Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1590–1597
Marseille, 11–16 May 2020

AlloSat: A New Call Center French Corpus for Satisfaction and Frustration Analysis

Manon Macary[1,2], Marie Tahon[1], Yannick Estève[3], Anthony Rousseau[2]
[1]LIUM / Le Mans Université, [2]Allo-Média, [3]LIA / Université d'Avignon
m.macary@allo-media.fr, marie.tahon@univ-lemans.fr, yannick.esteve@univ-avignon.fr, a.rousseau@allo-media.fr

Abstract

We present a new corpus, named AlloSat, composed of real-life call center conversations in French that is continuously annotated in frustration and satisfaction. This corpus has been set up to develop new systems able to model the continuous aspect of semantic and paralinguistic information at the conversation level. The present work focuses on the paralinguistic level, more precisely on the expression of emotions. In the call center industry, the conversation usually aims at solving the caller's request. As far as we know, most emotional databases contain static annotations in discrete categories or in dimensions such as activation or valence. We hypothesize that these dimensions are not task-related enough. Moreover, static annotations do not enable to explore the temporal evolution of emotional states. To solve this issue, we propose a corpus with a rich annotation scheme enabling a real-time investigation of the axis frustration / satisfaction. AlloSat regroups 303 conversations with a total of approximately 37 hours of audio, all recorded in real-life environments collected by Allo-Media (an intelligent call tracking company). First regression experiments, with audio features, show that the evolution of frustration / satisfaction axis can be retrieved automatically at the conversation level.
Keywords: Speech Corpus, Call center, Emotion Recognition, Frustration Satisfaction

1. Introduction

Information extraction is an important topic for research and industry. For instance, since in call center human agents receive hundreds of calls per day, it is interesting to provide indicators to help them analyzing this large amount of data. So we planned to work on emotion analysis in order to guide people into the analysis of massive conversations.

The present work is the first step of a bigger project located at the cross-domain between semantic and paralinguistic information modeling in call center spontaneous speech directly from audio signal. Call center speech, well-known for its negative emotions, is very interesting as these conversations are good representations of real-life expressions of speaker's states. It also has the advantage of producing massive amounts of data even if they have to be anonymized in order to protect the private information of the callers which costs time and money partially explaining why there are not a lot of available corpora in this field.

Nowadays, existing systems retrieve emotional information from the textual transcription of the conversations, in the field of sentiment analysis, opinion mining, or from audio signal, in the field of Speech Emotion Recognition (SER).

To better model emotions dimensions, we first studied the two main psychological models used in affect analysis to define emotions. The first one is composed of discrete emotional categories such as the "Big Six" (Ekman, 1999) which are joy, anger, surprise, sadness, disgust and fear; often added with a "neutral" class. The other one describes the complex nature of affect in speech with continuous dimensions, notably activation and valence (Russel, 1997), but also dominance, intention or conducive/obstructive axis (Scherer, 2005). Activation and valence are particularly convenient as most of the discrete emotional labels can be translated into these two dimensions, thus allowing multi-corpora approaches (Schuller, 2018).

We studied existing and available corpora for SER. Existing corpora are often acted one and usually not related to call center conversations. Even if many efforts are made to move from acted to real-life databases, there are still few available emotional spontaneous speech corpora. Most of these corpora aim at modeling social aspects of real-life or induced interactions such as laughter (Devillers et al., 2015) or disfluencies (Gilmartin and Campbell, 2016). In SER corpora, emotion is mainly represented with discrete categories, for instance anger, neutral and positive in call center conversations (Devillers et al., 2010), probably because "part of the reason for the dominance of discrete emotions is the ease of collecting training data" (Campbell, 2008).

In the course of AVEC challenges (Valstar et al., 2013), recent studies explored the prediction of continuous dimensions such as activation and valence in SEMAINE multimodal database (McKeown et al., 2012) and SEWA database (Kossaifi et al., 2019). SEMAINE is composed of simulated conversations between a human user and a machine through Sensitive Artificial Listener (SAL) scenarios (Douglas-Cowie et al., 2008) and SEWA consists of discussions on commercials between two persons, talking about the ads they saw. Call center corpora are usually domain-dependant: DECODA (Lailler et al., 2016) (parisian transportation operator) is annotated with named entities, CallSurf (Garnier-Rizet et al., 2008) (French energy operator) is partially annotated with emotion categories (Devillers et al., 2010) or NATURAL (Morrison et al., 2007) (Chinese electricity company) is annotated with two classes: anger and neutral. To our knowledge, no call center corpus gathers different domains with the same annotation scheme. Allo-Media company develops services for cross-domain call centers allowing us to collect data from various domains.

The main goals of call center conversations are either to pursue a person to sign a contract, or to solve some technical or financial problems. As a result, the question of the

evolution of frustration or satisfaction (called satisfaction dimension in the following) along the conversation, is crucial. State of the art representations of emotions (discrete or continuous) are able to model static speakers' states, however they are not goal-related. For these reasons, we intend to investigate the satisfaction dimension in the context of call center conversations.

Frustration and satisfaction are close to extreme categories of the Conducive / Obstructive axis as described by (Scherer, 2005), thus we consider this axis and the satisfaction dimension as comparable. Therefore, we propose a new corpus dedicated to the analysis of call centers conversations continuously annotated among the satisfaction dimension. We also add discrete valence annotations in order to be able to compare the performances of systems built on this corpus to the performance got on other existing databases.

In the remainder, the corpus design and the annotation protocol are introduced in section 2. Section 3 focuses on the analysis of the resulting annotation while section 4 explains the systems used for predicting the continuous satisfaction dimension, and section 5 shows the very first experiments results.

2. Corpus Design

2.1. General context

The corpus is composed of telephone conversations between speakers (the callers) and agents (the receivers) where speakers are French native adults. The information asked by the callers is various. It can be about contracts, global information on the company, complaints, etc. All conversations were recorded 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 recordings are sampled at 8 kHz.

2.2. Data collection

As we retrieved a huge number of calls, we had to decide which one had to be annotated. We could not annotate all calls owned by the company because of the cost and the time needed to treat such a huge amount of data. Moreover, we know there is no emotion expression in all conversations so we had to discard them. So we set up three criteria to select conversations:

- **Duration**: we decided to take only conversations longer than 30 seconds containing more than three speech turns.
- **Standard deviation (STD) of the fundamental frequency (F_0)**: extracted with YAPPT algorithm (Zahorian and Hu, 2008) (adapted to telephone signals), is a well known marker for emotion detection. It enabled us to only keep 500 conversations which were maximizing the F_0 standard deviation.
- **Valence score**: computed on conversations transcriptions using the French dictionary FAN (Monnier and Syssau, 2014). This dictionary contains a polarity value (between 0 and 10) for more than 1000 French words. We only use the words which have a polarity value to compute the score. The valence score is the mean of the word polarity value of each polarized words at the conversation level.

A manual check of automatically selected conversations enabled the selection of 253 recordings supposed to contain the expression of emotions.

To keep our corpus significant for phone conversation in call center, 50 neutral randomly selected conversations were added as we explained earlier that all conversations in this context does not always convey emotions. This procedure results in a database containing 303 conversations[1].

2.3. Audio preprocessing

The two audio channels (speaker and agent) were separated which allows us to have distinct documents for the caller and the agent. For ethical and commercial reasons the agent channel was discarded. As a result, the corpus contains callers' voice only without any overlapping speech. Because of the absence of agent response, there can be long moments of silence in our data. In order to minimize the annotator effort, we decided to replace these silences by 2 seconds of white noise, allowing the annotators to identify potentially longer silences. The resulting duration distribution of these conversations is represented in the Figure 1. Conversations last between 32 seconds to 41 minutes as reported in Table 1.

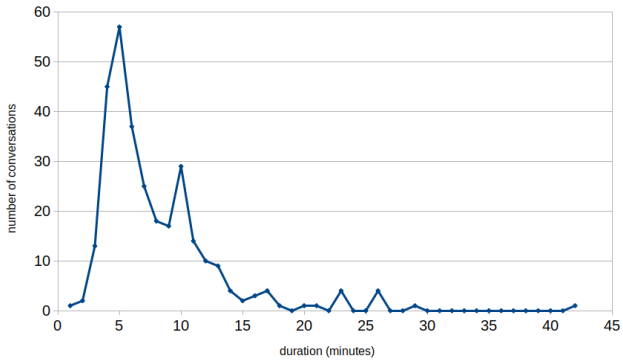

Figure 1: Conversation duration distribution in minutes.

There is generally a single speaker per conversation but we can also observe conversations where they are multiple speakers, when the caller switches with someone else. In total we have 308 speakers divided in 191 women and 117 men. The main characteristics of the corpus are summed up in the Table 1.

All our conversations also have automatic transcriptions thanks to a Kaldi based system (Povey et al., 2011) owned by Allo-Media. This ASR system is a variant for French language of the LIUM system detailed in (Garcia-Martinez et al., 2015). These transcriptions are partially manually checked.

[1] https://lium.univ-lemans.fr/allosat

Stat	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
automatic transcript	303

Table 1: Main characteristics of the corpus

2.4. Anonymization

In order to preserve the speakers' privacy, personal data (i.e. card number, IBAN number, phone number, first name, last name, spelled first or last name, e-mail address, path, zip code, city, etc) are obfuscated, respecting the General Data Protection Regulation (GDPR) recommendation. We also obfuscated everything that can identify the company: especially brands and products. All personal data is replaced with a jazzy sound of same duration, enabling the listener to know that there was private information at this very moment. The corpus was manually checked to search for those information and we also deleted all number series such as contract or client number.

Personal information was deleted from transcription and replaced by named entities, allowing us to know what kind of personal information was said.

2.5. Continuous satisfaction dimension annotation

In order to perform a continuous annotation, we have adapted CARMA (Girard, 2014), a toolkit derived from FeelTrace (Cowie et al., 2000) allowing us to make an annotation over one axis: frustration to satisfaction, using the arrows of a keyboard or a mouse. Annotators are able to see their own annotation as a curve while annotating. We customized the settings in order to match with the annotation scheme we produced. We put up a scale from 0 (extremely frustrated) to 10 (extremely satisfied). The continuous satisfaction dimension is initialized to 5, which is supposed to correspond to the neutral state. Emotions are mainly detectable within a second (Schuller and Devillers, 2010) unlike words which are usually studied by windows of 30ms. So we chose to retrieve the position of the cursor as an annotation every 0.25 seconds allowing us to have 4 values per second. The annotation were made by 3 annotators, 2 women and 1 man. They were given a guideline explained in section 2.7.

Two examples of continuous satisfaction dimension annotations are given in Figure 2 where a high inter annotator agreement is observed. In conversation A, we can observe that the caller is going from neutral state (5) to frustrated (almost 0) and stay relatively frustrated (1-2) at the end of the call. The conversation B corresponds to one of the neutral conversations randomly picked. We can see that the curve is relatively steady.

2.6. Discrete annotation

The corpus is also composed of discrete annotations corresponding to the speaker's emotional state at the beginning and at the end of the conversation, and also the temporal evolution between these states. Beginning and ending duration of the conversations are left to the appreciation of annotators. Speaker's states have been annotated according to two dimensions: Satisfaction and Valence. The discrete labels for these dimensions and their temporal evolution are the following:

- **Satisfaction dimension**: Very satisfied, satisfied, neutral, frustrated, very frustrated.
- **Valence**: very positive, positive, neutral, negative, very negative.
- **Temporal evolution** (for satisfaction and valence dimensions): rise, fall, stagnate, fluctuate, fluctuate considerably.

To summarize, each annotator has to complete 6 fields per conversation. Only satisfaction dimension ratings are investigated in the rest of this paper, valence will be used in future work.

2.7. Guidelines

The guidelines given to annotators contain the aim of the study and describes how their work is going to be used.
We required them to be as objective as possible in order to minimize the bias given by the annotator perceptions. In order to help them understand emotional dimension, we show them the self-assessment manikin aka SAM (Bradley and Lang, 1994) pictograms to grasp the concept of valence. In order to calibrate the perception of the satisfaction dimension, we extracted from the corpus two conversations that we thought were the extreme boundaries of frustration and satisfaction i.e. the conversations with the most frustrated caller and another one with the most satisfied caller.
The annotators could not go backward or forward in the document in order to keep a continuous annotation and they can annotate a conversation only once. We also ask them not to move the cursor when there are silences or non-expressive parts. We indicated that it was important to add the beginning and ending discrete annotations just after listening the whole conversation, in order to have the clear memory of what happened.

3. Data analysis

3.1. Annotation description

We choose to regroup "very satisfied" and "satisfied" in a single class because there were too few examples of "very satisfied". "Very frustrated" and "frustrated" labels were also merged. Annotators ratings were merged with a majority vote. In case of total disagreement, we trust the annotator (named a1, a2 and a3) who has the best correlation coefficient (i.e. a2 as seen section 3.3.2). We also merge for the temporal evolution "fluctuate" and "fluctuate considerably" for the same reason. The final distribution is summarized in Table 2.

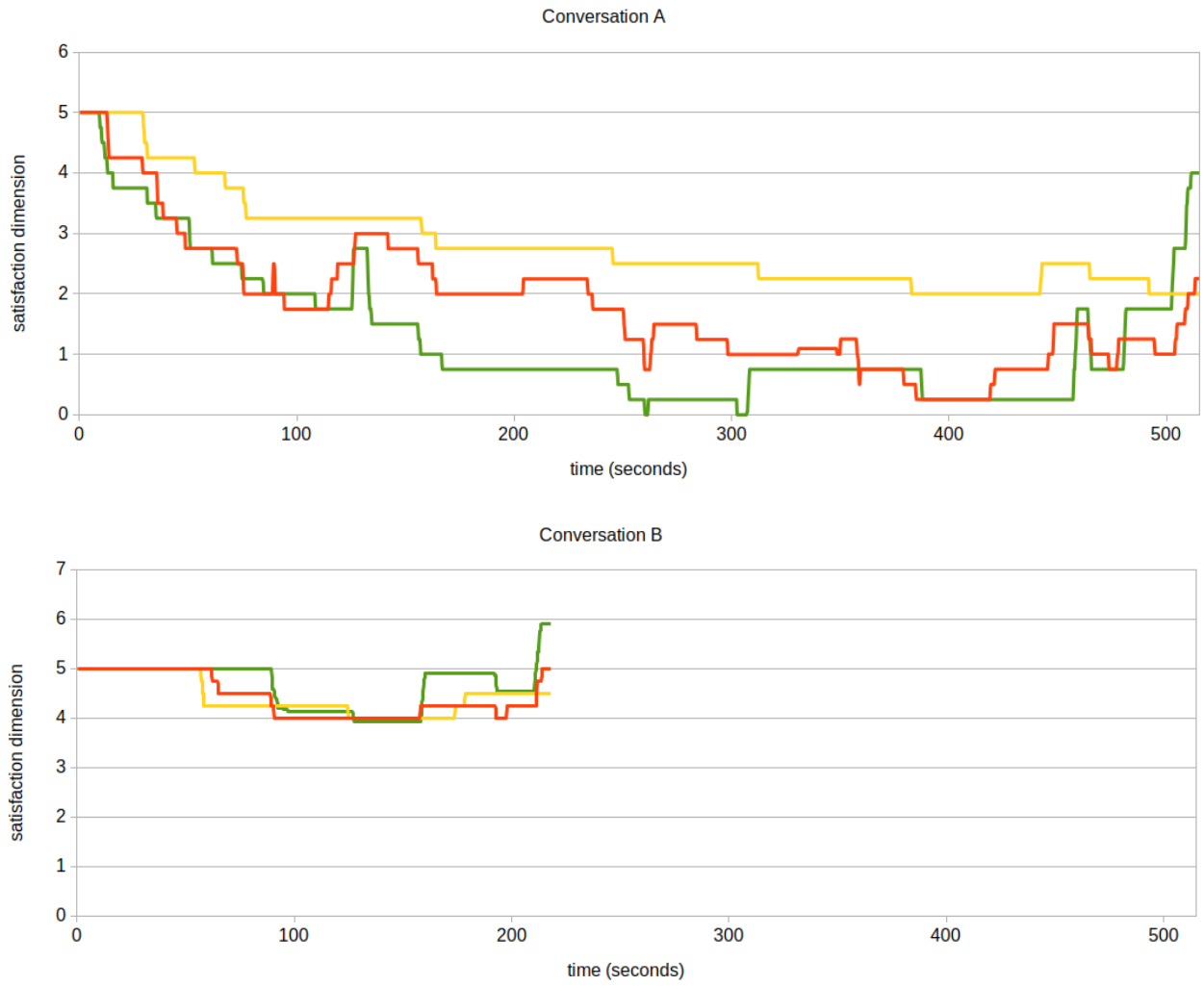

Figure 2: Temporal Satisfaction dimension annotations for the three annotators on the whole conversations. Discrete ending labels are "very frustrated" for A and "neutral" for B.

	Begin	End	Evolution	
satisfied	0	7	falling	190
neutral	299	107	stagnate	64
frustrated	4	189	rising	10
			fluctuate	39

Table 2: Distribution of the satisfaction's discrete annotations by majority voting.

We can observe that neutral state is over represented in conversation beginnings. It can be explained by two observations. Firstly, the continuous annotation is initialized to 5, meaning neutral. This can bias the perception of the annotator. But the main hypothesis is that the speaker is rarely frustrated at the beginning of the call: this emotion is provoked by the agent's response. The same goes to satisfaction. As expected, most conversations have been perceived with an increasing frustration, probably because the agent is not able to completely solve the speaker request.

3.2. Continuous annotation discretization

The annotation scheme was designed to verify if discrete annotation (beginning and ending annotation) matched with temporal satisfaction dimension for each annotator. To that end, we firstly normalize every annotation with a min/max normalization meaning that the annotation's value are now between [0, 1]. Then we discretized in time and in value annotators' satisfaction dimension called S. We called S_n the satisfaction dimension for the annotator n. First, we defined two thresholds to determine whenever the continuous normalized satisfaction dimension corresponds to frustration ($S < 0.45$), satisfaction ($S > 0.55$) or neutral state. These thresholds were defined by observing the annotations of the neutral conversations added to the corpus. Second, we defined the beginning (respectively ending) segment as the first (respectively last) 10% of the conversation.

For each annotator n, we calculated S_n^{begin} (resp. S_n^{end}) as the mean of S_n over the beginning (resp. ending) segment. The difference between S_n^{begin} and S_n^{end} is expected to be directly related to the satisfaction dimension evolution.

3.3. Annotation consistency

We cross-checked the different annotation levels using kappa scores and correlation coefficients.

3.3.1. Intra annotator consistency

To evaluate annotators consistency between continuous and discrete annotations, we compared the discrete annotation of beginning's and ending's satisfaction dimension to the discretized continuous satisfaction dimension annotation of each conversation as presented in the section 3.2. This protocol enables us to calculate a kappa value given by equation (1) for each annotator on the three classes frustrated / neutral / satisfied. P_0 is the relative observed agreement among raters while P_e represents random agreement. In case of an over-represented class (here neutral) P_e can be set to 1/3 (Callejas and López-Cózar, 2008), 3 representing the number of classes.

$$\kappa = \frac{P_0 - P_e}{1 - P_e} \quad (1)$$

Table 3 shows good consistency between discretized continuous and discrete labels on all beginnings ($\kappa_{avg} = 0.93$) and on endings ($\kappa_{avg} = 0.77$). We can conclude that continuous and discrete ratings are consistent through annotators.

3.3.2. Inter annotator agreement

In order to evaluate inter annotators agreement on temporal ratings, we used the correlation coefficient (R). This coefficient is computed at the conversation level on normalized continuous satisfaction dimension S between pairs of annotators. Final values reported in Table 3 show a good correlation between raters ($R_{avg} = 0.83$), thus meaning that continuous annotations are consistent between annotators.
Finally, the inter annotator agreement on discrete beginnings and endings are given with kappa values between pairs of annotators for the three classes (frustrated, neutral and satisfied). As shown in Table 3, annotators' agreement is very strong at the beginning of the conversations ($\kappa_{avg} = 0.91$) and lowers at the end ($\kappa_{avg} = 0.77$) but still relevant.

Intra annotator			**Inter annotator**			
Single	κ^{beg}	κ^{end}	Pairs	R	κ^{beg}	κ^{end}
a1	0.98	0.84	a1-a2	0.82	0.99	0.90
a2	0.88	0.72	a2-a3	0.87	0.88	0.69
a3	0.93	0.75	a1-a3	0.80	0.87	0.72
Avg.	0.93	0.77	Avg.	0.83	0.91	0.77

Table 3: Intra and inter annotator agreements per single/pairs annotator and average. a_i represents annotator i. R represents the correlation coefficient.

One of the reasons for this tendency is that the beginning of the conversation is nearly always neutral. The hypothesis for this phenomenon was explained above in Section 3.1.
Starting from these promising agreement results, all ratings can be merged and annotations can be used for analyses purposes. We compute a gold annotation for every conversation, by meaning the annotation's values of the 3 annotators for the continuous satisfaction dimension. This gold annotation is used in the following experiments.

4. Models for satisfaction prediction

As we said in Section 1, we want to help agents treating a large amount of conversations. In order to do so, having clues about the satisfaction dimension of the caller can be beneficial. Therefore, we define a task of satisfaction dimension prediction throughout the conversation. We compare two models for this prediction's task. One is the baseline model used in the 2018 AVEC challenge (Ringeval et al., 2018). Using this model will allow us to compare our result on this corpus to the result on the SEWA corpus described in the introduction. The second is a Deep Neural Network (DNN) models with biLSTM (bidirectional Long Short Term Memory) layers already used on the SEWA corpus (Schmitt et al., 2019). They are both using audio features as input extracted with the OpenSMILE framework (Eyben et al., 2010). We wanted to test different audio feature sets in order to find which one was the most suitable for our corpus. Both models and sets are explained in the following.

4.1. Acoustic features

To better compare our work with state-of-the-art in SER, we decided to use the well known eGeMAPS (Eyben et al., 2016) feature set. This feature set, implemented in OpenSMILE framework (Eyben et al., 2010), was designed for automatic voice analysis especially affect analysis. It contains 25 Low Level Descriptors (LLD) such as pitch, jitter, formants, loudness, etc. Arithmetics mean and standard deviation (STD) are computed every 0.1 seconds. Other LLD specific functionals are also extracted totaling a number of 88 features. In (Schmitt et al., 2019) f_eGeMAPS have been defined with 25 LLDs and functionnals (mainly mean and STD) extracted from eGeMAPS totalling 46 features. A last feature, behaving like voice activation detection (vad), denoting the speaker identity (0 or 1), is also included in f_eGeMAPS.
In our work, both sets of features has been extracted from our data every 0.25 seconds. Since we only keep the caller's signal, we modify the vad to denote if the caller is speaking (1) or not (0). The number of features in the 4 feature sets are summarized in Table 4.

Name	**Number features**
eGeMAPS	88
f_eGeMAPS	46
eGeMAPS+vad	89
f_eGeMAPS+vad	47

Table 4: Acoustic feature sets

4.2. DNN architectures

4.2.1. Input preprocessing

The neural network architecture we used requires a fixed input sequence size. As we have seen in Section 3, the conversations have variable lengths from 32 seconds to 41 minutes with a mean of 7m24s and a STD of 4m58s. Usually, the length is fixed to mean+STD in order to cover more than 95% of the corpus. Long sequences are cut at mean+STD while short sequences are padded. In our case,

		2 biLSTM		4 biLSTM	
		dev	test	dev	test
SEWA	eGeMAPS	0.112*	-	-	-
(valence)	f_eGeMAPS	-	-	0.517*	0.410*
AlloSat	eGeMAPS	0.607	0.424	**0.672**	**0.494**
(satisfaction)	f_eGeMAPS	0.591	0.399	0.615	0.445
	eGeMAPS&vad	0.596	0.423	0.657	0.477
	f_eGeMAPS&vad	0.601	0.421	0.618	0.399

Table 5: Results measured with CCC (1 means perfect prediction). * The results for SEWA corpus are coming from works for the AVEC Challenge 2018

the mean+STD is 12m22s. In order to reduce the effect of padding and the training duration costs, we decided to fix the length to 7 minutes. We applied a circular padding on short sequences.
We divided our corpora in three subsets in order to respect the distribution of neutral conversations: a train set (201 conversations), a development set (42 conversations) and a test set (60 conversations).

4.2.2. 2 biLSTM layers network

In order to be able to compare our results with the existing state of art, we made the choice to reproduce the system proposed in the AVEC 2018 challenge (Ringeval et al., 2018) on the Cross-cultural Emotion Sub-challenge. This model is composed of 2 biLSTM layers of respectively 64 and 32 units with tanh activation. A dropout of 0.1 is used to improve the performance. A single output neuron is used to predict the regression every 0.25 seconds.

4.2.3. 4 biLSTM layers network

This model is composed of 4 biLSTM layers as described in (Schmitt et al., 2019). The bidirectional architecture is used in order to prevent the problem of annotation delay. The layers are composed of respectively 200, 64, 32, 32 units with tanh activation. A single output neuron is also used to predict the regression every 0.25 seconds.

5. Experiments results

The DNNs are implemented with the Keras framework [2] using the Tensorflow backend [3]. Training is done on batches of 9 conversations using the ADAgrad optimiser. The learning rate is initialized at 0.001. The number of epochs was first fixed to 500. But after preliminary experiments, we observe that the network was not improving after the first 150 epochs, so we reduced our number of epochs to 200. We took the network weights of the epoch which had the best score on the development set to score on the test set. The concordance correlation coefficient (CCC) (Lin, 1989) was used as the loss function for training the network, and as evaluation metric to determine the best system. CCC score goes from 0 (chance level) to 1 (perfect) and is calculated thanks to the equation 2, where x correspond to the predicted value and y the label. μ_x and μ_y are the means for the two variables and σ_x and σ_y are the corresponding variances. ρ is the correlation coefficient between the two variables σ_x and σ_y therefore the covariance coefficient.

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

We compare two networks trained on two different emotional axis: satisfaction dimension with AlloSat and valence with SEWA. Table 5 gives a summary of the results obtained with the investigated models and datasets.
We can discuss that the result shows us that the corpora is relevant. In fact, we are able to retrieve good CCC's scores on it, comparable to the valence's result retrieved on the SEWA corpus. We have to take the CCC's score carefully because as we show in Figure 3, the system is able to do good predictions (conversation C) but also bad ones (conversation D).
In order to explain this phenomenon, we analyzed, at the conversation level, the CCC between the 3 pairs of annotators. In the case where, at least, two annotators have rated similarly satisfaction during the conversation, the CCC computed between the two annotation's values is close to 1. On the contrary, if satisfaction is completely differently rated by the three annotators, the CCC computed on each pair is close to 0. Consequently, the gold annotation, defined as the mean of the 3 annotation's values, is not consistent.
In our results, we observe that when maximum of the 3 CCCs computed on each pair is low, the predicted satisfaction is likely to be bad. On the contrary, if this maximum is high, the predicted satisfaction is likely to be good. Moreover, we may face an over-fitting situation because of the small amount of data.

6. Conclusion

In this paper, we introduce AlloSat, a new French call center speech corpus usable to explore the satisfaction dimension (from satisfaction to frustration) in real-life telephone conversations. This corpus contains 303 conversations for a total of more than 37 hours and can be obtained by following the procedure explained on the LIUM website [4]. The major objective of this work was to ensure the consistency of this new corpus. Good intra and inter annotator agreements validate the existence of satisfaction dimension in call center conversations. Our result also enables the use of the continuous manual ratings in regression experiments which are also distributed. The first experiments show that

[2] https://keras.io
[3] https://www.tensorflow.org/
[4] https://lium.univ-lemans.fr/allosat

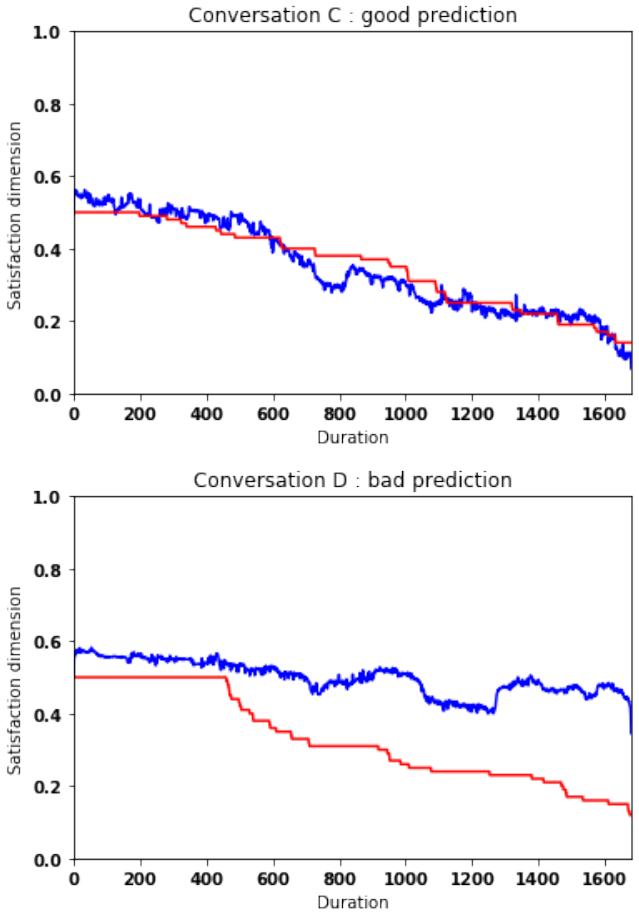

Figure 3: Continuous satisfaction dimension prediction on conversations from the test set. The label is in red, the prediction in blue.

biLSTM neural networks are able to predict satisfaction values.

As a next step, deeper investigations will be conducted to automatically retrieve the temporal evolution of the satisfaction dimension with better performances. In the future, we plan to go further in our experiments on both continuous and discrete annotations by using other regression protocols including regularization methods to reduce over-fitting. The addition of textual and semantic information combined with acoustic features is currently work in progress.

7. Acknowledgements

The authors thank coders and co-workers who participated in elaborating protocols and annotating emotional states. In particular Corinne Bignon and her staff for the implication in the annotation process.

8. Bibliographical References

Bradley, M. M. and Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59.

Callejas, Z. and López-Cózar, R. (2008). On the use of kappa coefficients to measure the reliability of the annotation of non-acted emotions. In *Perception in Multimodal Dialogue Systems*, pages 221–232, Berlin, Heidelberg. Springer Berlin Heidelberg.

Campbell, N., (2008). *Expressive/Affective Speech Synthesis*, pages 505–518. Springer Berlin Heidelberg, Berlin, Heidelberg.

Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., , and Schröder, M. (2000). Feeltrace: An instrument for recording perceived emotion in real time. In *ITRW on Speech and Emotion*, pages 19–24, Newcastle, UK.

Devillers, L., Vaudable, C., and Chasatgnol, C. (2010). Real-life emotion-related states detection in call centers: a cross-corpora study. In *Proc. of Interspeech*, pages 2350–2355, Makuhari, Chiba, Japan.

Devillers, L., Rosset, S., Duplessis, G. D., Sehili, M. A., Béchade, L., Delaborde, A., Gossart, C., Letard, V., Yang, F., Yemez, Y., Turker, B. B., Sezgin, M., Haddad, K. E., Dupont, S., Luzzati, D., Esteve, Y., Gilmartin, E., and Campbell, N. (2015). Multimodal data collection of human-robot humorous interactions in the joker project. In *Proc. of International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 348–354, Xian, China.

Douglas-Cowie, E., Cowie, R., Cox, C., Amier, N., and Heylen, D. (2008). The sensitive artificial listener: an induction technique for generating emotionally coloured conversation. In *LREC Workshop on Corpora for Research on Emotion and Affect*, pages 1–4, Marrakech, Morocco.

Ekman, P., (1999). *Basic Emotions*, pages 301–320. Wiley, New-York.

Eyben, F., Wöllmer, M., and Schuller, B. (2010). opensmile – the munich versatile and fast open-source audio feature extractor. In *Proc. of the ACM Multimedia 2010 International Conference*, pages 1459–1462, Firenze, Italy.

Eyben, F., Scherer, K., Schuller, B., Sundberg, J., André, E., Busso, C., Devillers, L., Epps, J., Laukka, P., Narayanan, S., and Truong, K. (2016). The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202.

Garcia-Martinez, M., Barrault, L., Rousseau, A., Deléglise, P., and Estève, Y. (2015). The LIUM ASR and SLT Systems for IWSLT 2015. In *12th International Workshop on Spoken Language Translation (IWSLT 2015)*, pages 50–54, Da Nang, Vietnam.

Garnier-Rizet, M., Adda, G., Cailliau, F., Guillemin-Lanne, S., Waast-Richard, C., Lamel, L., Vanni, S., and Waast-Richard, C. (2008). CallSurf: Automatic transcription, indexing and structuration of call center conversational speech for knowledge extraction and query by content. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*, pages 2623–2628, Marrakech, Morocco.

Gilmartin, E. and Campbell, N. (2016). Capturing chat: Annotation and tools for multiparty casual conversation. In *Proc. of the International Conference on Language*

Resources and Evaluation (LREC), pages 4453–4457, Portorož, Slovenia.

Girard, J. M. (2014). CARMA: Software for continuous affect rating and media annotation. *Journal of Open Research Software*, 2(1):e5.

Kossaifi, J., Walecki, R., Panagakis, Y., Shen, J., Schmitt, M., Ringeval, F., Han, J., Pandit, V., Schuller, B. W., Star, K., Hajiyev, E., and Pantic, M. (2019). Sewa db: A rich database for audio-visual emotion and sentiment research in the wild. *IEEE transactions on pattern analysis and machine intelligence*.

Lailler, C., Landeau, A., Béchet, F., Estève, Y., and Deléglise, P. (2016). Enhancing the RATP-DECODA corpus with linguistic annotations for performing a large range of NLP tasks. In *Proc. of Language Resources and Evaluation Conference (LREC)*, pages 1047–1050, Portorož, Slovenia.

Lin, L. I.-K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45(1):255–268.

McKeown, G., Valstar, M., Cowie, R., Pantic, M., and Schröoder, M. (2012). The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1):5–17.

Monnier, C. and Syssau, A. (2014). Affective norms for French words (FAN). *Behavior research methods*, 46 4:1128–1137.

Morrison, D., Wang, R., and De Silva, L. (2007). Ensemble methods for spoken emotion recognition in call-centres. *Speech Communication*, 49(2):98–112.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The kaldi speech recognition toolkit. In *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Hawaii, US.

Ringeval, F., Schuller, B., Valstar, M., Cowie, R., Kaya, H., Schmitt, M., Amiriparian, S., Cummins, N., Lalanne, D., Michaud, A., Ciftçi, E., Güleç, H., Salah, A. A., and Pantic, M. (2018). Avec 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition. In *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, AVEC'18, pages 3–13, New York, NY, USA. ACM.

Russel, J., (1997). *Reading emotions from and into faces: Resurrecting a dimensional-contextual perspective*, pages 295–360. Cambridge University Press, U.K.

Scherer, K. R., (2005). *What are emotions ? and how can they be measured ?*, chapter Social Science Information, pages 695–729.

Schmitt, M., Cummins, N., and Schuller, B. W. (2019). Continuous Emotion Recognition in Speech - Do We Need Recurrence? In *Proc. Interspeech 2019*, pages 2808–2812, Graz, Austria.

Schuller, B. and Devillers, L. (2010). Incremental acoustic valence recognition: An inter-corpus perspective on features, matching, and performance in a gating paradigm. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010*, pages 801–804, Makuhari, Chiba, Japan.

Schuller, B. W. (2018). Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the ACM*, 61(5):90–99.

Valstar, M. F., Schuller, B. W., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., Schnieder, S., Cowie, R., and Pantic, M. (2013). Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proc. of the AudioVisual Emotion Challenge*, Barcelon, Spain.

Zahorian, S. and Hu, H. (2008). A spectral/temporal method for robust fundamental frequency tracking. *The Journal of the Acoustical Society of America*, 123:4559–71.