Continuous emotion prediction from audio signal with acoustic and linguistic representations

Marie Tahon, Manon Macary, Yannick Estève

▶ **To cite this version:**

Marie Tahon, Manon Macary, Yannick Estève. Continuous emotion prediction from audio signal with acoustic and linguistic representations. 16ème Congrès Français d'Acoustique, CFA2022, Société Française d'Acoustique; Laboratoire de Mécanique et d'Acoustique, Apr 2022, Marseille, France. hal-03847806

HAL Id: hal-03847806
https://hal.science/hal-03847806v1

Submitted on 10 Nov 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

16ème Congrès Français d'Acoustique
11-15 Avril 2022, Marseille

Continuous emotion prediction from audio signal with acoustic and linguistic representations.

M. Tahon [a], M. Macary [a,b], Y. Estève [c]
[a] LIUM, Le Mans Université
[b] AlloMedia, Le Mans
[c] LIA, Avignon Université

The use of machine learning techniques has been a wide reference in many speech processing tasks. The power of neural networks (NN) allows to solve some very complex tasks such as automatic speech emotion prediction. Our works aim at continuously estimating the degree of satisfaction or frustration of speaker in call-center conversations. More precisely, we extract an acoustic representation directly from the audio signal and a linguistic representation from the automatic textual transcription, which will then be processed by a recurrent NN able to predict the level of satisfaction between 0 and 1 every 0.25s. Self-supervised learning allows to learn general contextualized speech representations with multi-layer convolutional networks from very large amount of unlabeled data. It is then possible to extract such representations (or embeddings) from new specific data as emotional speech. These representations have the advantage of capturing informations from a lot of data, what is not the case of models learnt on specific emotional speech only. We set up a protocol including different types of representation in input of the network : (i) cepstral coefficients (MFCCs), (ii) expert prosodic descriptors ; (iii) words pre-trained embeddings and (iv) signal pre-trained embeddings. Our results confirm the high potential of embedding representations for our task. More surprisingly, we show that the linguistic content seems to bring more emotional information than the single audio signal. A fine grained linguistic analysis will confirm this result.

1 Introduction

Affective computing is a field of research at the crossroads between artificial intelligence and human emotions. A lot of applications are derived from such a field. In social sciences, the characterization of the emotional content of a speaker can help to analyse political debates, understand social interactions in dialogues. From an industrial point of view, the knowledge of a speaker's emotional state can provide information to improve commercial relations with customers in call centers.

It is usually established that emotion in speech can be transmitted through linguistic messages conveyed by words and paralinguistic messages conveyed by the acoustic signal [1]. While activation (passive/active) is known to be well-recognized from acoustic features, valence (negative/positive) is known to be better recognized with linguistic ones [2]. According to the circumflex model of emotion from Scherer [3], satisfaction and frustration can be considered as a combination of activation and valence dimensions. Consequently, linguistic and acoustic features should both be highly relevant to retrieve such emotions.

For a long time, paralinguistic information in speech has been modelled with expert prosodic and acoustic features capturing intensity, intonation, rhythm or voice quality. These features are supposed to describe the voice production process usually on the basis of a source-filter model. Most of the expert feature sets [4, 5] intend to describe prosody in the signal, with low level descriptors (LLDs) such as fundamental frequency, loudness, spectral envelop features, rhythmic patterns, combined with additional statistical functionals. These features are still used in the Music Information Retrieval domain [6]. However, with the automatic processing of massive audio data, expert acoustic features has been replaced by standardized features such as Mel Frequency Cepstral Coefficients (MFCC) which are more robust to acoustic environment changes [7]. Emotion in speech is then captured by machine learning models trained on annotated audio databases in supervised manner.

Because manual discrete and continuous emotion annotation is a highly subjective perception task, it has to be done by multiple annotators to be relevant, and this explains the huge cost of the creation of large emotional speech datasets. Consequently, Speech Emotion Recognition (SER) databases are usually quite small (SEWA [8] : ∼44 h, RECOLA [9] : ∼2.8 h, AlloSat [10] : ∼37 h). That is one of the reasons why deep neural networks (DNN) have been used only recently in SER in comparison to Automatic Speech Recognition (ASR) where accessible databases are drastically bigger (*e.g.* TED-LIUM 3 [11] ∼450 h, LibriSpeech [12] : ∼960 h).

Transfer learning [13] within the deep learning paradigm, is a machine learning method where a DNN trained for a task is reused partially or entirely as the starting point to train or fine-tune a neural network on a second task. Such methods were also proposed to limit the impact of lack of data when only small databases are available to train a neural network for a specific domain or task. Transfer learning from ASR is widely used in Text Sentiment Analysis, where large databases are used to train generic cues which are fed into the training process, leading to better generalization abilities given limited training data [14].

As a variant of transfer learning, self-supervised learning of speech or language representations has been proposed in these last few years, for instance with the BERT system [15], used for textual representation. Such representations, computed by neural models trained on huge amounts of unlabelled data, have shown their effectiveness on some tasks under certain conditions, for instance for computer vision [16] and Natural Language Processing (NLP) tasks as described in [17].

For these reasons, we decided to study the impact of linguistic and acoustic features extracted with self-supervised pre-trained models in order to predict continuous emotion. In this paper, we investigate different speech and/or textual representations computed by models pre-trained through self-supervised learning for SER task. Since these already existing pre-trained models were initially designed for speech recognition ASR [18] or natural language understanding [15], it is not obvious that they are also relevant for SER. For instance, at the acoustic level ASR tends to focus on phone level that lasts about 30 ms while emotions are usually supported on about 1 s of speech.

2 Experimental protocol

This section describes the data used to train and evaluate our models, and the neural network designed to continuously predict the satisfaction.

2.1 Speech emotional data : AlloSat corpus

AlloSat corpus [10] is composed of real-life call-center conversations, annotated along the satisfaction axis. It 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 - and we intend to predict the satisfaction associated with such conversations.

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.

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 [10]. An automatic transcription was 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.

TABLEAU 1 – Summary of AlloSat characteristics. F/M : number of female/male speakers.

Statistics	Value
number of conversations	303
number of speakers (F/M)	308 (191/117
total duration	37h23m27s
min duration conversations	32s
max duration conversations	41m
mean duration conversations	7m24s

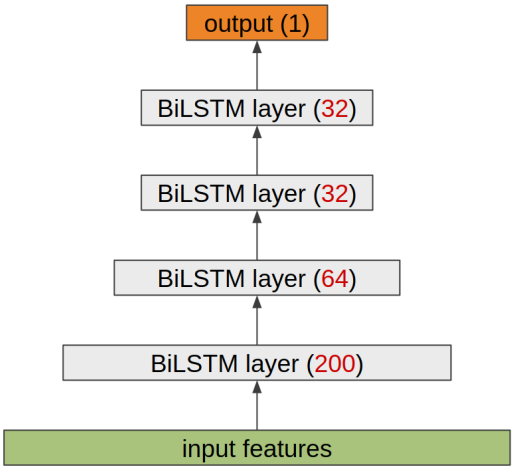

FIGURE 1 – Baseline network architecture. Number of neurons of each layer are written in red.

2.2 SER neural network model

2.2.1 Baseline architecture

We designed a regressive baseline neurol network to continuously predict the satisfaction along the conversation. To do so, a recurrent network, inspired from [19], is used for the prediction task using bidirectionnal 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. 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.

2.2.2 Loss and evaluation function

The concordance correlation coefficient (CCC) [20] 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.

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

In previous experiments on the prediction of emotional dimensions [21, 19], 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 [22].

2.2.3 Hyper-parameters

All networks are implemented under Pytorch framework [1]. 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.

2.2.4 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 experiments [2], the final CCC score of one of the experiments 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 reste of the article.

1. https://pytorch.org/
2. The results are not presented here

3 Signal representations

3.1 Acoustic modality

Baseline (MFCCs, eGeMAPS) 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 toolkit [3]. Mean and standard deviation of each coefficient are computed over the emotional segment in order to get a 48 dimensional vector each 250 ms.

The extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [4]. enables to capture prosodic features suitable for SER. 23 LLDs are extracted at the frame level. Mean and standard deviation of these 23 LLDs are computed over the emotional segment. This feature set is extracted with the toolkit OpenSmile [23]. 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.

Pre-trained Wav2Vec Self-supervised learning approaches have been designed in order to take benefit of huge amount of unlabelled data. Wav2Vec (1.0) [18] 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 [18], trained on Librispeech corpus [12] consisting of 960 hours of English audio book samples at 16 kHz. Our features were extracted on an upsampled version of AlloSat [4]. 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.

3.2 Linguistic modality

Baseline (Word2Vec) Word2Vec embeddings have been extensively used for sentiment analysis or opinion mining from text [24, 25], this motivated us to use such representation for the prediction of satisfaction. In the following experiments, a Word2Vec model has been trained

3. https://pytorch.org/audio/stable/index.html
4. We used FFMpeg resampling function with *sinc* interpolation function

with the toolkit GENSIM [26], 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 [27]. 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.

Pre-trained (CamemBERT) Inspired by BERT, CamemBERT [29] is a multi-layer bidirectional Transformer. CamemBERT is trained on the Masked Language Modeling 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 [30] 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 representations, embeddings are extracted either on the words pronounced during the current emotional segment (without context) or on the whole conversation input (with context).

4 Results

4.1 Satisfaction performances

Figure 2 presents the evolution of satisfaction according time. The three curves summarize the ground truth (average value over the three annotators in red), the values predicted with baseline features (green) and pre-trained features (blue). The predicted values are obtained with a late fusion of the predictions obtained with both modalities. We can seen that pre-trained features predicts a smooth curve which as very close to the ground truth, while baseline features predicts a noisy curve which is clearly distinct from the labels.

The obtained performances in terms of CCC on both Development and Test sets are summarized in Table 2. From this table, we can clearly claim the advantage of using pre-trained features for this task. Indeed for both acoustic and linguistic modalities, the pre-trained features get the best performances on both Development and Test sets. It is however surprising that wav2vec is best (CCC=.806) when the context (previous and next frames) is not included,

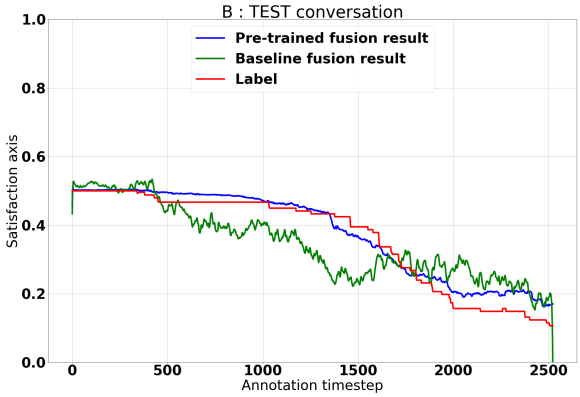

FIGURE 2 – Automatic prediction with the fusion of baseline features (MFCC + Word2Vec) and pre-trained features (Wav2Vec + CamemBERT

while CamemBERT is best (CCC=.924) when the context (previous and next words) is included. One of the reason which can explain this difference is the fact that anonymized words have been replaced in the speech signal by jazzy sounds which can disturb the surroundings of the current segment.

From the results, we can also conclude to the supremacy of the linguistic modality to predict satisfaction in our telephone conversations. Because this result was not expected as most SER models are based on the acoustic signal only. The automatic transcription has the advantage of providing some information at higher level that the signal itself. To better understand this insight, we have conducted a deep linguistic analysis on a small subset of conversations.

TABLEAU 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.

		Satisfaction	
Modality	**# size**	**Dev**	**Test**
AUDIO			
MFCC	48	.698	.513
eGeMAPS	46	.422	.354
Wav2Vec woc	512	**.844**	**.806**
Wav2Vec wc	512	.823	.656
TEXT			
Word2vec	40	.805	.569
Word2vec	100	.860	.759
CamemBERT woc	768	.916	.817
CamemBERT wc	768	**.917**	**.924**

4.2 Linguistic analysis

Through the linguistic analysis, we intend to provide elements that could explain the importance of linguistics to

Tableau 3 – Extract (137 - 166 sec.) from a conversation about a certified letter. Disfluencies : *italic* ; Hesitations, repairs, babbling : underline ; Semantic evidences of frustration : **bold** ; self-breaks : //

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

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,
- Usage of interrogations and negations,
- Semantic evidences of frustration or unhappiness,
- Amount of meaningfull 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.

5 Discussion and conclusion

This article investigates the use of different signal representations on both linguistic and acoustic modalities. The results suggest that pre-trained features are highly relevant for this kind of task for which the amount of data is relatively small. Indeed, pre-trained features are extracted with models which have seen a lot of diverse speech data during the training phase, thus are highly efficient when training SER models with unseen data. 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. This work raises the question of the place of acoustic cues, especially prosodic features. We suppose that DNN learn an implicit representation of prosody, however, it is a very hard task to turn these internal representations into interpretable cues. To pursue our investigation, we aim at applying the presented protocol on additional speech data, for instance broadcast news, political debates, etc.

Références

[1] 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, Pennsylvanie, USA, 2006, pp. 801–804.

[2] 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.

[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] 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.

[5] 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.

[6] B. McFee, R. Colin, L. Dawen, E. Daniel PW, M. McVicar, E. Battenberg, and O. Nieto, "librosa : Audio and music signal analysis in python," in *14th python in science conference*, 2015, pp. 18–25.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] F. Hernandez, V. Nguyen, S. Ghannay, N. Tomashenko, and Y. Estève, "TED-LIUM 3 : Twice as much data and corpus repartition for experiments on speaker adaptation," in *Proc. of Conference on Speech and Computer (SPECOM)*, Leipzig, Germany, 2018, pp. 198–208.

[12] 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.

[13] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[14] X. Dong and G. de Melo, "A helping hand : Transfer learning for deep sentiment analysis," in *Proc. of ACL*, Melbourne, Australia, 2018, pp. 2524–2534.

[15] 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.

[16] L. Nanni, S. Ghidoni, and S. Brahnam, "Handcrafted vs. non-handcrafted features for computer vision classification," *Pattern Recognition*, vol. 71, pp. 158–172, 2017.

[17] S. Evain, H. Nguyen, H. Le, M. Z. Boito, S. Mdhaffar, S. Alisamir, Z. Tong, N. Tomashenko, M. Dinarelli, T. Parcollet, A. Allauzen, 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.

[18] 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.

[19] 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.

[20] L.-K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, vol. 45, no. 1, pp. 255–268, 1989.

[21] 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.

[22] 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.

[23] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE – the munich versatile and fast open-source audio feature extractor," in *Proc. of the ACM Multimedia International Conference*, Savannah,Georgia,USA, 2010, pp. 1459–1462.

[24] 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".

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.