proc. of Multimedia Information Processing and Retrieval (MIPR)*, San Jose, California, USA, 2019, pp. 443–448.
- [38] S. Chen, Q. Jin, J. Zhao, and S. Wang, "Multimodal multi-task learning for dimensional and continuous emotion recognition," in *Proc. of the Audio/Visual Emotion Challenge and Workshop (AVEC)*, Mountain View, California, USA, 2017, p. 19–26.
- [39] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017.
- [40] C. Wu and W. Liang, "Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels," *IEEE Transactions on Affective Computing*, vol. 2, no. 1, pp. 10–21, 2011.
- [41] A. Barhoumi, N. Camelin, C. Aloulou, Y. Estève, and L. Hadrich Belguith, "Toward qualitative evaluation of embeddings for Arabic sentiment analysis," in *Proc. of Language Resources and Evaluation Conference (LREC)*, Virtual Conference, 2020, pp. 4955–4963.
- [42] B. T. Atmaja, K. Shirai, and M. Akagi, "Speech emotion recognition using speech feature and word embedding," in *Proc. of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Lanzhou, China, 2019, pp. 519–523.
- [43] P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, and *al.*, "Speech emotion recognition using spectrogram and phoneme embedding," in *Proc. of INTERSPEECH*, Hyderabad, India, 2018, pp. 3688–3692.
- [44] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Minneapolis, Minnesota, USA, 2019, pp. 4171–4186.
- [45] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu *et al.*, "Libri-light: A benchmark for ASR with limited or no supervision," in *Proc. of ICASSP*, Virtual Conference, 2020, pp. 7669–7673.
- [46] A. T. Liu, S. Yang, P. Chi, P. Hsu, and H. Lee, "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders," in *Proc. of ICASSP*, Virtual Conference, 2020, pp. 6419–6423.
- [47] H. Nguyen, F. Bougares, N. Tomashenko, Y. Estève *et al.*, "Investigating self-supervised pre-training for end-to-end speech translation," in *Proc. of the workshop on Self-supervision in Audio and Speech at the International Conference on Machine Learning (ICML)*, Virtual Conference, 2020.
- [48] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "Wav2vec: Unsupervised pre-training for speech recognition," in *Proc. of INTERSPEECH*, Graz, Austria, 2019, pp. 3465–3469.
- [49] P. Chi, P. Chung, T. Wu, C. Hsieh, S. Li *et al.*, "Audio ALBERT: A lite BERT for self-supervised learning of audio representation," in *Pre-print on arXiv/2005.08575*, 2020.
- [50] H. Le, L. Vial, J. Frej, V. Segonne, M. Coavoux *et al.*, "FlauBERT: Unsupervised language model pre-training for French," in *Proc. of Language Resources and Evaluation Conference (LREC)*, Virtual Conference, 2020, pp. 2479–2490.
- [51] S. Patel and K. R. Scherer, *Vocal Behaviour*. De Gruyter Mouton, 2013, pp. 167–204.
- [52] L. Devillers and L. Vidrascu, "Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs," in *Proc. of INTERSPEECH*, Pittsburgh, Pennsylvania, USA, 2006, pp. 801–804.
- [53] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Pre-print on arXiv/1301.3781*, 2013.
- [54] L. Martin, B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary *et al.*,

- "CamemBERT: a tasty French language model," in *Proc. of ACL*, Virtual Conference, 2020, pp. 7203–7219.
- [55] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *International Conference on Learning Representations*, New Orleans, Louisiana, USA, 2019, pp. 1–16.
- [56] L. Devillers, C. Vaudable, and C. Chasatgnol, "Real-life emotion-related states detection in call centers: a cross-corpora study," in *Proc. of INTERSPEECH*, Makuhari, Chiba, Japan, 2010, pp. 2350–2355.
- [57] K. M. Morrison, "Natural resources, aid, and democratization: A best-case scenario," *Public Choice*, vol. 131, no. 3–4, pp. 365–386, 2007.
- [58] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schröder, "The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 5–17, 2012.
- [59] F. Ringeval, A. Sonderegger, J. S. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pp. 1–8, 2013.
- [60] J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt *et al.*, "SEWA DB: A rich database for audio-visual emotion and sentiment research in the wild," *IEEE transactions on pattern analysis and machine intelligence*, pp. 1–1, 2019.
- [61] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, H. Kaya *et al.*, "AVEC 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition," in *Proc. of the Audio/Visual Emotion Challenge and Workshop (AVEC)*, Beijing, China, 2018, pp. 3–13.
- [62] S. Zahorian and H. Hu, "A spectral/temporal method for robust fundamental frequency tracking," *The Journal of the Acoustical Society of America*, vol. 123, no. 6, pp. 4559–4571, 2008.
- [63] L.-K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, vol. 45, no. 1, pp. 255–268, 1989.
- [64] J. J. Z. Liao and J. W. Lewis, "A note on concordance correlation coefficient," *PDA Journal of Pharmaceutical Science and Technology*, vol. 54, no. 1, pp. 23–26, 2000.
- [65] G. B. McBride, "A Proposal for Strength-of-Agreement Criteria for Lin's Concordance Correlation Coefficient," National Institute of Water & Atmospheric Research Ltd, Tech. Rep., 2005.
- [66] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems (NIPS)*, Vancouver, Canada, 2019, pp. 8024–8035.
- [67] M. Macary, M. Tahon, Y. Estève, and A. Rousseau, "On the use of Self-supervised Pre-trained Acoustic and Linguistic Features for Continuous Speech Emotion Recognition," in *IEEE Spoken Language Technology Workshop (SLT)*, Virtual, China, Jan. 2021.
- [68] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. of ICASSP*, South Brisbane, Queensland, Australia, 2015, pp. 5206–5210.
- [69] R. Řehůřek and P. Sojka, "Software framework for topic modelling with large corpora," in *Proc. of the LREC Workshop on New Challenges for NLP Frameworks*, Valletta, Malta, 2010, pp. 45–50.
- [70] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi *et al.*, "RoBERTa: A robustly optimized BERT pretraining approach," in *Pre-print on arXiv/1907.11692*, 2019.
- [71] P. J. Ortiz Suárez, B. Sagot, and L. Romary, "Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures," in *Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*, Cardiff, United Kingdom, 2019, pp. 9–16.
- [72] S. Evain, H. Nguyen, H. Le, M. Z. Boito, S. Mdhaffar, S. Alisamir, Z. Tong, N. Tomashenko, M. Dinarelli, T. Parcollet, A. Al-lauzen, Y. Estève, B. Lecouteux, F. Portet, S. Rossato, F. Ringeval, D. Schwab, and Laurent Besacier, "Task Agnostic and Task Specific Self-Supervised Learning from Speech with LeBenchmark," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.
- [73] B. Pallaud, R. Bertrand, P. Blache, L. Prévot, and S. Rauzy, "Suspensive and Disfluent Self Interruptions in French Language Interactions," in *Fluency and Disfluency across Languages and Language Varieties*, ser. Corpora and Language in use, P. U. de Louvain, Ed., 2019, no. 4.
- [74] A. Nasr, F. Béchet, J.-F. Rey, B. Favre, and J. Le Roux, "Macaon: An nlp tool suite for processing word lattices," in *Proc. of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Systems Demonstrations (HLT)*. USA: Association for Computational Linguistics, 2011, p. 86–91.

Manon Macary received the master's degree in Le Mans University (LIUM), where she is currently pursuing the Ph.D degree. She is also an employee of the Allo-Media company where she is active in the industrialization of her works. Her current research interests include emotion recognition from speech and spoken language understanding.

Marie Tahon is currently Associate Professor at Le Mans University and conducts her research at LIUM (France). She graduated in engineering from the Ecole Centrale de Lyon (France) in 2007 and received the M.S. degree in acoustics from the Ecole Centrale de Lyon, in 2007. She received the Ph.D. degree in computer science from the University of Paris-Sud (Orsay, France) in 2012. She has been with the LIMSI-CNRS (Orsay, France) and with the IRISA (Lannion, France). Her research interests concern automatic speech processing for expressive speech: recognition and synthesis.

Yannick Estève received the M.S. (1998) in computer science from the Aix-Marseilles University and the Ph.D. (2002) from Avignon University, France. He joined Le Mans Université (LIUM lab) in 2003 as an associate professor, and became a full professor in 2010. He moved to Avignon University in 2019 and is the head of the Computer Science Laboratory of Avignon (LIA) since 2020. He has authored and co-authored more than 150 journal and conference papers in speech and language processing.

Daniel Luzzati is currently Emeritus Professor at Le Mans University (LIUM). He performed two thesis, the first on spoken language (University of Paris 3), the second on human machine dialog (LIMSI-CNRS, Orsay). His research deals with these two areas, that affective computing both involves.