Prédiction continue de la satisfaction et de la frustration dans des conversations de centre d'appels

Manon Macary, Marie Tahon, Yannick Estève, Anthony Rousseau

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Prédiction continue de la satisfaction et de la frustration dans des conversations de centre d'appels

Manon Macary[1, 2] Marie Tahon[1] Yannick Estève[3] Anthony Rousseau[2]
(1) LIUM, Le Mans, France
(2) Allo-Media, Paris, France
(3) LIA, Avignon, France
m.macary@allo-media.fr, marie.tahon@univ-lemans.fr,
yannick.esteve@univ-avignon.fr, a.rousseau@allo-media.fr

RÉSUMÉ

Nous présentons un nouveau corpus, nommé AlloSat, composé de conversations en français extraites de centre d'appels, annotées de façon continue en frustration et satisfaction. Dans le contexte des centres d'appels, une conversation vise généralement à résoudre la demande de l'appelant. Ce corpus a été mis en place afin de développer de nouveaux systèmes capables de modéliser l'aspect continu de l'information sémantique et para-linguistique au niveau conversationnel. Nous nous concentrons sur le niveau para-linguistique, plus précisément sur l'expression des émotions. À notre connaissance, la plupart des corpus émotionnels contiennent des annotations en catégories discrètes ou dans des dimensions continues telles que l'activation ou la valence. Nous supposons que ces dimensions ne sont pas suffisamment liées à notre contexte. Pour résoudre ce problème, nous proposons un corpus permettant une connaissance en temps réel de l'axe frustration/satisfaction. AlloSat regroupe 303 conversations pour un total d'environ 37 heures d'audio, toutes enregistrées dans des environnements réels, collectées par Allo-Media (une société spécialisée dans l'analyse automatique d'appels). Les premières expériences de classification montrent que l'évolution de l'axe frustration/satisfaction peut être prédite automatiquement par conversation.

ABSTRACT

AlloSat : A New Call Center French Corpus for Affect Analysis

We present a new corpus, named AlloSat, composed of real-life call center conversations in French, 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 annotations in discrete categories or in dimensions such as activation or valence. We hypothesize that these dimensions are not task-related enough. To solve this issue, we propose a corpus 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 classification experimentsshow that the evolution of frustration / satisfaction axis can be retrieved automatically at the conversation level.

MOTS-CLÉS : Corpus, Reconnaissance des Émotions, Centre d'appels, Satisfaction / Frustration.

KEYWORDS: Speech Corpus, Emotion Recognition, Call center, Frustration / Satisfaction.

Volume 1 : Journées d'Études sur la Parole, pages 379–387.
HAL : hal-02798561.

1 Introduction

Aujourd'hui, alors que nous sommes capables de stocker de plus en plus de données et notamment des données audio, leur valorisation est un sujet urgent. De nombreuses études se penchent donc sur l'extraction d'information, par exemple sémantique et para-linguistique. Dans les centres d'appels, des conseillers humains reçoivent des centaines d'appels par jour et tentent de répondre au mieux aux problématiques des appelants. Nous avons donc décidé de travailler sur la reconnaissance des émotions, appelé Speech Emotion Recognition (SER), car dans un contexte de centre d'appel, fournir à posteriori l'évolution de la satisfaction et de la frustration lors d'une conversation dans des gros volumes d'appels, peut intéresser l'entreprise afin d'améliorer sa qualité de service.

Si nous nous référons aux travaux menés en psychologie, il y a plusieurs modèles utilisés pour caractériser une émotion. Une approche consiste à décomposer les émotions en catégories discrètes telles que les " Big Six " (Ekman, 1999) ou les 32 émotions de la roue de Plutchik (Plutchik, 1980). Une autre approche décrit l'émotion comme un état continu, décrit par plusieurs dimensions notamment l'activation et la valence, mais aussi la dominance, l'intention ou l'axe conducteur/obstructif (Scherer, 2005) dont l'extrême positif est la satisfaction et l'extrême négatif est proche de la frustration.

Si on se replace dans notre contexte, l'objectif d'un appel est soit d'ouvrir un contrat, soit de résoudre des problèmes techniques ou financiers. Du coup la question de l'évolution de la frustration et de la satisfaction (appelé satisfaction par la suite) est cruciale. Nous avons donc cherché des corpus disponibles pour la reconnaissance d'émotion dans notre contexte. Les corpus existants sont souvent actés et ne sont généralement pas liés à des conversations de centre d'appels. Même si de nombreux efforts sont faits pour passer de corpus actés à des corpus de parole spontanée, il existe encore peu de corpus de parole spontanée et encore moins qui sont annotés en émotion continue. Dans les corpus SER, l'émotion est principalement représentée par des catégories discrètes, par exemple la colère, la joie dans des conversations de centre d'appels d'urgence (Devillers *et al.*, 2010) probablement parce que la collecte d'émotions discrètes est plus aisée que celle d'émotions continues (Campbell, 2008). On retrouve néanmoins des corpus comportant des annotations continues : l'activation et la valence pour le corpus SEMAINE (McKeown *et al.*, 2012) (composé de conversations simulées entre un utilisateur humain et une machine) et SEWA (Kossaifi *et al.*, 2019) (composé de conversations entre deux locuteurs à propos de publicités visualisées en amont).

Nous avons donc cherché à proposer un schéma d'annotation et de procéder à l'annotation d'un corpus existant en émotions continues. Cependant les corpus des centres d'appels sont en général très dépendants d'un domaine d'activité, par exemple on retrouve DECODA (Lailler *et al.*, 2016) (opérateur de transport parisien) qui est annoté avec des entités nommées, ou NATURAL (Morrison *et al.*, 2007) (compagnie d'électricité chinoise) qui est annoté avec deux classes : colère et neutre. À notre connaissance, aucun corpus de centre d'appels n'est composé de domaines hétéroclites.

N'ayant pas trouvé de corpus de centre d'appels suffisamment diversifié et annoté continûment en satisfaction de l'appelant, nous avons collecté des données provenant de différents domaines sur lequel nous avons mis en place notre propre schéma d'annotation en collaboration avec la société Allo-Media. Nous proposons donc un nouveau corpus, dédié à l'analyse de conversations en centres d'appels, annoté continûment en satisfaction. Les premières prédictions réalisées sur ce corpus nous montre des résultats encourageants qui seront développés au cours de cet article.

Le reste de cet article est organisé comme suit. La construction du corpus est introduite dans la section 2. La section 3 se concentre sur l'analyse de la cohérence des résultats tandis que la section 4 montre la mise en place des systèmes de prédiction et les résultats des premières expériences.

2 Construction du corpus

2.1 Contexte général

Le corpus est composé de conversations téléphoniques en français entre des interlocuteurs (les appelants) et des agents (les conseillers) où les locuteurs sont des adultes qui demandent des informations. Diverses informations sont demandées par les appelants : il peut s'agir de création de contrats, de demande d'informations globales sur l'entreprise, de plaintes, etc. Toutes les conversations ont été enregistrées entre juillet 2017 et novembre 2018 dans des centres d'appels situés dans des pays francophones. Les conseillers sont employés de diverses sociétés dans différents domaines. On retrouve notamment des entreprises du secteur de l'énergie, de la santé, du voyage, de la vente immobilière et de l'assurance.

2.2 Collecte de données

Comme nous avons récupéré un très grand nombre d'appels, nous avons dû décider quelles conversations devaient être annotées. En effet, nous ne pouvions pas annoter tous les appels reçus pendant la période de captation en raison du coût et du temps nécessaires pour traiter une telle quantité de données. De plus, nous savons que toutes les conversations ne sont pas porteuses d'émotions et encore moins de satisfaction, nous avons donc dû sélectionner des conversations. Nous avons donc mis en place trois critères pour sélectionner les conversations :

— **La durée** : nous avons décidé de ne prendre que des conversations de plus de 30 secondes contenant au moins trois tours de parole ;

— **Écart type (STD) de la fréquence fondamentale (F_0)** : extraite avec l'algorithme YAPPT (Zahorian & Hu, 2008) (adapté aux signaux téléphoniques), elle est un marqueur utilisé pour la détection des émotions. Cela nous a permis de conserver 500 conversations qui maximisaient l'écart-type F_0.

— **Score de valence** : calculé sur les transcriptions des conversations à l'aide du dictionnaire français FAN (Monnier & Syssau, 2014). Ce dictionnaire contient une valeur de polarité (entre 0 et 10) pour plus de 1000 mots français. Le score de valence est la valeur moyenne calculée pour chaque conversation.

Une vérification manuelle des conversations sélectionnées automatiquement a permis de sélectionner 253 enregistrements susceptibles de contenir des informations émotionnelles.

Comme la plupart des conversations en centre d'appels ne véhiculent pas d'émotions, nous avons ajouté 50 conversations neutres, sélectionnées au hasard afin que notre corpus reste cohérent avec la réalité des centres d'appel. Cette procédure aboutit à une base de données contenant 303 conversations.

2.3 Pré-traitement audio

Les deux canaux audio (interlocuteur et agent) ont été séparés, ce qui nous permet d'avoir des documents distincts pour l'appelant et l'agent. Pour des raisons éthiques et commerciales, le canal de l'agent a été supprimé. Par conséquent, le corpus contient uniquement la voix des appelants sans aucun chevauchement de signal des locuteurs. Puisque nous n'avons pas conservé la réponse de l'agent, il peut y avoir de longs moments de silence dans nos données. Afin de minimiser l'effort des annotateurs, nous avons décidé de remplacer ces silences par 2 secondes de bruit blanc, permettant

aux annotateurs d'identifier des silences potentiellement plus longs. Les conversations durent entre 32 secondes et 41 minutes, comme indiqué dans le tableau 1.

Il n'y a généralement qu'un seul locuteur par conversation. Au total, nous avons 308 locuteurs répartis en 191 femmes et 117 hommes. Les principales caractéristiques du corpus sont résumées dans le tableau 1.

Statistiques	**Value**
nombre de conversations	303
nombre de locuteurs	308
nombre de femmes	191
nombre d'hommes	117
durée totale	37h23m27s
durée min conversation	32s
durée max conversation	41m
durée moyenne conv.	7m24s
transcription automatique	303

TABLE 1 – Caractéristiques principales du corpus

Toutes nos conversations ont également des transcriptions automatiques grâce à un système basé sur Kaldi (Povey *et al.*, 2011) appartenant à Allo-Media.

2.4 Anonymisation

Afin de préserver l'anonymat des locuteurs, les données personnelles sont obfusquées , afin de respecter le règlement général sur la protection des données (RGPD). Nous avons également anonymisé tout ce qui peut identifier une entreprise, notamment les marques et les produits. Les informations personnelles sont supprimées et remplacées par des entités nommées dans les transcriptions, ce qui nous permet de savoir de quel type de données personnelles il s'agissait, et par un son jazzy dans l'audio.

2.5 Annotation de la satisfaction

Afin d'effectuer l'annotation continue, nous avons adapté CARMA (Girard, 2014), un toolkit dérivé de FeelTrace (Cowie *et al.*, 2000) nous permettant de faire une annotation sur un axe : de la frustration à la satisfaction, en utilisant les flèches d'un clavier. Nous avons personnalisé les paramètres afin de les faire correspondre à notre schéma d'annotation : une échelle commençant à 0 (extrêmement frustré) allant jusqu'à 10 (extrêmement satisfait). L'axe est initialisée à 5, censée correspondre à l'état neutre. Les émotions sont principalement détectables dans la seconde (Schuller & Devillers, 2010) contrairement aux mots qui sont généralement étudiés par des fenêtres de 30ms. Nous avons donc choisi de récupérer la position du curseur sous forme d'annotation toutes les 0,25 secondes. L'annotation a été faite par 3 annotateurs, 2 femmes et 1 homme. Ils ont reçus un guide d'annotation afin d'homogénéiser les annotations et de réduire l'aspect subjectif de celle-ci. Une annotation discrète a également été réalisée pour le début et la fin de la conversation, afin de vérifier la cohérence des annotations continues.

Deux exemples d'annotations de la satisfaction sont donnés sur la Figure 1 où les valeurs observées d'accord inter-annotateur sont respectivement de 0,851 et de 0,732. Dans la conversation A, nous

pouvons observer que l'appelant passe d'un état neutre (5) à frustré (presque 0) et reste relativement frustré (1-2) jusqu'à la fin de l'appel. La conversation B correspond à l'une des conversations neutres choisies au hasard.

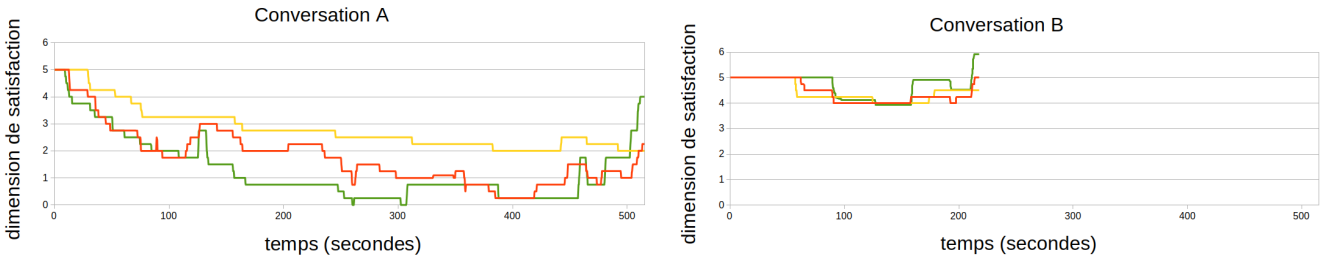

FIGURE 1 – Annotations continues de la satisfaction par les trois annotateurs pour deux conversations. Les étiquettes de fin discrètes sont "très frustrées" pour A et "neutres" pour B.

3 Analyse des données : cohérence des annotations

Afin d'évaluer l'accord inter-annotateur sur les annotations continues, nous avons utilisé le coefficient de corrélation. Ce coefficient est calculé au niveau de la conversation sur la satisfaction normalisée par rapport à l'ensemble des conversations entre les paires d'annotateurs. Les valeurs finales rapportées dans le tableau 2 montrent une bonne corrélation entre les annotateurs (en moyenne $R = 0,83$), ce qui signifie que les annotations continues sont cohérentes entre les annotateurs.

Paires d'annotateurs	Coefficient de corrélation R
a1-a2	0,82
a2-a3	0,87
a1-a3	0,80

TABLE 2 – Accords inter-annotateur par paire d'annotateur. a_i représente l'annotateur i.

L'une des raisons de ce fort accord est que le début de la conversation est presque toujours neutre. Cela peut s'expliquer de deux façons : d'abord, l'annotation continue est initialisée à 5, ce qui se traduit par un état neutre. Mais l'hypothèse principale est que l'interlocuteur est rarement frustré au début de l'appel : cette émotion est provoquée par les réponses de l'agent. Il en va de même pour la satisfaction. Comme nous le pensions, la plupart des conversations ont été perçues avec une frustration croissante, probablement parce que le conseiller n'est pas en mesure de donner une réponse suffisamment satisfaisante à l'interlocuteur.

En partant de ces résultats d'accord prometteurs, nous calculons une annotation de référence pour chaque conversation correspondant à la moyenne des trois annotations de la satisfaction et nous pouvons utiliser cette annotation de référence à des fins d'analyse et d'apprentissage. Cette annotation de référence est utilisée dans les expériences présentées par la suite.

4 Prédiction de la satisfaction

Comme nous l'avons dit dans l'introduction, notre objectif est de permettre de mieux comprendre l'état émotionnel des appelants dans un but d'analyse.Pour ce faire, il peut être utile d'avoir des indices sur la satisfaction de l'appelant tout au long de l'appel, et donc nous définissons une tâche de prédiction continue de cette dimension. Nous comparons deux modèles pour cette tâche. Le premier est le modèle de référence utilisé lors du challenge AVEC 2018 (Ringeval *et al.*, 2018). L'utilisation de ce modèle nous permettra de comparer les résultats obtenus sur AlloSat à ceux prédits sur SEWA (qui a été décrit en introduction), même si nous ne comparons pas exactement les mêmes dimensions, en effet AlloSat est annoté en satisfaction, alors que SEWA est annoté en valence. Le second est un modèle de réseau neuronal profond (DNN) avec des couches biLSTM (bidirectional Long Short Term Memory) déjà testé sur le corpus SEWA (Schmitt *et al.*, 2019). Ils utilisent tous deux des descripteurs audio comme entrée, que nous extrayons avec le framework OpenSMILE (Eyben *et al.*, 2010). Différents ensembles de descripteurs audio ont été testés afin de trouver celui qui était le plus adapté à notre corpus. Les modèles et les ensembles sont expliqués ci-dessous.

4.1 Descripteurs audio

Pour mieux comparer notre travail avec l'état de l'art dans le domaine du SER, nous avons décidé d'utiliser l'ensemble eGeMAPS (Eyben *et al.*, 2016). Cet ensemble a été conçu pour l'analyse automatique de la voix, en particulier l'analyse des émotions. Il contient 25 descripteurs de bas niveau (LLD) tels que le pitch, le jitter, les formants, etc. Une moyenne arithmétique et un écart-type (STD) sont calculés toutes les 0,1 secondes sur ces LLD. D'autres fonctions mathématiques calculées à partir de ces LLD sont également extraites pour un total de 88 descripteurs. Dans (Schmitt *et al.*, 2019) f_eGeMAPS a été défini avec 25 LLD et des fonctions mathématiques appliquées sur ces LLD (principalement moyenne et STD) extraits d'eGeMAPS totalisant 46 descripteurs. Un dernier descripteur, fonctionnant comme une détection de voix (voice activity detection i.e. vad), dénotant l'identité du locuteur (0 ou 1), est également incluse dans f_eGeMAPS.

Dans notre travail, les deux ensembles ont été extraits de nos données toutes les 0,25 seconde suivant le pas d'annotation d'AlloSat. Puisque nous ne gardons que le signal de l'appelant, nous modifions le vad pour indiquer si l'appelant parle (1) ou non (0). Le nombre de descripteurs des 4 ensembles est résumé dans le tableau 3.

4.2 Les architectures des DNN

4.2.1 Pré-traitement des inputs

Afin de s'aligner sur les architectures neuronales des articles de référence (Ringeval *et al.*, 2018; Kossaifi *et al.*, 2019), nous avons choisi d'utiliser une taille de séquence d'entrée fixe. Comme nous l'avons vu dans la section 3, les conversations ont des durées variant de 32 secondes à 41 minutes avec une moyenne (MOY) de 7m24s et un écart type (STD) de 4m58s. Habituellement, la taille d'entrée est fixée à $MOY + STD$ (ici 12m22s) pour couvrir plus de 95% du corpus. Les séquences longues sont alors coupées à la $MOY + STD$ tandis que les séquences courtes sont rallongées avec un padding. Afin de réduire l'effet du padding et la durée d'apprentissage, nous avons décidé de fixer la durée de la séquence d'entrée à 7 minutes. Nous avons appliqué un padding circulaire sur les courtes séquences.

Nous avons divisé notre corpus en trois sous-ensembles afin de respecter la répartition des conversations neutres : un apprentissage (201 conversations), un développement (42 conversations) et un test (60 conversations).

4.2.2 Les deux modèles neuronaux

Afin de pouvoir comparer nos résultats avec l'état de l'art, nous avons fait le choix de reproduire le système proposé dans le challenge AVEC 2018 (Ringeval *et al.*, 2018) sur la modalité "Cross-cultural Emotion". Le premier réseau neuronal est composé de 2 couches biLSTM de respectivement 64 et 32 unités. L'architecture bidirectionnelle est utilisée afin d'éviter les problèmes de délai d'annotation. En effet, il est possible que l'annotation présente des délais dans l'annotation, le temps que l'annotateur appuie sur les flèches du clavier ou qu'il décide s'il y a vraiment une variation à annoter.

Le second réseau est composé de 4 couches biLSTM comme décrit dans (Schmitt *et al.*, 2019). Les couches de biLSTM sont composées respectivement de 200, 64, 32, 32 unités.

Pour ces deux réseaux, la fonction d'activation utilisée est la fonction tangente hyperbolique. Un seul neurone de sortie est utilisé pour prédire une valeur toutes les 0,25 secondes.

4.3 Résultats des expériences

Les DNN sont implémentés avec le framework Keras [1] en utilisant Tensorflow [2]. L'apprentissage se fait par ensemble (batch) de 9 conversations en utilisant l'optimiseur ADAgrad. Le learning rate est initialisé à 0,001. Le nombre d'époques a d'abord été fixé à 500 avant d'être réduit à 200 puisque les réseaux ne s'amélioraient pas au delà d'environ 120 époques. Nous avons conservés les poids des réseaux donnant le meilleur score sur le développement afin de prédire les résultats sur le test. Le coefficient de corrélation de concordance (CCC) (Lin, 1989) a été utilisé comme fonction pour l'apprentissage du réseau et comme métrique d'évaluation pour déterminer le meilleur système. Ce score CCC varie de 0 (probabilité d'un tirage aléatoire) à 1 (corrélation parfaite).

Nous comparons deux réseaux appris sur deux axes émotionnels différents : la satisfaction avec AlloSat et la valence avec SEWA. Le tableau 3 donne un résumé des résultats obtenus avec les modèles et les sets de données étudiés.

		nb Descripteurs	**2 biLSTM**		**4 biLSTM**	
			dev	test	dev	test
SEWA	eGeMAPS	88	0.112*	-	-	-
(valence)	f_eGeMAPS	46	-	-	0,517*	0,410*
AlloSat	eGeMAPS	88	0,510	0,363	0,666	0,431
(satisfaction)	f_eGeMAPS	46	0,469	0,260	0,607	0,354
	eGeMAPS&vad	89	0,549	0,365	0,619	0,542
	f_eGeMAPS&vad	47	0,508	0,359	0,574	0,422

TABLE 3 – Résultats des expériences. *Ces résultats proviennent du challenge AVEC 2018 (Schmitt *et al.*, 2019)

Il semble que nous sommes en mesure de récupérer de bons scores de CCC pour notre corpus, comparables aux résultats de valence prédits sur le corpus SEWA. Le score CCC calculé sur l'ensemble

1. https ://keras.io
2. https ://www.tensorflow.org/

des données doit être pris avec précaution car, comme nous le montrons dans la Figure 2, le système est capable de faire de bonnes prédictions (conversation C) mais aussi de mauvaises prédictions (conversation D).

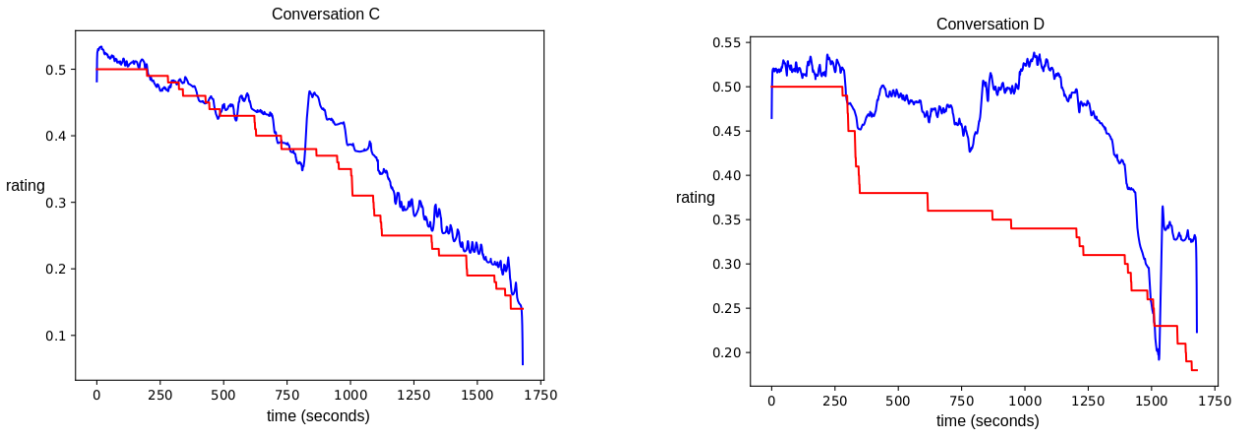

FIGURE 2 – Prédiction de la satisfaction sur des conversations issues du test. La référence est en rouge, la prédiction en bleu.

5 Conclusion

Dans cet article, nous présentons AlloSat, un nouveau corpus de conversations françaises en centre d'appels utilisable pour explorer la satisfaction (de la satisfaction à la frustration) dans de conversations téléphoniques réelles. Ce corpus contient 303 conversations pour un total de plus de 37 heures d'enregistrement et peut être obtenu en contactant les auteurs. L'objectif principal de ces travaux était de prédire la satisfaction tout au long d'une conversation. Cette prédiction ne pouvait être effectuée que si les annotations de ce nouveau corpus était cohérente, ce que nous avons vérifié. Les premières expériences montrent que les réseaux neuronaux biLSTM sont capables de prédire les valeurs de la satisfaction et donc de retracer cette dimension au cours d'un appel avec un score CCC correcte, comparable à celui calculé sur les prédictions de valence du corpus SEWA.

Par la suite, des investigations plus approfondies seront menées pour améliorer cette prédiction. Nous voulons notamment ajouter la modalité linguistique à nos descripteurs d'entrée. Nous prévoyons également d'aller plus loin dans nos expériences sur l'annotation continue et discrète en utilisant d'autres protocoles de classification (modèles, sets, niveaux de segmentation) tout en ajoutant des informations sémantiques supplémentaires.

Références

CAMPBELL N. (2008). *Expressive/Affective Speech Synthesis*, In *Springer Handbook of Speech Processing*, p. 505–518. Springer Berlin Heidelberg.

COWIE R., DOUGLAS-COWIE E., SAVVIDOU S., MCMAHON E. & AL. (2000). FEELTRACE : An instrument for recording perceived emotion in real time. In *ITRW on Speech and Emotion*, p. 19–24.

DEVILLERS L., VAUDABLE C. & CHASATGNOL C. (2010). Real-life emotion-related states detection in call centers : a cross-corpora study. In *Proc. of Interspeech*, p. 2350–2355.

EKMAN P. (1999). *Basic Emotions*, In *Handbook of Cognition and Emotion*, p. 301–320. Wiley, New-York.

EYBEN F., SCHERER K., SCHULLER B., SUNDBERG J. & AL. (2016). The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing*, **7**(2), 190–202.

EYBEN F., WÖLLMER M. & SCHULLER B. (2010). OpenSMILE – the munich versatile and fast open-source audio feature extractor. In *Proc. of the ACM Multimedia 2010 International Conference*, p. 1459–1462.

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. & AL. (2019). SEWA DB : A rich database for audio-visual emotion and sentiment research in the wild. *IEEE transactions on pattern analysis and machine intelligence (Early Access)*.

LAILLER C., LANDEAU A., BÉCHET F., ESTÈVE Y. & AL. (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)*, p. 1047–1050.

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. & AL. (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. & SYSSAU A. (2014). Affective norms for French words (FAN). *Behavior research methods*, **46 4**, 1128–1137.

MORRISON D., WANG R. & DE SILVA L. (2007). Ensemble methods for spoken emotion recognition in call-centres. *Speech Communication*, **49**(2), 98–112.

PLUTCHIK R. (1980). Chapter 1 - a general psychoevolutionary theory of emotion. In *Theories of Emotion*, p. 3 – 33. Academic Press.

POVEY D., GHOSHAL A., BOULIANNE G., BURGET L. & AL. (2011). The kaldi speech recognition toolkit. In *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*.

RINGEVAL F., SCHULLER B., VALSTAR M., COWIE R. & AL. (2018). AVEC 2018 workshop and challenge : Bipolar disorder and cross-cultural affect recognition. In *Proc. of the 2018 on Audio/Visual Emotion Challenge and Workshop*, p. 3–13.

SCHERER K. R. (2005). What are emotions ? and how can they be measured ? *Social science information*, **44**(4), 695–729.

SCHMITT M., CUMMINS N. & SCHULLER B. W. (2019). Continuous emotion recognition in speech - do we need recurrence ? In *Proc. Interspeech 2019*, p. 2808–2812.

SCHULLER B. & DEVILLERS L. (2010). Incremental acoustic valence recognition : An inter-corpus perspective on features, matching, and performance in a gating paradigm. In *Proc. Interspeech 2010*, p. 801–804.

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