



HAL
open science

Analyse de données massives en temps réel pour l'extraction d'informations sémantiques et émotionnelles de la parole

Manon Macary

► **To cite this version:**

Manon Macary. Analyse de données massives en temps réel pour l'extraction d'informations sémantiques et émotionnelles de la parole. Informatique et langage [cs.CL]. Le Mans Université, 2022. Français. NNT : 2022LEMA1014 . tel-03869572

HAL Id: tel-03869572

<https://theses.hal.science/tel-03869572v1>

Submitted on 24 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.

THÈSE DE DOCTORAT DE

LE MANS UNIVERSITÉ

ÉCOLE DOCTORALE N° 601
*Mathématiques et Sciences et Technologies
de l'Information et de la Communication*
Spécialité : *Informatique*

Par

Manon MACARY

Analyse de données massives en temps réel pour l'extraction d'informations sémantiques et émotionnelles de la parole

Thèse présentée et soutenue à LE MANS, le 24 juin 2022

Unité de recherche : LIUM

Thèse N° : 2022LEMA1014

Rapporteurs avant soutenance :

Martine ADDA-DECKER Directrice de recherche CNRS au LPP/CNRS-Sorbonne Nouvelle
Denis JOUVET Directeur de recherche INRIA à INRIA-LORIA/Université de Lorraine

Composition du Jury :

	Prénom NOM	Fonction et établissement d'exercice (<i>à préciser après la soutenance</i>)
Président :		
Examineurs :	Martine ADDA-DECKER	Directrice de recherche CNRS au LPP/CNRS-Sorbonne Nouvelle
	Denis JOUVET	Directeur de recherche INRIA à INRIA-LORIA/Université de Lorraine
	Fabien RINGEVAL	Maitre de Conférence au LIG/Université Grenoble Alpes, CNRS
	Marie TAHON	Maitre de Conférence au LIUM/Le Mans Université
	Damien LOLIVE	Directeur du département informatique à l'ENSSAT/Université de Rennes
Dir. de thèse :	Yannick ESTÈVE	Directeur du LIA au LIA/Université d'Avignon et des Pays de Vaucluse

Invité(s) :

Merouane Atig Directeur Technique à Allo-Media

TABLE DES MATIÈRES

Introduction	19
I État de l'art	24
1 Les émotions dans la parole	25
1.1 Définition de l'émotion	25
1.1.1 Point historique sur les théories de l'émotion	26
1.2 Représentation courante des émotions à but d'automatisation de traitement	31
1.2.1 Théorie des émotions discrètes	32
1.2.2 Théorie des émotions continues	35
1.3 L'émotion dans la parole	38
1.4 L'émotion dans le texte	40
1.5 Les autres marqueurs de l'émotion chez l'humain	41
1.5.1 Les marqueurs faciaux et comportementaux de l'émotion	41
1.5.2 Les marqueurs physiologiques de l'émotion	43
1.6 Conclusion	44
2 Apprentissage automatique pour le traitement de la parole	45
2.1 Apprentissage automatique : définition	45
2.1.1 Classification et régression	46
2.1.2 Les pré-requis pour un Apprentissage Automatique réussi	47
2.1.3 Apprentissage supervisé et non supervisé	47
2.2 Quelques familles d'apprentissage automatique	49
2.2.1 k-moyennes et k-plus proches voisins	50
2.2.2 Régression Linéaire	51
2.2.3 Machine à vecteurs de support	52
2.2.4 Modèle de Markov caché et prédiction de séquence	53
2.2.5 Modèle de mélange gaussien	53
2.2.6 Réseau de neurones	54

TABLE DES MATIÈRES

2.3	Réseau de neurones profonds	55
2.3.1	Du perceptron au multicouche	55
2.3.2	Algorithme d'apprentissage	58
2.3.3	L'initialisation et ses enjeux	59
2.4	Les architectures des réseaux dans le traitement du signal	61
2.4.1	Réseaux neuronaux convolutifs	62
2.4.2	Réseaux Neuronaux Récurrents	63
2.4.3	Réseaux long-short term memory	64
2.4.4	Encodeur Décodeur	65
2.4.5	Transformers	67
2.5	Modèles pré-entraînés	68
2.5.1	Représentation linguistique	68
2.5.2	Représentation acoustique	70
2.6	Conclusion	71
3	Reconnaissance automatique des émotions : corpus et méthodes	73
3.1	Le domaine de la reconnaissance automatique des émotions	73
3.2	Les corpus existants	73
3.2.1	Les différents types de corpus	74
3.2.2	Les différents types d'acquisition	75
3.2.3	Les différents types d'annotation	76
3.2.4	Synthèse	76
3.3	Les descripteurs	78
3.3.1	Descripteurs acoustiques	78
3.3.2	Descripteurs linguistiques	81
3.4	Évaluation des performances	86
3.4.1	Tâche de classification	86
3.4.2	Tâche de régression	88
3.5	Fusion de modalités	90
3.6	Notre référence : AVEC	92
3.7	Conclusion	93

II	Contributions	94
4	AlloSat un corpus pour la reconnaissance continue d'émotions	95
4.1	Motivation	95
4.2	Recueil des données	97
4.2.1	Sélection des données	97
4.2.2	Pre-traitement des données	99
4.3	Mise en place de l'annotation	103
4.3.1	Volonté d'annotation	103
4.3.2	Logiciel utilisé	105
4.3.3	Consignes	106
4.4	Analyse d'AlloSat	108
4.4.1	Accord intra-annotateur	109
4.4.2	Accord inter-annotateur	112
4.4.3	Calcul du Coefficient de Corrélacion de Concordanca entre annota- teurs	114
4.4.4	Étude empirique sur le corpus	116
4.5	Modalités de diffusion du corpus	116
4.6	Conclusion	118
5	Reconnaissance continue d'émotion à partir de représentations acous- tiques	119
5.1	Motivation	119
5.2	Construction d'une alerte sur les prédictions discrètes finales	120
5.2.1	Les descripteurs	121
5.2.2	Modèles et protocole d'apprentissage	121
5.2.3	Résultats et analyse	122
5.3	Reconnaissance continue de la dimension de satisfaction/frustration	123
5.3.1	Exploration des ensembles de descripteurs eGeMAPS	124
5.3.2	Comparaison eGeMAPS et MFCC	128
5.3.3	Comparaison CNN et biLSTM	129
5.4	Comparaison entre AlloSat et SEWA	132
5.4.1	Performances obtenues sur les deux corpus	132
5.4.2	Analyse des différences entre SEWA et AlloSat	134
5.5	Analyses annexes	137

TABLE DES MATIÈRES

5.5.1	Fonction de coût	137
5.5.2	Post-traitement : lissage des prédictions	138
5.6	Conclusion	139
6	Reconnaissance continue d'émotion à partir de représentations acoustiques et linguistiques pré-entraînées	141
6.1	Motivation	141
6.2	Représentation linguistique	142
6.2.1	Transcriptions automatiques	142
6.2.2	Synchroniser le linguistique et l'acoustique	143
6.2.3	Exploration des word2vec	144
6.3	Fusion des modalités acoustiques et linguistiques	146
6.4	Descripteurs pré-entraînés	149
6.4.1	Représentation linguistique	150
6.4.2	Représentation acoustique	151
6.4.3	Reproduction sur SEWA	152
6.5	Fusion des modalités acoustique et linguistique pré-entraînées	153
6.6	Conclusion	154
7	Analyse des annotations, explicabilité des modèles autour de la frustration	157
7.1	Motivations	157
7.2	Analyse de l'impact de chaque annotateur	158
7.2.1	Annotation moyenne ou 3 annotations?	158
7.2.2	Un modèle de reconnaissance par annotateur	159
7.3	Expliquer la frustration dans les conversations	162
7.3.1	Première écoute humaine : que retire-t-on de l'acoustique?	162
7.3.2	Études statistiques du contenu linguistique des conversations frustrées	163
7.3.3	Analyses conduites par un linguiste	166
7.4	Considération du genre	170
7.5	Conclusion	172
	Conclusion	173

8 Annexes	179
8.1 Guide d'installation et de configuration de CARMA	179
8.2 Guide d'annotations	183
8.3 Transcription d'une conversation	191
8.4 End User Licence Agreement	196
8.5 Intervalles de confiance statistiques des scores CCC	199
Bibliographie	203

TABLE DES FIGURES

1.1	Frise chronologique des grandes théories des émotions, distribuées selon leurs auteurs à la date de leur première date de parution.	26
1.2	Réactions physiologiques du corps devant une situation dangereuse. Image tirée du site www.comprendrelapeur.e-monsite.com	27
1.3	Traduction de la figure extraite de l'article de Paul Ekman et al. [ELF83] décrivant un arbre de décision pour déterminer l'émotion. On voit que la colère peut être détectée par un rythme cardiaque élevé et une température cutanée élevée.	34
1.4	Roue de Plutchik qui définit les émotions complexes à partir d'émotions basiques [R80].	35
1.5	Le modèle circumplex de Russell [Rus80].	36
1.6	La roue des émotions de Genève définie par Scherer [Sch05], qui nomme des émotions (primaires ou secondaires) dans des dimensions continues. On voit par exemple qu'une émotion de valence négative et d'activation passive peut être décrite comme de la fatigue.	37
1.7	Marqueurs faciaux des 6 émotions primaires d'Ekman. In Lie To Me. . . .	42
2.1	Représentation graphique d'un système d'apprentissage automatique avec ses entrées et ses sorties.	46
2.2	Les différents types d'apprentissage et des exemples d'utilisation. Tiré de https://www.coe.int/fr/web/artificial-intelligence/glossary et modifié. . . .	48
2.3	Représentation graphique d'une classification en trois classes par un algorithme k-moyenne. Ici k=3.	50
2.4	Représentation graphique d'un modèle de Markov caché à quatre états. X représente les états de l'automate, a représente les transitions entre les états. . . .	53
2.5	Représentation graphique d'un modèle de mélange de trois gaussiennes. . . .	54
2.6	Représentation schématique d'un neurone biologique. Illustration issue de https://www.schoolmouv.fr/definitions/neurones/definition	55

2.7	Représentation schématique d'un perceptron. x_i correspond aux entrées, les W_i correspondent aux poids associés à chacune des entrées x_i , ψ formalise le biais. Une fois sommée, la valeur est passée dans une fonction d'activation ϕ pour déterminer la valeur de sortie y . Dans le schéma la fonction d'activation correspond à un simple signal échelon.	56
2.8	Représentation schématique d'un perceptron multi-couche. Les neurones bleu correspondent à la couche d'entrée, les neurones oranges définissent plusieurs couches cachées et les neurones verts correspondent à la couche de sortie.	57
2.9	Représentation graphique de trois fonctions d'activation : la fonction non linéaire, sigmoïde et tanh. Illustration provenant de [Tho14].	57
2.10	Représentation schématique du traitement d'une image par un réseau de type convolutionnel. Image provenant de Wikipedia.	62
2.11	Représentation schématique d'un réseau neuronal récurrent. La partie gauche correspond à une récurrence sur une couche entière h . La partie droite explicite la récurrence au niveau de la couche h . Image provenant de Wikipedia.	63
2.12	Représentation schématique d'un réseau récurrent à mémoire court et long terme (LSTM). On observe qu'il y a deux entrées à l'unité neuronale : l'état de la cellule (cell state) en haut et la sortie classique d'un neurone en bas. Image provenant de http://colah.github.io/posts/2015-08-Understanding-LSTMs/	64
2.13	Représentation schématique d'un système encodeur décodeur. x_i correspondent aux entrées, y_i aux sorties. Image provenant de https://towardsdatascience.com/understanding-encoder-decoder-sequence-to-sequence-model-679e04af4346	66
2.14	Représentation graphique des deux étapes permettant d'obtenir le modèle BERT. Issue des travaux de Devlin et al. [Dev+19]	69
2.15	Wav2vec : Schéma du pré-entraînement à partir des données audio X qui sont encodées avec deux réseaux de neurones empilés. Le réseau <i>encodeur</i> donne la représentation Z et le réseau de <i>contexte</i> donne la représentation C . Issu de [Sch+19]	71
3.1	Transformation d'un corpus en représentation one-hot.	83

3.2	Représentation de plongements de mots en deux dimensions. Les mots de gauche correspondent au regroupement des termes associés au numérique. Les mots de droite correspondent aux termes associés à l'emploi. Issue des travaux de Turian et al. [TRB10].	85
4.1	L'axe de satisfaction va de la frustration à la satisfaction, en passant par le neutre. C'est donc sur cet axe que l'annotation a été effectuée.	96
4.2	Diagramme représentant la chaîne de traitement pour la sélection des conversations présentes dans le corpus AlloSat.	99
4.3	La répartition des conversations en fonction de leur durée. En bleu, nous avons les anciennes durées, en orange les nouvelles. Environ 76,6% des conversations duraient moins de 15 minutes (soit 900 secondes). Après réduction des silences, ce nombre augmente à environ 93,7%.	101
4.4	Exemple d'obfuscation d'un segment fictif. Les mots sont d'abord retrouvés dans la transcription, puis sont substitués par leur catégorie. Enfin le signal audio est modifié pour remplacer les informations personnelles par un son pré-enregistré.	103
4.5	Schéma des variations qui ont été données aux annotateurs afin de mieux comprendre la catégorie évolution	104
4.6	Capture d'écran de l'outil CARMA [Gir14]	105
4.7	Le Self Assessment Manikin (SAM). La ligne a correspond à l'activation, la ligne b correspond à la valence. Issu de [BL94]	107
4.8	Exemple d'annotation d'une conversation selon l'axe de satisfaction. L'annotation de référence correspond à la courbe en pointillée	114
5.1	Schéma de la configuration des systèmes neuronaux appelés biLSTM-2 (à gauche) et biLSTM-4 (à droite). Le nombre de neurones est indiqué en rouge et entre parenthèses sur chaque couche. Comme il s'agit de réseaux bidirectionnels, il faut multiplier par deux le nombre de neurones pour avoir le nombre de paramètres réels.	126
5.2	Prédiction de la satisfaction sur des conversations issues du test. La référence est en rouge, la prédiction en bleu.	127
5.3	Schéma de la configuration du système neuronal appelé CNN-4. Le nombre de neurones est en rouge et le filtre récepteur en bleu.	130

5.4 Evolution des prédictions (grises) et des références (rouge) de deux conversations provenant de l'ensemble de test d'AlloSat. $ccc(A) = 0.564$, $ccc(B) = 0.903$. La courbe bleue correspond au lissage des prédictions. Obtenu avec le système biLSTM-4 (l-rmse) et Mfcc-lib. 138

6.1 Représentation de l'apprentissage des word2vec selon deux algorithmes : le *CBOW* (gauche) et le *skip-gram* (droite). Ici on utilise un contexte de 5 mots, soit le mot cible entouré en rouge et deux mots de chaque côté, soulignés en rouge. 145

6.2 Représentation des quatre fusions utilisées pour la reconnaissance des émotions à partir des modalités acoustiques et linguistique. Les nombres entre parenthèses correspondent au nombre de neurones dans chaque couche. . . 147

7.1 Exemple d'annotation d'une conversation selon l'axe de satisfaction déjà présenté chapitre 4. L'annotation de référence correspond à la courbe en pointillée 159

7.2 Exemple de segment considéré comme une pente de frustration. Le rectangle bleu correspond au segment de pente de frustration. 164

7.3 Nuages de mots construits à partir des occurrences de bi-mots (gauche) et tri-mots (droite) extraites des transcriptions des segments de pentes de frustration. 166

7.4 Visualisation des données de segments émotionnels projetés sur les deux premières composantes obtenues par une ACP, sur laquelle une classification de type kmeans (k=3) est effectuée. 168

7.5 Analyse dynamique de la frustration de la conversation appelée *lettre certifiée*. Le nombre d'occurrences des sept caractéristiques linguistiques est tracé par rapport au temps. La référence de l'axe de satisfaction est représentée par la ligne en pointillée rouge. 170

7.6 Visualisation du pourcentage de conversations contenant des pentes de frustration en fonction du genre de la paire d'interlocuteurs. Le couple homme-homme n'apparaît pas puisqu'il n'y a pas de conversations avec une pente de frustration pour cette paire d'interlocuteurs dans les 81 conversations considérées. 171

LISTE DES TABLEAUX

1.1	Grille d'évaluation des premiers travaux de Scherer pour définir une émotion en fonction des cinq questionnements. La catégorie ouverte est utilisée lorsque l'évaluation peut être de différente catégorie. Par exemple, pour la joie, l'élément déclencheur peut être quelque chose de connu ou non (Familiarité). En rapport aux buts/besoins, les opportunités peuvent être obstruées (émotions négatives) ou facilitées (émotions positives). Tiré des travaux de P. Philippot [Phi07].	30
1.2	Définition des émotions basiques selon différents auteurs.	33
3.1	Principaux corpus utilisés dans la reconnaissance d'émotions dans la parole. Chaque corpus est caractérisé par la langue utilisée, si les enregistrements sont issus du domaine téléphonique ou non, s'il s'agit d'un corpus acté ou spontanée et si les émotions sont annotées en continue ou non.	77
3.2	Résumé des descripteurs de bas niveau (LLDs) utilisés dans les ensembles GeMAPS et eGeMAPS. * signale les descripteurs uniquement disponibles dans eGeMAPS. Légende des fonctions appliquées : M moyenne arithmétique, CV coefficient de variation, qui correspond à la variance normalisée par la moyenne, V variance, P percentiles : 20,50 et 80%, RP l'intervalle des percentiles 20 à 80%, SLOPE moyenne et variance des pente de montée/descente de signal, MUN moyenne arithmétique avec les parties sans parole, MCVV moyenne arithmétique et coefficient de variation des parties avec paroles.	82
3.3	Matrice de Confusion entre trois classes émotionnelles : la joie, le neutre et la colère. Les colonnes correspondent aux prédictions du système et les lignes correspondent aux références. On voit que sur 100 prédictions de la classe joie, seules 90 sont pertinentes et le système a mal prédit 13 segments qui ne devraient pas être dans la classe joie.	87

3.4 Compilation des scores de CCC sur l'ensemble de développement de SEWA sur les 3 dimensions : activation, valence et *liking*. L'acronyme BoAW signifie *Bag-of-audio-words*, BoTW signifie *Bag-of-text-words*, SVR signifie Support Vector Regression [SS04], proche des SVM mais applicable à des problèmes de régression et donc à une annotation continue. Les différents nombres associés aux features de type eGeMAPS dénotent de différentes configurations utilisées autour de ces sets : soit avec une sélection réduite des LLDs (47), soit avec l'ajout d'information sur le locuteur courant (89 et 176). 91

4.1 Arbre des différentes catégories de données personnelles anonymisées 102

4.2 Récapitulatif du schéma d'annotation discrète. 104

4.3 Découpage du corpus en ensemble de train, développement et test. Les durées sont données en heures, minutes, secondes. 108

4.4 Ensemble des annotations discrètes des trois annotateurs a1, a2 et a3. Vote majo correspond au vote majoritaire qui a conduit à l'annotation discrète de référence. 110

4.5 Pourcentage de désaccords sur les 303 conversations et kappa pour définir l'accord intra-annotateur pour chacun des trois annotateurs ainsi que la moyenne. a_i représente l'annotateur i 112

4.6 Accord inter-annotateur calculé entre les pairs d'annotateurs ainsi que la moyenne. a_i représente l'annotateur i . R représente le coefficient de corrélation, k représente le kappa de début et de fin de conversation. 113

4.7 Score CCC calculé entre les annotations des annotateurs a_i et l'annotation de référence. 115

4.8 40 scores calculés entre l'annotation de référence et l'annotation de l'annotateur a_i 117

4.9 Répartition de la durée des conversations en fonction des sets considérés . . 118

5.1 Description des trois ensembles de descripteurs utilisés pour réaliser la reconnaissance des émotions discrètes des fins de conversations. 121

5.2 Scores des systèmes de classification sur l'émotion finale de la conversation. UAP correspond à *unweighted average precision* et UAR à *unweighted average recall*. 122

5.3	Score CCC des systèmes de reconnaissance des émotions en utilisant quatre ensembles de descripteurs différents et deux architectures neuronales sur les ensembles de développement et de test d'AlloSat.	126
5.4	Comparaison de performance entre descripteurs experts et MFCCs. Score CCC rapporté sur les différents ensembles de descripteurs sur le dev et le test d'AlloSat. Le modèle utilisé est le biLSTM-4.	129
5.5	Comparaison des moyennes (maximum) des scores CCC de cinq systèmes différents sur les ensembles de développement d'AlloSat et de SEWA. . . .	131
5.6	Comparaison des scores moyens de CCC sur les corpus AlloSat et SEWA selon quatre dimensions émotionnelles : la satisfaction, l'activation, la valence et le liking. Nos scores correspondent à la moyenne des scores de 5 systèmes appris avec des initialisations aléatoires différentes et entre parenthèses nous retrouvons le score du meilleur des modèles. Nous reportons également les résultats inclus dans les papiers [Rin+18 ; Rin+19 ; SCS19], qui constituent notre base de comparaison. Ces scores correspondent au meilleur de leur expérimentation, c'est pour cela qu'ils sont notés entre parenthèses, puisqu'à comparer avec les scores des meilleurs modèles notés entre parenthèses dans nos résultats. *L'entraînement et les prédictions sont réalisés sur les conversations allemandes et hongroises. Sans le sigle, l'entraînement et les prédictions sont réalisés sur les conversations allemandes uniquement. Les couleurs permettent de repérer les différentes expérimentations comparables.	133
5.7	Architectures entraînées sur l'ensemble de développement d'AlloSat et de SEWA en prenant en entrée des segments émotionnels de 500 ms. Comparaison entre deux architectures neuronales : CNN et biLSTM-4 avec eGeMAPS-47.	135
5.8	Comparaison des scores de développement sur les annotations originales et lissées sur les corpus AlloSat et SEWA. Les descripteurs utilisés sont les eGeMAPS-47 features et le système utilisé est le CNN.	135
5.9	Comparaison des scores de développement entre les versions originales du corpus SEWA et les versions dégradées. Les dégradations sont le sous-échantillonnage des enregistrement audio et l'ajout de différents bruits. Nous utilisons les descripteurs eGeMAPS-47 ainsi qu'un modèle CNN. L'entraînement et les prédictions sont effectuées sur les conversations allemandes. .	136

5.10	Comparaison de l'utilisation de deux fonctions de coût $l\text{-ccc}$ et $l\text{-rmse}$. Score CCC des systèmes de reconnaissance des émotions sur l'ensemble de développement d'AlloSat. Le modèle utilisé est un biLSTM-4.	138
5.11	Scores CCC avec et sans lissage calculés sur les ensembles de développement et de test d'AlloSat. Score issu du meilleur modèle : système biLSTM-4, descripteurs Mfcc-lib et fonction de coût RMSE.	139
6.1	Exemple de la transcription d'un début de conversation au format ctm mise sous la forme d'un tableau pour expliquer les différentes colonnes. Conversation issue du corpus AlloSat.	143
6.2	Scores CCC des systèmes de reconnaissance des émotions d'une architecture neuronale bilstm à quatre couches en fonction des différents descripteurs d'entrée linguistiques.	146
6.3	Comparaison des scores CCC des systèmes de reconnaissance des émotions du biLSTM-4 en utilisant différents protocoles de fusion et différents descripteurs pour la modalité linguistique.	148
6.4	Scores CCC des systèmes de reconnaissance des émotions du biLSTM-4 en prenant en entrée des descripteurs linguistiques pré-entraînés.	151
6.5	Scores CCC des systèmes de reconnaissance des émotions du biLSTM-4 en prenant en entrée des descripteurs acoustiques pré-entraînés.	152
6.6	Scores CCC des systèmes de reconnaissance des émotions du biLSTM-4 en prenant en entrée des descripteurs acoustiques pré-entraînés sur l'ensemble de dev du corpus SEWA.	153
6.7	Comparaison des score CCC des systèmes de reconnaissance des émotions du biLSTM-4 en utilisant la fusion de décision avec des descripteurs pré-entraînés. Nous rappelons les scores obtenus précédemment avec la fusion des descripteurs baseline.	154
7.1	Résultats des systèmes de fusions pour chaque annotateur. Les modèles sont entraînés et évalués sur des annotations individuelles. AVG : moyenne sur les trois annotateurs. CV : coefficient de variation sur les trois annotateurs. Diff1 correspond à la différence relative entre CamemBERT et Wav2Vec.	160

7.2	Résultats des systèmes de fusions pour chaque annotateur. Les modèles sont entraînés sur des annotations individuelles et évalués sur les annotations moyennés de référence. AVG : moyenne sur les trois annotateurs. CV : coefficient de variation sur les trois annotateurs. Diff1 correspond à la différence absolue entre CamemBERT et Wav2Vec. La meilleure fusion est choisie sur le Dev.	161
7.3	Statistiques sur la présence d'événements retrouvés dans les 57 conversations écoutées.	163
7.4	Exemple de transcription de segments considérés comme pente de frustration.	165
7.5	Classification bi-mots des segments émotionnels correspondant aux pentes de frustration sur laquelle nous avons appliqué un TF-IDF.	167
7.6	Sept caractéristiques et leur nombre d'occurrences permettent de modéliser les indices supposés être responsables de la frustration dans les conversations. Le nombre total de mots et de segments de parole de la conversation appelée <i>lettre recommandée</i> , sont également indiqués.	169
7.7	Extrait (137 - 166 sec.) de la conversation <i>lettre recommandée</i> . Disfluences : <i>italic</i> ; Hesitations, bégaiements : <u>underline</u> ; Traces Semantiques de la frustration : bold ; auto-coupures : //	170
7.8	Présence de pentes de frustration en fonction du genre de la paire d'interlocuteur.	171
8.1	Intervalles de confiance pour les scores CCC obtenus sur le sous-ensemble de Dev. Ils sont calculés sur différentes prédictions d'expérimentations sur l'initialisation des poids du réseau de neurones.	200

INTRODUCTION

Contexte

Nous apprenons dès le plus jeune âge que la communication est l'outil par excellence pour vivre en société. Cette communication, pourtant si étudiée par de nombreux chercheurs, reste aujourd'hui pleine de mystère. Le nombre de domaines de recherche qui gravitent autour de la communication en est le marqueur plus qu'évident. Quand nous pensons à la communication, nous pensons tout d'abord à l'oral, à l'écrit, aux mots et aux phrases. Puis si nous creusons un peu, nous pensons aux gestes, aux visages, aux us et coutumes de chaque société. Mais la communication ne se réduit pas à ces quelques canaux. L'émotion est une part importante de tout être, elle nous définit à un instant t en tant qu'individu. Il est donc capital qu'elle puisse être comprise par autrui, qu'elle puisse être communiquée. Nous sommes tous capables de ressentir des émotions qui impactent nos moyens de communiquer et de ce fait qui peuvent être reconnues et interprétées par nos interlocuteurs.

La parole est un canal de communication où il est possible d'exprimer de nombreux messages. Il s'agit d'un de nos canaux principaux pour vivre en société, puisque c'est par elle que nous nous exprimons le plus et le plus clairement. La parole véhicule aussi les émotions qui nous traversent. Nous sommes capables de reconnaître la détresse d'un individu à son ton de voix et donc de nous montrer empathiques quand l'autre en a besoin. Cette reconnaissance des émotions d'autrui nous permet alors de mieux adapter notre discours, de mieux réagir et donc d'agir en tant que membre de la société.

Il est donc tout aussi important pour des industriels de pouvoir reconnaître l'état émotionnel de ses interlocuteurs afin d'adapter leur discours. En effet, toute entreprise privée a pour objectif de gagner de l'argent. Pour y parvenir, elle doit acquérir et conserver des clients qui vont lui acheter son ou ses produits. Il est donc primordial dans ce contexte d'avoir une connaissance, même sommaire, des émotions de ses clients. Mon sujet de thèse s'inscrivant dans un besoin industriel bien particulier, nous avons réfléchi aux émotions qui avaient une grande importance dans une relation client-entreprise.

Travaillant main dans la main avec des centres d'appels, il nous est paru évident que la plupart des clients n'appelle pas pour signifier leur joie, leur peine ou même leur

contentement. Non, la plupart d'entre nous, quand nous téléphonons à ces numéros généralement payants, c'est parce que nous avons besoin d'information ou alors parce que quelque chose ne va pas. Dans ce second cas, il nous semble que la frustration correspond à l'état émotionnel sur laquelle nous devons axer nos recherches. De même, il nous semblait important de ne pas recueillir uniquement le négatif, nous voulions aussi reconnaître les clients contents d'être entendus. C'est pour cela que nous avons également considéré la satisfaction. Ces deux émotions, que nous considérons comme deux facettes d'une même pièce, sont importantes dans la relation clientèle et elles sont étroitement liées à la perte ou au gain d'un client.

C'est dans ce contexte que nous avons voulu inscrire nos travaux : la reconnaissance des états émotionnels, notamment la satisfaction et la frustration, dans des échanges entre des clients et des représentants de diverses entreprises, c'est-à-dire des conseillers de centre d'appels.

Dans ce cas, pourquoi vouloir automatiser la détection de ces émotions alors que l'humain est tout à fait capable de les détecter ? Aujourd'hui les centres d'appels sont d'immenses plateformes et les conseillers qui travaillent dans ce secteur se comptent en milliers. Quant aux appels, ils se comptent en centaines par heure. Ces immenses quantités d'appels ne peuvent pas être traitées par l'humain sans devoir employer une quantité faramineuse de personnes. De même, ces conseillers ont beaucoup de travail, il ne serait pas pertinent de rajouter une charge d'annotation de l'émotion en plus. Surtout que ce n'est ni une tâche facile et ni une tâche sans ambiguïté. Même avec une formation adéquate, il est peu probable que tous les annotateurs se rangent à la même annotation.

D'aussi gros volumes de données nous font tout de suite penser au machine learning. Ces outils, dont le deep learning fait partie, sont spécialement équipés pour faire face à un tel problème. Depuis maintenant quelques années, nous avons appris à apprendre aux ordinateurs comment faire certaines de nos tâches. Dans le domaine de la parole, on pense notamment à l'ASR (Automatic Speech Recognition) qui utilise des systèmes intelligents pour traduire un signal audio en transcription textuelle. Dans notre étude, nous allons nous inscrire dans le domaine du SER (Speech Emotion Recognition), afin de traduire un signal audio en états émotionnels.

Cette traduction ainsi effectuée, les entreprises peuvent alors détecter les états émotionnels de leurs clients et ajuster leur discours. En extrapolant, on peut voir des campagnes promotionnelles de réduction envoyées aux clients frustrés afin de les satisfaire par exemple.

Qu'est ce qui est novateur dans nos travaux ? Tout d'abord nous travaillons sur deux états émotionnels peu dotés, surtout si on considère les données en langue française. Nous démontrons que nous sommes capables de les détecter automatiquement à l'aide de systèmes construits avec des réseaux de neurones. Nous étudions les meilleures représentations de la parole afin d'extraire la frustration et la satisfaction. Enfin, nous tentons d'apporter une justification au succès de ces systèmes.

Problématiques

La première problématique consiste à s'interroger sur le contenu émotionnel que l'on peut observer dans les conversations de centres d'appels et qui seront utiles aux industriels. En effet, peut-on appliquer les modèles d'émotions classiquement utilisées en affective computing ? [Chapitre 4] Une fois les émotions recherchées définies, quelles stratégies peut-on mettre en œuvre pour les annoter, évaluer la pertinence de ces annotations réalisées par plusieurs individus qui ont des perceptions émotionnelles différentes ? [Chapitre 4] Nous nous sommes également demandés dans quelle mesure il est possible de considérer chaque annotation individuelle comme une référence pour apprendre nos modèles. [Chapitre 7]

La seconde problématique consiste à proposer un système efficace pour la reconnaissance des émotions à partir d'un signal de parole. Nous avons questionné différentes architectures neuronales, ainsi que différentes représentations du signal acoustique. Pour cela, nous nous sommes interrogés sur la proposition d'un protocole expérimental complet depuis l'apprentissage des modèles à leur évaluation. [Chapitre 5]

Les corpus émotionnels contiennent généralement une faible quantité de données ce qui implique des modèles relativement légers et peu complexes en comparaison de ceux développés pour la reconnaissance automatique de parole. Nous nous sommes donc interrogés sur les stratégies permettant de compenser le manque de données d'apprentissage, notamment sur l'utilisation de représentations pré-entraînées avec des méthodes auto-supervisées. [Chapitre 6]

La question du choix des modalités pertinentes pour représenter les émotions dans la parole forme notre dernière problématique. En effet, la parole peut être représentée par son contenu acoustique, mais également son contenu linguistique. Nous avons donc questionné les possibilités de fusion de ces deux modalités. [Chapitre 6]

Et donc si le contenu linguistique est pertinent pour les conversations téléphoniques, nous pouvons nous interroger sur les marqueurs de satisfaction et de frustration présents dans le texte. [Chapitre 7]

Organisation du document

Ce document est divisé en deux parties. La première partie correspond à une présentation plus détaillée du contexte de cette thèse par un état de l'art des domaines proches de notre travail. La deuxième partie du document recense les contributions que nous avons apportées au sein de cette thèse.

Le premier chapitre traite de l'état de l'art des émotions. Celui-ci est à la croisée de nombreux domaines de recherche, puisque les émotions sont définies par un ensemble de concepts qui continuent d'évoluer. Nous commençons par une définition de l'émotion en nous basant sur les théories en psychologie qui ont eu pour but de décrire et de comprendre ce phénomène. Ensuite nous traitons de la représentation de ces émotions que nous utilisons en tant qu'informaticien afin d'en automatiser le traitement. Nous discutons alors des marqueurs des émotions, que ce soit dans la parole, dans le texte ou dans les traits du visage. Ce chapitre nous permet de nous placer dans un contexte clair en ce qui concerne les émotions et la définition à laquelle nous nous confortons.

Le deuxième chapitre traite de l'état de l'art de l'apprentissage automatique. Comme nous travaillons sur la parole, nous avons surtout insisté sur les ensembles de méthodologies utilisées pour traiter la parole. Après avoir défini l'apprentissage automatique, nous passons en revue quelques grandes familles d'algorithmes qui permettent cet apprentissage. Nous nous intéressons particulièrement aux réseaux de neurones, à ses spécificités et à quelques architectures établies dans le domaine et très usitées. Enfin nous discutons des systèmes permettant de mettre en place des modèles pré-appris afin de mieux décrire les données contenues dans la parole.

Notre troisième et dernier chapitre de la première partie se concentre sur notre domaine d'étude, à savoir le Speech Emotion Recognition. Après avoir défini les contours de ce domaine, nous présentons les différents corpus utilisés dans ce domaine, tout en définissant les caractéristiques communes à ces corpus. Ensuite nous nous intéressons aux descripteurs qui correspondent à la transformation du signal audio en un ensemble de données qui sont utilisables et compréhensibles par les systèmes de reconnaissance. Pour finir, nous parlons de la fusion des modalités acoustiques et linguistiques, et nous passons en revue une partie des campagnes AVEC (Audio-Visual Emotion Challenge), qui sert de point de comparaison aux travaux de cette thèse.

Le quatrième chapitre, qui ouvre la seconde partie, se concentre sur la construction du corpus AlloSat permettant l'étude de la satisfaction et de la frustration dans des conversations entre des clients et des conseillers de centres d'appels. Nous commençons

par relater les techniques de recueil de données et notre mise en place de l'annotation. Une fois la création de ce corpus effectuée, nous avons fait des analyses afin de jauger sa pertinence et son utilisabilité. Enfin nous rapportons les modalités de sa diffusion.

Dans le cinquième chapitre, nous détaillons la mise en place des premiers systèmes de reconnaissances de l'émotion. Nous commençons par détailler les protocoles de reconnaissance que nous avons mis en place que ce soit pour les émotions discrètes ou les émotions continues. Nous détaillons les représentations acoustiques et les architectures neuronales mises en place puis nous discutons de la pertinence de la fonction de coût avant de proposer un post-traitement afin d'obtenir nos scores de référence pour le corpus AlloSat et pour le corpus SEWA.

Dans le sixième chapitre, nous entamons des discussions sur la performance de la fonction d'évaluation, puis nous détaillons nos travaux sur la modalité linguistique. Nous détaillons ensuite les protocoles utilisés pour la fusion des modalités. Et enfin, nous présentons des travaux utilisant des descripteurs pré-entraînés, qui donnent les meilleurs résultats de reconnaissance sur le corpus AlloSat.

Enfin, dans le septième et dernier chapitre, nous décrivons les travaux effectués sur la pertinence de systèmes de reconnaissance appris sur les annotations de chaque annotateur. Nous revenons également sur le succès de la modalité linguistique, que nous essayons d'expliquer en utilisant à la fois des procédés statistiques et des analyses linguistiques humaines.

PREMIÈRE PARTIE

État de l'art

LES ÉMOTIONS DANS LA PAROLE

«Tout le monde sait ce qu'est une émotion, jusqu'à ce que vous lui demandiez de la définir. A ce moment-là, il semble que plus personne ne sache.» (Fehr & Russell, 1984)

1.1 Définition de l'émotion

L'étude de l'émotion humaine est à la croisée de plusieurs domaines dont notamment la psychologie, la physiologie et la linguistique. Sa définition et sa caractérisation est encore aujourd'hui source d'études. En effet, il n'y a pas de consensus clair et établi sur une définition et une théorie qui priment sur les autres [KK81 ; Str96]. Néanmoins les scientifiques s'accordent à dire que les émotions sont des facteurs explicatifs des comportements de l'humain [Gor03]. Ce qui explique pourquoi les chercheurs persistent à vouloir révéler leurs secrets.

La définition de l'émotion est exprimée différemment en fonction des domaines d'étude. Pour le grand public, le dictionnaire Le Robert¹ définit trois sens du mot émotion :

- État affectif intense, caractérisé par des troubles divers (pâleur, accélération du pouls, etc.). Par exemple : Être paralysé par l'émotion ; Tu nous as donné des émotions, tu nous as fait peur (familier).
- État affectif, plaisir ou douleur, nettement prononcé.
- Sensibilité. Par exemple : Interpréter une œuvre avec émotion.

Au sein de cette thèse, nous considérons l'émotion selon la deuxième définition. L'émotion est un état temporaire dans lequel se trouve une personne, elle est causée par un sentiment vif ressenti habituellement en réponse à une stimulation de l'environnement. Ce concept assez général regroupe une multitude d'états qui peuvent aller de la joie à la tristesse en passant par la peur et la colère. De nombreuses théories ont été présentées au fil des siècles pour définir l'émotion.

1. <https://dictionnaire.lerobert.com/definition/emotion>

1.1.1 Point historique sur les théories de l'émotion

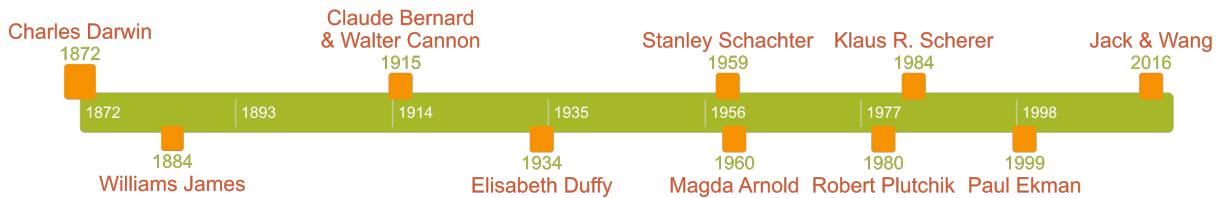


FIGURE 1.1 – Frise chronologique des grandes théories des émotions, distribuées selon leurs auteurs à la date de leur première date de parution.

L'émotion et les états émotionnels, non content de ne pas avoir une unique définition, ont évolué dans leur caractérisation au fil du temps.

Les premières mentions importantes de l'émotion nous viennent d'Aristote qui, au IV^e siècle avant J-C, définit l'Être comme une combinaison d'émotion et de raison. Bien plus tard, au XVII^e siècle, d'autres philosophes se sont intéressés à l'émotion en la mettant en relation avec la raison. Spinoza théorise que les états émotionnels ont une influence sur le raisonnement humain, tandis que Descartes pense que ces deux notions sont décorrélées.

Les prémices de Darwin

L'étude contemporaine des états émotionnels a réellement débuté avec les travaux de Charles Darwin (1872) [Dar72], qui a défini ses sept principes régissant l'émotion :

- les émotions sont innées : elles sont dues à l'Évolution et présentes dès la naissance. Elles se complexifient lorsque la personne grandit.
- Elles suivent une continuité phylogénétique : les émotions sont aussi présentes chez les animaux proches de l'être humain, par exemple les primates.
- Elles sont dénombrables : on peut caractériser chaque émotion par une des huit catégories définies par Darwin (souffrances, abattement, joie, mauvaise humeur, haine, mépris, surprise et honte) ou par une combinaison de ces dernières.
- Elles sont analysables : on peut les caractériser en fonction de l'activité musculaire du visage.
- Elles sont reconnaissables : les témoins reconnaissent naturellement l'émotion d'une personne et la traitent en tant qu'information.
- Elles sont universelles : comme elles viennent de l'Évolution, elles sont multi-culturelles et leur manifestation est reconnaissable par tous.

- Et enfin, elles sont actionnables : *le simple acte de simuler une expression tend à la faire naître dans notre esprit* : nous pouvons ressentir de la joie en nous persuadant que nous sommes heureux.

Les émotions servent à la survie de l'espèce et sont définies comme adaptatives. En effet, elles permettent d'adopter une réaction appropriée à un stimulus de l'environnement.

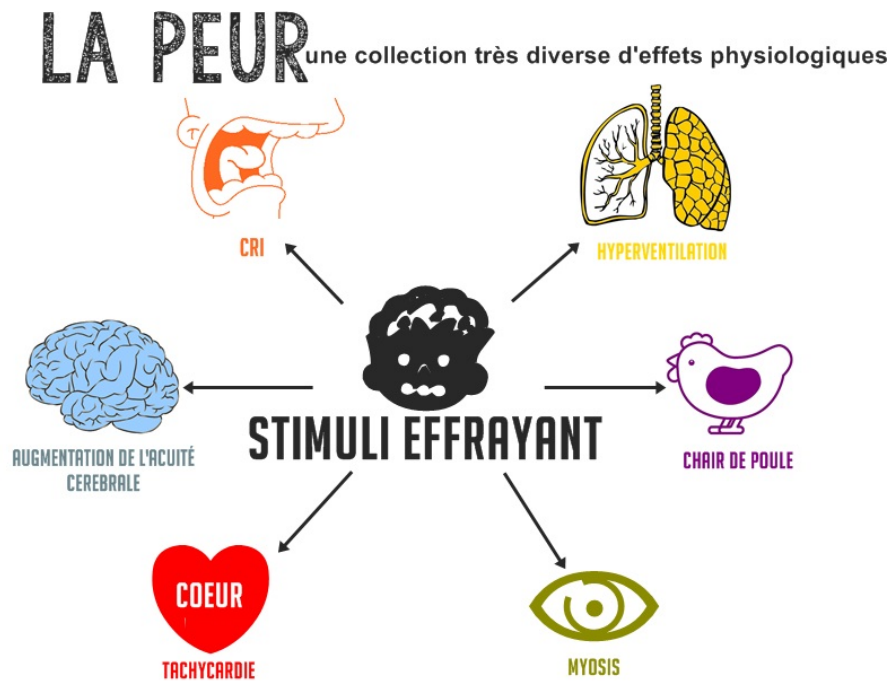


FIGURE 1.2 – Réactions physiologiques du corps devant une situation dangereuse. Image tirée du site www.comprendrelapeur.e-monsite.com.

Par exemple prenons le cas d'une situation dangereuse : la rencontre avec un animal dangereux, tel un hippopotame. L'homme ressent une émotion en réponse à ce changement d'environnement : la peur. Celle-ci va activer toute une chaîne de réponses biologiques afin d'augmenter les chances de survie. Parmi ces réponses biologiques présentées dans la figure 1.2, l'augmentation du rythme cardiaque permettant de mieux oxygéner les muscles, et l'agrandissement de l'iris de l'oeil, permettant de mieux voir les mouvements brusques. Ces réponses faciliteront la survie de l'humain qui choisira entre l'attaque et la fuite.

Darwin avance également que certaines émotions, les émotions basiques, sont universelles et innées. Elles sont donc à la fois ressenties et comprises de tous. Cette notion d'émotion fondamentale va longtemps accompagner les théories visant à expliquer l'émotion, mais chaque contributeur définira ses propres émotions primaires.

L'émotion dite périphérique

Avec la naissance de la psychologie au XIXe siècle, de nombreuses théories ont émergé pour définir et caractériser l'émotion. Williams James (1884) définit l'émotion comme une conséquence de la réponse physiologique à un stimulus de l'environnement [Jam84]. L'émotion est dite *périphéraliste*, ce qui était considéré auparavant comme la conséquence de l'émotion est ici avancé comme la cause.

Si une personne nous insulte, on ne crie pas parce qu'on est en colère, on est en colère parce qu'on crie. L'émotion se résumerait donc à la prise de conscience des changements qui s'opèrent dans notre corps, que ce soit au niveau de nos muscles, notre respiration, nos viscères... Cela implique que les émotions sont contrôlables, on peut les accentuer ou les inhiber par le simple exercice de sa volonté.

Bien que cette théorie soit en totale rupture avec les conceptions classiques de l'époque, elle va trouver un écho dans le siècle qui suit. Elle sera notamment partiellement validée par les travaux de James Douglas Laird (1974) [Lai74] qui traite de l'impact des expressions faciales simulées dans le ressenti des émotions, ou encore par les travaux de Sabine Stepper et Fritz Strack (1993) sur l'impact de la posture [SS93] ou encore par les travaux de Pierre Philippot (2002) [PCB02] sur l'impact de la respiration.

L'émotion dite centrale

En opposition à la théorie de l'émotion périphérique, Claude Bernard et Walter Cannon (1915, 1927) vont définir la notion d'homéostasie [Can15 ; Can27]. Étudiant à l'époque le mouvement des viscères, ils ont observé que ces dernières se contractent lorsque le sujet est soumis à de vives émotions, pouvant aller jusqu'à l'arrêt de la digestion lorsque le sujet est soumis à une émotion suffisamment intense. L'émotion est donc un processus qui permet au corps d'interrompre son fonctionnement normal. Cette interruption permet de concentrer les ressources du corps afin d'opérer une réponse adaptée au changement de l'environnement, principalement l'attaque ou la fuite. L'émotion est donc centrale dans le corps, servant de système de mise en alerte de l'organisme.

Ces travaux permettent également de situer la partie du cerveau responsable de l'émotion : les régions sous-corticales [CR33]. Ces régions sont donc responsables des réponses viscérales, c'est-à-dire des réponses homéostatiques d'urgence mais également du ressenti émotionnel de l'individu en tant qu'expérience subjective. Les conclusions de ces travaux lanceront l'exploration du cerveau pour trouver les régions responsables des différentes

émotions, amenant les neurosciences à s'emparer du domaine de l'émotion [Bar34].

L'émotion sous forme d'activation

En parallèle de ces explorations, la notion d'activation va émerger avec Élisabeth Duffy (1934, 1941) [Duf34 ; Duf41]. En effet, maintenant que les émotions sont détectables dans le cerveau humain, elles peuvent être traduites par des mesures du potentiel d'activité. D'un point de vue biologique, les informations sont perçues par le cerveau sous forme de messages nerveux. Ces messages sont en réalité des signaux électriques, encore appelés influx nerveux, qui transitent de neurone en neurone. Ces influx libèrent un potentiel d'activité, que nous savons mesurer.

La théorie des émotions en catégories discrètes comme définie par Darwin est donc réévaluée. En effet, si toutes les émotions peuvent être mesurées en potentiel d'activité de certaines zones du cerveau, la frontière entre les émotions peut être plus perméable que préalablement définie. Cette théorie inscrit donc un tournant dans l'étude des émotions, en introduisant des émotions plus complexes qui ne sont plus définies par un terme, une catégorie ; mais par une mesure de potentiel sur une échelle donnée. Il s'agit là des prémices des théories dimensionnelles aussi appelées théories continues. Cette théorie sera néanmoins plus difficile à démontrer que prévu, puisque les époux Lacey (1958) ont constaté que toutes les mesures (électro-encéphalogramme, activité des viscères, tension musculaires) ne covarient pas et sont différentes d'un individu à un autre [LL58].

L'émotion en tant que théorie cognitivo-physiologique

En réponse à ces constatations, Stanley Schachter (1959) a proposé sa théorie cognitivo-physiologique [Sch59 ; SS62]. Selon lui, nous définissons nos émotions en fonction de la situation dans laquelle nous nous trouvons. En effet, nous devons raisonner pour définir nos propres émotions, en *attribuant* celle-ci à un contexte interne et/ou externe. L'émotion naît donc de deux facteurs : l'activation physiologique et l'attribution cognitive. L'état émotionnel n'est donc plus une simple réponse à un stimulus de l'environnement, elle implique également une part de raisonnement.

L'émotion avec la théorie de l'évaluation

En s'appuyant sur ces travaux, Magda Arnold (1960) va amorcer la théorie de l'évaluation cognitive (appraisal en anglais) [Arn60], qui va prendre de l'ampleur avec les travaux

Dimension d'évaluation émotionnelle	Colère/Rage	Peur	Tristesse	Joie
Nouveauté				
Soudaineté	haut	haut	bas	bas
Familiarité	bas	bas	bas	ouvert
Prévisibilité	bas	bas	ouvert	moyen
Valence				
Intrinsèque	ouvert	bas	ouvert	haut
Rapport aux buts/besoins				
Pertinence	haut	haut	haut	moyen
Degré de certitude dans la prédiction des conséquences	très haut	haut	très haut	très haut
Congruence avec les attentes	dissonant	dissonant	ouvert	consonnant
Opportunité	obstruction	obstruction	obstruction	facilitation
Urgence	haut	très haut	bas	très bas
Potentiel de maîtrise				
Causalité : agent	autrui	autrui/naturel	ouvert	ouvert
Causalité : motivation	intentionnel	ouvert	hasard	intentionne
Contrôle	haut	ouvert	très bas	ouvert
Puissance	haut	très bas	très bas	ouvert
Ajustement	haut	bas	moyen	haut
Accord avec les normes				
Standards externes	ouvert	ouvert	ouvert	ouvert
Standards internes	bas	ouvert	ouvert	ouvert

TABLE 1.1 – Grille d'évaluation des premiers travaux de Scherer pour définir une émotion en fonction des cinq questionnements. La catégorie ouvert est utilisée lorsque l'évaluation peut être de différente catégorie. Par exemple, pour la joie, l'élément déclencheur peut être quelque chose de connu ou non (Familiarité). En rapport aux buts/besoins, les opportunités peuvent être obstruées (émotions négatives) ou facilitées (émotions positives). Tiré des travaux de P. Philippot [Phi07].

de Scherer (1984) [Sch84]. Ce dernier considère que l'évaluation d'une situation est définie par cinq questionnements :

- Est-ce que la situation est nouvelle ? (nouvelle/ancienne)
- Est-ce qu'elle suscite du plaisir intrinsèque ? (agréable/désagréable)
- Est-ce qu'elle est pertinente ? (aidante/gênante)
- Est-ce qu'on sait y faire face ? (on a le contrôle/ on n'a pas le contrôle)
- Est-ce que c'est compatible avec les normes ? (compatible avec les normes sociales et ses propres convictions)

C'est en réponse à ces cinq questions que Scherer va proposer cinq dimensions pour représenter l'émotion : la nouveauté, la valence, le rapport aux buts, le potentiel de maîtrise et l'accord avec les normes. La grille d'évaluation, présentée dans le tableau 1.1 , permet de caractériser l'état émotionnel d'un individu en quatre émotions : la colère/rage, la peur, la tristesse et la joie. On voit sur ce tableau la description de la joie par exemple. Pour Philippot [PCB02], elle se caractérise par une situation peu soudaine et assez prévisible, qui suscite un fort plaisir intrinsèque, qui est une situation plutôt aidante (donc elle a une forte relation avec les attentes du sujet), sur laquelle on a plus ou moins de contrôle et qui est plus ou moins compatible avec les normes.

Toutes ces théories sont encore aujourd'hui étudiées et affinées en fonction des situations d'étude des émotions. Mais aucune n'est reconnue comme étant une vérité absolue. C'est en se basant sur toutes ces théories et ces travaux que le domaine de l'informatique va alors se mettre au service de l'étude de l'émotion.

1.2 Représentation courante des émotions à but d'automatisation de traitement

Comme nous l'avons vu précédemment, il n'y a pas de consensus sur une définition ou même une caractérisation des états émotionnels [KK81]. Cela peut être expliqué par notamment, la multitude de domaines qui sont intéressés par l'émotion (psychologie, physiologie, linguistique, phonologie, sciences cognitives, informatique...) ou encore par la part non négligeable de la subjectivité du domaine et donc de la grande variabilité inter-sujet.

Néanmoins dans le domaine de l'informatique, nous faisons principalement la distinction entre deux grands courants : l'émotion définie par des états discrets et/ou par des états continus. En effet, afin d'automatiser le traitement des émotions, il est nécessaire de se tenir à une théorie algorithmiquement descriptive. Il faut donc définir des repré-

sentations émotionnelles en adéquation avec les besoins exprimés pour une tâche donnée. Ce domaine d'étude est appelé informatique affective ou affective computing en anglais. Décrit dans les travaux de Rosalind Picard en 1995 [Pic00], ce domaine reste très récent et en plein développement.

Au sein de cette thèse, notre tâche est la détection et la caractérisation automatique d'émotions contenues dans la parole. Nous proposons donc de rappeler la définition de ces deux ensembles de théorie qui sont *dans une guerre centenaire* l'une contre l'autre [Lin+13], donnant naissance à la multitude de définitions que nous lui connaissons.

1.2.1 Théorie des émotions discrètes

Comme a dit Izard : *... people need the category label of joy (or its equivalent) to explain the pride of achievement, sadness to explain the experience of a life-changing loss, anger to explain the frustration of blocked goal responses, and fear to explain flight to one another for safety* [Iza07]. Nous avons donc besoin de pouvoir regrouper des états émotionnels pour définir des catégories, ce qui nous aide à mieux expliciter nos émotions.

La théorie des émotions discrètes considère qu'il existe un nombre défini d'émotions qui peuvent être formellement caractérisées. Certaines d'entre elles sont définies comme étant *basiques*, *primaires* ou *basales* et permettent, en se combinant, d'exprimer des émotions plus *complexes*. Elles sont caractérisées par un ressenti, des expressions comportementales et physiologiques qui sont spécifiques à chaque individu mais reconnaissables par tous. Ces trois aspects permettent de caractériser chaque émotion. Certaines émotions sont admises par tous (colère, dégoût, joie, peur, surprise, tristesse), d'autres sont plus discutées [Cos94].

Historiquement Charles Darwin est l'un des premiers à définir ces émotions basiques, au nombre de cinq (la colère, le dégoût, la joie, la peur et la tristesse), dans sa théorie de l'évolution. Même s'il y a des théories qui ont plus d'adeptes que d'autres, il n'y a pas de théorie qui fait consensus au sein de la communauté qui permettrait de définir ces émotions basiques, de les dénombrer ou d'expliquer leur combinaison en émotions complexes. D'autant plus que les émotions et leurs manifestations n'ont eu de cesse d'évoluer en même temps que nos sociétés.

De nombreux auteurs ont proposés des émotions primaires : Paul Ekman propose les populaires six émotions faciales basiques (colère, dégoût, joie, peur, surprise, tristesse) qu'on appellera par la suite les *Big Six* [Ekm99], tandis que Robert Plutchik en propose huit (anticipation, colère, confiance, dégoût, joie, peur, surprise, tristesse) [R80]. On peut constater que la définition de ces émotions basiques est encore aujourd'hui source de

Auteurs	Émotions basiques
Darwin (1872)	colère, dégoût, joie, peur, tristesse
James (1884)	amour, douleur/chagrin, peur, rage
Arnold (1960)	amour, aversion, colère, courage, découragement, désespoir, désir, espoir, haine, peur, tristesse
Tomkins (1962)	angoisse, dégoût, honte, intérêt, joie, peur, rage, surprise
Izard (1971)	auto-hostilité, colère, culpabilité, dégoût, honte, intérêt, joie, mépris, peur, surprise, timidité, tristesse
Plutchik (1980)	acceptation, anticipation, colère, dégoût, joie, peur, surprise, tristesse
Frijda (1986)	bonheur, désir, intérêt, surprise
Oatley & Jonhson-Laird (1987)	bonheur, colère, dégoût, inquiétude, tristesse
Gray (1990)	anxiété, joie, rage, terreur
Ekman (1999)	colère, dégoût, joie, peur, surprise, tristesse
Jack (2016), Gu (2015) et Wang (2016)	colère, joie, peur, tristesse

TABLE 1.2 – Définition des émotions basiques selon différents auteurs.

discussions : Rachael Jack, Simeng Gu et Fushun Wang [Jac+16; Gu+15; WP16] ont proposé quatre émotions basiques (la colère, la joie, la peur et la tristesse) en 2015 et 2016. Le tableau 1.2 liste les catégories des émotions primaires selon leurs auteurs. On peut remarquer néanmoins que l'on considère de quatre à douze émotions basiques, et que sur ces onze auteurs, on a une moyenne de six à sept émotions basiques. On retrouve également les émotions colère, joie, peur et tristesse dans presque toutes les théories.

Paul Ekman a joué un grand rôle dans la diffusion de cette théorie grâce, entre autre, à la création de la méthode *Facial Action Coding System* (FACS) [EF78]. Utilisée dans des domaines grand public tel que la télévision avec la série *Lie to Me*, cette méthode permet de définir l'émotion à partir de l'expression faciale d'un individu. Il a notamment décrit les neuf caractéristiques d'une émotion basique [Ekm99], inspirée de toutes les autres théories, qui seront globalement repris par tous les autres auteurs :

- Les signaux émotionnels sont universels : les émotions de base sont reconnues par tout le monde, quelque soit sa culture ou son origine.
- Il existe des expressions communes aux hommes et aux autres primates : nous sommes capables d'identifier la colère chez les Bonobos par exemple.
- Chaque émotion est caractérisée par un ensemble de comportements physiologiques comme exprimé par la figure 1.3. Par exemple, un individu ayant un rythme car-

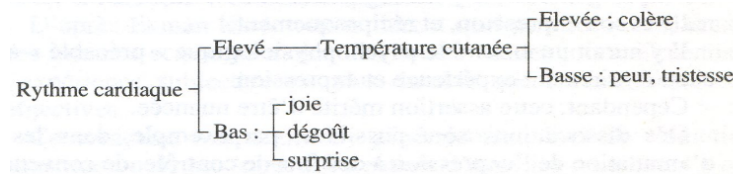


FIGURE 1.3 – Traduction de la figure extraite de l’article de Paul Ekman et al. [ELF83] décrivant un arbre de décision pour déterminer l’émotion. On voit que la colère peut être détectée par un rythme cardiaque élevé et une température cutanée élevée.

diacque élevé et une température cutanée élevée ressent de la colère.

- Il existe des éléments déclencheurs d’émotion qui sont universels : controversée, cette caractéristique stipule que des types de situations donnés provoqueront la même émotion donnée chez tout sujet.
- Les réactions émotionnelles sont cohérentes : il y a un lien établi et connu entre une expérience émotionnelle et son expression physiologique et réciproquement.
- L’émotion engendre une réaction rapide : une fraction de seconde pour les réactions physiologiques et quelques millisecondes pour les mimiques [EF78].
- L’émotion est limitée dans le temps : la durée d’une émotion n’excède pas la minute.
- L’émotion n’est pas contrôlée : elle frappe soudainement, n’étant ni volontaire ni raisonnée. Un individu peut essayer de contrôler les manifestations de l’émotion : il a été observé que la posture et les mimiques sont plus contrôlables que la voix. Les réactions viscérales sont quant à elles très peu contrôlables.
- L’émotion est spontanée : elle n’est pas choisie et ne peut pas vraiment être évitée. Toutefois son anticipation peut réduire son intensité. Par exemple devant un film d’horreur, le sujet s’attend à avoir peur, ce qui permet de réduire l’intensité de cette peur.

Ces émotions primaires peuvent également se combiner pour donner des émotions dites *complexes*, qui permettent de nuancer les états émotionnels. C’est notamment le cas avec la roue de Plutchik, présentée dans la figure 1.4, qui dénombre 32 émotions issues en huit émotions primaires selon leur intensité [R80]. Par exemple une joie très intense correspond à un sentiment d’extase, alors qu’une colère de faible intensité correspond à de la contrariété.

De plus, on parle également d’émotions secondaires afin de décrire des émotions qui ne sont pas innées, mais qui sont apprises pendant le développement de la personne. Paul Ekman définit ainsi neuf émotions secondaires, plus complexes et plus difficiles à iden-

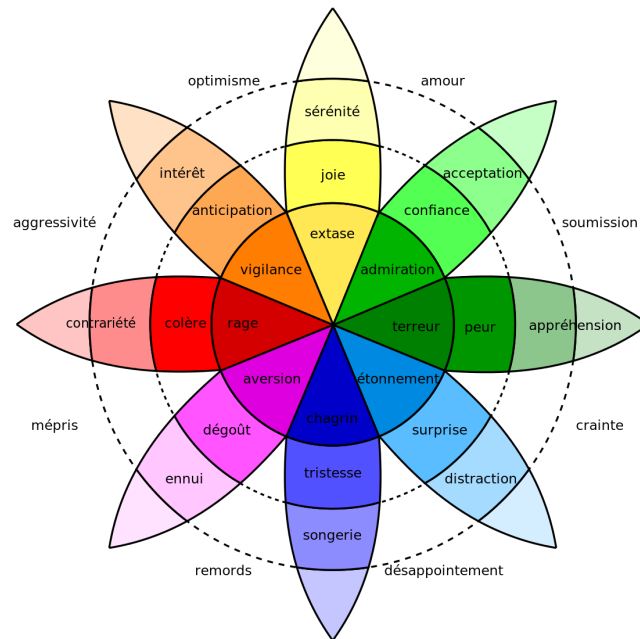


FIGURE 1.4 – Roue de Plutchik qui définit les émotions complexes à partir d'émotions basiques [R80].

tifier (la culpabilité, l'embarras, le mépris, la complaisance, l'enthousiasme, la fierté, le plaisir, la satisfaction et la honte). Elles marquent des différences d'expression en fonction des cultures et des individus. Louis Charland (1995) [Cha95] et Antonio Damasio (1999) [Dam99] ont notamment contribué à analyser les différences culturelles de ces émotions secondaires.

Pour résumer, ces théories discrètes caractérisent des émotions par des catégories telles que la joie et la peur. Elles sont observables par tous et limitées dans le temps. Elles peuvent se combiner pour définir des émotions plus complexes. D'autres théories ont émergé au fil des années, pour définir les émotions. En effet, certains ont trouvé que les émotions en catégories discrètes étaient trop limitantes pour tout représenter.

1.2.2 Théorie des émotions continues

Dans la vie courante, il est bien plus facile et coutumier de se référer à des catégories d'émotion telles que la joie ou la tristesse pour décrire ce que l'on ressent. Par exemple, lorsque l'on veut communiquer son état émotionnel à un interlocuteur, on aura tendance à le nommer. Toutefois, une autre façon de définir les émotions est de les inscrire dans

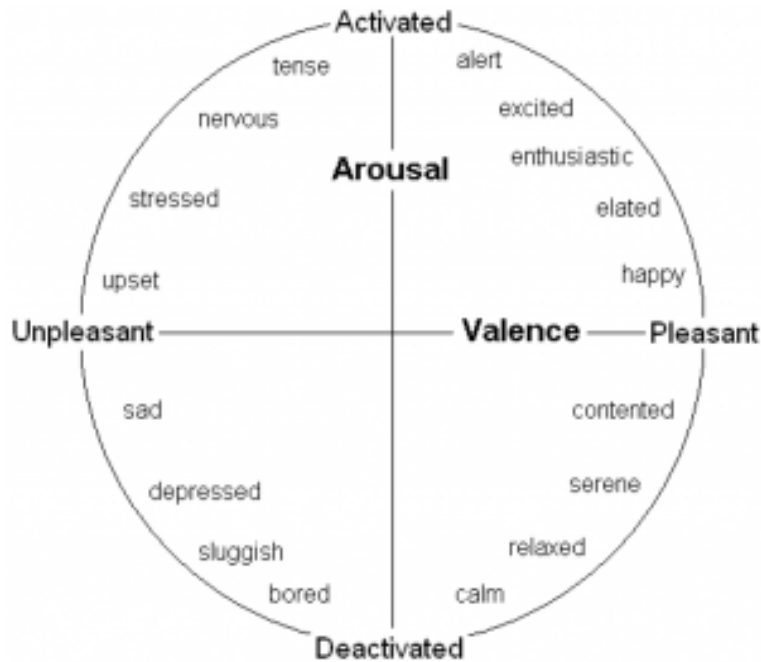


FIGURE 1.5 – Le modèle circumplex de Russell [Rus80].

des espaces continus, permettant de s'affranchir des contraintes des catégories définies. Il est difficile de déterminer des frontières entre les différentes émotions : les frontières sont ressenties différemment par tout le monde [BBN12]. Le manque de frontière dure peut expliquer l'émergence des recherches sur des théories continues pour définir l'émotion. Si l'on s'en réfère aux travaux de Lisa Feldman-Barrett (2006) [Fel06], les catégories «représente(nt) maintenant un obstacle majeur à la compréhension de ce que sont les émotions et de comment (les émotions) fonctionnent». En effet, les catégories limitent les possibilités pour exprimer des émotions.

La théorie des émotions continues, aussi appelées dimensionnelles, a été introduite par les travaux de Wilhelm Wundt (1902) [WJ02] et Harold Schlosberg (1954) [Sch54]. Les émotions peuvent être décrites selon trois dimensions indépendantes nommées en fonction de leur extremum : agréable-désagréable, tendu-détendu et agité-calme. Grâce à ces trois dimensions, chaque émotion ressentie par un individu peut être décrite comme une combinaison pondérée de ces trois axes. Cependant certaines émotions peuvent se placer au croisement des axes de différentes manières. La joie est agréable, plutôt détendue mais peut être agitée ou calme. Comme ces dimensions se chevauchent, ces dernières ont laissé place au modèle du Circumplex, présenté dans la figure 1.5, de James Russell en 1980 [Rus80], devenant la théorie principale permettant de décrire les émotions de façon

avec les théories discrètes.

En ce qui concerne les études menées sur les émotions continues dans le domaine de l’informatique, il y a une forte prévalence des dimensions de valence et d’activation quel que soit le support utilisé (la voix, le texte, la vidéo ou des données physiologiques). En effet, elles sont facilement identifiables par l’humain et reposent sur des caractéristiques automatiquement reconnaissables. Toutes ces différentes théories permettant de décrire les états émotionnels se rejoignent sur un fait : les émotions sont perceptibles chez les individus de différentes manières.

Au sein de cette thèse, nous avons décidé de nous appuyer sur la théorie des émotions continues, afin de répondre à la problématique industrielle qui est sensible à l’évolution temporelle de l’intensité de la satisfaction et la frustration. Le terme frustration vient du latin pour « frustra » = en vain ou « frustration » = « tromperie d’une attente ». En psychologie, on définit la frustration comme sentiment de déception et d’impuissance qui survient lorsqu’un événement attendu, planifié ou souhaité n’est pas possible ou se déroule complètement différemment de ce qui était prévu. Elle est souvent associée à d’autres sentiments tels que la colère, l’amertume, la déception, le ressentiment ou l’offense [Maslow2013].

1.3 L’émotion dans la parole

Un des supports universels vecteur de l’émotion est la voix humaine, et par extension la parole. La voix est produite par l’appareil phonatoire de l’émetteur et elle est reçue par l’appareil auditif du récepteur. Ce processus est considérée comme un canal de communication selon la théorie de Claude Shannon (1948) [Sha48]. Ce canal permet de faire passer un message entre un émetteur (ici le locuteur) et un récepteur (ici l’écouter).

Les appareils phonatoires et auditifs sont la base de la communication orale entre deux individus. L’appareil phonatoire décrit l’ensemble des phénomènes anatomiques qui sont impliquées dans la production des vibrations acoustiques donnant naissance à la parole. Ces phénomènes sont, par exemple, la création d’un son par la vibration des cordes vocales et le contrôle du souffle ou encore la modulation de ce son par la bouche et par le nez.

La parole peut être découpée en sous-unités phoniques que l’on appelle des phonèmes. Ces derniers, à ne pas confondre avec des syllabes, permettent de composer l’ensemble des sonorités nécessaires à l’énonciation de parole. Par exemple, le mot *émotion* utilise six phonèmes : /é/, /m/, /o/, /s/, /i/ et /on/. Il est important de noter que chaque langue

possède sa propre palette de phonèmes. Le français par exemple utilise 36 phonèmes, tandis que l'anglais en utilise 44.

L'appareil auditif, situé dans la boîte crânienne, est composé de tout ce qui permet à l'individu d'entendre et d'écouter la parole. Pour ce qui est de l'oreille, elle est constituée de trois parties :

- l'oreille externe : un conduit qui permet de transmettre les vibrations acoustiques,
- l'oreille moyenne : qui permet de transformer le signal pour qu'il puisse être conduit dans un environnement liquide en utilisant les osselets,
- et l'oreille interne : qui transforme les vibrations en influx nerveux grâce à la cochlée.

C'est cet appareil qui permet à un interlocuteur de recevoir un message.

Les informations communiquées par ce canal peuvent être notamment divisées en deux catégories : le linguistique et le para-linguistique. Le linguistique va décrire le langage en tant que succession de mots tandis que le para-linguistique va définir tout ce qui n'est pas du message verbal. Par exemple, dans la vie de tous les jours, *Bonjour, je voudrais une baguette de pain* correspond au message linguistique, tandis que l'hésitation vocale ou le raclement de gorge font partie du domaine para-linguistique.

Shirley Weitz (1974) [Wei74] explique que le para-linguistique s'intéresse à *la façon dont quelque chose est dit, pas à ce qui est dit*. Globalement, on considère du domaine du para-linguistique dans la parole, l'accent, la hauteur de la voix, le volume, la vitesse de parole, la modulation, la prosodie et la fluidité de l'élocution. La définition du paralangage étant évolutive, il est naturel que sa définition reste imprécise comme l'indique Peter Matthews dans son dictionnaire de la linguistique [MM14].

La prosodie est définie par l'ensemble des phénomènes acoustiques qui accompagnent le discours, tout en n'étant pas du discours. C'est un élément de la partie para-linguistique de la parole. La prosodie est généralement définie par trois paramètres très importants [SM03 ; Doh+04], bien qu'ils ne fassent pas consensus :

- L'intonation de la parole. Elle peut être analysée par le contour de la fréquence fondamentale (F0) de la parole, qui correspond à l'inverse de la période d'un son périodique. On utilise également l'évolution temporelle de la F0 pour définir l'intonation.
- Le rythme de la parole est lié au temps nécessaire à l'émission d'un segment de parole, et donc à la durée. On peut également parler de débit.
- L'intensité de la parole est définie par l'énergie par unité de temps contenue dans

le signal.

L'étude para-linguistique d'un discours par le biais de la prosodie nous permet de percevoir notamment l'état émotionnel d'un locuteur de façon naturelle. Comme il s'agit du sujet principal de cette thèse, le chapitre 3 revient plus en détail sur la relation entre la voix et les émotions. Mais la voix n'est pas la seule modalité véhiculant des indices émotionnels. Grâce à des outils de plus en plus performants, la parole peut être transformée en texte.

1.4 L'émotion dans le texte

Avec le développement des nouvelles technologies et le nombre de communications écrites toujours grandissant, il devient de plus en plus utile de mieux cerner les informations contenues dans ces messages. En plus de l'aspect sémantique, c'est-à-dire de la compréhension du sujet et du discours tenu dans les messages textuels, il est également possible d'exprimer des états émotionnels [HLS07 ; Sch15]. Par exemple, sur les réseaux sociaux, il n'est pas rare de retrouver des messages aux opinions très tranchées sur des sujets, dans lesquels l'individu va montrer son engagement, sa colère ou son soulagement.

Le texte écrit vient avec ces propres codes, en fonction du média utilisé, pour exprimer les états émotionnels. Traditionnellement, l'émotion va se retrouver dans la morphologie du discours, le lexique employé, la syntaxe utilisée et l'aspect figuratif ou non de l'énoncé (par exemple le sarcasme) [Sai+18].

En plus de la construction sémantique des mots et des phrases, la ponctuation ou les émoticônes peuvent apporter leur concours pour exprimer des émotions. Par exemple, la répétition de plusieurs points d'interrogation peut traduire la stupéfaction ou l'incompréhension [TP13]. L'émoticône :) (deux points, parenthèse fermante) va permettre d'exprimer la joie ou l'approbation selon le contexte [PSM07].

La détection d'émotion dans le texte peut être réalisée principalement selon les méthodes suivantes :

- L'approche basée sur les mots clés [MPI05 ; LLS03]. Pour détecter la joie, on recherche tous les mots qui sont du champ lexical de la joie (joie, heureux, content...) dans le texte. Cette approche est basée sur l'hypothèse que tous les mots sont indépendants. Le principal inconvénient vient de cette hypothèse. En effet, entre les phrases *je suis heureux* et *je ne suis pas heureux*, cette approche détecte de la joie dans les deux cas, ce qui est faux.
- L'approche basée sur des règles [API07 ; Cha07]. Pour détecter la joie, on va mettre

en place une série de règles en utilisant ou non le champ lexical de la joie et une série de règles. Par exemple, pas de négation dans le segment ou de mot appartenant au champ lexical de la tristesse. Cette approche peut vite devenir complexe, car il faut de nombreuses règles qui ne se contredisent pas pour arriver à une reconnaissance performante.

- L'approche basée sur l'apprentissage. Cette approche sera développée dans le chapitre 3.

De nombreuses applications peuvent découler de l'analyse des émotions en partant du texte. En effet, le nombre d'interactions entre un humain et une machine ne cesse d'augmenter tous les jours. Permettre aux machines de mieux comprendre et identifier les besoins de l'humain reste donc une préoccupation majeure de notre époque. Les émotions ont donc un rôle à jouer dans cette communication [Pic00]. Outre la parole et le texte, il existe d'autres comportements reconnaissables qui accompagnent l'émotion.

1.5 Les autres marqueurs de l'émotion chez l'humain

Outre la présence de marqueurs émotionnels dans la voix, d'autres indicateurs peuvent être relevés dans les expressions faciales et dans les réponses physiologiques. Il est donc possible d'étudier l'état émotionnel d'une personne à partir d'une vidéo ou de relevés physiologiques.

1.5.1 Les marqueurs faciaux et comportementaux de l'émotion

Il est possible de capter beaucoup d'informations à partir du comportement et du corps de l'humain. Nos expressions faciales notamment permettent de véhiculer les émotions primaires telles que la peur ou la colère comme montré dans la figure 1.7 en mettant en jeu le positionnement des sourcils, la forme de la bouche, le plissement du front... Par exemple, la tristesse se caractérise par des paupières tombantes, un regard absent et une bouche légèrement inclinée vers le bas selon le *Facial Action Coding System* de Paul Ekman [EF78].

Bien que ces marqueurs ont été affirmés comme étant universels par Paul Ekman [EF78], de nombreuses études ont mis en doute ce postulat [Ley10; Gen+14]. En effet, certains indicateurs étant communs entre plusieurs émotions, des interlocuteurs peuvent se tromper dans l'analyse de l'émotion de la personne. Par exemple, la bouche ouverte peut signifier

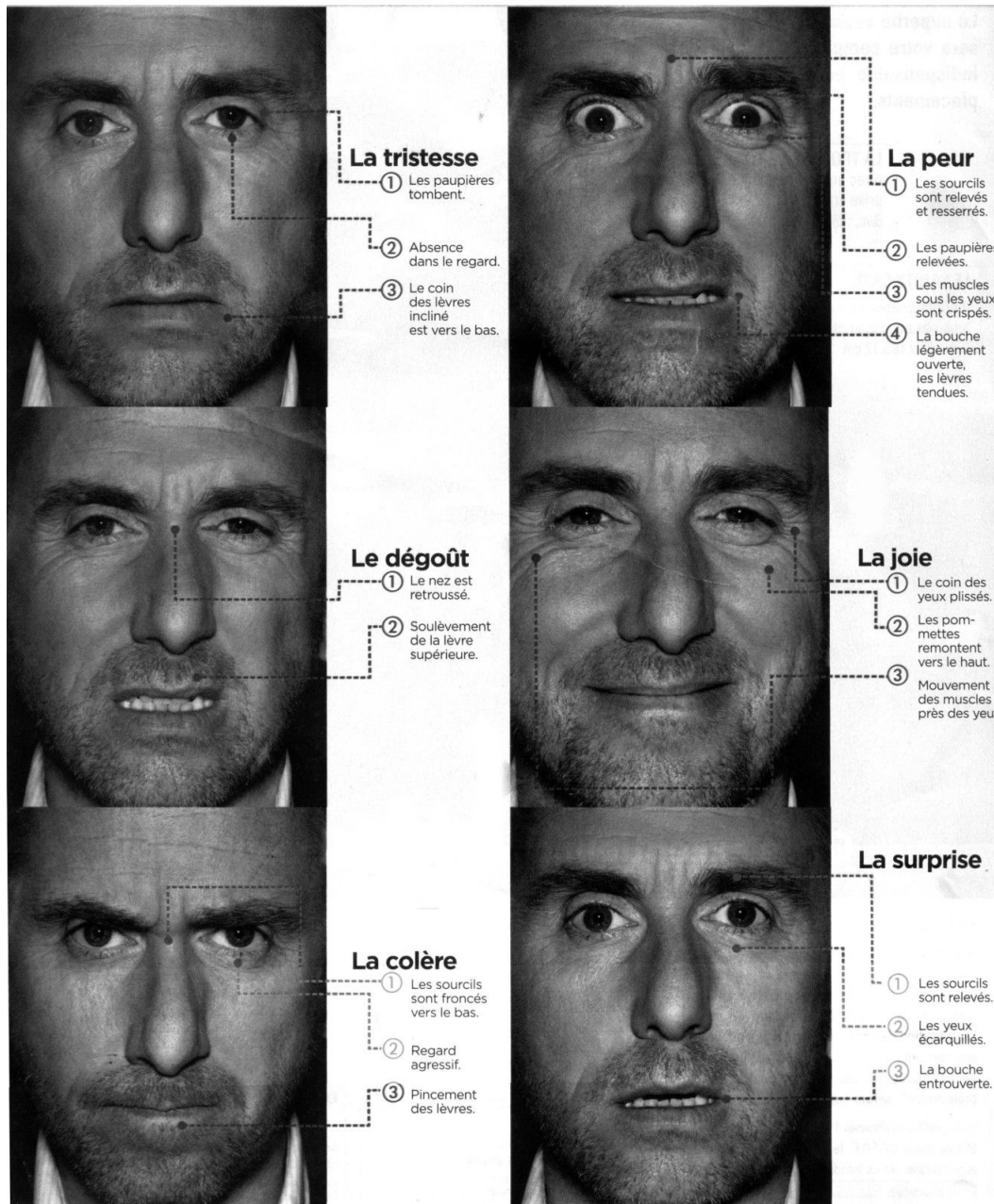


FIGURE 1.7 – Marqueurs faciaux des 6 émotions primaires d'Ekman. In Lie To Me.

la peur mais également la surprise. Les expressions faciales sont contrôlables. En effet, il est tout à fait possible de simuler une émotion ou dans une moindre mesure, d'en cacher une en fonction de l'entraînement d'un individu. C'est le cas notamment des acteurs professionnels qui sont capables de simuler un grand nombre d'émotions.

Les expressions faciales sont, aujourd'hui encore, l'une des modalités les plus observées pour définir et reconnaître des émotions. Elles peuvent également servir à détecter des états dépressifs ou des mensonges par exemple [SJA01 ; Owa+12].

Les positionnements du corps, qu'on peut appeler *l'attitude* d'une personne permet également de reconnaître ses émotions. En effet, les comportements adoptés par un individu, notamment la posture, peuvent refléter de l'état émotionnel interne d'une personne. Ce phénomène, pressenti par Darwin [Dar72], est expliqué dans les travaux de Frijda (1987) [Fri87] qui indique que les émotions sont présentes pour aider l'homme à agir : l'homme est caractérisé par une tendance à l'action qui se traduit par des émotions. Ainsi lors de la colère, la position du buste sera en avant, tandis que pour la peur, elle sera plus retraits. Mais ces caractéristiques faciales et comportementales ne sont pas les seules permettant de reconnaître l'état émotionnel d'un individu.

1.5.2 Les marqueurs physiologiques de l'émotion

Comme nous l'avons vu précédemment, l'expression de l'émotion est étroitement liée au cerveau et donc au système nerveux [Dan02]. En étudiant ce dernier, nous pouvons retrouver des marqueurs émotionnels. La peur par exemple commence par un influx nerveux qui engendre une augmentation du rythme cardiaque, qui va augmenter l'afflux sanguin, la sudation, l'apport en oxygène par une respiration plus rapide [Ste02].

De manière non-exhaustive, nous pouvons citer le rythme cardiaque [WMK00], la sudation, la température de la peau (à l'origine des joues rouges), l'activité cérébrale, le suivi du regard, le rythme de la respiration comme des marqueurs physiologiques de l'émotion [MP10 ; Lev03]. Certains de ces marqueurs physiologiques peuvent être quantifiés par des outils de mesure. L'électrocardiogramme (EG) permet de suivre l'évolution du rythme cardiaque, l'électroencéphalogramme (EEG) permet de détecter les zones d'activité du cerveau. Des capteurs de sudation et des thermomètres peuvent permettre de suivre la production de sueurs au niveau des mains ou la température cutanée. Tous ces indicateurs peuvent caractériser la présence ou l'absence d'émotion chez un individu.

1.6 Conclusion

Dans ce chapitre, nous avons défini l'émotion et les différentes théories qui proposent de la définir, en proposant notamment un point de vue chronologique. Nous avons détaillé les différentes représentations des émotions qui sont adaptées à du traitement automatique.

Nous avons également défini les différentes modalités où l'émotion s'exprime. En revenant sur la demande industrielle à l'origine de cette thèse, nous n'avons que la voix comme vecteur d'émotion, puisqu'il s'agit d'analyser des enregistrements audio de conversation issues de centres d'appels. Dans le chapitre 3, nous allons explorer plus en détail la mise en place de solutions de reconnaissance automatique de l'émotion depuis la modalité vocale.

APPRENTISSAGE AUTOMATIQUE POUR LE TRAITEMENT DE LA PAROLE

«La tristesse de l'intelligence artificielle est qu'elle est sans artifice, donc sans intelligence.» (Jean Baudrillard, 1987)

2.1 Apprentissage automatique : définition

Dans le domaine de l'intelligence artificielle (IA), l'apprentissage automatique (machine learning, abrégé ML en anglais) regroupe des méthodes permettant à un système d'apprendre un comportement. Selon le Journal Officiel¹, l'apprentissage se définit par un *processus par lequel un algorithme évalue et améliore ses performances sans l'intervention d'un programmeur, en répétant son exécution sur des jeux de données jusqu'à obtenir, de manière régulière, des résultats pertinents*. Généralement, l'apprentissage automatique permet, à partir de données plus ou moins massives, d'apprendre à caractériser de nouvelles données de même nature par une classification ou une régression. On apprend des faits à partir de données connues, pour les appliquer sur des nouvelles situations. On dit alors qu'un système automatique a été entraîné avec des données d'entraînement (première étape) afin de permettre la prédiction des caractéristiques de nouvelles données similaires (seconde étape) appelées données de développement, de test ou de validation.

Concrètement, si on regarde la figure 2.1, nous prenons en exemple la caractérisation de l'image d'entrée par la présence ou l'absence d'hippopotame. L'apprentissage sert à construire un modèle mathématique et apprendre les différentes variables θ_i de celui-ci à partir de données pré-traitées que l'on nomme d'entrées x_i . Les différents paramètres sont actualisés au fur et à mesure de l'apprentissage par minimisation de la fonction de coût. Les sorties y_i de cette fonction mathématique peuvent être analysées afin de caractériser les données d'entrées.

1. <https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000037783813>

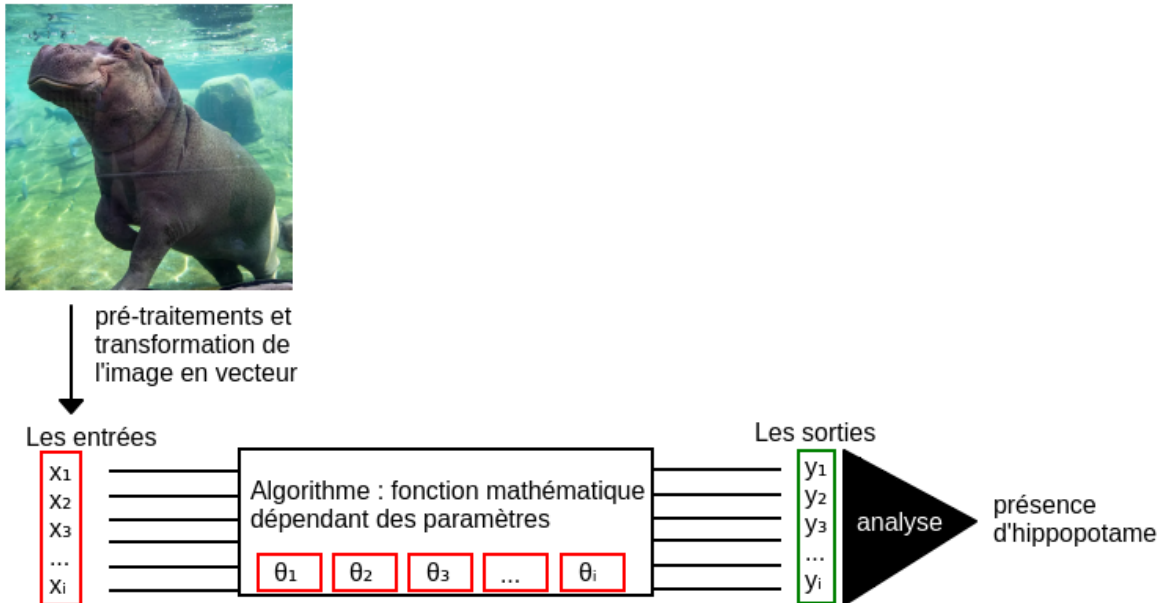


FIGURE 2.1 – Représentation graphique d'un système d'apprentissage automatique avec ses entrées et ses sorties.

2.1.1 Classification et régression

La classification et la régression sont les deux grandes tâches d'apprentissage automatique. La classification correspond à résoudre un problème d'affectation de classes. Par exemple, on peut apprendre à détecter la présence d'hippopotame au sein d'une image. Nous avons alors deux réponses possibles : la présence ou l'absence de l'animal. Ces réponses se transforment donc en classes, qui seront apprises par un apprentissage automatique de type classification. Nous avons un nombre défini et connu de catégories, donc un espace discret, qui sera utilisé par le système pour catégoriser chacune des données.

La régression quant à elle ne permet pas de catégoriser les données en classes, elle les inscrit dans un espace continu. Par exemple, on peut apprendre à prédire la valeur d'un stock en bourse à partir des valeurs des jours précédents. On peut également prédire une valeur représentant le niveau de satisfaction d'un client à partir d'un questionnaire. Les réponses sont des nombres réels qui peuvent être positifs ou négatifs. Nous avons donc une échelle de valeurs dans laquelle le système va inscrire sa prédiction.

Afin d'avoir des systèmes performants, il est important de bien configurer leur apprentissage.

2.1.2 Les pré-requis pour un Apprentissage Automatique réussi

Un facteur crucial garantissant la qualité de ces systèmes automatiques réside dans les données d'entraînement. Plus elles sont qualitatives, c'est-à-dire qu'elles sont les plus proches des données réelles que nous voulons analyser, et plus le système sera performant dans sa tâche de prédiction. Mais la qualité ne fait pas tout dans ce contexte, la diversité et l'exhaustivité sont tout aussi important. Par exemple dans une tâche de reconnaissance d'hippopotame au sein d'une image, si on présente uniquement des animaux noirs alors le système pourra mal classifier des hippopotames marrons.

Mais les données ne font pas tout, le choix de l'algorithme mis en place pour apprendre le système automatique est tout aussi important. Il existe de nombreuses méthodes mathématiques permettant de modéliser un système automatique. Ces méthodes d'apprentissage automatique sont diverses et les réseaux de neurones, appelées Deep Neural Network en anglais (DNN) ou Deep Learning en font partie. Lorsque nous cherchons à mettre en place un apprentissage automatique ces deux facteurs sont décisifs : les données et le choix de l'algorithme.

Ces choix doivent être en adéquation avec le type de caractérisation recherchée. En effet, il existe deux principales approches d'apprentissage : l'apprentissage supervisé et l'apprentissage non-supervisé.

2.1.3 Apprentissage supervisé et non supervisé

Il existe de nombreuses approches d'apprentissage dont les principaux sont présentés dans la figure 2.2. Dans le contexte de cette thèse, nous ne parlerons que d'apprentissage supervisé, non-supervisé et auto-supervisé.

Dans le cas de l'apprentissage supervisé, l'humain va *guider* l'algorithme en fournissant des exemples concrets qui sont étiquetés avec les résultats attendus, comme dans la figure 2.1. Concrètement, pour apprendre à discerner des images contenant des hippopotames, on va collecter un ensemble d'image contenant ou non un hippopotame, et pour chaque image, on va lui attribuer une étiquette *présence* ou *absence*. Cette étiquette peut également être appelée référence, annotation ou label. Ainsi nous avons un ensemble d'exemples concrets que nous pouvons soumettre à notre algorithme d'apprentissage. Le système va alors apprendre à partir de chaque exemple ses paramètres de façon à diminuer l'écart entre le résultat obtenu et le résultat attendu. La marge d'erreur se réduit avec chaque itération de l'apprentissage, avec pour but d'être capable de généraliser son

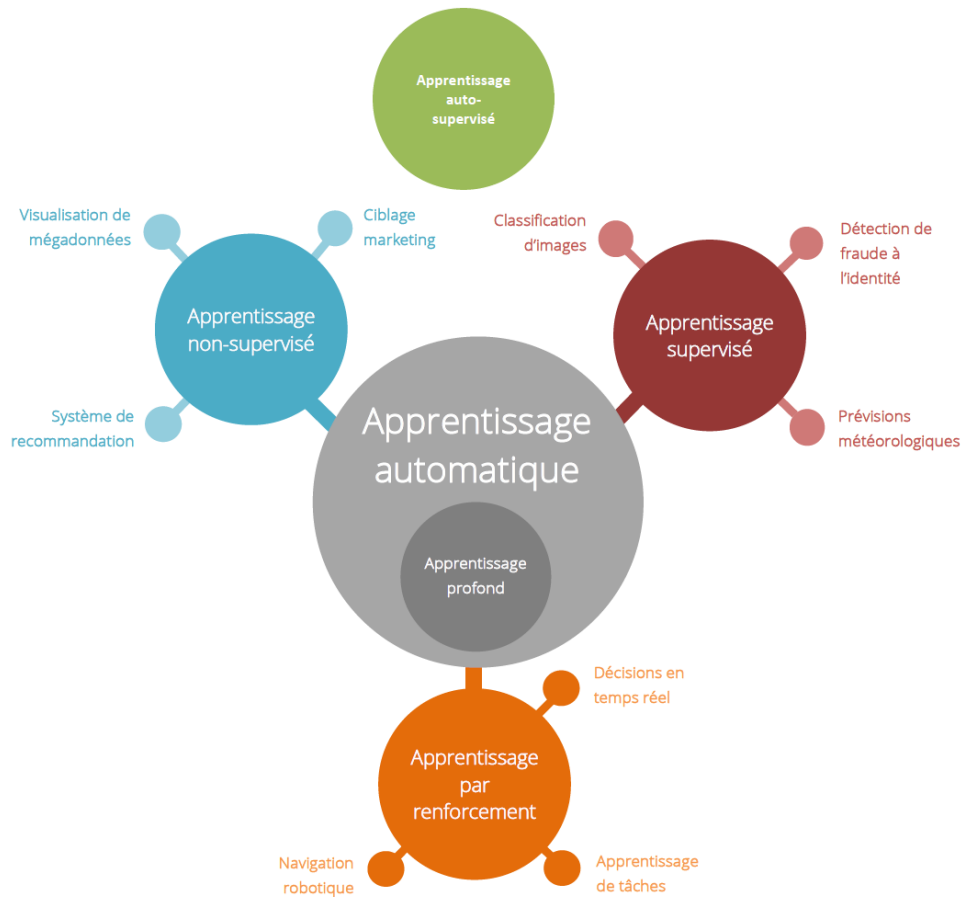


FIGURE 2.2 – Les différents types d'apprentissage et des exemples d'utilisation. Tiré de <https://www.coe.int/fr/web/artificial-intelligence/glossary> et modifié.

apprentissage à de nouveaux cas. Ce type d'apprentissage ne peut donc être mis en place que lorsque nous avons des données d'entraînement qui sont annotées. Cette annotation peut être effectuée soit par l'humain soit par la machine.

Dans l'apprentissage non supervisé, l'objectif du système automatique est d'inférer une caractérisation des données d'entrée : il doit capturer de lui-même les structures sous-jacentes aux données. Il n'y a pas de label pour *guider* l'apprentissage. Nous spécifions le nombre de classes au système et il va essayer d'apprendre tout seul ce qui différencie les images en entrée, ici la présence ou l'absence de l'animal.

Cet apprentissage se base principalement sur deux méthodes : la méthode par partitionnement et la méthode de regroupement. Pour le partitionnement, on va chercher à diviser les entrées en un nombre k de partitions. Pour cela, on peut comparer la dis-

tance entre les différents échantillons d'entrée, par le calcul d'une distance euclidienne par exemple. Le regroupement, appelé clustering en anglais, va chercher à minimiser l'inertie intra-classe, c'est-à-dire les distances entre les échantillons d'une même classe. Elle va en même temps chercher à maximiser l'inertie inter-classe, c'est-à-dire les distances entre les centres de chaque classes.

Un nouveau type d'apprentissage est de plus en plus utilisé depuis ces trois dernières années, notamment dans le domaine de traitement des langages naturels (TALN). On l'appelle apprentissage auto-supervisé, self supervised learning en anglais (SSL). Comme pour l'apprentissage non-supervisé, il s'agit d'un apprentissage où l'on n'a pas de références pour nos données d'apprentissage. Pour compenser leur absence, il y a plusieurs méthodes qui peuvent être utilisées. Par exemple on peut prendre une très grande masse de données et on en cache une partie. Le système devra alors retrouver les parties cachées à partir des parties disponibles. On peut également demander au système d'apprendre à prédire la suite des données. Ainsi le système crée à la volée des étiquettes qui lui permettront d'apprendre. Il s'agit en fait de mélanger les avantages de l'apprentissage supervisé et non-supervisé.

L'apprentissage par renforcement consiste à laisser le système inférer ses propres décisions et le récompenser positivement ou négativement en fonction de sa réponse. Le système va alors chercher, en répétant les expériences, la meilleure stratégie qui maximise la somme des récompenses au cours du temps. Cet apprentissage est très utilisé dans les jeux vidéos ou de société, où la récompense peut se traduire par la victoire ou la défaite du système par exemple. On peut citer AlphaZéro [Sil+18], un système automatique qui a battu les meilleurs systèmes automatiques ayant eux-même battu les champions du monde humains de Go, de shogi et d'échecs.

Il existe de nombreux algorithmes permettant l'apprentissage de ces systèmes, que la tâche soit de nature supervisée, non-supervisé ou auto-supervisé.

2.2 Quelques familles d'apprentissage automatique

Dans le contexte de la reconnaissance d'émotion, de nombreuses méthodes d'apprentissage automatique sont utilisées, que ce soit pour faire de la classification lorsque l'on considère une émotion comme discrète ou de la régression quand on considère une émotion comme continue. Dans les sections suivantes, nous détaillerons certaines d'entre elles.

2.2.1 k-moyennes et k-plus proches voisins

a revoir, pas de corrections encore Les k-moyennes [Llo82], appelé k-means en anglais, et les k-plus proches voisins [FH51 ; CH67], appelé k-nearest neighbors (KNN) sont des algorithmes de classification.

Soit un ensemble de points X_i avec $i = 1, 2, \dots, n$ que l'on souhaite classer en k classes C_i avec $i = 1, 2, \dots, k$. L'algorithme k-moyennes cherche le meilleur partitionnement afin de minimiser la distance entre les centres des classes et tous les points qui leur sont affectés. L'ensemble des distances D entre le centre C de la classe C_n et tous les points X_i appartenant à cette classe C_n est définie par l'équation 2.1.

$$D(C_n) = \sum_{X_i \in C_n} \|x_i - C\|^2 \quad (2.1)$$

Les k premières données X_1, \dots, X_k sont définies comme les centres C des classes C_1, \dots, C_k puis on considère la distance entre les nouvelles données X_{k+1}, \dots, X_n et ces centres pour leur affecter une classe. Le centre de la classe C est donc recalculé à chaque nouvelle donnée assignée à la classe.

En ce qui concerne l'algorithme du k-plus proches voisins, il est très semblable au k-moyennes. Le centre de chaque classe est cependant fixé par l'expérimentateur et ne sera pas recalculé.

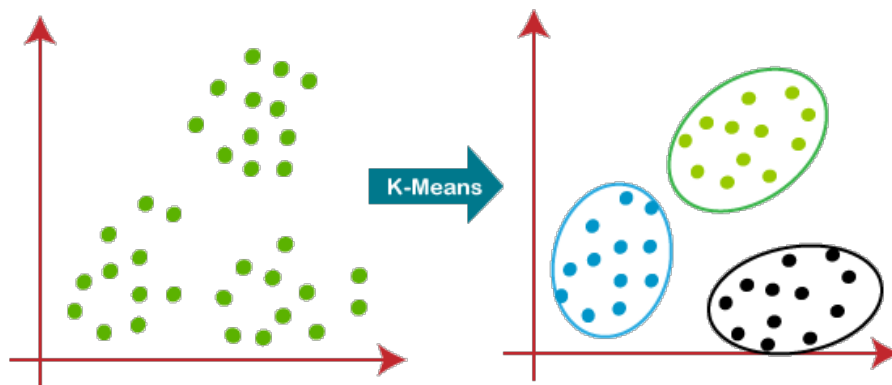


FIGURE 2.3 – Représentation graphique d'une classification en trois classes par un algorithme k-moyenne. Ici $k=3$.

Le k-moyennes permet de diviser les données en k groupes distincts, appelés clusters. Par exemple, dans la figure 2.3, on considère trois classes. Nous sommes donc dans un contexte d'apprentissage non supervisé.

Le k plus proches voisins est, quant à lui, un algorithme utilisé en classification supervisée : on connaît les références attendues pour les centres de classes. Le nombre de classe k est lui aussi connu et on cherche à caractériser des données en fonction de leur distance avec les données contenues dans les k classes. On décide donc de la classe d'une donnée en fonction de celle de ces plus proches voisins.

Ces deux algorithmes sont donc très dépendants de l'initialisation du système : le nombre de classes ainsi que l'ordre de présentation des données dans l'apprentissage sont des paramètres critiques.

On retrouve des utilisations de ces algorithmes en NLP et en parole, par exemple avec Kamper et al. qui utilise l'algorithme du k -moyennes pour de la segmentation de parole non annotée [KLG17], la classification d'actes de parole dans des jeux éducatifs [Rus+12] ou l'utilisation de l'algorithme k plus proche voisin pour la classification de texte [ZW15].

2.2.2 Régression Linéaire

La régression linéaire permet de catégoriser des sous-espaces en fonction de la proximité des données. Ce type de système apprend pour chaque classe une fonction de régression linéaire qui est optimale pour représenter les données de la classe. Son objectif est d'expliquer une variable y à l'aide d'une ou plusieurs variables x . La régression linéaire peut être simple, représentée par une fonction affine $y = ax + b$ avec y la variable à expliquer, x correspondant à une caractéristique de la donnée, a le coefficient associé à la caractéristique et b une constante. Elle peut également être multiple, représentée par l'équation 2.2 lorsque plusieurs caractéristiques des données sont prises en compte. Ici, on dispose de n observations appelées y_i que l'on souhaite expliquer par les caractéristiques $x_{i,j}$.

$$\begin{pmatrix} y_1 \\ \dots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{1,1} & \dots & x_{1,p} \\ \dots & \dots & \dots \\ x_{n,1} & \dots & x_{n,p} \end{pmatrix} \begin{pmatrix} a_0 \\ \dots \\ a_p \end{pmatrix} + \begin{pmatrix} b_1 \\ \dots \\ b_n \end{pmatrix} \quad (2.2)$$

Par exemple pour prédire la présence d'un hippopotame dans un zoo (y_1), les données peuvent contenir la présence d'un zoo dans une ville ($x_{1,1}$), le nombre d'animaux qui le compose ($x_{1,2}$), la présence d'un bassin ($x_{1,3}$), ...

Afin de classer une nouvelle donnée, le système va la comparer avec les différentes courbes de régression apprises en calculant l'écart de cette donnée à ces courbes. On appelle cette méthode, l'estimateur des moindres carrés.

Cette méthode éprouvée depuis plus de deux cents ans [Gau09; AM05; ADR08] est

explicable et facile à mettre en place. L'apprentissage est également très rapide. Mais elle ne fonctionne pas pour toutes les tâches et toutes les données. En effet, elle implique que les caractéristiques des données permettent d'expliquer la caractéristique recherchée. Dans notre exemple, la présence d'un bibliothèque dans la ville ($x_{1,4}$) ne permettra probablement pas de prédire la présence d'un hippopotame dans un zoo.

Cet algorithme est utilisé dans des problématiques de NLP, dans des problématiques concrètes telles que la détection de vraie ou fausse note de suicide [Pes+10], la classification de rapport d'accident de site de construction [Zha+19a] pour déterminer la cause des incidents en combinant différents algorithmes ou encore il était utilisé dans la reconnaissance de la parole afin de réorganiser les meilleures hypothèses de transcription [CR01].

2.2.3 Machine à vecteurs de support

La méthode, appelée Support Vector Machine (SVM) en anglais [CV95], est principalement utilisée pour résoudre des tâches de classification supervisée. Elle consiste en la séparation linéaire des données projetées dans un espace en utilisant un hyperplan optimal. Cette séparation est effectuée en maximisant la marge entre les données de chaque classe : l'hyperplan optimal doit être le plus éloigné des données des différentes classes tout en les séparant.

Pour avoir des résultats pertinents il faut que les données soient linéairement séparables dans le plan de projection. Si ce n'est pas le cas, l'utilisation de différents noyaux va permettre de modifier l'espace et donc la répartition des données dans cet espace. On peut par exemple voir la projection de données 2D en 3D, permettant de rendre certaines données linéairement séparables.

Cette méthode est très utilisée en classification, étant assez rapide à apprendre. De plus, elle est très performante lorsque l'on possède peu de données d'apprentissage.

On retrouve des utilisations de SVM dans les domaines de l'ASR, notamment couplé à d'autres algorithmes dans [Sol+07], dans la reconnaissance d'émotion, des SVM sont utilisés dans les arbres de décisions de [Roz+12] pour déterminer si une phrase énoncée fait partie des quatre catégories (joie, colère, tristesse et neutre) ou encore dans la tâche de reconnaissance d'entités nommés dans la langue arabe [BDR08].

2.2.4 Modèle de Markov caché et prédiction de séquence

Le modèle de Markov caché, appelé hidden Markov model (HMM) en anglais [RJ86], est un modèle probabiliste à base d'automates. Ces automates se composent d'états qui sont reliés par des transitions, comme montré sur la figure 2.4.

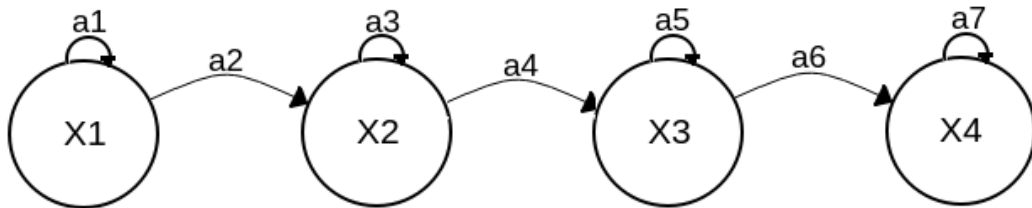


FIGURE 2.4 – Représentation graphique d'un modèle de Markov caché à quatre états. X représente les états de l'automate, a représente les transitions entre les états.

Chaque transition boucle sur lui-même et va vers l'avant et correspond à la probabilité de changer de l'état courant pour le suivant. Comme le changement d'état n'est pas obligatoire, une transition circulaire est mise en place pour boucler sur l'état courant. Utilisé en classification principalement, en apprentissage supervisé ou non, il va considérer chaque élément de nos données d'entrée de manière séquentielle pour prédire son état final.

Les modèles de Markov cachés ont été massivement utilisés notamment en ASR, en reconnaissance d'écriture manuscrite [HBT96], en intelligence artificielle [GY08], en traitement automatique du langage naturel [CSR06] ou en segmentation et regroupement de locuteurs [Ajm+02].

2.2.5 Modèle de mélange gaussien

Le modèle de mélange gaussien, appelé Gaussian Mixture Model (GMM) en anglais, est très souvent associé aux HMM. En effet, il permet d'associer un ensemble de densités de probabilités aux différents états, selon un mélange de lois gaussiennes. Les paramètres de ces dernières sont appris lors de la phase d'apprentissage par la maximisation de la vraisemblance. Le nombre de gaussiennes est un hyper-paramètre fixé par l'humain. Sur la figure 2.5, on peut voir un modèle à trois gaussiennes.

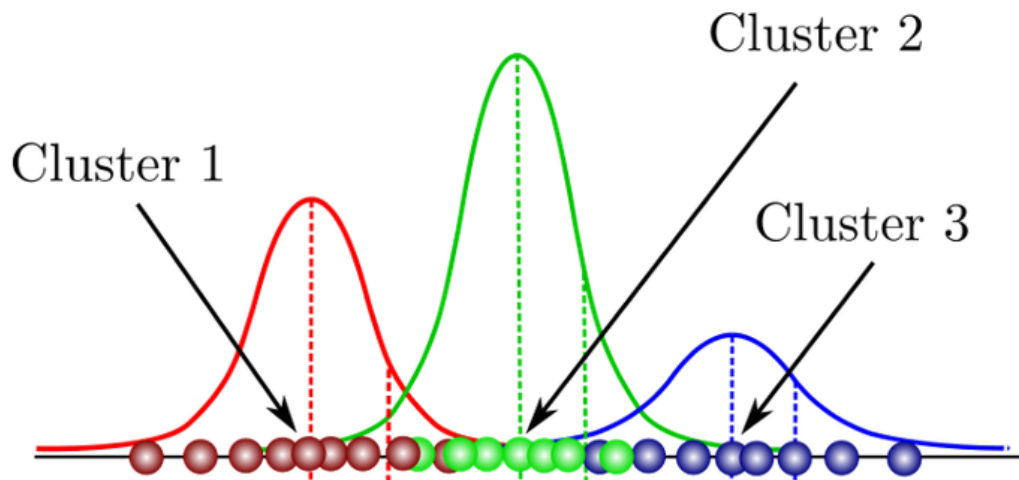


FIGURE 2.5 – Représentation graphique d'un modèle de mélange de trois gaussiennes.

Les GMM sont notamment utilisés dans des tâches de NLP tel que le résumé automatique de texte [FR09] ou encore pour la vérification de locuteur [BRS05]. Couplés aux HMM, ils ont longtemps été les systèmes à l'état de l'art majoritaires pour la reconnaissance de la parole, jusque dans les années 2010 [Hin+12]. Aujourd'hui, les réseaux de neurones sont de plus en plus utilisés pour améliorer notamment la reconnaissance de la parole.

2.2.6 Réseau de neurones

Les réseaux de neurones ont été conçus en s'inspirant du fonctionnement du cerveau humain. Ils sont composés d'une succession de neurones qui sont interconnectés entre eux pouvant ainsi propager des signaux. Afin de mieux comprendre leur fonctionnement, il est intéressant de les mettre en relation avec la biologie du système nerveux de l'homme.

Au début du XXe siècle, les avancées de la biologie ont permis de mettre en lumière les méthodes de fonctionnement de notre système nerveux [Ram06]. On définit alors les neurones en tant que corps cellulaire muni d'un axone et de dendrites. Cette cellule, illustrée dans la figure 2.6 est traversée par des influx nerveux de type électrique, qui entrent par les dendrites et ressort par les synapses de son axone. Chaque neurone peut être relié à un ou plusieurs neurones, que ce soit en entrée ou en sortie. C'est en 1943 que Warren Sturgis McCulloch et Walter Pitts proposent une représentation mathématique des neurones [MP43], que l'on appelle neurone logique dans la suite de ce document. Ils

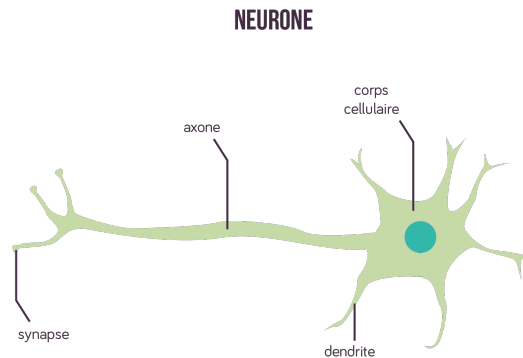


FIGURE 2.6 – Représentation schématique d'un neurone biologique. Illustration issue de <https://www.schoolmouv.fr/definitions/neurones/definition> .

définissent plusieurs principes :

- un neurone biologique peut être représenté par un neurone logique,
- les entrées du neurone logique sont comparables aux dendrites,
- le neurone logique possède une seule sortie qui représente l'axone,
- les connexions entre les neurones se font par le clonage de la sortie en autant de liens qu'il y a de prochains neurones, représentant les connexions synaptiques,
- la fonction d'activation correspond à une prise de décision au niveau du neurone logique qui représente un potentiel d'activation : le neurone biologique émet un influx nerveux ou non.

C'est à partir de ces règles que les réseaux de neurones vont être définis.

2.3 Réseau de neurones profonds

Les réseaux de neurones profonds sont de nos jours de plus en plus utilisés. Pourtant, ils ne sont pas récents : le premier algorithme d'apprentissage utilisant un réseau de neurones a plus de 50 ans. Il s'agit du perceptron de Frank Rosenblatt [Ros58].

2.3.1 Du perceptron au multicouche

Le perceptron, schématisé dans la figure 2.7, est un modèle qui permet de discriminer les données en deux classes. Il s'agit donc d'un modèle qui permet de faire de la classification et de la régression de manière supervisée. Ce modèle binaire utilise un neurone pour

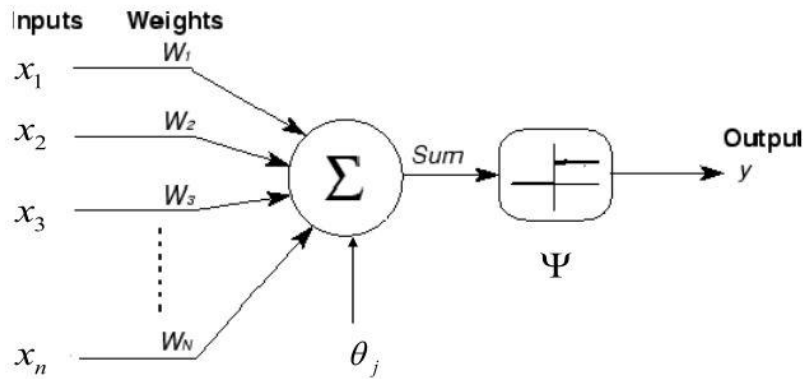


FIGURE 2.7 – Représentation schématique d'un perceptron. x_i correspond aux entrées, les W_i correspondent aux poids associés à chacune des entrées x_i , ψ formalise le biais. Une fois sommée, la valeur est passée dans une fonction d'activation ϕ pour déterminer la valeur de sortie y . Dans le schéma la fonction d'activation correspond à un simple signal échelon.

l'apprentissage et la prédiction de la classe de chaque document. Un document est défini par un nombre n de caractéristiques, chacune d'entre elles étant considérées comme une entrée du perceptron.

Lors de l'apprentissage, des poids w_i , initialisés de façon aléatoire, sont mis à jour pour chacune des entrées. Cette mise à jour est effectuée en fonction du taux d'apprentissage, appelé learning rate en anglais, afin de retrouver les étiquettes déjà connues des données d'apprentissage. En plus du poids, le biais est également défini lors de l'apprentissage. Il correspond à un unique poids qui sera utilisé pour la prédiction des étiquettes des données. Les entrées sont ainsi agrégées en les pondérant selon leur poids. La fonction d'agrégation est explicitée par l'équation 2.3, où x_i correspond à l'entrée i , w_i correspond au poids associé à l'entrée i et b correspond au biais.

$$z = \sum_{i=1}^n (w_i * x_i) - b \quad (2.3)$$

Une fonction d'activation est ensuite appliquée. Dans le cas du perceptron de Rosenblatt, soit le tout premier perceptron, la fonction d'activation est définie par une fonction échelon qui prend une valeur de 0 si z est inférieur ou égal à 0 et une valeur de 1 s'il est strictement supérieur à 0.

Le perceptron est une architecture très simple qui est de moins en moins utilisée. En effet, elle permet de résoudre uniquement des problèmes binaires linéairement séparables.

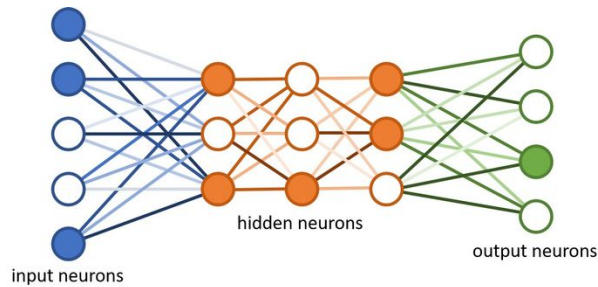


FIGURE 2.8 – Représentation schématique d'un perceptron multi-couche. Les neurones bleu correspondent à la couche d'entrée, les neurones oranges définissent plusieurs couches cachées et les neurones verts correspondent à la couche de sortie.

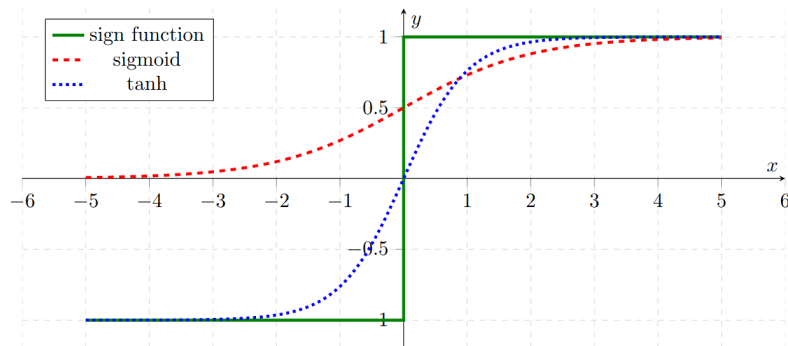


FIGURE 2.9 – Représentation graphique de trois fonctions d'activation : la fonction non linéaire, sigmoïde et tanh. Illustration provenant de [Tho14].

Le perceptron multicouche est notamment capable de palier à ces inconvénients.

Ce dernier, appelé multilayer perceptron (MLP) en anglais, assemble plusieurs neurones entre eux comme illustré dans la figure 2.8. Un ou plusieurs neurones sont assemblés les uns à la suite des autres. On a forcément une couche de neurones (ou un neurone seul) en entrée, qui reçoit les caractéristiques de chaque donnée et une couche de neurones (ou un neurone seul) en sortie qui prend la décision finale. Entre les deux, on trouve éventuellement des couches intermédiaires dites cachées, qui vont prendre en entrées les sorties de la couche précédente (soit la couche d'entrée, soit une autre couche cachée) et en sortie la couche suivante (soit une autre couche cachée soit la couche de sortie).

Les informations vont donc être propagées de la couche d'entrée à la couche de sortie. On appelle cette propagation, propagation directe avant, ou forward propagation en anglais. Le nombre de couches et le nombre de neurones par couche doivent être définis par l'utilisateur afin d'être en adéquation avec la tâche visée. Par exemple, le nombre de neurones de sortie permet de définir le nombre de classes dans une classification.

Il existe plusieurs fonctions d'activation qui peuvent être utilisées. On utilise le plus souvent, en plus de la fonction échelon, les fonctions sigmoïde et tangente hyperbolique qui sont décrites dans la figure 2.9. La fonction sigmoïde est définie par l'équation 2.4 et la fonction tangente hyperbolique, abrégé \tanh , est définie par l'équation 2.5. Le choix de ces différentes fonctions influence le résultat de la classification ou de la régression.

$$f(x) = \frac{1}{1 + e^{-x}} \quad (2.4)$$

$$\tanh(x) = \frac{1 - e^{-2x}}{1 + e^{-2x}} \quad (2.5)$$

Le perceptron multicouche correspond à l'une des premières architectures dites profondes. En effet, il est constitué d'au moins deux couches. Les chercheurs admettent que le nombre de couches est croissant en fonction de la difficulté de la tâche à accomplir [GBC16]. Mais plus le système est complexe (plus il a de couches et de neurones par couche), plus la phase d'apprentissage sera compliquée car elle demandera beaucoup de données et de puissance de calcul. En partant du principe de couches successives, de nombreuses architectures particulières ont vu le jour. Les architectures utilisées dans le Speech Emotion Recognition sont décrites dans le prochain chapitre. Mais avant d'explorer les architectures, nous allons nous intéresser aux méthodes d'apprentissages.

2.3.2 Algorithme d'apprentissage

L'apprentissage des réseaux de neurones correspond à apprendre différents paramètres, dont notamment des poids et des biais. Il est important de dissocier les paramètres des hyper-paramètres. Ces derniers sont définis au préalable de l'apprentissage, par l'évaluateur. Par exemple pour un réseau de neurones, on retrouve comme hyper-paramètres le nombre de couches, le nombre de neurones par couches, le taux d'apprentissage. Il est également essentiel de définir la présentation des données, quelle soit à l'unité, soit une par une ou en lot (batch). La présentation par lot permet généralement d'avoir des systèmes ayant un pouvoir de généralisation plus grand, puisqu'on laisse le système voir plusieurs données avant d'actualiser ces paramètres. De même le nombre d'époques, soit le nombre de fois que le système voit les données, est un hyper-paramètre essentiel dans la configuration de l'apprentissage.

En apprenant les paramètres, le système peut alors s'en servir pour prédire une référence, que ce soit une classe (classification) ou une valeur numérique (régression). Cette

prédiction est obtenue à partir d'un ensemble de caractéristiques, appelés *features* en anglais, d'une donnée. Ces paramètres sont mis à jour au fur et à mesure de la présentation des données d'apprentissage au réseau, afin de trouver des paramètres optimaux qui permettent de maximiser les bonnes prédictions. Il est donc essentiel de mettre en place cet apprentissage sur des données cohérentes avec la tâche visée.

Afin de réaliser cette phase d'apprentissage, il est nécessaire de définir une fonction de coût. La fonction de coût correspond à une fonction que l'apprentissage va chercher à minimiser et qui correspond à une distance entre les valeurs prédites par le système et les valeurs attendues (les références).

Fonction de coût

Il existe de nombreuses fonctions de coût qui permettent de calculer l'écart entre les valeurs prédites et les valeurs de références. Par exemple l'erreur quadratique moyenne, appelée *mean square error* (MSE) en anglais, est une fonction de coût très utilisée qui a pour avantage d'être simple et rapide à calculer. Elle consiste à calculer la différence quadratique entre l'observation, notée O , et la prédiction, notée p , d'une donnée selon l'équation 2.6 :

$$MSE = \frac{1}{n} \sum_{i=1}^n (O - p)^2 \quad (2.6)$$

n correspond aux nombres d'observation lors des différentes époques. Parmi les fonctions de coût les plus utilisées, on peut notamment citer l'entropie croisée [Ste+02], la Classification Temporelle Connexionniste (CTC) [Gra+06] ou encore la somme des carrés des résidus (SCR). C'est donc cette fonction de coût que l'on va dériver pour récupérer la valeur du gradient, que l'on utilise pour apprendre un système automatique.

2.3.3 L'initialisation et ses enjeux

Comme nous l'avons dit précédemment, l'initialisation des poids d'un modèle neuronal est très importante pour garantir de bonnes performances et une convergence rapide du système. Nous allons détailler dans cette section, quelques unes des techniques d'initialisation des poids utilisées de nos jours.

Initialisation aléatoire

La plus simple de ces initialisations est d'utiliser des valeurs de poids tirées au hasard comprises entre -1 et 1. Facile à mettre en place, elle ne favorise pas un apprentissage efficace. Cependant elle est toujours utilisée de nos jours pour sa rapidité d'exécution et sa performance toute relative. En fonction de la loi de probabilité utilisée, elle peut conduire à des performances satisfaisantes voir optimales des réseaux de neurones. Ce tirage aléatoire suit généralement des distributions définies par des lois statistiques telles que la loi normale ou la loi de poisson par exemple.

La méthode de Xavier, introduite par Glorot et Bengio [GB10], garde le principe de l'initialisation aléatoire tout en la contraignant. En effet, elle considère des poids aléatoires mais dont la moyenne est égale à 0 et dont la variance doit être constante entre les couches neuronales.

En pratique, il faut effectuer plusieurs apprentissages avec des initialisations différentes, pour trouver plusieurs minima et retenir le système le plus performant. Cette stratégie permet également de quantifier l'impact de l'initialisation sur le système considéré. Il est néanmoins très difficile voir impossible de s'assurer que le minimum trouvé correspond au minimum global. En effet, Bishop démontre dans son livre [C06] que pour un système à n neurones, le nombre de minima locaux est de $2^n n!$.

Il existe d'autres méthodes qui ne s'appuient pas sur une initialisation aléatoire, mais sur un apprentissage préalable de poids pour le réseau considéré.

Pré-apprentissage

La méthode de transfert d'apprentissage (transfer learning en anglais) est une méthode de plus en plus utilisée [PY10; WKW16] qui propose que les poids des neurones soient appris en amont de l'apprentissage du système sur la tâche visée. Ce premier apprentissage se fait sur des données plus nombreuses et ont pour but de mieux représenter les données utilisées, sur une tâche proche de celle visée. Par exemple, pour effectuer de la reconnaissance d'entités nommées, on va dans un premier temps apprendre à reconnaître automatiquement la parole. Donc on va utiliser les poids appris pour faire de la reconnaissance automatique de la parole pour initialiser les poids d'un système de reconnaissance des entités nommées par exemple. Comme les entités nommées sont formées de mots, ces deux tâches sont différentes mais proches. Ainsi, nous pouvons initialiser notre système avec des poids qui permettent de reconnaître des mots avant de reconnaître des entités

nommées.

Comme nous avons effectué un apprentissage en amont, les poids des neurones ont pu se stabiliser une première fois, afin d'atteindre une première représentation intermédiaire. Les avantages de cette technique sont multiples. Dans un premier temps, elle permet d'accélérer la convergence d'un système. Comme nous avons déjà des poids stables et considérés comme assez proches de leur optimum, on se libère de toute une partie de la phase d'apprentissage pendant laquelle le système cherche une représentation correcte des données. Dans un second temps, l'utilisation de ces poids aide à palier au manque de données concernant la tâche visée. Puisque les poids ont été initialisés sur d'autres données, proche de celle de la tâche considérée, le pré-apprentissage peut être considéré comme une méthode d'augmentation de données.

Elle n'a pas que des avantages. En effet, il faut posséder d'autres données proche de celles de notre tâche. Il faut également que la tâche visée par le pré-apprentissage soit cohérente avec la tâche courante. En tant qu'observateur humain, on peut définir deux tâches comme étant proches alors que le réseau de neurones ne les considère pas en tant que tel, ce qu'on appelle du transfert d'apprentissage négatif. De même, l'utilisation de ces poids pré-appris peut aboutir à des situations de sur-apprentissage, dont nous parlerons dans une prochaine section.

2.4 Les architectures des réseaux dans le traitement du signal

Dans le cadre de cette thèse, nous travaillons sur les informations contenues dans le signal. Cependant l'utilisation du perceptron multi-couche (MLP) n'est pas la plus performante des approches pour représenter les données de type parole. Il existe d'autres architectures de couches neuronales bien plus pertinentes avec ces données. En effet, le principal inconvénient du MLP, c'est que les segments de parole sont considérés un à un, il est donc difficile de gérer les relations temporelles par exemple : on ne prend pas en compte le contexte. Le passé et le futur n'influencent pas directement les systèmes d'apprentissage. Une entrée est donc considérée seule et sans contexte.

Or nous savons que la parole est caractérisée par l'entièreté de sa séquence. Si nous faisons un parallèle avec un texte, une lettre seule ne veut rien dire. Accompagnée d'autres lettres, elles forment un mot. Ces mots sont eux-mêmes agencés avec d'autres mots pour donner un sens à une phrase.

Comme le contenu d'un segment de parole à un instant t dépend des instants précédents, il est nécessaire d'utiliser des réseaux permettant de prendre en compte les segments précédents voir les segments suivants. Dans cette section, nous allons décrire les architectures neuronales permettant de conserver l'historique des états précédents, qui sont par conséquent utilisées dans le traitement de la parole.

2.4.1 Réseaux neuronaux convolutifs

Les réseaux neuronaux convolutifs (Convolutional Neural Network en anglais), abrégés CNN [Le +89] sont prépondérants dans l'analyse d'image. En effet les CNNs traitent des données représentées en n dimensions comme des images (disposition des pixel en 2D), permettant des résultats probants sur des tâches de reconnaissance d'écriture manuscrite de chiffres et de nombres [Lec+98], de reconnaissance d'objets au sein d'une image [TKT18], ou de reconnaissance de visages [LL16]. Ils sont également utilisés en parole, lorsque l'on considère le signal comme un spectrogramme par exemple [Abd+14]. Récemment, cette architecture a été utilisée pour répondre à des tâches concernant la reconnaissance d'émotions, aussi bien à partir d'images de visage [Pit+17; Meh20] qu'à partir de signaux de parole [ZQR16].

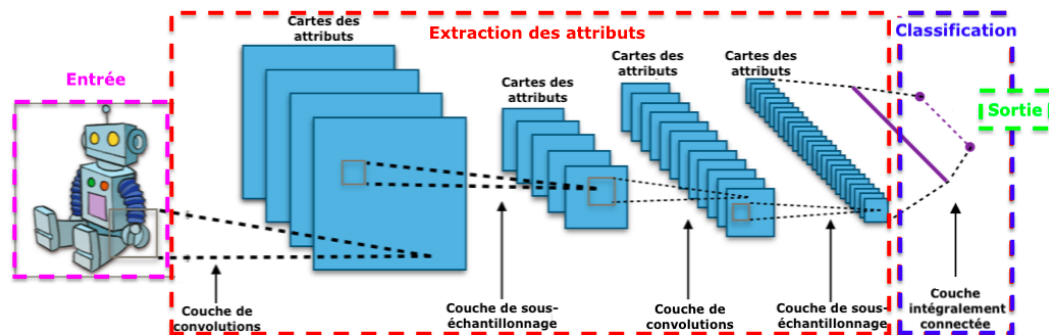


FIGURE 2.10 – Représentation schématique du traitement d'une image par un réseau de type convolutif. Image provenant de Wikipedia.

Comme nous l'avons vu précédemment, ils sont pertinents dans le traitement des images, puisque ces dernières se représentent sous la forme d'une addition de tableaux de 2D (trois pour des images RGB, quatre si on considère la transparence de l'image, un seul pour une image en noir et blanc). Le CNN va découper les tableaux en fenêtres. Il va ensuite chercher ces fenêtres dans les nouvelles données qui lui sont présentées, il va donc opérer un filtrage. Ce filtrage est effectué pour chaque fragment de la donnée.

On peut décrire le fonctionnement d'un réseau convolutif par une succession d'étapes, présentée dans la figure 2.10 :

- l'étape de convolution consiste à faire glisser le filtre sur l'ensemble du tableau et à calculer le produit de la convolution entre chaque fenêtre et le filtre. Cette étape est répétée pour couvrir tout le tableau et utilise un filtre de même taille pour chaque fenêtre.
- l'étape de sous-échantillonnage (pooling en anglais) va permettre de réduire la taille totale du tableau tout en conservant les informations pertinentes pour le système. Différentes méthodes de sous-échantillonnage existent mais les plus usitées sont les fonctions mathématiques de type minimum, maximum ou moyenne du tableau. Cette étape est importante, puisqu'elle permet de garantir une réduction du nombre de paramètres, qui peut augmenter très rapidement avec les couches de convolution.
- optionnellement, on peut ajouter des couches denses.

Le CNN a pour avantage de couvrir la plupart du temps l'intégralité des données pour retrouver une caractéristique. Ainsi, même si cette dernière n'est pas présente au même endroit dans les nouvelles données, elle sera retrouvée. De plus, grâce à l'étape de pooling, le nombre de paramètres se voit fortement réduit, ce qui permet d'obtenir de bonnes performances pour un temps et un usage mémoire réduit.

Cependant le CNN a besoin de séquences d'entrée de taille fixe, ce qui n'est pas très adapté quand on pense à de la parole.

2.4.2 Réseaux Neuronaux Récurrents

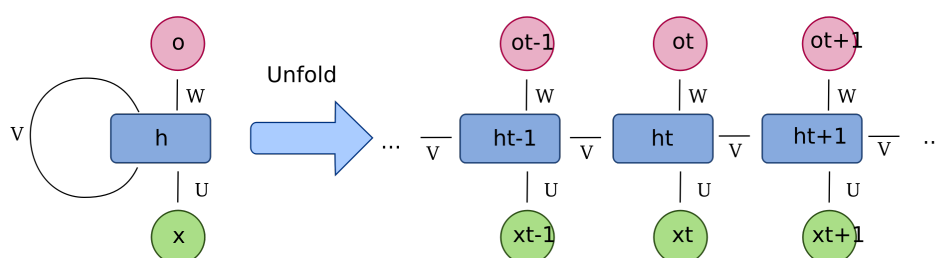


FIGURE 2.11 – Représentation schématique d'un réseau neuronal récurrent. La partie gauche correspond à une récurrence sur une couche entière h . La partie droite explicite la récurrence au niveau de la couche h . Image provenant de Wikipedia.

Les réseaux neuronaux récurrents (RNN) [Jor86] décrivent une architecture neuronale qui est très utilisée de nos jours pour résoudre de nombreuses tâches. En effet, ils permettent de modéliser des séquences (que ce soit des segments de parole, ou des vecteurs de descripteurs par exemple) dont la taille est variable, en présentant toutes les données de manière séquentielle au système. Il est donc possible de faire correspondre plusieurs entrées à plusieurs sorties (appelé *many-to-many*) ainsi que plusieurs entrées à une seule sortie (appelé *many-to-one*).

Ils permettent également de prendre en compte l'aspect chronologique, ou antéchronologique si nécessaire, de la séquence et met en avant les dépendances temporelles dans les séquences. Pour modéliser cet aspect mémoriel, une deuxième entrée est ajoutée au neurone, correspondant à la sortie précédente dans la séquence, comme explicité dans le schéma 2.11. Cette boucle va donc permettre de conserver des informations entre les différentes itérations du système.

Même si en théorie leur mémoire devrait pouvoir contenir tous les éléments de la séquence visualisée, ce n'est pas le cas dans les faits. Bengio et al. [BSF94] ont montré que ces systèmes ont beaucoup de mal à modéliser des dépendances éloignées. On pourra difficilement modéliser une dépendance entre un début de conversation et sa fin si la séquence est trop longue par exemple.

C'est pour pallier ce problème qu'un autre type de réseau récurrent a été développé.

2.4.3 Réseaux long-short term memory

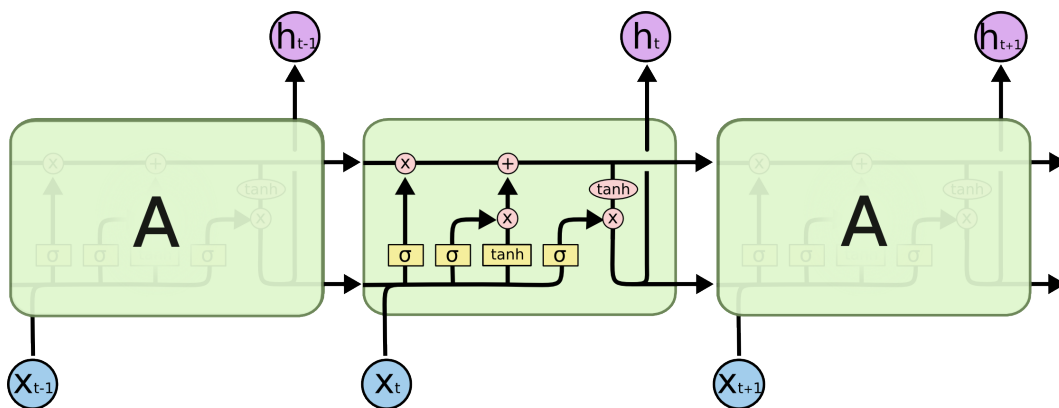


FIGURE 2.12 – Représentation schématique d'un réseau récurrent à mémoire court et long terme (LSTM). On observe qu'il y a deux entrées à l'unité neuronale : l'état de la cellule (cell state) en haut et la sortie classique d'un neurone en bas. Image provenant de <http://colah.github.io/posts/2015-08-Understanding-LSTMs/>.

Les réseaux *long-short term memory* (LSTM) [HS97] ont été créés pour répondre à la perte de la mémoire longue des RNN. Présentés dans la figure 2.12, ils sont spécialisés dans la modélisation de dépendances éloignées dans une séquence, en fonctionnant avec une mémoire interne à chaque neurone (*cell state*) qui permet de modéliser des dépendances à des instants éloignés dans la séquence.

Lorsque la sortie de l'unité neuronale précédente h_{t-1} arrive dans l'unité courante t , elle va passer par trois portes, utilisant des sigmoïdes et des tangentes hyperboliques. La première porte, *la porte de l'oubli*, est en charge de la suppression d'informations non pertinentes stockées dans la mémoire interne (*cell-state* en anglais). La deuxième porte, la porte d'entrée, va décider des nouvelles informations à ajouter dans la mémoire interne. Elle est alors mise à jour avec les informations ajoutées et retirées. La dernière porte, la porte de sortie, est en charge de la sortie effective. Elle est calculée en fonction de la mémoire interne mais aussi de l'état courant en utilisant une tangente hyperbolique pour forcer les valeurs entre -1 et 1 puis une dernière sigmoïde afin de ne transmettre que les informations activées.

Cette architecture existe également de façon bidirectionnelle : la séquence est lue dans l'ordre chronologique et antéchronologique. Pour avoir ces deux directions, on multiplie le nombre de couches du système par deux : une prenant les informations dans l'ordre et une autre prenant les informations dans l'ordre inverse. Enfin les sorties des deux couches sont concaténées pour donner la sortie finale de chaque élément. Cela permet de mettre en évidence les dépendances passées et futures.

Cette architecture est pertinente pour toutes les données qui ont besoin de contexte. C'est pour cela qu'elle est très utilisée dans le domaine du traitement de la parole notamment. Cependant, le temps d'apprentissage est long puisqu'il y a beaucoup de paramètres à apprendre et qu'il est difficile de paralléliser les calculs au vu de son caractère séquentiel. De plus, dans le cas du LSTM bidirectionnel, nous avons besoin de la séquence en entier pour avoir une prédiction, ce qui n'est pas forcément compatible avec du temps réel.

2.4.4 Encodeur Décodeur

Conçu dans les années 2010, l'encodeur décodeur (Encoder-Decoder en anglais) [Cho+14] se compose, comme son nom l'indique, de deux parties distinctes. Ces deux parties sont utilisées conjointement pour prédire une ou plusieurs sorties y_i comme illustré par la figure 2.13.

Un encodeur correspond à un réseau de neurones récurrent qui transforme une séquence

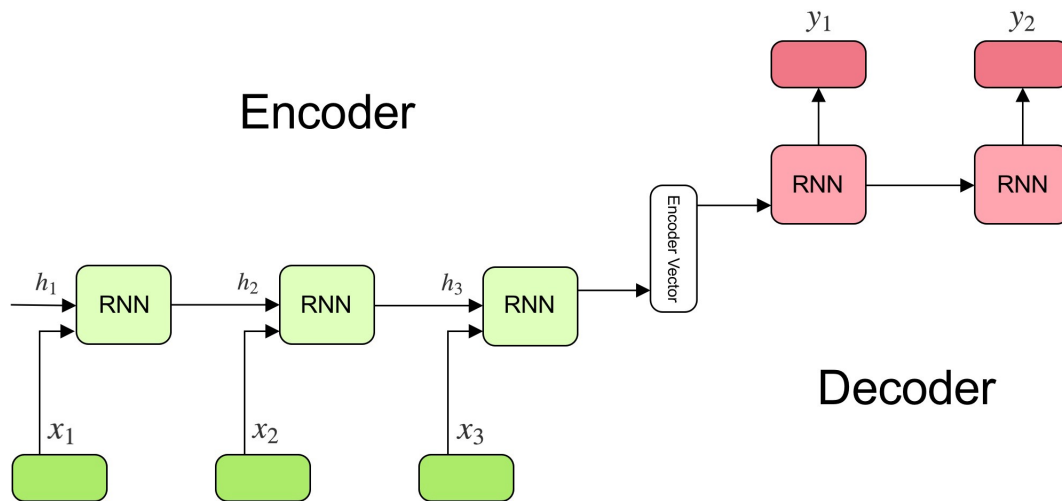


FIGURE 2.13 – Représentation schématique d'un système encodeur décodeur. x_i correspondent aux entrées, y_i aux sorties. Image provenant de <https://towardsdatascience.com/understanding-encoder-decoder-sequence-to-sequence-model-679e04af4346>.

d'entrée X en une représentation vectorielle de taille fixe X' . Une fois cette première transformation effectuée, le décodeur, correspondant lui aussi à un RNN, transforme cette représentation intermédiaire X' en une séquence de sortie Y . Cette transformation finale est calculée en fonction de la distribution de probabilité de toutes les sorties possibles $P(Y|X')$.

Il existe un cas particulier de cette architecture, que l'on nomme les auto-encodeurs. Ce sont des réseaux qui possèdent exactement le même nombre de neurones sur leur couche d'entrée et leur couche de sortie. L'objectif de ces réseaux est d'avoir une sortie la plus proche possible de l'entrée. Ils permettent notamment de débruiter des données d'entrée ou de créer des nouvelles données dans le cadre de l'augmentation de données.

L'encodeur décodeur a notamment prouvé sa pertinence dans des tâches de reconnaissance de la parole [Chi+18], dans des tâches de NLP [Hu19] ou dans des tâches de traduction [Cho+14] par exemple.

Néanmoins cette architecture peut avoir des difficultés à modéliser les dépendances qui sont trop éloignées dans la séquence courante. Un moyen de contourner ce problème est d'utiliser des mécanismes d'attention [Bah+16]. En effet, le modèle peut concentrer son attention sur les parties qu'il juge pertinentes, où qu'elles soient dans la séquence d'entrée.

Mécanismes d'attention

Les mécanismes d'attention ont d'abord été utilisés dans le cadre de la traduction automatique [LPM15], avant d'être rapidement étendus à toutes tâches utilisant des séquences.

Les mécanismes d'attention apportent plusieurs concepts aux encodeurs décodeurs. Tout d'abord, ils étendent les informations contenues dans les états cachés : au lieu de ne fournir que le dernier état caché de l'encodeur, l'ensemble des états cachés pour chacune des entrées précédentes est fourni.

De plus, le vecteur de contexte est pondéré en fonction de la pertinence des informations qu'il contient pour chacun des états émis par l'encodeur. Cette pondération est produite par un score d'attention, compris entre 0 et 1. De ce fait, le vecteur de contexte est composé de la somme des états cachés calculés par l'encodeur pondérés par les scores d'attention.

Enfin le vecteur de contexte devient dynamique : il est recalculé afin de mieux cibler les informations pertinentes dans la séquence d'entrée.

2.4.5 Transformers

Les Transformers sont issus des travaux de Vaswani et al. [Vas+17]. Il s'agit d'un modèle qui utilise une architecture encodeur décodeur mais qui met en place des mécanismes d'attention différents. Elle correspond à un empilement de plusieurs encodeurs et décodeurs. Le nombre d'encodeurs et de décodeurs est identique au départ. Maintenant ce n'est plus forcément le cas.

Dans le cas des Transformers, un encodeur est composé d'un bloc d'auto-attention (*self-attention*), suivie d'une couche linéaire. Le décodeur est lui aussi complété par une couche d'attention. De plus, les Transformers ont introduit les mécanismes d'attention à plusieurs têtes (*multi-head attention*).

Ils sont très performants mais également très coûteux à l'apprentissage. En effet, ils sont composés de beaucoup de paramètres du fait de l'architecture, ce qui se traduit par un long temps de convergence, et un besoin de très grandes quantités de données. Cependant ces inconvénients peuvent être en partie gommé par un apprentissage parallélisable. Ils ont notamment contribué à la création de modèles pré-appris.

2.5 Modèles pré-entraînés

La méthode de transfert d'apprentissage (*transfer learning* en anglais) [GBC16] dans le paradigme de l'apprentissage profond est une méthode d'apprentissage automatique dans laquelle un réseau de neurones entraîné pour une tâche ou un domaine est réutilisé partiellement ou entièrement comme point de départ pour entraîner ou affiner un réseau de neurones sur une autre tâche ou un autre domaine, comme nous l'avons vu précédemment. Cette méthode permet notamment de limiter l'impact du manque de données sur des réseaux de neurones. Elle est largement utilisée dans l'analyse d'émotions, où de grandes bases de données sont utilisées pour pré-apprendre les poids des réseaux, conduisant à de meilleures capacités de généralisation compte tenu des données d'entraînement limitées [DM18].

Parmi les techniques de transfert d'apprentissage, l'apprentissage auto-supervisé des représentations de la parole ou du langage (*self-supervised learning of speech or language representations*) a émergé ces dernières années avec l'introduction du modèle BERT [Dev+19].

2.5.1 Représentation linguistique

Parmi les représentations linguistiques, les plongements de mots (*word embedding*) sont l'une des représentations continues de données textuelles les plus populaires. Ces plongements sont des représentations vectorielles d'un mot particulier. Pour les calculer, on peut utiliser Word2vec [Mik+13] qui est un des modèles les plus utilisés pour apprendre des plongements de mots dans les tâches d'analyse des sentiments telles que la polarité ou la classification des états émotionnels [Pas+19; DM18]. Les plongements de mots obtenus avec Word2vec sont statiques : un mot a toujours la même représentation vectorielle, quel que soit le sens du mot et le contexte de son apparition. Ce qui est problématique pour les mots polysémiques par exemple, qui sont fréquents en français [Pus+96].

BERT

D'autres modèles sont sortis ces dernières années, comme BERT [Dev+19] (entraîné sur des données anglaises puis sur des données multilingues) ou Ernie [Zha+19b] (entraînés sur des données chinoises ou des données anglaises) qui est dépendant du langage et du contexte. Ces modèles ont besoin de beaucoup de données pour être entraînés : ils utilisent notamment le corpus de Wikipédia totalisant plus de 2 500 millions de mots et le Book Corpus [Zhu+15] totalisant plus de 800 millions de mots.

BERT est l'acronyme de *Bidirectional Encoder Representations from Transformers*. Il existe deux modèles BERT disponibles, le *BASE* et le *LARGE*. Comme leur nom l'indique, ils sont composés d'un nombre plus ou moins important de blocs d'encodeurs de type Transformers pour un total de 110 millions de paramètres pour le modèle *BASE* et 340 millions de paramètres pour le modèle *LARGE*.

Ce modèle est pré-appris sur deux tâches distinctes : la tâche *Masked LM* où certains mots sont masqués dans les données et le système apprend à les prédire, ainsi que la tâche *Next Sentence Prediction* où le système apprend à prédire la phrase suivante. Une fois ce pré-apprentissage effectué, le modèle BERT est appris. Ces deux étapes sont illustrées par la figure 2.14. Dans l'article de Devlin et al. [Dev+19], 11 tâches de NLP ont été effectuées pour montrer les performances du modèle BERT ainsi appris.

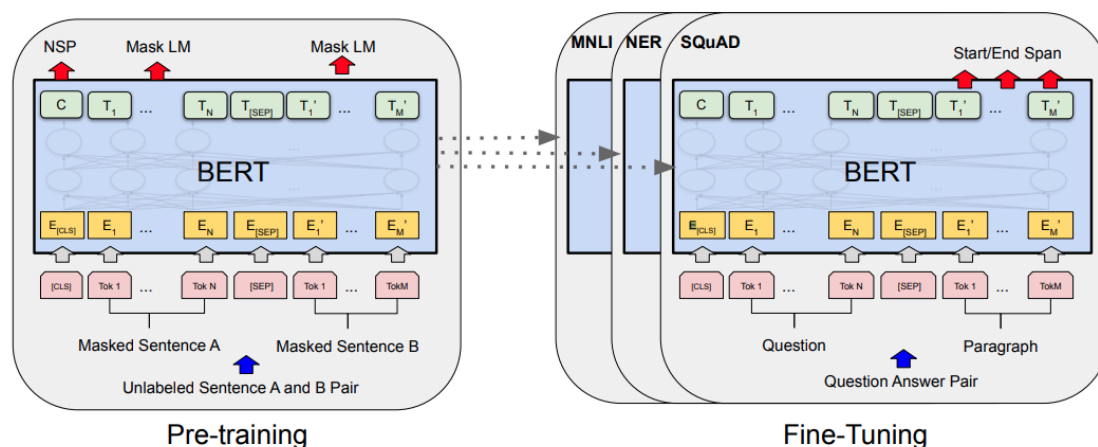


FIGURE 2.14 – Représentation graphique des deux étapes permettant d'obtenir le modèle BERT. Issue des travaux de Devlin et al. [Dev+19]

Ce modèle a montré son efficacité sur certaines tâches, par exemple les tâches de traitement du langage naturel (NLP) telles que la classification de phrases, la similarité de texte, ou le classement par pertinence [Liu+19a; You+18; Yan+19]. De même cette approche montre de très bons résultats dans les domaines de l'ASR [Kah+20; Liu+20] ou de traduction depuis la parole [Ngu+20].

CamemBERT

Si nous nous replaçons dans notre contexte, nous travaillons avec des données françaises. Nous nous sommes donc intéressés aux dérivées de BERT adapté pour la langue

française. CamemBERT [Mar+20] et FlauBERT [Le+20] sont deux modèles appris sur des données françaises. Ces deux modèles sont accessibles^{2,3}, permettant de mettre en place des représentations linguistiques adaptées à la langue française.

CamemBERT est appris en utilisant le système RoBERTa [Liu+19b], un système BERT amélioré avec moins d'étapes d'apprentissage mais de plus gros lot d'apprentissage et de plus gros volumes de données. Cet apprentissage est réalisé sur les données OSCAR [OSR19]. Ce modèle donne des résultats plutôt performants sur plusieurs tâches : l'étiquetage morphosyntaxique (*part of speech* ou POS), l'analyse des dépendances sémantiques, la reconnaissance des entités nommées (NER) et l'inférence du langage naturel (NLI).

FlauBERT est très similaire à CamemBERT. Leurs principales différences proviennent des corpus d'entraînement et des pré-traitements qui sont différents.

Ces modèles linguistiques ont inspiré d'autres chercheurs qui ont étendu le concept à la représentation acoustique.

2.5.2 Représentation acoustique

Récemment, wav2vec [Sch+19] et Audio ALBERT [Chi+20] ont été introduits dans le domaine de l'ASR et de l'identification du locuteur comme les premières approches pré-entraînées pour extraire des représentations acoustiques contextuelles de signaux sonores bruts.

Wav2vec est entraîné sur la tâche de prédiction des futurs échantillons à partir d'une analyse de la fenêtre courante. Il est composé de deux réseaux de neurones convolutifs distincts : le réseau dit *encodeur* et celui dit *contexte* utilisés ensemble. Comme nous pouvons le voir sur la figure 2.15, les entrées X sont d'abord replacés dans un espace latent et transformées en Z . Puis le réseau de *contexte* va transformer ces représentations Z en C en combinant plusieurs pas de temps de l'encodeur précédent pour ajouter du contexte à la représentation finale.

Depuis la publication du modèle BERT, d'autres modèles pré-entraînés permettant d'obtenir une représentation acoustique ont été proposés en utilisant les blocs de Transformers et la technique de maskage, comme le modèle récent Wav2vec 2.0 [Bae+20]. Il existe également d'autres modèles tels que Mockingjay [Liu+20] ou HuBERT [Hsu+21].

2. <https://camembert-model.fr/>

3. <https://github.com/getalp/Flaubert>

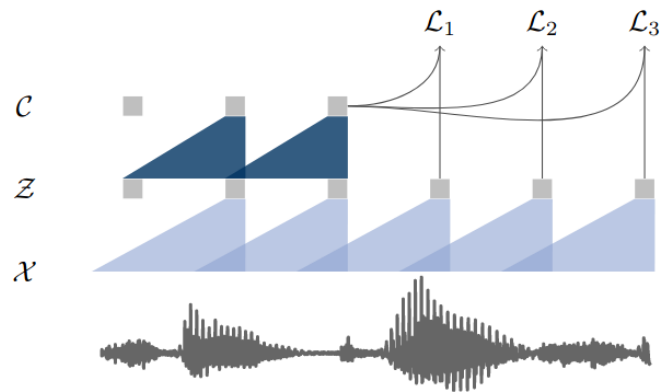


FIGURE 2.15 – Wav2vec : Schéma du pré-entraînement à partir des données audio X qui sont encodées avec deux réseaux de neurones empilés. Le réseau *encodeur* donne la représentation Z et le réseau de *contexte* donne la représentation C . Issu de [Sch+19]

2.6 Conclusion

Dans ce chapitre, nous avons défini les grands principes de l'apprentissage automatique notamment appliqués pour le traitement de la parole. Nous avons vu un éventail de possibilités permettant à une machine d'apprendre à catégoriser des données. Avec l'émergence des réseaux de neurones dans le traitement de la parole, nous nous sommes focalisés sur leur fonctionnement et la mise en place de leurs paramètres afin d'obtenir des systèmes performants.

Dans le chapitre 3, nous allons explorer plus en détail les différentes architectures et paramètres d'apprentissage qui sont mis en place pour la reconnaissance automatique de l'émotion depuis la modalité vocale.

RECONNAISSANCE AUTOMATIQUE DES ÉMOTIONS : CORPUS ET MÉTHODES

«Si vous voulez être libre de vos émotions il faut avoir la connaissance réelle, immédiate de vos émotions.» (Arnaud Desjardins, 1925-2011)

3.1 Le domaine de la reconnaissance automatique des émotions

Le domaine de la reconnaissance automatique des émotions, ou Speech Emotion Recognition (SER) se concentre sur les tâches de reconnaissance des différents états émotionnels d'un ou de plusieurs locuteurs. Ce domaine est en pleine expansion grâce notamment à l'utilisation de nouveaux systèmes neuronaux empruntés à d'autres domaines du Machine Learning. Afin de mettre en place des expérimentations sur ces tâches de reconnaissance et de caractérisation des émotions, il est important de mettre en place des données pertinentes sur lesquelles s'appuyer, des méthodes pour représenter efficacement ces données ainsi que de mettre en place des systèmes d'évaluation robustes permettant de comparer les différentes expérimentations.

3.2 Les corpus existants

Aujourd'hui, les tâches de reconnaissance d'émotion sont principalement traitées en tant que tâches supervisées : le système apprend à reconnaître des étiquettes émotionnelles associées à des segments de parole à partir de corpus de parole annotés en émotions.

3.2.1 Les différents types de corpus

Corpus actés

Comme nous l'avons vu dans le premier chapitre de cette thèse, la caractérisation de l'état émotionnel d'une personne peut être assez délicate, puisqu'elle est en grande partie subjective. De plus, il est communément admis que l'expression des états émotionnels sont ponctuels et rares dans la parole. Pour pallier ce problème, on peut construire des corpus dit *actés*. Il s'agit de corpus où l'état émotionnel n'est pas naturel.

On fait alors appel à des acteurs, qui vont simuler des états émotionnels dictés par le responsable de l'élaboration du corpus. Par exemple, les acteurs devront dire le mot *hippopotame* de façon joyeuse, triste puis avec dégoût. Cette pratique est très utilisée pour avoir différents états émotionnels d'un même locuteur, le tout rapidement. Elle est à privilégier quand on recherche l'exhaustivité des états émotionnels chez un sujet. Ces corpus sont aujourd'hui toujours majoritaires. On peut notamment citer les corpus Emo-Db [Bur+05] et DES [Eng+97]. L'inconvénient principal d'un tel protocole est que les émotions obtenues sont prototypiques dans leur manifestation, les rendant difficile à comparer à des émotions dites *naturelles*. De plus, l'utilisation d'acteurs peut se révéler coûteuse.

Corpus induits

Un autre type de corpus utilisé dans la reconnaissance d'émotion est le corpus induit. On utilise différentes méthodes pour induire les états émotionnels que l'on souhaite étudier.

Une des méthodes pour construire ce genre de corpus est d'utiliser un *magicien d'Oz* abrégé WoZ (*Wizard of Oz*). Cette méthode consiste à simuler un dialogue homme-machine, où la machine est en fait complètement commandée par un opérateur humain, qui va simuler une réponse de machine. Ainsi l'humain pense avoir affaire à un serveur vocal ou un robot, alors que c'est l'opérateur qui est responsable des réponses engendrées par le système.

On peut également parler de la recherche de stress par laquelle on peut placer les sujets dans des situations stressantes comme des montagnes russes [HB97] ou une prise de parole en public [Gir+13]. Quant à la tristesse, on peut utiliser des extraits de films par exemple [Sch+10]. Ces émotions sont plus proches des émotions réelles, elles sont donc moins manifestes.

Les corpus actés et induits ont pour avantage d'être créés dans des environnements contrôlés et permettent de rendre les données facilement accessibles. Ces données sont, la plupart du temps, produites à des fins d'analyse. Il est donc facile de recueillir le consentement des participants avant la mise en place de l'expérience, de contrôler la qualité des enregistrements et la cohérence des données en contraignant les participants sur un sujet ou à exprimer un type d'état émotionnel défini. Néanmoins ils ne représentent pas la vie réelle, sont souvent de petite taille et sont difficilement fusionables pour permettre d'avoir une plus grande quantité de données.

Corpus naturels

Les corpus non actés, aussi appelés naturels ou *real-life* sont obtenus directement à partir de données issues de la vraie vie. Ils proviennent d'environnements peu ou pas contrôlés où les participants n'ont pas connaissance de leur implication dans l'expérimentation a priori. On peut travailler par exemple avec des enregistrements provenant de débats télévisés, de reportages, de vidéos internet ou de centres d'appels. Ces données sont donc constituées d'états émotionnels spontanés. Les émotions y sont souvent moins marquées et moins prototypiques que des émotions actées. De plus la parole spontanée contient une grande part de parole neutre non expressive, et la présence d'émotions reste peu fréquente.

Ces corpus sont difficiles à construire et à diffuser : soit les participants doivent être retrouvés et ils doivent donner leurs consentements a posteriori, soit les données doivent être anonymisées pour respecter la réglementation générale sur la protection de données (RGPD). De même, il est difficile d'avoir tous les états émotionnels de tous les locuteurs et de garantir une qualité d'enregistrement identique entre tous les documents. De plus, l'étiquetage de ces données est souvent plus complexe, puisque le cadre expérimental n'a pas été expliqué aux participants.

3.2.2 Les différents types d'acquisition

L'acquisition des enregistrements est un paramètre important dans la catégorisation des corpus. En effet, un ensemble d'enregistrements captés par un micro spécifique ne donnera pas la même qualité acoustique qu'une voix téléphonique.

On retrouve plusieurs types d'acquisition :

— Acquisition en labo : le milieu d'acquisition est contrôlé par les responsables du

corpus. Cela permet une homogénéité entre les différents enregistrements. On peut notamment citer le corpus SEWA [Kos+19].

- Acquisition en studio : les données sont acquises par des studios de radio ou de télévision. Par exemple MSP-Podcast [LB19] regroupe des enregistrements de podcast, réalisés en studio.
- Acquisition en conditions réelles : le milieu d’acquisition est peu contrôlé par les responsables du corpus. Ce genre d’acquisition est prépondérante dans les corpus naturels. On peut citer notamment les corpus comportant des conversations de centres d’appels comme CallSurf [Gar+08] ou Natural [Mor07], de conversations téléphoniques d’urgence [LCC10] ou bien de micro trottoir ou d’enregistrements de conversations dans des lieux publics comme ESLO [Esh+11].

3.2.3 Les différents types d’annotation

Dans le chapitre précédent, nous avons vu différentes théories émotionnelles. La plupart des travaux du domaine se basent principalement soit sur les catégories discrètes d’émotions, soit sur les dimensions émotionnelles, soit une combinaison des deux. Ainsi suivant la théorie utilisée, les données émotionnelles présentes dans les corpus collectés se fait donc généralement soit suivant une annotation discrète, soit sur une annotation continue, en respectant un schéma d’annotation qui lui est propre. Cette variabilité dans les protocoles d’annotation ne facilite pas l’utilisation jointe de plusieurs corpus pour construire des systèmes de reconnaissance d’émotion appris sur de plus grandes quantités de données.

Pour garantir la qualité de l’annotation, plusieurs indicateurs sont à notre disposition, notamment le kappa (qui mesure un accord annotateur sur plusieurs classes) ou le coefficient de corrélation (qui mesure un accord annotateur sur une dimension). Ces deux méthodes sont décrites dans le prochain chapitre.

3.2.4 Synthèse

Une liste non exhaustive des principaux corpus actés et non actés, utilisés dans la reconnaissance automatique d’émotion, est donnée dans le tableau 3.1.

Les corpus les plus utilisés de nos jours sont le Berlin Emotional database, aussi appelé EMO-DB et IEMOCAP. Il s’agit de deux corpus annotés selon des catégories discrètes. EMO-DB [Bur+05] est composé de phrases courtes en allemand prononcées par 10 acteurs

Nom du Corpus	Langue	Tél	Acté	Continue	Domaine
EMO-DB [Bur+05]	Allemand	x	o	x	Mot isolés
DES [Eng+97]	Danois	x	o	x	Mots isolés
INTERFACE [Hoz+02]	Multi	x	o	x	Mots isolés
SUSAS [HB97]	Anglais	x	o	x	Stress induit
IEMOCAP [Bus+08]	Anglais	x	o	x	Conversations Scriptées
CallSurf [Gar+08]	Français	o	x	x	Energie
Natural [Mor07]	Chinois	o	x	x	Energie
Conversation d'urgence [LCC10]	Français	o	x	x	Centre d'urgence
RECOLA [Rin+13]	Français	x	x	o	Vidéo conférence
SEMAINE [McK+12]	Anglais	x	o	o	Conversation SAL
SEWA [Kos+19]	Multi	x	x	o	Commentaire de publicité

TABLE 3.1 – Principaux corpus utilisés dans la reconnaissance d'émotions dans la parole. Chaque corpus est caractérisé par la langue utilisée, si les enregistrements sont issus du domaine téléphonique ou non, s'il s'agit d'un corpus acté ou spontanée et si les émotions sont annotées en continue ou non.

et est annoté en peur, colère, joie, tristesse, dégoût, ennui et neutre. IEMOCAP [Bus+08] est composé de conversations scriptées entre deux acteurs et est annoté en peur, colère, joie, tristesse, dégoût, frustration, surprise, excitation, neutre et *autres* pour toutes les autres émotions. Joué par 10 acteurs également, cette base de données est composée de 12 heures d’audio, ce qui lui permet d’être compatible avec des approches neuronales profondes, bien qu’on préfère généralement travailler avec un plus gros volume de données.

Nous pouvons également citer en particulier le corpus RECOLA et le corpus SEWA. Ces deux corpus sont particulièrement adaptés à notre tâche, puisqu’ils sont annotés selon des émotions continues. RECOLA [Rin+13] est constitué de conversations dyadiques effectuées en visioconférence pendant laquelle les deux participants doivent compléter une tâche qui leur demande de coopérer. La base de données est notamment constituée des 5 premières minutes de l’enregistrement audio des 23 binômes ainsi formés, totalisant 3 heures et 50 minutes d’audio. 6 annotateurs ont mesuré les états de valence et d’activation des participants. SEWA [Kos+19] est constitué de conversations entre deux locuteurs concernant des publicités visualisées en amont. Le corpus est notamment constitué des enregistrements audio de ces conversations réalisées en 6 langues différentes et réunissant 398 participants pour un total de 44 heures. Ces deux corpus sont disponibles pour les membres d’institution de recherche, en faisant des corpus de plus en plus utilisés par la communauté scientifique.

Dans le cadre de cette thèse, nous avons décidé de comparer nos résultats à ceux obtenus avec le corpus SEWA, puisqu’il se rapproche sur plusieurs points de notre problématique.

3.3 Les descripteurs

3.3.1 Descripteurs acoustiques

Afin de valoriser les données que nous avons dans les différents corpus, il est important de mettre en place une transformation pertinente des données brutes en descripteurs, aussi appelés caractéristiques ou *features* en anglais. Les descripteurs proviennent des domaines de l’acoustique notamment du modèle source-filtre de la voix, mais aussi de la musique. Ces descripteurs ont été conçus pour décrire le timbre, l’intonation, le rythme, et l’intensité des signaux de paroles. Les trois derniers éléments sont généralement regroupés sous le terme de prosodie. Le but est d’obtenir les caractéristiques phonatoires et articulatoires,

ainsi que les évolutions prosodiques des locuteurs [Sch86] afin d'extraire les informations linguistiques et para-linguistiques du discours.

Spectrogramme

Un spectrogramme est une représentation du signal audio mettant en avant l'intensité en fonction de la fréquence et du temps. Il s'agit d'une représentation tridimensionnelle du signal audio. Cette transformation est possible en utilisant la transformation de Fourier qui permet de passer de l'espace temporel à l'espace fréquentiel. Les zones des énergies les plus fortes sont appelées des formants.

Cette représentation temps/fréquence peut être utilisée comme une image. Il est alors possible d'utiliser des systèmes de reconnaissance d'images afin de traiter des problématiques touchant au domaine de la parole [Sto+17].

Coefficients Cepstraux en Fréquence Mel

Les MFCCs, Mel-Frequency Cepstral Coefficients en anglais, sont des descripteurs spectraux utilisés très largement en traitement du signal audio, que ce soit en reconnaissance de la parole, en identification du locuteur ou en détection de concepts sémantiques par exemple. Ils sont issus d'un ensemble de traitements qui est appliqué sur le signal audio traditionnellement sur des fenêtres temporelles de 30 ms tous les 10 ms. Afin de modéliser l'évolution temporelle du signal, on utilise les dérivées premières et secondes de ces coefficients. Contrairement aux Linear Predictive Coding (LPC) [RH93], ces coefficients perceptifs sont adaptés à l'audition humaine puisqu'ils suivent l'échelle de perception de Mel. En effet, notre perception des sons n'est pas linéaire : nous percevons plus de différence entre des sons de 1000 et 2000 Hz qu'entre des sons de 7000 et 8000 Hz.

Robustes au bruit, ces coefficients permettent de représenter le spectre de façon compacte en éliminant les redondances entre les différents coefficients.

Descripteurs prosodiques

Dans le domaine du SER, il n'y a pas de consensus sur le meilleur ensemble de descripteurs à utiliser pour effectuer des tâches de reconnaissance d'émotions. C'est pour cela qu'il existe un grand nombre d'ensembles de descripteurs regroupant les différents indices acoustiques qui seront mis en relation avec les émotions. Ces indices sont soit extraits au

niveau de fenêtres de courtes durées (30 ms le plus souvent), soit sur des fenêtres plus longues (plus d'une seconde) pour capturer des phénomènes para-linguistiques.

Traditionnellement, on prend un très grand nombre de descripteurs, comme par exemple l'ensemble de base extrait avec OpenSMILE [EWS10], qui compte 988 descripteurs acoustiques, puis on les filtre pour ne conserver que les plus pertinents par une sélection de descripteurs.

La sélection des descripteurs réduit la dimensionnalité de leur espace, supprime les données redondantes, non pertinentes ou bruitées. Elle apporte des effets bénéfiques directs aux systèmes : l'accélération du temps de traitement puisqu'il y a moins de données à traiter, l'amélioration de la qualité des données et donc de la performance des systèmes. De plus, elle peut permettre de rendre les résultats plus compréhensibles. Cette sélection peut s'effectuer de plusieurs manières :

- selon des méthodes de ranking, en comparant les descripteurs les uns aux autres.
- en utilisant des *wrappers* : en utilisant le modèle de classification et en testant tous les descripteur les uns après les autres. Pour ce faire, on les enlève ou on les ajoute un à un et on sélectionne l'ensemble de descripteurs qui donne au modèle sa meilleure performance.

Néanmoins l'ordre de grandeur des corpus annotés en émotion peut vite poser problème si la dimension des descripteurs est trop importante.

Pour pallier ce problème, de plus petits ensembles de descripteurs, réalisés par des experts du domaine ont été proposés. On peut notamment citer l'ensemble INTERSPEECH 2009 [SSB09] ainsi que *The Geneva Minimalistic Acoustic Parameter Set* (GeMAPS) et sa version étendue (eGeMAPS) [Eyb+16].

L'ensemble INTERSPEECH 2009 contient 384 descripteurs qui ont été utilisés comme référence lors du premier challenge en reconnaissance d'émotions en 2009. Il est composé de 16 descripteurs de bas niveau (Low Level Descriptors en anglais) :

- *zero crossing rate* : taux de passage par 0 du signal sur une fenêtre temporelle donnée.
- *RMS energy* ou énergie moyenne quadratique : modélise la variation de l'énergie du signal à chaque fenêtre d'analyse.
- *F0* ou fréquence fondamentale.
- *Harmonic to noise ratio* ou rapport Harmonique-Bruit : modélise le bruit contenu dans le signal.
- *MFCC* : les 12 premiers coefficients MFCCs sont utilisés.

Sur ces descripteurs sont calculées 12 fonctions statistiques (la moyenne, la variance, le coefficient de Pearson, de dissymétrie, le maximum, le minimum, la position relative, la plage de valeur et la régression linéaire). Soit, in fine, 16 descripteurs et leurs 16 dérivées sur lesquels on applique 12 fonctions pour un total de 384 descripteurs.

GeMAPS est composé de 18 descripteurs de bas niveau représentant des propriétés de fréquence, d'énergie, d'amplitude et des propriétés spectrales sur lesquels sont appliquées des fonctions statistiques pour un total de 62 descripteurs. Nous reprenons la liste complète établie dans l'article de Eyben et al. [Eyb+16] dans le tableau 3.2.

Comme la version courte de l'ensemble minimaliste ne contient aucun paramètre cepstral et très peu de paramètres dynamiques, on ajoute sept LLD (MFCCs, Flux spectral, Bande passante des formants 2,3) pour construire l'ensemble d'extension. En lui appliquant des fonctions statistiques, eGeMAPS contient un total de 88 descripteurs résumés dans le tableau 3.2.

Bag-of-Audio-Words (BoAW)

Comme nous l'avons déjà indiqué, la sélection de la représentation de l'audio est un choix qui va directement influencer la qualité de la reconnaissance des émotions. C'est pour cela qu'il existe de nombreuses représentations, dont les sacs de mots-audio, ou BoAW. Inspiré des sacs de mots utilisés en NLP, il s'agit d'utiliser les LLDs sélectionnés pour former un lexique appelé *codebook* de toutes les valeurs possibles, puis de les coder par un vecteur. Ce vecteur est alors utilisé en tant qu'entrée du système.

Cette solution présente comme avantage de renforcer la robustesse du système, vu que les LLDs en entrée sont en quelque sorte normalisées par ce processus. Ces features sont utilisées dans de nombreuses tâches reliées à la parole : la classification d'évènements sonores [PA12 ; SRS16] ou la détection de plagiat [Liu+10] par exemple. Ils ont également été utilisés en reconnaissance d'émotions continues [SRS16 ; Han+18].

3.3.2 Descripteurs linguistiques

Nous avons vu que lorsque l'on cherche à déterminer l'état émotionnel d'un locuteur à partir d'un enregistrement, l'approche la plus naturelle consiste à extraire des descripteurs acoustiques directement à partir du signal. On peut cependant ajouter des informations linguistiques, syntaxiques, phonémiques et sémantiques. Ces informations peuvent être extraites directement à partir d'une transcription automatique de la parole.

Paramètres	fonctions appliquées	nb paramètres
Paramètres fréquentiels : 20 ou 24 paramètres		
Hauteur de la voix (Pitch) F0	M, CV, P, RP, SLOPE	10
Micro-variations de F0 de la voix (Jitter)	M, CV	2
Frequence des Formants 1,2,3	M, CV	6
Bande passante (Bandwidth) du premier Formant	M, CV	2
Bande passante des Formants 2,3 *	M, CV	4
Paramètres d'énergie et d'amplitude : 14 paramètres		
Micro-variation d'énergie (Shimmer)	M, CV	2
Énergie perçue (Loudness)	M, CV, P, RP, SLOPE	10
Rapport Harmonique-Bruit (Harmonics-to-Noise Ratio)	M, CV	2
Paramètres spectraux : 22 ou 43 paramètres		
Ratio Alpha	M, CV, MUN	3
Index de Hammarberg	M, CV, MUN	3
Pente spectrale : 0-500Hz	M, CV, MUN	3
Pente spectrale : 500-1500Hz	M, CV, MUN	3
Energie relative des Formants 1,2,3	M, CV	6
Différence harmonique H1-H2	M, CV	2
Différence harmonique H1-A3	M, CV	2
MFCC 1 à 4 *	M, CV, MCVV	16
Flux spectral *	M, CV, MUN, MCVV	5
Paramètres temporels : 6 paramètres		
Taux des pics d'énergie (Rate of loudness peaks)		1
Durée des segments parlés ($F0 > 0$)	M, V	2
Durée des segments non-parlés ($F0 < 0$)	M, V	2
Nombre de zones continues parlées par secondes		1
Intensité sonore *		1

TABLE 3.2 – Résumé des descripteurs de bas niveau (LLDs) utilisés dans les ensembles GeMAPS et eGeMAPS. * signale les descripteurs uniquement disponibles dans eGeMAPS. Légende des fonctions appliquées : **M** moyenne arithmétique, **CV** coefficient de variation, qui correspond à la variance normalisée par la moyenne, **V** variance, **P** percentiles : 20,50 et 80%, **RP** l'intervalle des percentiles 20 à 80%, **SLOPE** moyenne et variance des pente de montée/descente de signal, **MUN** moyenne arithmétique avec les parties sans parole, **MCVV** moyenne arithmétique et coefficient de variation des parties avec paroles.

Les domaines du Sentiment Analysis et de l'Opinion Mining cherchent à identifier une opinion à partir d'un texte écrit. Il y a ici une différence sémantique significative entre opinion et état émotionnel, cependant les approches peuvent se rejoindre. On peut également considérer la transcription automatique comme un élément associé à la parole, et donc y retrouver des marqueurs de l'émotion. Nous détaillons dans cette partie certaines méthodes utilisées dans ces domaines pour transformer du texte (dans notre cas la transcription automatique) en descripteurs.

Représentation en one-hot

La représentation en *one-hot*, comme illustrée sur la figure 3.1, correspond à associer à chaque mot un vecteur de binaires d'une taille fixe. On rassemble tous les mots-types (appelés token en anglais) utilisés dans ce que l'on nomme un vocabulaire. Puis pour chaque mot-type de ce vocabulaire, on associe un vecteur binaire unique. Cette méthode présente l'avantage d'être exhaustive et facile à mettre en œuvre, cependant la représentation est très volumineuse : le vecteur sera de la taille du vocabulaire et elle contiendra principalement des zéros.

Corpus : Les hippopotames passent presque toute leur vie dans l'eau cependant les hippopotames ne savent pas nager
 Vocabulaire : Les hippopotames passent presque toute leur vie dans l'eau cependant ne savent pas nager
 Taille du vocabulaire : 15

Représentation en one-hot

Les	0000000000000001
Hippopotames	0000000000000010
passent	0000000000000100
....	
nager	1000000000000000

FIGURE 3.1 – Transformation d'un corpus en représentation one-hot.

Cette représentation permet de réaliser des opérations numériques directement sur les vecteurs *one-hot*. Cependant elle ne porte aucune information sur le contexte (la position du mot-type dans la séquence), la sémantique ou sur le nombre d'occurrence de chaque mot-type.

Représentation statistique

Parmi les représentations statistiques des données textuelles, les plus courantes et les plus mises en place sont les méthodes TF (Term Frequency) et TF-IDF (Term Frequency-

Inverse Document Frequency) qui prennent en compte la distribution des tokens dans les documents. Le TF est défini comme “the frequency of occurrence of the terms in the document or query texts” [SB88]. C’est-à-dire que le TF d’un mot-type t présent dans un document d correspond au nombre d’occurrences de t dans d divisé par le nombre total de tokens dans le document (Eq. 3.1).

$$\text{TF}(d, t) = \frac{\text{Nombre de } t \text{ dans } d}{\text{Nombre de tokens dans } d} \quad (3.1)$$

Toujours selon Salton [SB88], le facteur IDF varie “inversely with the number of documents n to which a term is assigned in a collection of N documents. A typical idf factor may be computed as $\log N/n$ ”. Le facteur IDF correspond à l’inverse du rapport entre le nombre de documents où le token t apparait sur le nombre total de documents N (Eq. 3.2).

$$\text{IDF}(t) = -\log \left(\frac{N}{\text{Nombre de documents où } t \text{ est présent} + 1} \right) \quad (3.2)$$

Le TF-IDF est alors défini comme le produit des deux termes précédents comme indiqué dans l’équation 3.3. Pour permet d’éviter les cas où aucun document ne contiendrait le token t et donc la division par zéro. Cette approche permet que les tokens qui sont soit trop fréquents, soit présents dans tous les documents, soient moins mis en avant par la représentation.

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t) \quad (3.3)$$

Ces méthodes statistiques ont fait leurs preuves [MF09 ; Cam+13 ; PRJ20], même si on leur préfère maintenant des méthodes qui intègrent des aspects sémantiques notamment.

Plongement de mots

Les plongements de mots (word embeddings en anglais) ont été présentés par Bengio et al. [Ben+03]. A l’origine, un modèle neuronal est appris pour une tâche de reconnaissance automatique de la parole sur un grande nombre de données. Une fois le modèle entraîné, on peut prendre le réseau dans l’autre sens et supprimer les couches basses (proches de l’audio) et figer les poids afin de créer un extracteur capable de calculer des embeddings, c’est-à-dire les représentations internes du réseau de neurones, à partir d’une entrée textuelle. Ces embeddings permettent d’avoir accès à une nouvelle représentation des mots

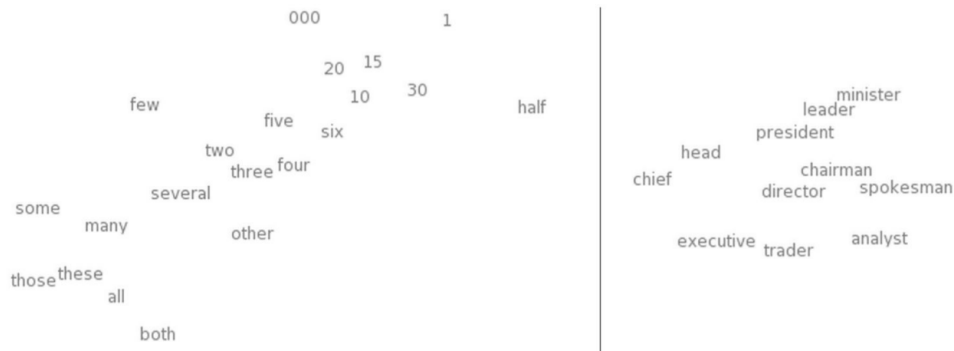


FIGURE 3.2 – Représentation de plongements de mots en deux dimensions. Les mots de gauche correspondent au regroupement des termes associés au numérique. Les mots de droite correspondent aux termes associés à l’emploi. Issue des travaux de Turian et al. [TRB10].

dans un espace dense et à valeurs réelles. On projette les mots dans un espace de faible dimension, tout en isolant ensemble les mots qui ont des similarités sémantiques et syntaxiques [Gha17]. Cela permet d’associer à chaque token un vecteur de valeurs réelles, de dimension bien inférieure à celle utilisée pour la représentation *one-hot*. Chaque vecteur est ensuite inscrit dans un dictionnaire, et on peut alors remplacer chaque mot par le vecteur le représentant.

Leur utilisation et leur pertinence a été démontré dans de nombreuses tâches, notamment des tâches de TALN : l’étiquetage morphosyntaxique, la reconnaissance d’entités nommées, la détection de mention [TRB10; BGL14] et de compréhension de la parole [Mes+13; Yao+14; LL16].

Parmi les plongements lexicaux les plus utilisées, on trouve Word2vec [Mik+13] et GloVe [PSM14]. Plus précisément, les plongements de Word2Vec sont appris soit avec un algorithme de sac de mots continus (continuous bag of words, CBOW) qui prédit un mot sachant son contexte lexical, soit l’algorithme Skip-gram qui prédit les mots du contexte sachant le mot d’entrée.

Il est utile de noter que ce genre de représentation peut être visualisée dans un espace plus restreint, typiquement en deux dimensions en utilisant une réduction de dimensions (par exemple une Analyse en Composantes Principales). Ainsi les expérimentateurs peuvent observer les rapprochements sémantiques ou syntaxiques détectés par le système, comme l’illustre la figure 3.2.

3.4 Évaluation des performances

De nombreuses métriques ont été utilisées au fur et à mesure de l'avancée du domaine pour évaluer les performances des systèmes de reconnaissance d'émotions dans la parole suivant le type de tâche : classification ou régression.

Pour les tâches de classification les métriques d'*accuracy*, de précision et de rappel pondérés ou non pondérés sont les métriques les plus utilisées traditionnellement. On peut notamment se référer aux challenges INTERSPEECH en émotions de 2009 et 2011 [SSB09; Sch+11] qui utilisent le rappel moyen non pondéré comme métrique afin de compenser le biais de données souvent très déséquilibrées en nombre d'instances par classe.

En ce qui concerne les tâches de régression, l'erreur quadratique moyenne est une métrique efficace qui a été largement utilisée avant que des campagnes d'évaluation [Rin+17] ne standardisent l'utilisation du Coefficient de Corrélacion de Concordance (CCC) comme la mesure de référence pour l'évaluation de systèmes de reconnaissance de l'émotion continue.

3.4.1 Tâche de classification

Matrice de confusion et scores associés

La matrice de confusion est une matrice permettant de mesurer la performance d'un système de classification. Elle permet d'associer pour chaque classe de référence, les classes prédites par le système. À chaque segment émotionnel est associée une classe qui définit l'état émotionnel de référence de la personne. Cette matrice est donc mise en place en tant qu'évaluation lorsque les émotions sont de nature discrètes. Grâce à elle, on peut retrouver les différentes erreurs du système et les quantifier. Un exemple de matrice de confusion est donné par le tableau 3.3. Les lignes correspondent aux références, et les colonnes aux prédictions d'un système.

Cette présentation permet notamment de visualiser si une classe est mieux prédite que d'autres. Il est facile de relever si le système est performant en se basant sur la diagonale, qui regroupe les vrais positifs et donc de calculer l'*accuracy* définie par le nombre de segments correctement prédits sur le nombre total de segments N (eq. 3.4.1), tandis que les autres cases correspondent à des erreurs du système.

$$Ac = \frac{\text{Nb de prédictions vraies}}{N} \quad (3.4)$$

		Prédiction				précision	rappel
		joie	neutre	colère	total		
Réf	joie	90	11	2	103	0.900	0.874
	neutre	4	80	10	94	0.889	0.851
	colère	6	9	20	35	0.625	0.571

TABLE 3.3 – Matrice de Confusion entre trois classes émotionnelles : la joie, le neutre et la colère. Les colonnes correspondent aux prédictions du système et les lignes correspondent aux références. On voit que sur 100 prédictions de la classe joie, seules 90 sont pertinentes et le système a mal prédit 13 segments qui ne devraient pas être dans la classe joie.

Cette matrice de confusion permet également de calculer le rappel et la précision de chaque classe. La précision P_i de la classe c_i , correspond au nombre de segments de la classe c_i correctement prédits parmi tous les segments prédits comme étant de classe c_i . (eq. 3.5). Le rappel R_i de la classe c_i est donné par l'équation 3.6 et correspond au nombre de segments correctement prédits parmi tous les segments de la classe c_i

$$P_i = \frac{\text{nb prédictions vraies}_i}{\text{nb prédictions vraies}_i + \text{nb prédictions fausses}_i} \quad (3.5)$$

$$R_i = \frac{\text{nb prédictions vraies}_i}{\text{nb segments de classe } c_i} \quad (3.6)$$

Bien que pratique, la matrice de confusion ne permet pas de donner un score unique pour le système.

Précision ou rappel pondéré et non-pondéré

Afin d'avoir un score global de la classification des émotions, on peut moyenner les précisions et rappels par classe en pondérant ou non en fonction du nombre de segments par classe. La précision moyenne non pondérée (unweighted average precision, UAP) est obtenu en prenant simplement la moyenne des précisions par classe. La précision moyenne pondérée (weighted average precision, WAP) est calculée en prenant la moyenne des précisions par classe suivant l'équation 3.7 où N est le nombre de segments total et n_i le nombre de segments dans la classe c_i . Pour éviter des possibles confusions entre les métriques de précision et de rappel, comme dans les travaux de Lee et Tashev [LT15] ou de Han et al. [HYT14] (où la définition de l'*accuracy* par classe est ambiguë) nous indiquerons systématiquement sur quelle métrique s'applique la pondération.

$$\text{WAP} = \frac{1}{N} \sum_{i=1}^n n_i \cdot P_i \quad (3.7)$$

La mesure de précision moyenne pondérée permet de calculer la performance d'un système de reconnaissance, mais elle ne rend pas bien compte des différences de performance entre les classes. Notamment si le corpus est déséquilibré et que le système a tendance à favoriser la classe majoritaire, on aura toujours des taux de WAP importants. Dans l'exemple du tableau 3.3, la classe colère n'est pas bien reconnue. Pourtant si on calcule la UAP ($\frac{0.9+0.851+0.571}{3}$) on trouve 0,774 soit une performance élevée de 77,4% de bonne prédiction. Or les prédictions de la classe colère ont une précision de 57,1%. L'utilisation d'une moyenne non pondérée permet de mettre en évidence ce phénomène et d'apporter une métrique plus équitable entre les différentes classes.

Nous pouvons également utiliser le rappel moyen non-pondéré, comme lors des challenges INTERSPEECH [SSB09; Sch+11]. Il est calculé selon l'équation 3.8 et permet de déterminer combien d'éléments pertinents ont été retrouvés.

$$\text{UAR} = \frac{1}{n} \sum_{i=1}^n R_i \quad (3.8)$$

F-mesure

La F-mesure, ou F-score en anglais, combine à la fois la précision et le rappel selon l'équation 3.9. Elle est comprise entre 0 et 1. Plus elle est grande, plus le système évalué est performant. Elle peut être calculée par classe sur les précisions P_i et rappels R_i .

$$F = 2 \left(\frac{\text{precision.rappel}}{\text{precision} + \text{rappel}} \right) \quad (3.9)$$

Néanmoins toutes les reconnaissances d'émotions ne se font pas sur des émotions discrètes, il existe donc d'autres indicateurs utilisés pour mesurer la performance des systèmes.

3.4.2 Tâche de régression

Erreur quadratique moyenne

L'erreur quadratique moyenne est une métrique qui est utilisée dans l'évaluation des systèmes de reconnaissance d'émotions continues [Rin+17]. Elle permet de calculer un

score d'accord entre deux séries temporelles, ici les annotations de références et les prédictions du système. Afin que cette métrique soit de la même dimension que les valeurs de référence, on utilise principalement la racine de l'erreur quadratique moyenne, RMSE pour *Root Mean Square Error*. Cette métrique se calcule selon l'équation 3.10 entre les valeurs prédites x_i et les valeurs de références y_i . n correspond au nombre de valeurs de la série temporelle.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - x_i)^2} \quad (3.10)$$

Elle est généralement comprise entre 0 et 1 (en fonction des valeurs x_i et y_i). Comme il s'agit d'une mesure d'erreur, les systèmes à haute performance se rapprochent de 0, ce qui indique un ajustement parfait entre les références et les prédictions. Cette métrique n'est pas exempte d'inconvénients. En effet, elle est très sensible aux valeurs extrêmes et elle est difficilement comparable à d'autres scores calculés sur des valeurs d'ordres de grandeur différents.

Coefficient de Corrélation de Concordance

Le coefficient de corrélation de concordance (CCC) [Lin89] a été établi comme un standard d'évaluation lors notamment des trois précédents Audio/Visual Emotion Challenge and Workshops (AVEC) [Rin+17; Rin+18; Rin+19]. Cette métrique évalue l'accord entre deux séries temporelles selon l'équation 3.11, où x et y sont les deux séries temporelles, dans notre cas la prédiction et la référence. μ_x , μ_y correspondent à leur moyenne, σ_x , σ_y à leur écart-type et ρ le coefficient de corrélation entre ces deux variables aléatoires.

$$CCC = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2 + \epsilon} \quad (3.11)$$

Plus le CCC s'approche de 1 et plus le système est considéré comme performant. À l'inverse, plus le score s'approche de 0 et moins il y a de corrélation entre les prédictions et les références, dénotant un système peu performant. Cette métrique n'est pas définie dans le cas où les deux séries temporelles x et y sont constantes et de mêmes moyennes. On ajoute un ϵ au dénominateur pour pallier à ce problème lors de l'implémentation.

Cette métrique sera utilisée dans les travaux de cette thèse pour évaluer la performance des systèmes de prédiction continue de la satisfaction.

3.5 Fusion de modalités

Afin de pouvoir profiter à la fois des informations acoustiques et linguistiques, il est pertinent de fusionner ces deux modalités pour avoir un système performant et plus robuste [Wöl+13; AR14; Atr+10; Liu+18]. La fusion de modalités est assez vaste : il est possible également d’inclure des informations issues de vidéos ou de capteurs physiologiques, dont nous avons parlé au premier chapitre.

Dans le cadre de cette thèse, nous sommes principalement intéressés par les modalités acoustiques et linguistiques, puisque les autres des modalités ne peuvent pas être récupérées depuis les centres d’appels.

Type de fusion

La fusion peut s’effectuer de différentes manières.

- Fusion des features [Wöl+13; AR14; Atr+10] : La fusion s’opère au niveau des features, on concatène les vecteurs représentant les différentes modalités. Cette méthode augmente le nombre de features en entrée du système et peut donner des résultats très différents en fonction de la stratégie de normalisation des données. En effet, il peut être compliqué pour le système de comprendre que les données représentent deux espaces différents. Donc il est courant de normaliser les données afin de se référer à un seul espace.
- Fusion des modèles [Atr+10; Liu+18] : Plusieurs apprentissages sont faits de façon distincts pour chaque modalité jusqu’à une certaine couche dans le réseau de neurones. Les couches sont alors fusionnées, et l’apprentissage reprend. Plus la fusion arrive tôt et plus le système doit en théorie avoir un bon pouvoir de généralisation.
- Fusion de décision [Wöl+13; Atr+10] : Plusieurs apprentissages sont faits de façon distincts pour chaque modalité. On prend la prédiction de chacun des modèles et on les fusionne. S’il s’agit d’une classification, on peut fusionner par vote majoritaire par exemple. S’il s’agit d’une régression, on peut faire la moyenne des sorties. De plus, on peut facilement mettre plus d’importance sur une des modalités en faisant une moyenne pondérée des sorties.

Models	Modalité	Features	SEWA		
			activation	valence	liking
AVEC 2017 [Rin+17] : Sur les conversations allemandes					
SVR	audio	BoAW [SRS16]	.344	.351	.081
SVR	audio	BoTW	.373	.390	.314
Huang et al. [Hua+17] : Sur les conversations allemandes					
LSTM	audio	eGeMAPS-88	.506	.455	.193
LSTM	audio	IS10	.465	.440	.227
LSTM	audio	Bottle-neck [Fér+15]	.533	.466	
LSTM	audio	Mfcc	.341	.421	
LSTM	texte	BoTW	.451	.518	.473
AVEC 2018 [Rin+18] : Sur les conversations allemandes					
biLSTM-2	audio	eGeMAPS-88	.124	.112	.001
biLSTM-2	audio	Mfcc	.253	.217	.136
Huang et al. [Hua+18] : Sur les conversations allemandes					
LSTM	audio	eGeMAPS-88	.497	.438	.281
LSTM	audio	eGeMAPS-89	.520	.461	.335
LSTM	audio	eGeMAPS-176	.514	.493	.217
LSTM	texte	Word2Vec-300	.597	.600	.454
LSTM-2	texte	BoW			.407
LSTM-2	texte	Word2Vec			.480
LSTM-2	texte	GloVe			.413
AVEC 2019 [Rin+19] : Sur les conversations allemandes et hongroises					
biLSTM-2	audio	eGeMAPS-88	.371	.286	.159
biLSTM-2	audio	Mfcc	.326	.187	.144
Schmitt et al. [SCS19] : Sur les conversations allemandes					
CNN	audio	eGeMAPS-47	.571	.517	
biLSTM-4	audio	eGeMAPS-47	.568	.561	

TABLE 3.4 – Compilation des scores de CCC sur l’ensemble de développement de SEWA sur les 3 dimensions : activation, valence et *liking*. L’acronyme BoAW signifie *Bag-of-audio-words*, BoTW signifie *Bag-of-text-words*, SVR signifie Support Vector Regression [SS04], proche des SVM mais applicable à des problèmes de régression et donc à une annotation continue. Les différents nombres associés aux features de type eGeMAPS dénotent de différentes configurations utilisées autour de ces sets : soit avec une sélection réduite des LLDs (47), soit avec l’ajout d’information sur le locuteur courant (89 et 176).

3.6 Notre référence : AVEC

Si nous nous replaçons dans le contexte de la thèse, nous cherchons à construire et donc à évaluer un système de reconnaissance des émotions continues depuis la parole. Dans ce cadre, nous avons recherché dans la littérature, des systèmes et des expérimentations qui soient comparables à nos recherches. Nous avons choisi de nous référer aux campagnes AVEC, *Audio/Visual Emotion Challenge and Workshop*.

Ce challenge, qui en était à sa huitième itération en 2018, vise à comparer les méthodes de traitement multimédia et d'apprentissage automatique pour l'analyse automatique de la santé et des émotions dans les modalités audio et visuelles. Ces campagnes sont divisées en différents objectifs qui gravitent autour des émotions : de la détection de dépression, de bipolarité, d'état d'esprit ou d'émotions issues de différentes cultures par exemple. Comme ces campagnes s'appuient sur des corpus multimodaux comme RECOLA [Rin+13] et SEWA [Kos+19] notamment, les émotions peuvent être détectées à partir de différentes modalités, notamment depuis la parole et les expressions faciales.

Afin de pouvoir nous comparer à l'état de l'art, nous avons décidé de comparer nos résultats à ceux obtenus sur le corpus SEWA, dont nous avons parlé dans ce chapitre. Ce corpus considère trois dimensions émotionnelles : la valence, l'activation et le *liking* qui correspond à l'appréciation par les participants des clips visionnés en amont. Il est utilisé notamment dans les campagnes AVEC depuis 2017 [Rin+17 ; Rin+18 ; Rin+19] et sert actuellement de baseline dans la communauté. Nous nous intéressons donc à la tâche de régression audio uniquement.

Nous résumons dans le tableau 3.4, différents résultats obtenus qui utilisent la partie Allemande et Hongroise du corpus, soit celle à laquelle nous avons accès. Comme nous pouvons le voir, de nombreux ensembles de features et types d'algorithmes ont été utilisés pour la reconnaissance des émotions depuis la parole pour le corpus SEWA. Pour ce qui est de la modalité acoustique, nous pouvons voir que les systèmes les plus performants ont des scores de 0.571 pour l'activation, 0.561 pour la valence et 0.335 pour le liking. Nous voyons que la plupart des participants ont utilisé des systèmes CNN ou LSTM (détaillés dans le chapitre 5) pour résoudre la régression. En regardant les deux études de Huang et al. [Hua+17 ; Hua+18], nous pouvons remarquer que le choix de l'ensemble de descripteurs influence fortement sur les scores des systèmes de régression. On remarque également une prédominance de l'utilisation des eGeMAPS, qui donne les meilleurs scores.

Pour ce qui est de la modalité linguistique, on peut notamment remarquer que l'on

retrouve principalement l'utilisation de plongements de mots (word2vec et GloVe) et le même type d'architecture que pour la modalité acoustique, avec une variation du réseau LSTM, qui possède deux couches pour les expérimentations sur le liking.

Si on compare les résultats à ceux trouvés par la modalité acoustique, les scores maximum sont assez similaires : 0.597 contre 0.571 pour l'activation, 0.600 contre 0.561 pour la valence. On observe également une nette amélioration pour le liking : 0.480 au lieu de 0.335. En général, on trouve des scores un peu plus élevés en utilisant la modalité linguistique [GS13]. Cela peut être dû au fait que les descripteurs utilisés sont plus adaptés ou que les architectures neuronales sont plus adaptées à ce type de données.

Dans le cadre de cette thèse, nous cherchons à mettre en place des solutions neuronales profondes. Nous avons donc fait le choix de nous comparer aux résultats obtenus par Schmitt et al. [SCS19] correspondant aux lignes rouges du tableau.

3.7 Conclusion

Dans ce chapitre, nous avons résumé les principaux composants de la reconnaissance des émotions depuis la parole, sans oublier la reconnaissance depuis le texte. A partir de ces connaissances, nous avons pu établir un référentiel sur la tâche que nous cherchons à accomplir. En effet, nous avons fait le choix de nous comparer au corpus SEWA et aux différents systèmes et features utilisés avec celui-ci. De plus, nous avons introduit le principe de fusion des modalités, qui apparaîtra dans les contributions de cette thèse.

Dans la prochaine partie, nous allons nous concentrer sur les contributions de cette thèse : de la construction d'un corpus répondant à nos besoins, à la mise en place de systèmes de reconnaissance des émotions performants et l'analyse des résultats obtenus.

DEUXIÈME PARTIE

Contributions

ALLOSAT UN CORPUS POUR LA RECONNAISSANCE CONTINUE D'ÉMOTIONS

«Vos clients les plus mécontents sont votre meilleure source d'apprentissage» (Bill Gates, 1955-)

4.1 Motivation

L'un des objectifs de cette thèse est de reconnaître l'état émotionnel de personnes en relation téléphonique avec des agents de centres d'appels. Les motivations de ces conversations sont diverses, mais en général elles concernent soit l'achat d'un produit (service d'abonnement, service d'achat), soit la recherche d'une solution à un problème donné (plainte, service après vente, service de réclamation), soit la demande d'informations diverses sur une entreprise ou leurs produits. Selon l'exigence industrielle, les émotions sont primordiales dans la relation clientèle [Tum11] :

- la satisfaction notamment est un facteur de fidélité et de diffusion de la marque (un client satisfait peut à son tour promouvoir la marque à son entourage),
- la frustration favorise l'attrition du client, c'est-à-dire le rejet de la marque au profit de ses concurrents, et la *mauvaise publicité*.

Ces deux émotions, comme nous avons pu le voir dans le chapitre 1, peuvent s'inscrire dans différentes théories discrètes ou continues. Pour rappel, il s'agit de deux émotions qui peuvent s'inscrire dans le circumplex de Russell [Rus80] (cf Figure 1.6), la satisfaction étant polarisée de façon positive avec une activation plutôt neutre et la frustration étant polarisée de façon plutôt négative, avec une activation assez forte.

En accord avec les besoins de l'entreprise Allo-Média, partenaire de cette thèse, nous avons choisi de les inscrire dans un espace continu, en créant la dimension de satisfaction

comme explicitée dans la figure 4.1. En effet, en plus de la catégorie de l'émotion, la variation de son intensité et de sa durée, permettent de visualiser une dynamique des émotions. Celle-ci est un indicateur crucial pour améliorer l'expérience du client. De plus, Allo-Média, notre partenaire industriel, cherche à définir une satisfaction et une frustration en synchrone. C'est-à-dire que l'entreprise souhaite pouvoir afficher à la fois la catégorie et l'intensité de l'émotion au fur et à mesure de la conversation. Tandis que des émotions discrètes ne contiennent pas autant d'informations. Il est donc plus logique de nous inscrire dans une théorie émotionnelle continue.

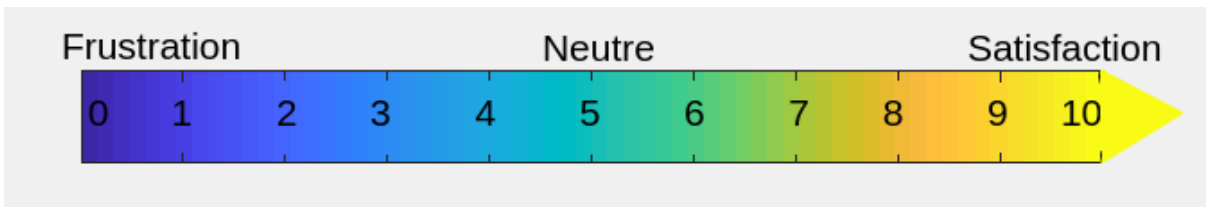


FIGURE 4.1 – L'axe de satisfaction va de la frustration à la satisfaction, en passant par le neutre. C'est donc sur cet axe que l'annotation a été effectuée.

Pour étudier ces émotions dans ce contexte, nous avons besoin de données et donc d'un corpus. Nos contraintes sont les suivantes :

- les conversations doivent provenir de conversations téléphoniques. Elles doivent être assimilables à une relation clientèle où un utilisateur recherche des informations auprès d'un conseiller,
- les conversations doivent être en français,
- une séparation entre le canal de l'appelant et de l'appelé est souhaitable, pour ne pas avoir de problèmes de chevauchements et donc permettre l'annotation d'un seul locuteur sans être influencé par l'autre locuteur,
- l'annotation doit être faite selon l'axe de satisfaction et de frustration. Elle doit également être continue.

Nous avons recherché un corpus dans la littérature, notamment parmi ceux décrits dans le chapitre 3, qui répond à toutes ces contraintes. Malheureusement aucun ne correspond parfaitement à notre demande, d'autant plus que peu de corpus comportant des conversations réelles issues de centre d'appel sont disponibles pour la recherche. Nous avons donc décidé de construire un corpus adapté à nos besoins.

4.2 Recueil des données

Avec l'entreprise Allo-Média, nous avons pu recueillir des données provenant de différents centres d'appels français, sous la forme de conversations impliquant des appelants (les clients) et des agents. Comme nous ne voulons pas nous restreindre dans un domaine d'activité donné, nous avons récupéré une cinquantaine de conversations de chaque domaine d'activité des entreprises. Parmi eux, nous avons sélectionné des conversations issues des domaines de l'assurance, de la distribution d'énergie, des agences de voyages, des agences immobilières et de la téléphonie. Ces appels ont eu lieu entre juillet 2017 et novembre 2018.

Ces conversations étant séparées dans l'enregistrement entre le canal du client et le canal de l'agent, nous n'avons donc pas de chevauchement, des *overlaps*, de signal entre les locuteurs. De plus, pour des contraintes éthiques et commerciales, la partie concernant l'agent ne peut être diffusée et a donc été supprimée des données collectées. En effet, Allo-Média ne s'inscrit pas dans une logique de contrôle et de notation des agents. La partie du client quant à elle, est principalement constituée d'un locuteur unique qui ne sera pas retrouvé dans d'autres conversations. Ici, nous avons fait le choix d'étudier les émotions d'un ensemble de locuteurs, plutôt que des émotions d'un seul locuteur : on considère la reconnaissance d'émotion comme indépendante du locuteur. Toutefois, il existe également des conversations où il peut y avoir plusieurs locuteurs, comme par exemple quand une personne passe le téléphone à un autre membre de sa famille. Tous les enregistrements sont issus du canal téléphonique, échantillonné à 8kHz.

Face à cette masse de données, et conscients que nous ne pouvons pas tout annoter, nous avons mis en place un processus pour sélectionner les données qui constitueront le corpus.

4.2.1 Sélection des données

Il est communément admis que toute parole n'est pas forcément marquée par des états émotionnels, d'où la présence d'un état neutre. C'est d'autant plus vrai pour des conversations issues de centres d'appels. En effet, lors d'un appel pour une précision sur la livraison d'un produit ou sur le suivi d'un abonnement par exemple, il est rare que des émotions soient exprimées.

A priori la frustration peut se reconnaître par une modification du timbre, une parole plus rapide et des hésitations plus importantes. On peut également noter l'augmentation

de la présence du discours para-linguistique avec des soupirs, des bruits de bouches ou des rires nerveux pour la frustration. Le discours se teinte également de mots à polarité négative et présente des répétitions. Dans certains cas, on peut observer une augmentation des tournures négatives. Pour la satisfaction, on assiste à un phénomène inverse, avec une parole moins rapide et plus posée. Le discours se teinte de mots à polarité positive et présente moins de répétitions et moins de prise de parole longue.

En se basant sur ces observations, afin de réduire le coût d'annotation, nous avons donc choisi de sélectionner automatiquement des conversations où la présence de la satisfaction ou de la frustration peut être détectée par l'humain. Pour ce faire, nous avons mis en place plusieurs critères :

- La durée de la conversation : celle-ci doit être d'au moins 30 secondes, pour avoir le temps d'exprimer une émotion et pour éviter les appels ratés ou manqués. Comme la frustration et la satisfaction semble être induite par une interaction avec l'agent, nous ne conservons que les conversations composées d'au moins trois tours de parole, soit trois prises de parole de la part du client et du conseiller.
- Une forte variation de la fréquence fondamentale. En effet, comme nous l'avons vu précédemment dans le troisième chapitre, la variation de la fréquence fondamentale est une indication permettant de caractériser la prosodie de la parole. Afin d'extraire la fréquence fondamentale, nous avons utilisé l'algorithme appelé Yet Another Algorithm for Pitch Tracking (YAPPT) [ZH08] qui permet de pallier le contexte téléphonique. En effet, le signal téléphonique possédant une bande passante limitée, la fréquence fondamentale peut en être absente. Cet algorithme cherche donc à restaurer la fréquence fondamentale des signaux dégradés.
- La conversation doit être polarisée : pour ce faire, nous avons calculé un score de polarité en partant des transcriptions de la partie du client. En utilisant le dictionnaire French Affective Norms (FAN) [MS14] que nous avons présenté dans le troisième chapitre, nous avons calculé un score de valence correspondant à la moyenne des scores de chaque mot polarisé présent dans la transcription. Les autres mots sont pris en compte avec le score de 5, considéré comme le neutre. Ce score de valence de la conversation varie donc entre 0 et 10, 0 étant le plus négatif et 10 le plus positif.

Prenons la phrase *c' est une supercherie, votre cupidité est sans limite* : les termes *supercherie* et *cupidité* ont des scores respectivement de 3.32 et 3.35. Les autres mots n'ont pas de scores de polarité. Cette phrase a un score de $\frac{3.32+3.35+6\times 5}{8} = 4.58$.

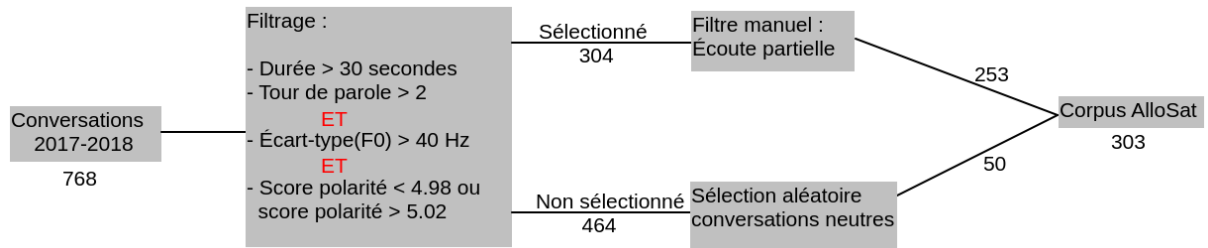


FIGURE 4.2 – Diagramme représentant la chaîne de traitement pour la sélection des conversations présentes dans le corpus AlloSat.

L'application de ces critères a permis d'isoler 304 conversations présentant des caractéristiques intéressantes sur un total de 768 conversations. On peut voir sur le diagramme 4.2 les différents filtres qui ont été appliqués aux conversations. Elles sont d'une durée suffisante, ont un écart type de leurs fréquences fondamentales supérieures à 40 Hz et un score de polarité non compris entre 4.98 et 5.02. Ces scores sont très resserrés de par la part importante de mots neutres. Ces 304 conversations ont été écoutées afin de garder les conversations où la manifestation de la dimension de satisfaction était la plus flagrante. Par ce procédé, nous avons conservé 253 conversations. Afin de mieux respecter la répartition des états émotionnels dans un contexte de centres d'appels, nous avons également sélectionné au hasard 50 conversations qui n'étaient pas retenues par le filtre mis en place et donc considérées comme neutre.

Nous avons donc une sélection finale contenant 303 conversations. Une fois cette sélection effectuée, nous avons dû traiter les données pour qu'elles puissent être annotées par la suite.

4.2.2 Pre-traitement des données

Les données en l'état ne peuvent pas être immédiatement annotées, des pré-traitements sont nécessaires : la réduction des silences et l'anonymisation des données personnelles.

Réduction des silences

Comme les deux canaux (client et agent) sont séparés en amont, nous avons conservé uniquement les documents provenant du canal client, afin de ne pas traiter la parole de l'agent. L'absence de la réponse du conseiller ajoute de longs moments de silence dans le signal audio. Afin de réduire l'effort d'annotation, nous avons décidé de réduire les silences

de plus de deux secondes. Cette réduction suit le protocole suivant :

- Nous détectons les silences automatiquement en utilisant l'outil *silencedetect* de *ffmpeg* [Tom06]. Cet outil détecte les zones de conversation dont le volume est inférieur à un seuil de tolérance donné, dans notre cas -40dB, pendant une durée donnée, dans notre cas deux secondes. Ce seuil a été déterminé empiriquement, pour supprimer les silences et quelques bruits parasites, sans détruire des séquences de parole. On liste ainsi les silences qui sont présents dans le document.
- On supprime le signal audio dès lors que l'on trouve un silence d'une durée supérieure à deux secondes, en suivant la liste établie précédemment. Si le silence dure moins de deux secondes, on ne le supprime pas.
- On réassemble les différents fragments du signal, en intercalant un signal audio de bruit blanc d'une durée exacte de deux secondes. Un bruit blanc correspond à une génération automatique d'un signal audio par le tirage aléatoire de fréquences suivant la même densité spectrale de puissance. Dans notre cas, nous avons créé un bruit blanc en suivant la loi normale. Ce choix a été motivé par une volonté de confort pour les annotateurs.

Ce traitement nous permet de passer d'un corpus de 57 heures majoritairement composé de silences à un corpus de 37 heures où les silences sont contrôlés. Nous avons donc une répartition des durées de conversation plus homogènes, comme le montre la figure 4.3. Les conversations ont une durée variant de 32 secondes à 41 minutes, avec une moyenne d'environ 7 minutes.

Une fois ce traitement effectué, nous avons mis en place l'anonymisation du corpus.

Anonymiser les données personnelles

Afin de respecter la vie privée des personnes, la France et les autres pays européens ont mis en place une réglementation sur la collecte, le stockage et le traitement des données personnelles de tout individu. Le règlement général sur la protection des données (RGPD) établit des règles relatives à la protection des utilisateurs vis à vis du traitement de leur données personnelles, pour s'assurer que les utilisateurs conservent leurs libertés et leurs droits fondamentaux. Cette réglementation, contrôlée par la Commission Nationale de l'Informatique et des Libertés (CNIL) a été mise en application en 2018, nous avons donc dû traiter les données du corpus afin de respecter cette réglementation.

Un des points les plus importants de cette réglementation est l'obfuscation de toutes les données personnelles. Les catégories de données personnelles que nous avons anonymi-

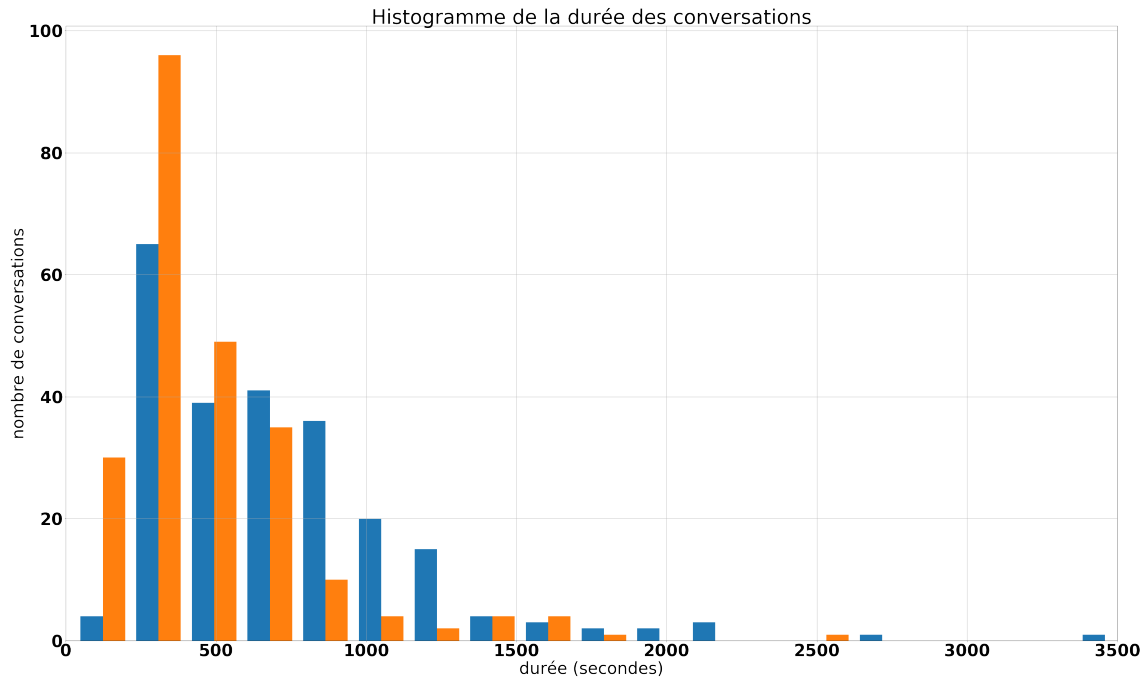


FIGURE 4.3 – La répartition des conversations en fonction de leur durée. En bleu, nous avons les anciennes durées, en orange les nouvelles. Environ 76,6% des conversations dureraient moins de 15 minutes (soit 900 secondes). Après réduction des silences, ce nombre augmente à environ 93,7%.

sées sont présentées dans l'arbre 4.1. Ce sont celles définies par l'entreprise Allo-Média, en accord avec les réglementations en vigueur. Il est à noter que l'anonymisation ne se fait pas exclusivement par les feuilles de l'arbre, tous les nœuds peuvent être utilisés. L'anonymisation des nœuds est exceptionnelle, elle obstrue les données indirectes qui permettent de reconnaître un individu. Par exemple, dans le cadre de la localisation, un nom donné à une maison *l'hirondelle* est anonymisé en catégorie *localisation*. À ces dernières s'ajoutent aussi les données permettant d'identifier des entreprises. Les marques et les produits sont obfusqués. Ainsi, les conversations finales permettent de reconnaître le domaine d'activité de l'entreprise mais pas son identité.

Pour procéder à cette obfuscation, nous avons utilisé un obfuscatrice automatique, détenu par la société Allo-Média, appliqué sur les transcriptions, puis répercuté sur les fichiers audio. Ce parseur, dans un premier temps, recherche les mots appartenant aux groupes présentés dans l'arbre 4.1, dans les transcriptions des conversations. Il substitue dans le texte les mots désignés par la catégorie. Enfin il substitue le segment audio correspondant au code temporel des données personnelles détectées par un signal audio

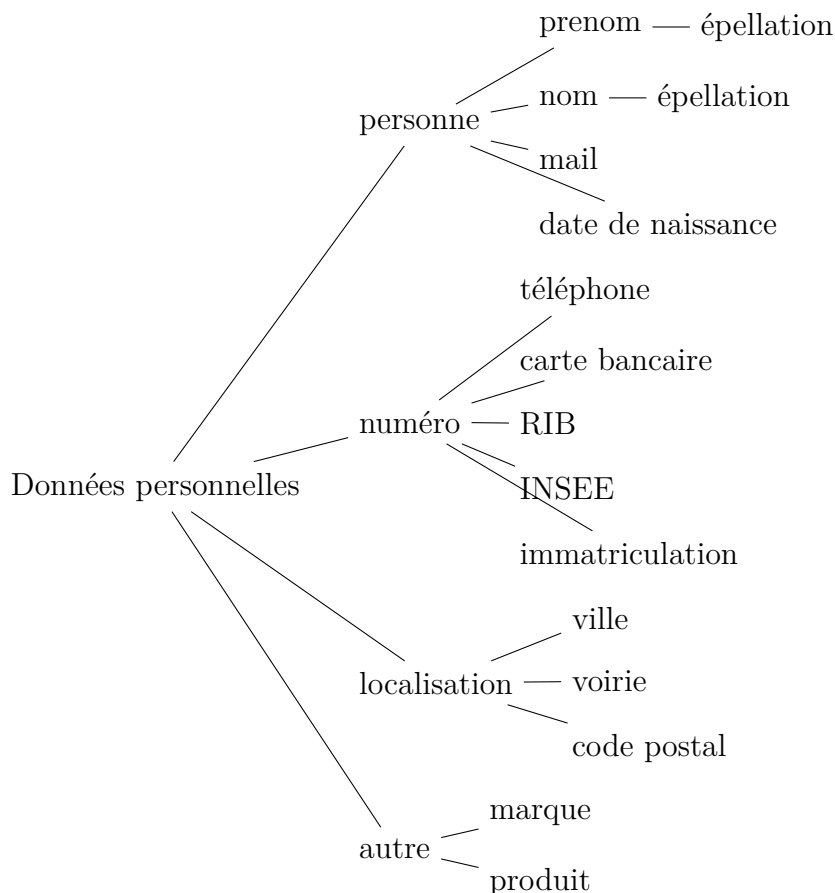


TABLE 4.1 – Arbre des différentes catégories de données personnelles anonymisées

pré-enregistré. Ce dernier se compose d'un air de percussion de type jazzy. Un exemple fictif de cette obfuscation est illustré dans la figure 4.4.

Une deuxième passe, humaine cette fois ci, permet de garantir l'obfuscation de toutes les données personnelles. En utilisant l'outil Transcriber, chaque conversation a été écoutée et lue, permettant d'identifier et de segmenter les données personnelles restantes. C'est lors de cette deuxième phase que les données permettant d'identifier les entreprises ont été isolées. Nous avons également choisi de supprimer tous les numéros permettant d'identifier un contrat ou une date significative. Les segments identifiés sont alors substitués par le même signal audio jazzy. Ce choix de remplacement est motivé par des retours des annotateurs : nous avons proposé plusieurs sons pour masquer les données personnelles (bruits blancs, son de guitare, de chant, de basse, de percussion...) et ils ont voté pour le moins fatigant.

Lors de cette passe, les annotateurs ont également effectué une correction partielle de

Protection des données personnelles

Bonjour, je m'appelle **Andréa Sullivan** et j'habite à **Marseille**.

Bonjour, je m'appelle **[pers.pre]** **[pers.nom]** et j'habite à **[loc.ville]**.



FIGURE 4.4 – Exemple d’obfuscation d’un segment fictif. Les mots sont d’abord retrouvés dans la transcription, puis sont substitués par leur catégorie. Enfin le signal audio est modifié pour remplacer les informations personnelles par un son pré-enregistré.

la transcription : ils ont eu pour consigne de corriger les segments de parole effectifs, et de ne pas modifier les erreurs de transcription lors que le locuteur ne parle pas au conseiller.

4.3 Mise en place de l'annotation

4.3.1 Volonté d'annotation

Comme nous l’avons dit précédemment, nous souhaitons analyser l’axe de satisfaction de manière continue. Pour cela, nous avons mis en place un axe dont les extrema sont la frustration (0) et la satisfaction (10) qui passe par un état neutre (5) situé à mi-chemin entre ces deux émotions comme montré dans la figure 4.1. La valeur de satisfaction est automatiquement extraite en continu toutes les 250 ms, ce qui nous permet d’avoir quatre valeurs par seconde. Cette extraction est cohérente avec notre tâche puisque les émotions s’expriment sur un temps long, généralement de l’ordre de la minute [SD10].

Pour enrichir l’annotation, nous avons décidé de mettre en place, en plus de l’annotation continue, une annotation discrète de la dimension de satisfaction. Cette annotation discrète est effectuée au niveau de la conversation. La catégorie émotionnelle du début

et de la fin d'une conversation, comprise entre très frustré, frustré, neutre, satisfait, très satisfait est annotée. La durée de ce début et de cette fin de conversation sont laissées à l'appréciation des annotateurs. En plus, une caractérisation de l'évolution de l'émotion est annotée selon les catégories suivante : montante, descendante, stagnante, varie, varie fortement. Pour mieux comprendre ces catégories, elles ont été explicitées par les schémas de la figure 4.5, qui ont été montré aux annotateurs.

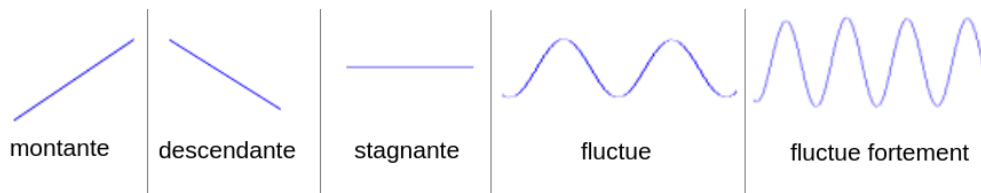


FIGURE 4.5 – Schéma des variations qui ont été données aux annotateurs afin de mieux comprendre la catégorie évolution

Afin de comparer AlloSat aux corpus existants, nous avons également mis en place une annotation de la valence sur le même principe que l'annotation discrète de la dimension de la satisfaction. Elle relève la valence de début et de fin de conversation, comprise entre très négative, négative, neutre, positive, très positive. On y ajoute la même caractérisation de l'évolution de la valence. Le tableau 4.2 récapitule les annotations discrètes.

Satisfaction & Frustration	
debut	très frustré, frustré, neutre, satisfait, très satisfait
fin	très frustré, frustré, neutre, satisfait, très satisfait
evolution	montante, descendante, stagnante, varie, varie fortement
Valence	
debut	très négative, négative, neutre, positive, très positive
fin	très négative, négative, neutre, positive, très positive
evolution	montante, descendante, stagnante, varie, varie fortement

TABLE 4.2 – Récapitulatif du schéma d'annotation discrète.

La différence entre la dimension de satisfaction et la valence a fait l'objet de plusieurs séances d'explication auprès des annotateurs, afin que ces deux notions ne soient pas confondues par les annotateurs.

Pour récapituler, nous nous sommes concentrés sur deux notions :

- L'évolution d'une émotion : tous les appels commencent au 'neutre' et évoluent en fonction du temps entre frustration et satisfaction. Cette annotation est continue.

- Une évaluation de l'émotion à posteriori : une fois l'appel terminé, nous voulons avoir un retour sur l'évolution de l'émotion. Pour cela nous voulons savoir comment était l'appelant au début de la conversation, comment il était à la fin et comment était cette évolution. Cette annotation est donc discrète.

4.3.2 Logiciel utilisé

Nous avons dans un premier temps considéré l'utilisation de FeelTrace [Cow+00], l'outil le plus utilisé pour réaliser de l'annotation continue. Cependant l'outil est optimisé pour annoter deux dimensions à la fois, la valence et l'activation, afin de placer l'annotation dans un contexte bi-dimensionnel qui ne correspond pas avec notre définition de l'axe de satisfaction. De plus, cet outil est assez difficile à prendre en main, et il requiert l'utilisation d'un joystick.

Nous avons donc fait le choix d'utiliser CARMA (Continuous Affect Rating And Media Annotation) [Gir14]. CARMA permet d'annoter de façon continue l'émotion selon une dimension définie en amont en utilisant un clavier et une souris. La figure 4.6 montre l'interface visible par les annotateurs lors de l'annotation. Un guide d'installation et de configuration a été fourni à l'administrateur Système de l'entreprise. Il est disponible dans l'annexe 8.1



FIGURE 4.6 – Capture d'écran de l'outils CARMA [Gir14]

Pour l'annotation discrète, les annotateurs ont rempli un modèle pré-construit vide de tableau avec un logiciel bureautique de type Excel. Comme nous l'avons dit précédemment, il y a six catégories à remplir par conversation (état émotionnel de début, de fin et la forme de l'évolution entre les deux). Nous avons également demandé l'annotation en genre des appelants et un espace était mis à disposition pour d'éventuels commentaires.

4.3.3 Consignes

L'annotation a été réalisée par une équipe de trois annotateurs (deux femmes et un homme), employés d'une société de transcription manuelle de la parole basée en France. Ils avaient déjà collaboré avec l'entreprise Allo-Média pour des tâches de transcription manuelle de la parole et d'annotation de données sémantiques, notamment de l'annotation d'entités nommées et des résumés des conversations.

Ces personnes ont donc reçu en amont des formations sur ces différentes tâches : la reconnaissance d'une entité nommée, l'utilisation de l'outil Transcriber, des formations sur l'orthographe et sur la conjugaison notamment.

Afin d'aider les annotateurs et de guider au mieux l'annotation, un guide d'annotation a été mis à leur disposition. Ce guide est disponible dans l'annexe 8.2. Ce dernier explique le contexte de l'étude et les consignes à respecter. Comme l'émotion possède une part non négligeable de subjectivité, il fallait que les consignes soient les plus objectives possibles, tout en assurant l'homogénéité des annotations.

Objectivité des annotations

- Pour réduire la subjectivité de l'annotation, les annotateurs ne doivent pas prendre parti pour le locuteur durant la conversation.
- Les annotateurs ne doivent pas échanger entre eux au sujet des conversations écoutées.
- Pour ce qui est des dimensions discrètes, elles doivent être annotées tout de suite après l'écoute de la conversation, afin que les états émotionnels ne soient pas oubliés ou pollués par l'écoute d'autres conversations. De plus le genre des locuteurs et des observations diverses peuvent être rajoutés par les annotateurs.
- Chaque conversation est annotée une seule fois par chaque annotateur.

Homogénéité des annotations

- Pour expliquer la notion de valence, nous avons utilisé le Self-Assessment Manikin (SAM) [BL94] qui donne une description visuelle de la valence, que l'on retrouve sur la figure 4.7.
- Pour leur permettre d'avoir une meilleure compréhension de la satisfaction et de la frustration que nous voulons annoter, nous avons fourni deux conversations en tant que borne de satisfaction et borne de frustration, afin d'aider à étalonner les émotions. Ainsi les annotateurs peuvent appréhender l'amplitude potentielle de la dimension de satisfaction de façon homogène. Ces bornes ont été établies de façon empirique, lorsque nous avons sélectionné les données du corpus.

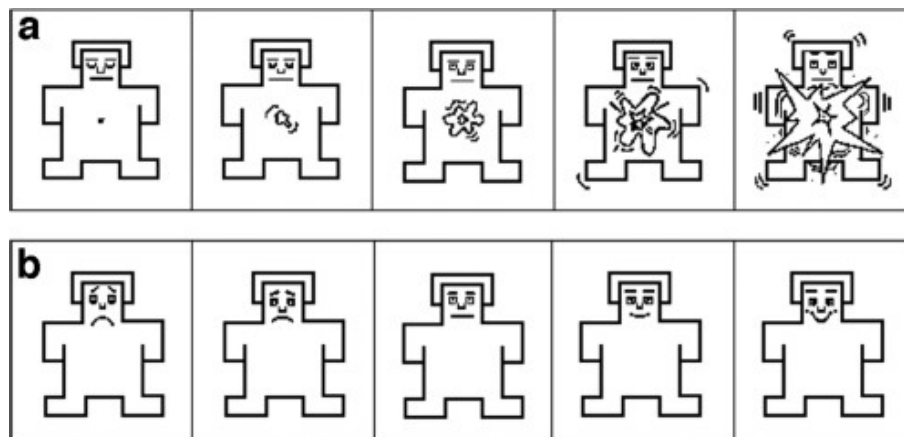


FIGURE 4.7 – Le Self Assessment Manikin (SAM). La ligne a correspond à l'activation, la ligne b correspond à la valence. Issu de [BL94]

L'annotation de la dimension de satisfaction a donc été réalisée de façon continue selon les consignes suivantes :

- La barre d'annotation va de 0 (Frustration) à 10 (Satisfaction), et elle est graduée par palier de 1. Le 5 correspond à l'état neutre et à l'état de départ de l'annotation.
- Le curseur d'annotation peut être contrôlé par la souris ou par les flèches du clavier. Les pas du clavier sont de 0.1 tandis que la souris peut avoir une granularité plus fine.
- Si aucun état émotionnel n'est constaté, ou qu'il ne varie pas, alors l'annotateur ne doit pas toucher à l'annotation. Cela est valable également lors des silences. Les annotateurs ont été notifiés de la présence de conversations où aucun état émotionnel n'a été constaté en amont.

- Un document ne doit être annoté qu'une seule fois par la même personne, sans possibilité de revenir en arrière. Nous voulons la réaction immédiate des annotateurs et non une réaction plus réfléchie, qui tend à minimiser les émotions détectées, selon des travaux préliminaires que nous avons réalisés. Il n'est pas non plus possible d'avancer la conversation, l'annotation doit être faite en temps réel de l'écoute de la conversation.
- Un document doit être annoté en une seule fois. On ne doit pas revenir à l'annotation d'un fichier après une longue pause (plusieurs heures ou un jour). Cela évite à l'annotateur de ne plus être dans le contexte émotionnel de la conversation.

Ces consignes ont pour but d'aider l'annotateur à effectuer une annotation la plus objective possible. En plus de ces consignes, la structuration des documents a été expliquée. Les bruits blancs indiquent qu'un silence de plus de 2 secondes s'est produit. Les silences ont été retirés du document afin fluidifier l'écoute et l'annotation du document. Ce son est donné à titre informatif pour aider dans l'annotation. Nous avons signalé également la présence de bruits jazzy qui sont utilisés pour l'anonymisation des conversations. Ils remplacent les parties de conversation qui permettent l'identification de personnes ou d'entreprise.

4.4 Analyse d'AlloSat

	Nombre de conversations	Durée	Durée de parole
TRAIN	201 (70%)	25h26m35	15h48m50
DEV	42 (15%)	05h55m46	03h22m14
TEST	60 (15%)	05h58m41	03h28m57
TOTAL	303	37h23m27	22h40m01

TABLE 4.3 – Découpage du corpus en ensemble de train, développement et test. Les durées sont données en heures, minutes, secondes.

AlloSat est composé de 303 conversations, d'une durée totale de 37 heures 23 minutes et 27 secondes, de 308 locuteurs distincts dont 191 femmes et 117 hommes (voir analyse en genres dans le chapitre 7). Une répartition semi aléatoire en trois sous-ensembles a été définie. Nous avons veillé à ce que chaque partition comprenne des conversations choisis aléatoirement qui peuvent être dépourvues de contexte émotionnel. L'ensemble d'entraînement (train) est composé de 201 conversations, l'ensemble de développement

(dev) de 42 conversations et l'ensemble de test (test) de 60 conversations. De plus amples détails sont disponibles dans le tableau 4.3. On peut noter que la durée du dev et du test sont très similaires. On a donc une répartition approximative de 25 heures et demi des données dans le train et douze heures des données dans le dev et le test (avec six heures chacun). Comme nous l'avons vu au chapitre 2, le découpage d'un corpus en sous-ensembles permet d'assurer la cohérence des scores notamment.

Le tableau 4.4 regroupe l'ensemble des annotations discrètes du corpus. Nous pouvons observer une sur-représentation de l'état neutre, surtout en début de conversation, ce qui ne nous a pas surpris, puisque l'on retrouve peu de passage marqué par un état émotionnel dans la parole. Comme nous le pensions, la plupart des conversations ont été perçues avec une frustration croissante, probablement parce que la plupart des clients appelle quand ils ont un problème ou par exemple parce que le conseiller n'est pas en mesure de donner une réponse suffisamment satisfaisante à l'interlocuteur.

Comme nous pouvons le voir dans le tableau 4.4, peu de conversations ont été annotées en très satisfait. Nous avons donc choisi de regrouper les catégories très satisfaites et satisfaites. Afin de rester symétrique dans nos annotations, nous avons également regroupé les catégories très frustrées et frustrées. Les annotations discrètes sont ensuite fusionnées par vote majoritaire. En cas d'égalité, nous choisissons le neutre que ce soit pour la valence ou la dimension de satisfaction. Pour l'évolution des deux axes, le vote majoritaire n'est pas suffisant pour fusionner les données : il y a trop de cas d'égalité. Nous avons considéré que si l'évolution était différentes selon les annotateurs, elle correspond à la catégorie *fluctue*.

Nous avons mis en place des mesures d'accord intra-annotateurs, permettant de mesurer la pertinence d'une annotation vis à vis des autres annotations du même annotateur. Et nous avons également mis en place des mesures d'accord inter-annotateurs, permettant de mesurer la pertinence de l'annotation d'un document par rapport à une autre annotation du même document.

4.4.1 Accord intra-annotateur

Cet accord est mesuré entre les annotations continues et discrètes d'une même conversation par un même annotateur. Les annotations continues ont été normalisées en suivant la méthode de la normalisation standard selon l'équation 4.1 avec une annotation x issue de toutes les annotations X , remettant les annotations dans un espace allant de zéro à un.

Axe de Satisfaction	Début de la conversation					
		neutre	frustré	satisfait	très frustré	très satisfait
	a1	300	3	0	0	0
	a2	284	19	0	0	0
	a3	298	5	0	0	0
	vote majo.	299	4	0	0	0
	Fin de la conversation					
		neutre	frustré	satisfait	très frustré	très satisfait
	a1	111	123	21	48	0
	a2	92	91	15	105	0
	a3	56	48	53	135	11
	vote majo.	85	195	23	0	0
	évolution					
		stagne	descend	monte	fluctue	fluctue fort
	a1	87	165	10	39	2
	a2	50	193	8	51	1
a3	64	193	19	27	0	
vote majo.	55	184	7	57	0	
Valence	Début de la conversation					
		neutre	néгатif	positif	très négatif	très positif
	a1	298	5	0	0	0
	a2	280	22	1	0	0
	a3	298	5	0	0	0
	vote majo.	299	4	0	0	0
	Fin de la conversation					
		neutre	néгатif	positif	très négatif	très positif
	a1	93	136	19	55	0
	a2	86	105	20	92	0
	a3	55	50	53	135	10
	vote majo.	91	188	24	0	0
	évolution					
		stagne	descend	monte	fluctue	fluctue fort
	a1	89	164	8	40	2
	a2	44	185	14	57	3
a3	56	196	22	29	0	
vote major.	52	184	8	59	0	

TABLE 4.4 – Ensemble des annotations discrètes des trois annotateurs a1, a2 et a3. Vote majo correspond au vote majoritaire qui a conduit à l'annotation discrète de référence.

$$x' = \frac{x - \text{Minimum}(X)}{\text{Maximum}(X) - \text{Minimum}(X)} \quad (4.1)$$

Nous avons discrétisé les valeurs continues, appelée S_n dans la suite, suivant les trois niveaux utilisés dans l'annotation discrète. Pour ce faire, nous avons défini empiriquement deux seuils en observant l'annotation des conversations neutres contenues dans le corpus. En effet, l'annotation de ces dernières restent globalement entre les bornes 0.45 et 0.55 de l'axe de satisfaction. Notre seuil permet de déterminer si une valeur continue correspond à un état de frustration ($S_n < 0.45$), à un état neutre ($0.45 < S_n < 0.55$) ou à un état de satisfaction ($S_n > 0.55$).

Nous avons également défini la notion de début et fin de conversation comme étant 10% de la durée de la conversation, dc . Le début correspond donc au segment commençant à 0 et finissant à $0.1 \times dc$ et la fin correspond au segment commençant à $dc - 0.1 \times dc$, jusqu'à la fin de la conversation, dc .

Nous avons alors fait la moyenne des valeurs de satisfaction sur le début de la conversation et nous appliquons le seuillage pour déterminer quelle étiquette caractérise cet intervalle de temps. Nous faisons exactement la même chose avec les annotations de fin de conversation. Ainsi pour chaque conversation, nous avons l'annotation discrète effectuée par l'annotateur et une autre annotation discrète correspondant à la discrétisation de son annotation continue. La différence entre l'annotation discrétisée du début et de la fin de la conversation devrait être en adéquation avec l'évolution annotée.

Nous avons également décidé de calculer un kappa par annotateur. Le kappa (κ) est une mesure permettant de quantifier l'accord entre deux observations. Elle mesure le degré de concordance entre deux observations selon l'équation 4.2 :

$$\kappa = \frac{P_0 - P_e}{1 - P_e} \quad (4.2)$$

P_0 correspond à l'accord relatif des observations entre les annotations et P_e représente la probabilité d'un accord aléatoire. Comme nous avons un cas de sur-représentation d'une classe, ici le neutre, nous avons fixé P_e à $1/3$ suivant les recommandations de Callejas et al. [CL08].

Le kappa est compris entre 0 et 1. Plus le kappa est proche de 1, plus l'accord entre les deux observations est forte. En revanche, plus le kappa se rapproche de 0 et moins on a d'accord entre les observations. De plus, le kappa permet d'indiquer la difficulté de la tâche à effectuer, dans le cas d'une tâche subjective comme la nôtre. Généralement, on

considère le kappa comme suffisant lorsqu'il est proche de 0.8. D'après McHugh [ML12], si on rapproche le kappa à l'accuracy, un coefficient de 0.8 correspond à environ 64% d'accuracy, étant qualifié d'accord fort.

Annotateur	Pourcentage de désaccord		kappa κ	
	début	fin	début	fin
a1	1,32%	10,23%	0.98	0.84
a2	7,59%	18,15%	0.88	0.72
a3	4,29%	16,50%	0.93	0.75
Moyenne	4,40%	14,96%	0.93	0.77

TABLE 4.5 – Pourcentage de désaccords sur les 303 conversations et kappa pour définir l'accord intra-annotateur pour chacun des trois annotateurs ainsi que la moyenne. a_i représente l'annotateur i .

Les résultats de ces calculs sont disponibles dans le tableau 4.5. Nous pouvons observer une forte concordance entre les annotations discrètes et les annotations continues discrétisées dans les débuts de conversation avec un kappa moyen de 0.93. Si on se concentre sur le nombre de cas où l'annotation discrète est différente de l'annotation continue discrétisée, on voit que le premier annotateur (a1) a environ 1% de désaccord et qu'au maximum, l'annotateur 2 (a2) est en désaccord de moins de 8%. Pour les fins de conversation, on observe de moins bons scores, avec un kappa moyen de 0.77 mais qui reste suffisant pour exprimer une cohérence des annotations. En regardant le nombre d'annotations différentes, on observe de nouveau que l'annotateur 1 (a1) a le moins de désaccord avec moins de 10%, tandis que l'annotateur 2 (a2) culmine à presque 17% de désaccord. Nous avons conclu que, même si l'accord intra-annotateur n'est pas parfait, il est suffisant pour certifier de la cohérence des annotations continues et discrètes produites par un même annotateur.

4.4.2 Accord inter-annotateur

Afin d'évaluer l'accord inter-annotateur sur les annotations continues, nous avons utilisé le coefficient de corrélation linéaire. Ce coefficient est calculé au niveau de la conversation sur la dimension de satisfaction normalisée par rapport à l'ensemble des conversations entre les paires d'annotateurs comme défini dans la section précédente.

Le coefficient de corrélation linéaire donne une mesure de l'intensité et du sens de la relation linéaire entre deux variables, ici les deux annotations des deux annotateurs. Son

calcul est défini par l'équation suivante 4.3 :

$$R_{12} = \frac{Cov(x_1, x_2)}{\sigma_{x_1} * \sigma_{x_2}} \quad (4.3)$$

où $Cov(x_1, x_2)$ désigne la covariance entre les variables x_1 et x_2 , ici l'ensemble des annotations de deux annotateurs a_1 et a_2 . σ_{x_1} , σ_{x_2} désignent leur écart type. Ce coefficient est compris entre -1 et 1. Plus il est proche de 1, plus la relation linéaire positive entre les variables est forte. Plus il est proche de -1, plus la relation linéaire négative entre les variables est forte. Si le coefficient est proche de 0, on ne peut pas établir de relation linéaire. Nous avons également calculé le kappa entre paires d'annotateur sur les valeurs discrètes de début et de fin de conversation.

	R	$\kappa_{début}$	κ_{fin}
a1-a2	0.82	0.99	0.90
a2-a3	0.87	0.88	0.69
a1-a3	0.80	0.87	0.72
Moyenne	0.83	0.91	0.77

TABLE 4.6 – Accord inter-annotateur calculé entre les pairs d'annotateurs ainsi que la moyenne. a_i représente l'annotateur i . R représente le coefficient de corrélation, k représente le kappa de début et de fin de conversation.

Les valeurs rapportées dans le tableau 4.6 montrent une bonne corrélation entre les annotateurs (un coefficient de corrélation moyen de 0,83), ce qui signifie que les annotations continues sont cohérentes entre les annotateurs. On observe toutefois que l'annotateur 1 (a1) et l'annotateur 3 (a3) sont moins enclin à donner les mêmes annotations que les paires d'annotateurs a1-a2 et a2-a3.

On remarque également que le kappa de début de conversation est très élevé. 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. Tout d'abord, l'annotation continue est toujours initialisée à cinq, ce qui se traduit par un état neutre. Nous avons donc un biais introduit par cet état initial, qui permet à toutes les annotations de commencer de la même manière. Mais l'hypothèse principale est que l'interlocuteur est rarement frustré ou satisfait en début de l'appel : ces émotions sont provoquées par les réponses de l'agent.

En partant de ces résultats d'accord prometteurs, nous avons défini 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

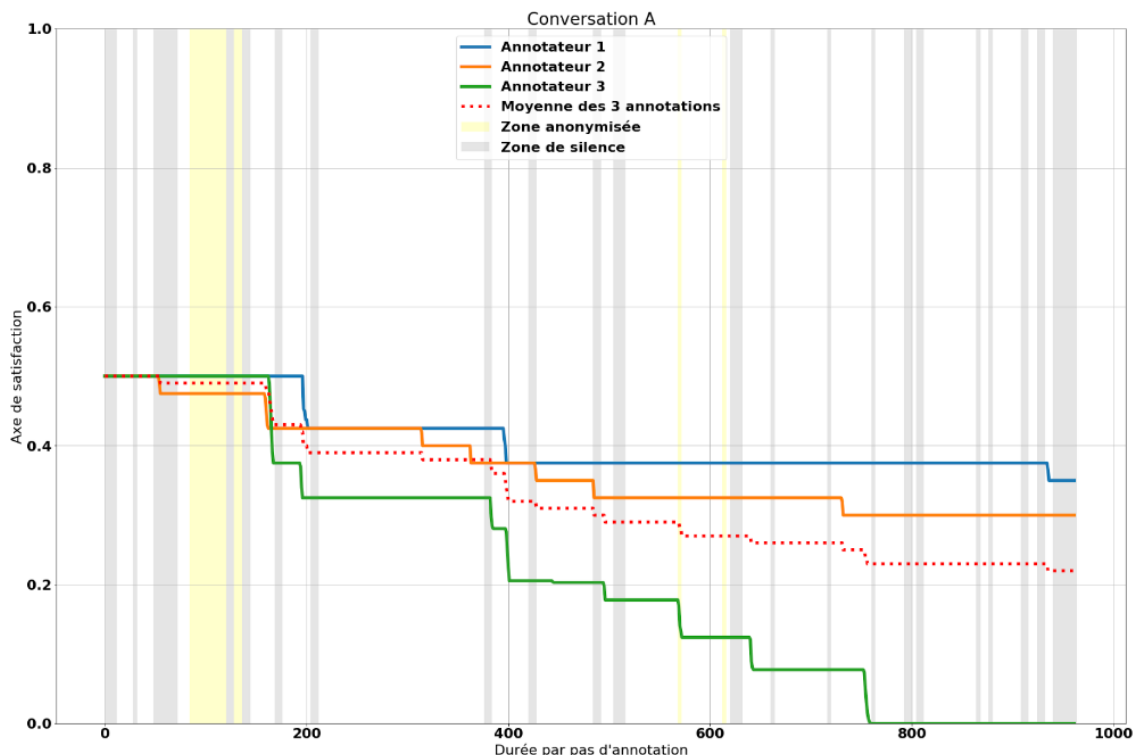


FIGURE 4.8 – Exemple d'annotation d'une conversation selon l'axe de satisfaction. L'annotation de référence correspond à la courbe en pointillée

d'apprentissage. Cette annotation de référence, aussi appelée *annotation de ref*, est utilisée dans les expériences présentées par la suite. Un exemple d'annotation de conversation et de son *annotation de ref* est présenté dans la figure 7.1, qui illustre également les moments de silence et d'anonymisation dont la transcription est disponible en annexe 8.3.

D'autres stratégies de fusion d'annotations existent, permettant de mettre un poids plus important à un annotateur ou à lisser les grands écarts d'annotation, mais devant nos résultats d'accord intra et inter-annotateurs, nous avons décidé d'utiliser la fusion d'annotation la plus simple, pour ne pas influencer les futures systèmes de reconnaissance automatique.

4.4.3 Calcul du Coefficient de Corrélation de Concordance entre annotateurs

Afin de confirmer notre analyse de l'homogénéité de l'annotation, nous avons également décidé de calculer le score du Coefficient de Corrélation de Concordance (CCC) entre les

annotateurs et l'annotation de référence correspondant à la moyenne de leurs observations. Comme nous l'avons expliqué dans le chapitre 3, le CCC est la métrique principale pour l'évaluation de la performance des systèmes de reconnaissance automatique des émotions continues. Nous avons donc fait l'hypothèse que l'évaluation de nos annotations avec cette métrique nous permettrait de mettre en place une comparaison entre la performance d'un système automatique et la performance d'un humain.

Ce score CCC a été calculé de deux façons. Dans un premier temps, nous avons déterminé un score global de l'annotateur. Puis nous avons voulu aller plus loin et calculer un score pour chacune des conversations afin de repérer les documents où les annotateurs ne sont pas d'accord entre eux.

Annotateur	Score global de CCC
a1	0.815
a2	0.944
a3	0.918
Moyenne	0.892

TABLE 4.7 – Score CCC calculé entre les annotations des annotateurs a_i et l'annotation de référence.

Les scores CCC entre annotateurs sont disponibles dans le tableau 4.7. On peut remarquer que les scores sont très bons, allant de 0.815 à 0.944 selon les annotateurs. Ces bons scores sont tout à fait logiques puisque l'annotation de référence est issue de l'annotation de ces trois annotateurs. On peut alors prendre plusieurs positionnements :

- On peut considérer qu'un score supérieur à 0.815 correspond à un bon score de reconnaissance puisqu'il est au niveau du *plus mauvais* de nos humains.
- On peut également considérer qu'un score de 0.892, moyenne de ces trois scores, correspond au score atteignable en moyenne par l'humain, et que donc si la reconnaissance automatique dépasse ce score, elle est au moins autant performante que l'humain.
- Mais également, on peut considérer que si le système de reconnaissance a un score supérieur à 0.944, il est plus performant que *le meilleur* de nos humains et donc qu'il est meilleur que l'homme pour annoter la satisfaction et la frustration.

Nous avons décidé de suivre la deuxième conjecture. Ainsi, si le système atteint un score supérieur à 0.892, on peut considérer qu'il est aussi performant que l'humain dans la tâche de reconnaissance continue de l'axe de satisfaction.

En regardant les scores de chaque conversation, dont un extrait est disponible dans le

tableau 4.8, nous avons pu constater que certaines conversations posent problème avec des scores moyens inférieurs à 0.2. Il sera intéressant par la suite de regarder les scores de reconnaissance automatique de l’axe de satisfaction sur ces conversations, dans le chapitre 7.

4.4.4 Étude empirique sur le corpus

Nous avons étudié la répartition de nos sous-ensembles. Nous avons observé qu’il y avait plus de fichiers courts en durée dans notre set de test. En effet, en moyenne un fichier du set de test dure 363 secondes alors qu’un fichier de train dure 464 secondes et un fichier de développement 492 secondes comme explicité dans le tableau 4.9.

Nous avons également pris en considération les silences que nous avons rajoutés artificiellement lors de la création du corpus pour vérifier leurs répartitions dans les sous-ensembles. Nous avons constaté que cet ajout était plutôt bien équilibré dans nos sous-ensembles : il y a en moyenne 42 sections de bruit blanc par conversation dans le train, 41 sections dans le développement et 36 sections dans le test. Les trois ensembles ont donc un traitement des silences compris dans le même ordre de grandeur.

4.5 Modalités de diffusion du corpus

Le corpus est distribué à toute personne affiliée à un institut public de recherche sur simple demande. AlloSat peut être demandé par mail aux personnes responsables de sa diffusion, à savoir Marie Tahon (marie.tahon@univ-lemans.fr) et moi-même (m.macary@allo-media.fr). Une charte de diffusion a été établie en partenariat avec DeepPrivacy, une entreprise spécialisée dans le traitement des données personnelles, en collaboration avec Allo-Média. Ainsi, un End User Licence Agreement (EULA) doit être rempli par toute personne qui souhaite accéder au corpus. Une copie de cette licence est disponible en annexe 8.4.

Le corpus est distribué en l’état, les responsables ayant mis en place tout ce qu’ils pouvaient pour garantir l’anonymat des participants tout en conservant des données automatiquement exploitables. Néanmoins, si une personne se reconnaît ou reconnaît un de ses proches, il peut demander à ce que sa participation soit retirée du corpus. Toutes les personnes ayant en leur possession le corpus reçoivent alors une notification leur stipulant qu’un document doit être détruit. Ce cas de figure ne s’est, pour le moment, jamais

Conversation	a1	a2	a3	moyenne
Conversation 1	0.682	0.911	0.930	0.841
Conversation 2	0.704	0.91	0.841	0.818
Conversation 3	0.726	0.931	0.789	0.816
Conversation 4	0	0.341	0	0.114
Conversation 5	0.903	0.976	0.970	0.950
Conversation 6	0.841	0.482	0.201	0.508
Conversation 7	0.255	0.247	0.377	0.293
Conversation 8	0.909	0.924	0.928	0.920
Conversation 9	0.767	0.929	0.772	0.823
Conversation 10	0.002	0.598	0.002	0.201
Conversation 11	0	0.515	0.563	0.359
Conversation 12	0.910	0.978	0.981	0.956
Conversation 13	0	0.241	0.539	0.260
Conversation 14	0.831	0.986	0.900	0.906
Conversation 15	0.746	0.942	0.919	0.869
Conversation 16	0.404	0.915	0.817	0.712
Conversation 17	0	0.501	0	0.167
Conversation 18	0.724	0.955	0.913	0.864
Conversation 19	0.761	0.968	0.858	0.862
Conversation 20	0.624	0.962	0.855	0.814
Conversation 21	0.582	0.974	0.814	0.790
Conversation 22	0.847	0.649	0.741	0.746
Conversation 23	0.741	0.940	0.934	0.872
Conversation 24	0.716	0.879	0.942	0.846
Conversation 25	0.825	0.821	0	0.549
Conversation 26	0.774	0.962	0.853	0.863
Conversation 27	0.637	0.868	0.972	0.825
Conversation 28	0.827	0.858	0.909	0.865
Conversation 29	0.154	0.680	0.810	0.548
Conversation 30	0.640	0.937	0.769	0.782
Conversation 31	0.678	0.952	0.908	0.846
Conversation 32	0.338	0.867	0.843	0.683
Conversation 33	0.512	0.929	0.872	0.771
Conversation 34	0.268	0.771	0.731	0.590
Conversation 35	0.583	0.892	0.917	0.798
Conversation 36	0	0.244	0	0.081
Conversation 37	0	0.687	0.686	0.458
Conversation 38	0.969	0.969	0.971	0.970
Conversation 39	0	0.627	0.815	0.480
Conversation 40	0	0.258	0.019	0.092

TABLE 4.8 – 40 scores calculés entre l’annotation de référence et l’annotation de l’annotateur a_i .

	Train	Dev	Test
< 7min	99	18	38
> 7min	101	24	22
< 5min	40	4	17
> 5min	160	38	43

TABLE 4.9 – Répartition de la durée des conversations en fonction des sets considérés

présenté. Depuis sa mise à disposition et jusqu’au mois d’août 2021, douze demandes d’accès ont été reçu par les responsables de sa diffusion avec une diffusion effective à quatre établissements.

Une discussion est toujours en cours pour l’ouverture de ce corpus aux chercheurs du secteur privé. Les informations sur sa diffusion sont mises à jour sur une page dédiée hébergée par l’université du Mans (<https://lium.univ-lemans.fr/allosat/>).

Ce corpus sera utilisé pour toutes les expériences présentées dans la suite de cette thèse.

4.6 Conclusion

Dans ce chapitre, nous avons décrit la construction du corpus AlloSat. Nous avons d’abord justifié les différents choix que nous avons mis en place, avant de parler de la sélection des données et de leur annotation. Enfin, nous avons donné quelques pistes d’analyse pour rendre compte de la qualité du corpus et de la difficulté de la tâche de reconnaissance des émotions continues.

Grâce à ce corpus, nous avons pu mettre en place des systèmes de reconnaissance qui seront détaillés dans les prochains chapitres.

RECONNAISSANCE CONTINUE D'ÉMOTION À PARTIR DE REPRÉSENTATIONS ACOUSTIQUES

5.1 Motivation

Dans ce chapitre, nous allons détailler les différentes procédures et tous les questionnements que nous avons eu lors de la réalisation de systèmes de reconnaissance d'émotions dans la parole. Nous utilisons principalement le corpus AlloSat dont nous avons détaillé la création et les caractéristiques dans le chapitre 4.

Dans le cadre de cette thèse, nous nous inscrivons dans un contexte industriel. L'entreprise Allo-Media, partenaire de cette CIFRE, a exprimé des besoins spécifiques quant aux émotions que nous devons traiter. Spécialisée dans le traitement d'informations issues de conversations téléphoniques entre un client et un conseiller, nous nous sommes inscrits dans ce domaine d'analyse. Comme nous l'avons expliqué dans le chapitre 4, nous avons choisi de nous focaliser sur deux émotions : la satisfaction et la frustration. Ces deux émotions, considérées comme deux émotions opposées dans le cadre de la relation clientèle, sont d'autant plus significatives si l'on analyse leur évolution au fil de la conversation. De plus, il nous semblait pertinent de nous concentrer sur l'aspect dimensionnel de l'émotion, afin de mieux alerter et rapporter les points ayant soulevé le plus de satisfaction et de frustration lors de la conversation. Nous rappelons ici que ce type de données en français n'existant pas à notre connaissance, nous avons collecté le corpus AlloSat que nous utilisons.

Plusieurs facteurs ont guidé les choix que nous avons mis en place dans nos systèmes de reconnaissance. Tout d'abord, nous faisons face à du signal de qualité médiocre, puisque provenant d'une source téléphonique. Ensuite, nous avons les problématiques de données personnelles dont nous avons parlé au chapitre 4. De plus, nous avons des conversations

tenues en français. Et enfin il faut considérer le temps de traitement qui ne doit pas être trop long pour pouvoir être retourné au client rapidement après la fin de la conversation.

La problématique de reconnaissance de ces émotions précises dans un espace continu en considérant ces contraintes nous ont conduit à mener plusieurs expérimentations sur le choix de l'architecture neuronale, des ensembles de descripteurs acoustiques pertinents, de la fonction de coût, tout en gardant à l'esprit les solutions industrielles qui peuvent en découler. Nous avons également validé nos résultats en nous comparant à l'état de l'art.

5.2 Construction d'une alerte sur les prédictions discrètes finales

Dans un premier temps, nous nous sommes posés la question de la pertinence d'une prédiction en classe discrète ou en dimension continue dans le cadre industriel. En effet, étant donné qu'une catégorie émotionnelle est définie au niveau d'un tour de parole (de l'ordre de la seconde), il y a généralement moins d'étiquettes discrètes que de valeurs (toutes les 250 ms) à prédire. De plus la diversité des classes étant plus faible que la diversité des valeurs continues à prédire, on peut admettre que les modèles auront raisonnablement moins de paramètres à apprendre avec le discret qu'avec le continu. Ainsi, un système de reconnaissance portant sur les étiquettes discrètes semble plus facile à mettre en place et moins coûteux en ressources. Dans notre contexte, cela restera complexe de comparer les deux approches : quel que soit la durée de la conversation, seules trois étiquettes sont à prédire : le niveau de frustration/satisfaction de début, de fin de conversation et son type d'évolution entre les deux, alors que la dimension de satisfaction couvre l'ensemble de la conversation.

Nous avons donc considéré dans un premier temps une prédiction d'étiquettes discrètes. En effet, cela nous permet d'avoir une première approche sur la satisfaction et la frustration qui peut être utilisée dans un contexte industriel. Nous avons imaginé un système d'alerte permettant de mettre en avant les conversations dites *préoccupantes*, que nous détaillons dans la prochaine section. On appelle préoccupantes, les conversations concluant sur annotation en frustration.

Nous avons donc mis en place une classification supervisée de reconnaissance de la satisfaction et de la frustration discrète. Nous avons tout d'abord cherché à prédire l'étiquette de fin de conversation. Comme les étiquettes de début de conversation ne contiennent quasiment que des états neutres, il ne nous semblait pas constructif de

construire un système de classification autour de ces données.

5.2.1 Les descripteurs

Pour mettre en place cette alerte, nous avons fait le choix de travailler avec plusieurs ensembles de descripteurs. En effet, les états émotionnels sont détectables à la fois dans les modalités acoustique et linguistique. Ces descripteurs sont décrits dans le tableau 5.1.

Modalité	Contient	Nombre features avant réduction	Nombre features après réduction
Audio	IS2009 [EWS10]	384	100
Texte	TF-IDF	4000	100
Audio+Texte	IS2009 et TF-IDF	200	200

TABLE 5.1 – Description des trois ensembles de descripteurs utilisés pour réaliser la reconnaissance des émotions discrètes des fins de conversations.

La modalité acoustique est représentée par l'ensemble de référence IS2009 contenant 384 descripteurs, décrit dans le chapitre 3 [EWS10] et la modalité linguistique par un TF-IDF contenant 4000 descripteurs aussi décrit au chapitre 3. Comme le nombre de descripteurs est très grand pour une classification de 303 documents, nous avons décidé de faire de la sélection de descripteurs afin d'éviter le sur-apprentissage notamment [TD16]. Cette sélection de type ranking est effectuée en amont de l'apprentissage avec l'outil scikit-learn [Ped+11] selon un seuil de variance (`feature_selection.VarianceThreshold`) et un classement de *mutual information* (`feature_selection.RFE`).

Nous avons également fait le choix de considérer les deux modalités en même temps, en fusionnant les deux ensembles de descripteurs. Pour la fusion, nous avons concaténé les deux vecteurs de descripteurs en testant différentes modalités de normalisation et nous avons choisi de conserver la normalisation standard (à chaque valeur on soustrait la moyenne et on divise par l'écart type) puisqu'elle donne les meilleurs résultats.

5.2.2 Modèles et protocole d'apprentissage

En ce qui concerne les systèmes de Machine Learning mis en place, nous avons décidé d'utiliser des modèles simples adaptés à notre tâche : la régression logistique et un SVM que nous avons décrit au chapitre 2. Ces choix proviennent de précédentes expérimentations que nous avons mis en place sur une classification antérieure qui ne sera

Modèle	Modalité	# descripteurs	UAP	UAR
3 classes : satisfait, neutre et frustré				
LR	audio	100	0.31	0.35
	texte	100	0.52	0.52
	fusion	200	0.42	0.46
SVM	audio	100	0.36	0.39
	texte	100	0.44	0.48
	fusion	200	0.40	0.45
2 classes : neutre et frustré				
LR	audio	100	0.46	0.52
	texte	100	0.78	0.78
	fusion	200	0.63	0.68
SVM	audio	100	0.54	0.58
	texte	100	0.67	0.72
	fusion	200	0.61	0.67

TABLE 5.2 – Scores des systèmes de classification sur l’émotion finale de la conversation. UAP correspond à *unweighted average precision* et UAR à *unweighted average recall*.

pas présentée dans ce manuscrit. Malheureusement, suite à une mise en conformité avec la RGPD, nous avons dû supprimer des données et les expérimentations qui avaient été menées avec. Ces deux classifieurs sont implémentés en utilisant l’outil scikit-learn également.

La classification se fait en validation croisée (*k-folds*) avec $k = 5$, c’est-à-dire que nous avons divisé les conversations du train et du dev en k échantillons. Pour chaque pli, les ensembles d’apprentissage et de validation changent et un nouvel apprentissage est lancé. Chaque expérience consiste donc en cinq apprentissages. Le score final correspond à la moyenne de ces cinq scores de performance sur l’ensemble de test. Ce procédé est très utilisé pour éviter le sur-apprentissage dans le cas où peu de données sont disponibles.

Nous rappelons ici que les fins de conversations ont été annotées satisfaites (23), neutre (85) ou frustrées (195). Les modèles ont été évalués avec les mesures de précision et rappel non pondérées, que nous avons introduites au chapitre 3.

5.2.3 Résultats et analyse

Les performances des différents modèles sont indiquées dans le tableau 5.2. Nous pouvons observer que la régression logistique donne des meilleurs résultats sur la modalité linguistique et la fusion des deux modalités. Nous voyons également que la modalité

linguistique donne toujours de meilleurs résultats. Comme il y a peu de conversations satisfaites dans nos jeux de données, nous avons fait le choix de supprimer entièrement la classe satisfaite et de ré-apprendre un système avec deux classes : neutre et frustré¹. Ce modèle est donc appris avec moins de documents que le précédent. Par ce procédé, nous obtenons un score tout à fait correcte avec la modalité linguistiques de $UA = 78\%$.

Bien que simple, cette première classification nous a permis de confirmer l'importance de la modalité linguistique, dont nous reparlerons dans le chapitre 6. La prédiction de la classe de fin de conversation reste une tâche simpliste et préliminaire. Nous souhaitons mettre en place à présent une reconnaissance des émotions à chaque instant de la conversation.

5.3 Reconnaissance continue de la dimension de satisfaction/frustration

Pour mettre en place cette reconnaissance de l'émotion continue, nous avons choisi dans un premier temps de nous comparer à l'état de l'art et donc de reproduire les expérimentations des articles de référence [SCS19; Kos+19] et des campagnes AVEC [Rin+17; Rin+18; Rin+19].

Ces travaux portent sur le traitement de séquences audio de taille fixe. La sous partie de SEWA (Allemand et Hongrois) à laquelle nous avons accès ne comporte que des conversations de durée inférieure à trois minutes. De plus, un padding circulaire est utilisé pour ramener toutes ces conversations à une durée de trois minutes dans la plupart des travaux. Ce positionnement, souvent adopté, permet notamment d'utiliser toutes les architectures neuronales (notamment les CNN). Il est adapté lorsque nous avons des données assez homogènes, dans notre cas en terme de durée, comme c'est le cas avec le corpus SEWA.

Comme nous l'avons vu dans le chapitre 4, les conversations d'AlloSat ont des durées très variables entre 32 secondes et 41 minutes avec une moyenne (MOY) de 7m24s et un écart type (STD) de 4m58s. Une bonne pratique courante du domaine est de fixer la taille d'entrée à $MOY + STD$ (ici 12m22s) pour couvrir statistiquement plus de 95% du corpus. Les séquences longues sont alors coupées à $MOY + STD$ tandis que les séquences courtes sont rallongées avec un padding. Dans notre cas, cela reviendrait à traiter des conversations de 12m22s.

1. fusionner la satisfaction et le neutre donne des résultats presque identiques.

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 sept minutes, soit environ à la moyenne. Nous avons appliqué un padding circulaire sur les courtes séquences.

En résumé nous avons trois cas possibles :

- pour une conversation de durée inférieure à sept minutes : on prend la conversation et on fait un padding circulaire, c’est-à-dire qu’à la fin de la séquence, on concatène le début de la conversation. Par exemple, pour un document de cinq minutes, nous concaténons les deux premières minutes du document à la suite de la fin de ce dernier,
- pour une conversation de durée égale à sept minutes : on prend la conversation en entier, sans traitement,
- pour une conversation de durée supérieure à sept minutes : on supprime tout ce qui est au delà des sept minutes.

Une fois ce pré-traitement des données mis en place, nous avons commencé à travailler sur les descripteurs acoustiques.

5.3.1 Exploration des ensembles de descripteurs eGeMAPS

Pour mieux comparer notre travail avec l’état de l’art dans le domaine du SER, nous avons décidé d’utiliser l’ensemble eGeMAPS [Eyb+16], que nous avons détaillé dans le chapitre 3. Pour rappel, cet ensemble est constitué de descripteurs de bas niveau et de fonctions statistiques appliquées sur ces derniers permettant ainsi de capturer des informations prosodiques et spectrales. Cet ensemble regroupe 88 descripteurs.

Dans les travaux de Schmitt et al. [SCS19], un sous ensemble, appelé f_eGeMAPS, a été défini à partir de 23 LLD et des fonctions statistiques appliquées sur ces LLD (principalement moyenne et STD). Il totalise 46 descripteurs. Un dernier descripteur, fonctionnant comme une détection de voix (*voice activity detection* i.e. vad), dénotant l’identité du locuteur (zéro ou un), est également inclus dans f_eGeMAPS, portant le nombre de descripteurs à 47. Ce descripteur est directement lié aux choix de concept du corpus SEWA. En effet, il est composé de conversations entre deux personnes qui sont enregistrés dans le même canal, ce qui signifie que pour analyser séparément les deux interlocuteurs, une même conversation est présentée dans deux documents distincts. Pour dissocier ces deux documents, on utilise le descripteur de vad.

Dans notre travail, eGeMAPS et f_eGeMAPS ont été extraits de nos données toutes les 250 ms de manière synchronisée avec le pas d’annotation d’AlloSat et l’outil Ope-

nEAR [EWS09], qui s’appuie sur OpenSMILE [EWS10]. Puisque nous ne gardons que le signal de l’appelant, nous ajoutons un descripteur, appelé *vad*, pour indiquer si l’appelant parle (1) ou non (0). Ce descripteur est déduit des transcriptions automatiques.

Nous avons donc comparé un total de quatre ensembles de descripteurs :

- **eGeMAPS-88** : L’ensemble de descripteurs eGeMAPS, extrait avec OpenEAR par pas de 250 ms,
- **eGeMAPS-89** : Même ensemble que précédemment mais avec l’ajout du descripteur *vad* qui modélise la présence ou l’absence de parole de la part du locuteur.
- **eGeMAPS-46** : Le sous-ensemble de descripteurs utilisé par Schmitt et al. [SCS19], extrait avec OpenSMILE par pas de 250 ms.
- **eGeMAPS-47** : Même ensemble que précédemment mais avec l’ajout du descripteur *vad* qui modélise la présence ou l’absence de parole de la part du locuteur.

Pour réaliser cette reconnaissance, nous utilisons deux réseaux de neurones illustrés dans la figure 5.1. 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 [Rin+18] sur la modalité *Cross-cultural Affect*. Ce réseau neuronal, correspondant à la figure 5.1 gauche, est composé de deux couches biLSTM de respectivement 64 et 32 neurones. L’architecture bidirectionnelle est utilisée afin notamment d’éviter les problèmes de différences de délai entre les annotateurs. En effet, il est possible que l’annotation présente des délais, 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 nombre d’époques doit être déterminé par l’expérimentation en elle-même, nous avons donc vérifié à posteriori qu’il n’y avait plus d’amélioration du score après 200 époques. Nous avons contrôlé jusqu’à 500 époques et fixé le nombre d’époques maximum à 200.

Le second réseau, correspondant à la figure 5.1 à droite, est composé de quatre couches biLSTM comme décrit dans [SCS19]. Les couches de biLSTM sont composées respectivement de 200, 64, 32 et 32 neurones. Ce réseau est plus profond, nous avons donc augmenté le nombre d’époques à 500.

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 250 ms. Comme il s’agit d’une régression et pour être comparable aux articles de référence, on utilise le coefficient de corrélation de concordance (CCC) [Lin89] comme fonction de coût pour l’apprentissage du réseau et comme métrique d’évaluation pour déterminer le meilleur système. Pour rappel, ce score CCC varie de 0 (probabilité d’un tirage aléatoire)

à 1 (corrélation parfaite).

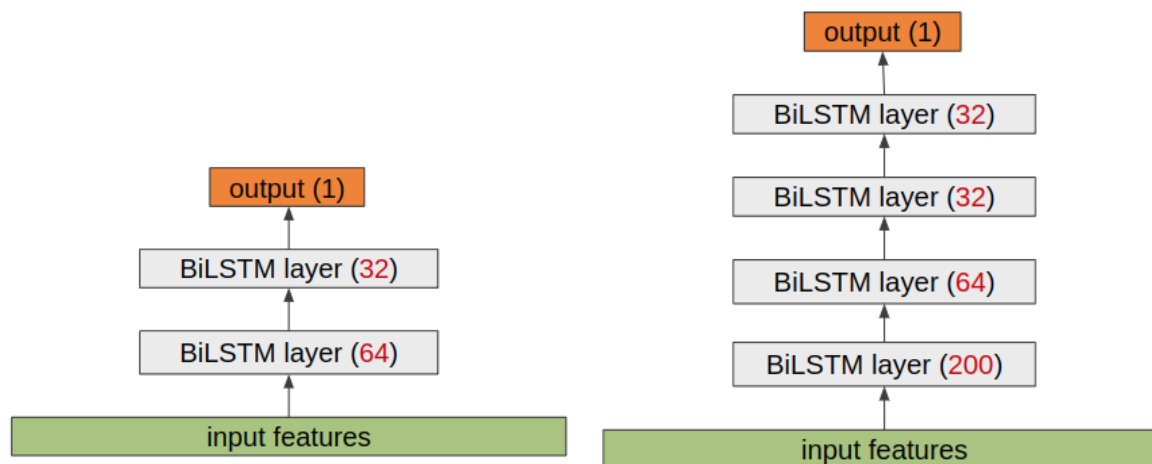


FIGURE 5.1 – Schéma de la configuration des systèmes neuronaux appelés biLSTM-2 (à gauche) et biLSTM-4 (à droite). Le nombre de neurones est indiqué en rouge et entre parenthèses sur chaque couche. Comme il s’agit de réseaux bidirectionnels, il faut multiplier par deux le nombre de neurones pour avoir le nombre de paramètres réels.

Ces premiers réseaux sont implémentés avec le framework Keras² en utilisant Tensorflow³. L’apprentissage se fait par batch de neuf conversations en utilisant l’optimiseur ADAgrad [DHS11] dont nous avons parlé au chapitre 2. Les conversations sont mélangées aléatoirement entre chaque époque. Le learning rate est initialisé à 0,001.

Nous avons conservé les poids des réseaux donnant le meilleur score sur le développement afin de prédire les résultats sur le test.

Ensemble de descripteurs	biLSTM-2		biLSTM-4	
	dev	test	dev	test
eGeMAPS-88	0.510	0.363	0.666	0.431
eGeMAPS-89	0.549	0.365	0.619	0.542
f_eGeMAPS-46	0.469	0.260	0.607	0.354
f_eGeMAPS-47	0.508	0.359	0.574	0.422

TABLE 5.3 – Score CCC des systèmes de reconnaissance des émotions en utilisant quatre ensembles de descripteurs différents et deux architectures neuronales sur les ensembles de développement et de test d’AlloSat.

2. <https://keras.io>

3. <https://www.tensorflow.org/>

Le tableau 5.3 donne un résumé des résultats obtenus avec les modèles et les ensembles de descripteurs étudiés. Nous pouvons remarquer que les ensembles eGeMAPS-88 et 89 donnent de meilleurs résultats sur les deux architectures neuronales que f_eGeMAPS-46 et 47. Nous remarquons également que l’ajout d’un descripteur *vad* permet de mieux généraliser puisque les scores sur le test sont toujours meilleurs lorsque l’on a ce descripteur. De plus, l’architecture qui donne les meilleurs résultats est biLSTM-4 quel que soit l’ensemble de descripteurs utilisé.

La meilleure configuration pour ce système de reconnaissance est donc le suivant :

- Ensemble de descripteurs eGeMAPS-89 (avec la vad),
- Architecture neuronale biLSTM-4 comportant quatre couches de biLSTM.

Ces premières expériences montrent que les réseaux neuronaux biLSTM sont capables de prédire les valeurs de l’axe de satisfaction et donc de retracer cette dimension au cours d’un appel avec un score CCC correct. Néanmoins le score CCC calculé sur l’ensemble des données doit être pris avec précaution car, comme nous le montrons dans la Figure 5.2, le système est capable de faire de bonnes prédictions (conversation C) mais aussi de mauvaises prédictions (conversation D). Ces résultats ont fait l’objet d’une publication à la conférence LREC (Language Resources and Evaluation Conference) de 2020 [Mac+20a].

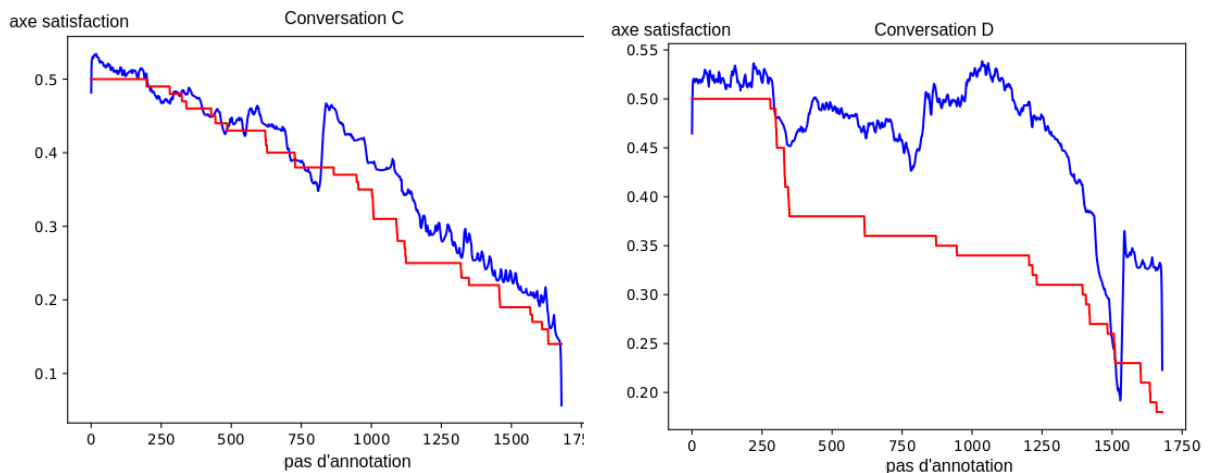


FIGURE 5.2 – Prédiction de la satisfaction sur des conversations issues du test. La référence est en rouge, la prédiction en bleu.

5.3.2 Comparaison eGeMAPS et MFCC

En reconnaissance des émotions, la plupart du temps, la représentation acoustique se concentre principalement sur l’extraction de la prosodie. C’est pour cette raison que les ensembles de descripteurs dits *experts*, comme par exemple GeMAPS ou IS2009, sont majoritairement utilisés.

Cependant, dans notre contexte nous traitons des conversations téléphoniques, où le signal audio peut être plus ou moins dégradé et l’extraction de ces ensembles de descripteurs *experts* peut contenir beaucoup d’erreurs. Par exemple la détection des formants est généralement peu robuste au bruit, et le descripteur f_0 peut être également fortement dégradé.

Il peut donc être intéressant de comparer l’ensemble eGeMAPS à des descripteurs plus robustes aux signaux dégradés. Nous avons fait le choix de les comparer aux MFCC, décrit au chapitre 3, qui sont plus robustes : leurs extracteurs sont plus fiables dans des contextes de signaux dégradés. De plus, comme nous l’avons dit précédemment dans le chapitre 3, les MFCCs permettent une représentation compacte du signal avec un maximum d’informations non redondantes.

Nous avons décidé de comparer les MFCCs issus de deux outils d’extraction différents : OpenSMILE (utilisé pour extraire les ensembles IS2009 ou dans notre cas GeMAPS et toutes ses dérivées) ainsi que librosa⁴. En effet, il est possible d’observer des variations entre plusieurs extracteurs [GFK05] qui peuvent être liées à la normalisation, au choix de l’échelle Mel, à la fenêtre d’analyse ou encore aux paramètres de FFT. Une fois ces extractions effectuées, nous avons mis en place deux protocoles pour les adapter à nos segments émotionnels de 250 ms :

- **Mfcc-Os** correspond aux MFCC 1 à 13 ainsi qu’à leurs dérivés premières et secondes pour qualifier la dynamique du signal. Ces 39 descripteurs ont été extraits avec l’outil OpenSMILE. Un segment émotionnel est alors représenté par la moyenne et l’écart type de ces 39 descripteurs. Ainsi nous avons un ensemble de 78 descripteurs pour chaque segment émotionnel.
- **Mfcc-lib** correspond aux MFCC 1 à 24 qui sont extraits tous les 10 ms sur des fenêtres de 30 ms. Ces 24 descripteurs ont été extraits avec l’outil librosa. Comme pour les descripteurs précédents, nous avons calculé la moyenne et l’écart type de ces descripteurs au niveau du segment émotionnel. Ainsi nous avons un ensemble de 48 descripteurs pour chaque segment émotionnel. A noter que nous n’avons pas

4. <https://librosa.github.io/librosa/>

de dérivées dans cet ensemble contrairement à l'ensemble Mfcc-Os.

Nous utilisons la meilleure des deux architectures neuronales de la section précédente pour faire notre comparatif, à savoir le biLSTM-4. Les résultats sont regroupés dans le tableau 5.4.

Features	AlloSat	
	DEV	TEST
eGeMAPS-88	0.666	0.431
eGeMAPS-46	0.607	0.354
Mfcc-lib 48	0.675	0.510
Mfcc-Os 78	0.382	0.299

TABLE 5.4 – Comparaison de performance entre descripteurs experts et MFCCs. Score CCC rapporté sur les différents ensembles de descripteurs sur le dev et le test d'AlloSat. Le modèle utilisé est le biLSTM-4.

Nous pouvons observer tout d'abord une forte différence de score entre les deux types de MFCCs. Outre la différence d'implémentation de l'extraction, nous avons mis en place deux protocoles différents pour ces descripteurs (présence ou absence des dérivées, nombre différents de paramètres,...), ce qui peut expliquer cette différence de score. On observe également que les Mfcc-lib ont une meilleure performance que l'ensemble eGeMAPS-89 sur le dev. Ce score semble confirmer que les MFCCs sont plus adaptés au signal téléphonique, pour notre corpus, sur la tâche de reconnaissance des émotions.

5.3.3 Comparaison CNN et biLSTM

Dans les travaux menés par Schmitt et al. [SCS19], les architectures CNN et biLSTM sont comparées. Le postulat de ces travaux est que nous avons trop tendance à complexifier nos architectures neuronales dans les expériences, et qu'à complexité égale, un système à base de CNN peut faire aussi bien voir mieux qu'un système à base de réseaux récurrents. Le résultat de leur expérimentation confirme cette hypothèse : à complexité égale, leur CNN a de meilleures performances que leur biLSTM sur le corpus SEWA.

Nous avons donc voulu confronter leur conclusion à notre propre corpus, pour voir si nous pouvions améliorer nos résultats avec des architectures moins complexes que des réseaux récurrents LSTM bidirectionnels. Cela permettrait d'avoir moins de paramètres à apprendre et semble donc pertinent lorsque l'on a des bases de données en quantité limitée, comme avec le corpus AlloSat. En plus des deux architectures biLSTM précédemment expliquées, nous avons mis en place une architecture à base de quatre couches

convolutionnelles, avec une activation de type ReLU comme illustré dans lma Figure 5.3. Comme pour les autres systèmes, un unique neurone en sortie permet de faire la prédiction continue de la valeur de satisfaction.

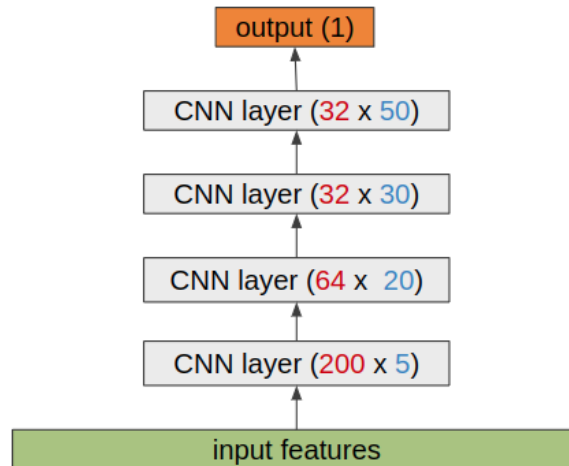


FIGURE 5.3 – Schéma de la configuration du système neuronal appelé CNN-4. Le nombre de neurones est en rouge et le filtre récepteur en bleu.

Nous avons décidé de comparer nos résultats avec ceux des travaux réalisés sur le corpus SEWA. Pour cela, nous avons utilisé l’ensemble de descripteurs eGeMAPS-47, utilisé par Schmitt et al. [SCS19], sur les conversations de taille fixe, donc paddés à 7 minutes. Nous avons donc refait leur expérimentation sur le corpus SEWA et nous l’avons adapté au corpus AlloSat, en partant de l’implémentation de la baseline du challenge AVEC disponible sur github⁵.

Nous avons eu quelques difficultés à faire converger le système convolutif sur nos données. Nous avons fait varier le learning rate, l’initialisation des poids et le nombre d’époques mais nous avons remarqué que certaines initialisations de poids ne permettent pas de faire converger le système. Nous avons donc fait le choix de présenter dans le tableau 5.5 la moyenne de cinq modèles dont les poids ont été initialisés différemment (contrôlé par le paramètre *seed*), ainsi que le meilleur des cinq meilleurs modèles sur l’ensemble de développement.

Comme nous pouvons l’observer, nous avons un grand écart entre la moyenne des scores et le score maximal pour le CNN sur le corpus AlloSat. En effet, sur les cinq systèmes, seuls deux ont fini par converger aux alentours de la 80ème époque. Avec un

5. <https://github.com/AudioVisualEmotionChallenge/AVEC2019>

Modèles	Descripteurs	AlloSat	SEWA		
		satisfaction	activation	valence	liking
Reproduction des résultats avec nos systèmes					
CNN	eGeMAPS-47	.178 (.458)	.528 (.541)	.515 (.527)	.304 (.321)
biLSTM-4	eGeMAPS-47	.437 (.458)	.487 (.527)	.428 (.468)	.258 (.346)

TABLE 5.5 – Comparaison des moyennes (maximum) des scores CCC de cinq systèmes différents sur les ensembles de développement d’AlloSat et de SEWA.

score de 0.437 de moyenne sur la satisfaction, le système biLSTM-4 donne une meilleure performance sur la moyenne des modèles appris. Cela nous montre que cette architecture est plus stable et moins sujette à la différence de l’initialisation : on a une différence de 0.021 de score entre le meilleur modèle et la moyenne des scores de cinq modèles différents. Alors que le CNN, même si le meilleur modèle nous donne un score similaire au biLSTM-4, on observe qu’il est bien plus sujet à la différence d’initialisation et donc bien moins stable. De plus, nous n’avons pas réussi à reproduire parfaitement les scores sur le corpus SEWA. Notamment parce que, une fois encore, l’initialisation des réseaux joue un très grand rôle dans le score final.

Pour conclure, nous avons montré que les systèmes à base de couches convolutionnelles ne sont pas tout à fait adaptés à nos données. Ces travaux ont été présentés à la conférence SPECOM (Conference on Speech and Computer) [Mac+20b]. Ces résultats sont à nuancer notamment à cause de la difficulté de convergence des réseaux convolutifs. Il n’est pas à exclure une mauvaise implémentation de cet apprentissage ou une erreur lors du lancement des expérimentations, malgré tout le soin apporté.

Notre objectif, à terme, est de travailler sur des séquences de taille non fixe, ce qui ne peut pas être réalisé avec des CNN. Cette volonté correspond aux besoins de l’entreprise qui a besoin d’une analyse de l’émotion sur les conversations entières et non sur une partie uniquement.

Afin de comprendre les différences de scores entre les corpus SEWA et AlloSat, nous avons décidé d’enquêter sur les différences inhérentes à ces deux corpus.

5.4 Comparaison entre AlloSat et SEWA

5.4.1 Performances obtenues sur les deux corpus

Avant d’étudier les différences intrinsèques aux deux corpus, nous rappelons tout d’abord les scores obtenus par les campagnes AVEC 2018 et 2019 ainsi que ceux obtenus par Schmitt et al., que nous comparons aux scores obtenus sur AlloSat. Le tableau 5.6 présente l’ensemble des résultats comparables que nous avons obtenus sur le corpus AlloSat et le corpus SEWA avec les différents modèles ainsi que les performances obtenues sur SEWA dans l’état de l’art.

L’initialisation étant un facteur important dans le score final des systèmes, nous avons fait le choix de présenter à la fois la moyenne de cinq systèmes initialisés aléatoirement et le meilleur de ces cinq systèmes (score indiqué entre parenthèses dans le tableau). Comme nous pouvons l’observer, les résultats de nos modèles sont comparables à ceux obtenus dans les différentes baselines sur le corpus SEWA. Ils sont meilleurs que ceux diffusés lors de la campagne de 2018 mais moins performants que ceux de la campagne 2019. Les auteurs de la campagne ont eux-même reconnu que des variations de scores peuvent être observées en fonction de l’initialisation.

On remarque également que nous n’avons pas réussi à bien reproduire les résultats obtenus par Schmitt et al. sur l’activation et la valence. Ceci peut être dû notamment à l’initialisation mais également à la gestion du délai des annotations qui n’est pas pris en compte dans nos expérimentations. Nos systèmes sont plus performants sur la dimension du *liking*, quelle que soit les architectures utilisées.

Nous observons également, en comparant les meilleures performances et les performances moyennes, que le système biLSTM-2 entraîné avec les descripteurs Mfcc-Os sur les conversations allemandes et hongroises (en violet) semble être le moins sensible à l’initialisation des poids. De façon générale, le plus grand écart type est observé sur la dimension du *liking*.

Les performances obtenues sur l’axe de satisfaction sont comparables à celles obtenues sur les autres dimensions dans le cas de systèmes à base de réseaux récurrents. Les réseaux convolutifs, comme nous l’avons énoncé auparavant, ne semblent pas convenir à notre corpus. La prédiction de l’axe de satisfaction est plus performante avec des réseaux récurrents que des réseaux convolutifs. Il est donc plus prudent d’utiliser des réseaux récurrents, puisqu’ils semblent plus robustes aux changements d’axe émotionnel et aux changements de corpus. De plus, nous avons eu des difficultés à faire converger les systèmes convolutifs

Modèles	Descripteurs	AlloSat-Dev	SEWA-Dev		
		satisfaction	activation	valence	liking
Nos systèmes : pour SEWA Train et Dev sur les conversations allemandes					
CNN	eGeMAPS-47	.178 (.458)	.528 (.541)	.515 (.527)	.304 (.321)
biLSTM-4	eGeMAPS-47	.437 (.458)	.487 (.527)	.428 (.468)	.258 (.346)
biLSTM-2	eGeMAPS-88	.480 (.564)	.280 (.357)	.174 (.212)	.095 (.171)
biLSTM-2	Mfcc-Os	.364 (.439)	.395 (.438)	.325 (.373)	.158 (.208)
biLSTM-2	eGeMAPS-88*	.480 (.564)	.244 (.273)	.118 (.155)	.082 (.132)
biLSTM-2	Mfcc-Os*	.364 (.439)	.325 (.326)	.186 (.192)	.125 (.126)
biLSTM-4	eGeMAPS-88	.564 (.634)	.316 (.429)	.237 (.309)	.119 (.188)
biLSTM-4	Mfcc-lib	.666 (.675)	.489 (.501)	.449 (.464)	.124 (.133)
biLSTM-4	Mfcc-Os	.374 (.382)	-	-	-
AVEC 2018 : Train et Dev sur les conversations allemandes [Rin+18]					
biLSTM-2	eGeMAPS-88		(.124)	(.112)	(.001)
biLSTM-2	Mfcc-Os		(.253)	(.217)	(.136)
AVEC 2019 : Train et Dev sur les conversations allemandes et hongroises [Rin+19]					
biLSTM-2	eGeMAPS-88*		(.371)	(.286)	(.159)
biLSTM-2	Mfcc-Os*		(.326)	(.187)	(.144)
Schmitt et al. : Train et Dev sur les conversations allemandes [SCS19]					
CNN	eGeMAPS-47		(.571)	(.517)	
biLSTM-4	eGeMAPS-47		(.568)	(.561)	

TABLE 5.6 – Comparaison des scores moyens de CCC sur les corpus AlloSat et SEWA selon quatre dimensions émotionnelles : la satisfaction, l’activation, la valence et le liking. Nos scores correspondent à la moyenne des scores de 5 systèmes appris avec des initialisations aléatoires différentes et entre parenthèses nous retrouvons le score du meilleur des modèles. Nous reportons également les résultats inclus dans les papiers [Rin+18 ; Rin+19 ; SCS19], qui constituent notre base de comparaison. Ces scores correspondent au meilleur de leur expérimentation, c’est pour cela qu’ils sont notés entre parenthèses, puisqu’à comparer avec les scores des meilleurs modèles notés entre parenthèses dans nos résultats. *L’entraînement et les prédictions sont réalisés sur les conversations allemandes et hongroises. Sans le sigle, l’entraînement et les prédictions sont réalisés sur les conversations allemandes uniquement. Les couleurs permettent de repérer les différentes expérimentations comparables.

selon les différents paramètres et hyper-paramètres du système.

En se concentrant sur la prédiction de l’axe de satisfaction, le modèle appris avec eGeMAPS-88 ($CCC = 0.480$) semble être plus performant que celui appris avec Mfcc-Os ($CCC = 0.364$) quand on considère l’architecture biLSTM-2. Cependant, en utilisant le système biLSTM-4, les descripteurs eGeMAPS-47 atteignent également de bons résultats ($CCC = 0.437$). Pour conclure sur le meilleur nombre de couches, nous effectuons une

dernière expérience avec le système biLSTM-4 et eGeMAPS-88 ($CCC = 0.564$) qui obtient un très bon résultat. Cette architecture semble être la mieux adaptée pour l’axe de satisfaction. Nous avons également reporté les résultats avec les descripteurs Mfcc-lib et Mfcc-Os pour le corpus AlloSat : notre meilleur modèle correspond au système biLSTM-4 et Mfcc-lib ($CCC = 0.666$).

De cette expérience, nous concluons que le système biLSTM-4 est la meilleure des trois architectures comparées en ce qui concerne la robustesse à la variabilité induite par l’utilisation de différents corpus d’émotions et d’axe émotionnel. Nous confirmons que l’ensemble de descripteurs Mfcc-lib est celui qui représente le mieux les données contenus dans AlloSat. Dans la section suivante, nous détaillons les différences entre les deux corpus qui pourraient expliquer la différence de performance entre les systèmes sur ces deux corpus.

5.4.2 Analyse des différences entre SEWA et AlloSat

Nous souhaitons comprendre pourquoi certaines architectures et certains ensembles de descripteurs fonctionnent sur SEWA et non sur AlloSat et inversement. Pour cela, nous cherchons à rendre les corpus plus semblables.

Variation de l’annotation : pas de l’annotation, taux d’échantillonnage

Nos premières hypothèses ont porté sur la durée du pas d’annotation et sur la différence de taux d’échantillonnage entre les deux corpus. Nous avons donc décidé de changer les pas d’annotation des deux corpus pour qu’ils soient comparables. Comme nous avons un corpus annoté toutes les 100 ms (SEWA) et l’autre annoté toutes les 250 ms (AlloSat), nous avons choisi de considérer l’annotation toutes les 500 ms. Pour cela, nous avons utilisé une moyenne glissante, afin de rester cohérent avec les annotations d’origines. Nous avons dans le même temps sur-échantillonné les conversations issues d’AlloSat, afin de se ramener à deux corpus échantillonnés selon le même taux soit 44,1kHz. Nous avons choisi de représenter ces échantillons avec l’ensemble eGeMAPS-47 pour pouvoir être comparable à nos précédentes expérimentations.

Nous souhaitons voir des scores du même ordre de grandeur pour trouver des justifications à la différence de scores pointée entre les deux corpus dans la section précédente en fonction de l’architecture et des descripteurs utilisés.

Les résultats, présentés sur le tableau 5.7 ne permettent de valider aucune des deux

	AlloSat-Dev	SEWA-Dev		
	satisfaction	activation	valence	liking
CNN	.046 (.151)	.368 (.401)	.375 (.424)	.074 (.089)
biLSTM-4	.503 (.517)	.457 (.476)	.411 (.420)	.248 (.267)

TABLE 5.7 – Architectures entraînées sur l’ensemble de développement d’AlloSat et de SEWA en prenant en entrée des segments émotionnels de 500 ms. Comparaison entre deux architectures neuronales : CNN et biLSTM-4 avec eGeMAPS-47.

hypothèses. En effet, les résultats sont comparables aux précédents.

Variation de l’annotation : intensité de la dynamique

En comparant les deux corpus, nous remarquons que l’axe de satisfaction varie très lentement dans le temps par rapport à l’activation, la valence et au liking. Cela peut être dû au protocole d’annotation (la souris ou le joystick), aux guidelines et au contenu affectif annoté.

Pour déterminer si la différence de dynamique dans l’annotation est responsable des mauvais résultats des systèmes convolutifs sur la satisfaction, nous effectuons des expériences supplémentaires avec des références lissées pour l’activation et la valence. Pour mettre en place le lissage, nous avons utilisé une moyenne glissante sur des fenêtres de durées un multiple entier de la durée du pas d’annotation, soit 200, 500 ms et 1s sur SEWA, ainsi que de 500 ms et 1s sur AlloSat. Les résultats, rapportés dans le tableau 5.8 montrent qu’aucun de ces procédés ne permet d’améliorer les performances des systèmes.

satisfaction	AlloSat-Dev	SEWA-Dev		
	activation	valence	liking	
original	.178 (.458)	.528 (.541)	.515 (.527)	.304 (.321)
200ms	–	.512 (.549)	.484 (.492)	.340 (.368)
500ms	.177 (.467)	.514 (.537)	.491 (.526)	.321 (.379)
1000ms	.185 (.475)	.519 (.548)	.505 (.518)	.329 (.359)

TABLE 5.8 – Comparaison des scores de développement sur les annotations originales et lissées sur les corpus AlloSat et SEWA. Les descripteurs utilisés sont les eGeMAPS-47 features et le système utilisé est le CNN.

On peut en conclure que la dynamique faible de l’annotation présente dans le corpus AlloSat n’est pas responsable de la faible performance du système à base de réseaux convolutionnels.

Impact du bruit téléphonique

Une autre piste que nous avons suivie pour expliquer la dégradation des scores d’AlloSat avec un CNN provient de la qualité téléphonique. En classification d’images, des études [Roy+18; DK16] ont montré que l’entraînement de modèles à base de CNN avec des images de mauvaise qualité peut considérablement dégrader les performances des modèles de classification d’images. Comme AlloSat est composé de conversations téléphoniques, l’enregistrement audio est échantillonné en 8 kHz et il est rempli de bruits de fond, coupures de son, changement de haut-parleur, etc. Nous émettons donc l’hypothèse qu’AlloSat est trop dégradé pour être bien traité par des réseaux convolutifs. Pour confirmer notre hypothèse, nous avons sous-échantillonné les données SEWA de 44 kHz à 8 kHz.

Nous proposons également d’ajouter du bruit au jeu de données SEWA, afin de le rendre plus comparable avec le contexte AlloSat. Les bruits d’appels téléphoniques sont principalement composés de bruits de fond tels que les bruits de la rue ou de la circulation.

Pour injecter du bruit dans nos enregistrements, nous avons utilisé la base de données SoundJay⁶ et nous avons suivi le processus décrit dans l’article [ANG16]. Trois types de bruits (voiture, rue et conversations de foule) sont ajoutés de manière aléatoire au signal à différents niveaux de volumes (9, 15, 21 dB). En plus de ces bruits, nous avons fait le choix d’ajouter des voix d’enfants, puisque l’on en entend assez souvent dans le corpus. Les enregistrements résultants ont été vérifiés manuellement pour s’assurer que les perturbations auditives sont comparables à celles observées dans AlloSat.

	activation	valence	liking
44kHz	.528	.515	.304
8kHz	.486	.495	.187
8kHz + bruit	.471	.464	.204

TABLE 5.9 – Comparaison des scores de développement entre les versions originales du corpus SEWA et les versions dégradées. Les dégradations sont le sous-échantillonnage des enregistrements audio et l’ajout de différents bruits. Nous utilisons les descripteurs eGeMAPS-47 ainsi qu’un modèle CNN. L’entraînement et les prédictions sont effectuées sur les conversations allemandes.

Les résultats, présentés dans le tableau 5.9, montrent qu’en dégradant les signaux SEWA (en sous-échantillonnant à 8 kHz et en ajoutant du bruit), nous diminuons les performances. Mais nous sommes loin d’atteindre les mauvaises performances des réseaux convolutifs sur AlloSat.

6. <https://www.soundjay.com/>

Toutes nos hypothèses ne nous permettent pas de justifier la différence de score entre AlloSat et SEWA lorsque l'on utilise des réseaux convolutifs. Les bonnes performances obtenues sur SEWA avec le CNN dans nos expérimentations nous font douter que le problème vienne de l'implémentation en elle-même. Le changement de la dimension annotée, les conditions d'enregistrement, le contenu sémantique, la langue, le choix des hyperparamètres ou encore le protocole mis en œuvre sont autant d'explications possibles quant à la différence de performance de ces réseaux sur les deux corpus.

5.5 Analyses annexes

5.5.1 Fonction de coût

Comme nous l'avons vu précédemment dans le chapitre 3, le CCC est un outil d'évaluation très utilisé dans la reconnaissance d'émotions continues. Les challenges AVEC [Rin+18] ont notamment contribué à la standardisation de l'utilisation de cette métrique pour évaluer les systèmes de reconnaissance continue d'émotion. Pour rappel, le CCC se calcule selon l'équation 5.1.

$$CCC = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2 + \epsilon} \quad (5.1)$$

Si on regarde l'équation, quand la référence est constante sur la totalité de la conversation, $\sigma_y = 0$ et donc le coefficient est égal à zéro. De façon plus générale, quand la référence varie peu, le CCC va s'approcher de zéro, même si la prédiction est quasiment parfaitement synchronisée à la référence.

On peut donc en conclure que la fonction de coût va pénaliser les conversations où la référence varie peu ($\sigma_y \simeq 0$) et que le système ainsi entraîné aura du mal à prédire correctement de telles références.

Nous avons comparé les scores obtenus en utilisant la RMSE comme fonction de coût que nous avons introduit au chapitre 3 pour pallier ce problème. Le tableau 5.10 résume les différences de scores calculés sur le dev.

Comme nous pouvons l'observer, la fonction de coût *l-ccc* a un léger avantage en moyenne sur la *l-rmse*. On remarque cependant que dans le cas des descripteurs Mfcc-lib, l'utilisation de la fonction de coût *l-rmse* permet d'améliorer les résultats. Ne permettant pas une amélioration significative des résultats, nous choisissons de continuer à utiliser la fonction de coût CCC sauf pour les expériences utilisant Mfcc-lib. Cependant lorsque l'on

Descripteurs	AlloSat	
	$l\text{-ccc}$	$l\text{-rmse}$
eGeMAPS-47	.437	.381
eGeMAPS-88	.564	.514
Mfcc-lib	.675	.698
Mfcc-Os	.382	.405
Moyenne	.515	.500

TABLE 5.10 – Comparaison de l’utilisation de deux fonctions de coût $l\text{-ccc}$ et $l\text{-rmse}$. Score CCC des systèmes de reconnaissance des émotions sur l’ensemble de développement d’AlloSat. Le modèle utilisé est un biLSTM-4.

regarde la prédiction effective sur les conversations, on se rend compte que les prédictions ne sont pas homogènes en terme de qualité, comme en témoigne la Figure 5.4 où à gauche nous retrouvons une prédiction peu satisfaisante avec un score CCC de 0.564 alors qu’à droite nous retrouvons une très bonne prédiction avec un score CCC de 0.903.

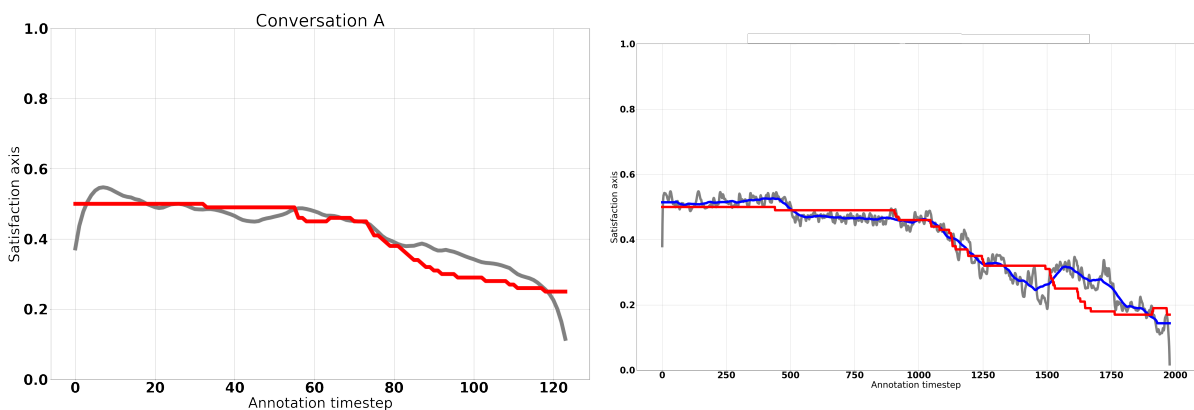


FIGURE 5.4 – Evolution des prédictions (grises) et des références (rouge) de deux conversations provenant de l’ensemble de test d’AlloSat. $ccc(A) = 0.564$, $ccc(B) = 0.903$. La courbe bleue correspond au lissage des prédictions. Obtenu avec le système biLSTM-4 ($l\text{-rmse}$) et Mfcc-lib.

Nous avons donc décidé de mettre en place un post-traitement afin de lisser les courbes de prédiction, pour avoir une meilleure représentation des émotions.

5.5.2 Post-traitement : lissage des prédictions

Nous avons décidé de mettre en place un lissage des prédictions. Pour cela, nous avons utilisé l’algorithme de lissage Savistky-Golay [SG64] avec un degré polynomial de zéro.

Bien qu'ancien, cet algorithme est toujours autant utilisé dans le traitement des signaux. C'est une extension de la moyenne glissante, qui est utilisée pour atténuer les pics des signaux en approximant un polynôme pour chaque fenêtre de la moyenne glissante, ici un polynôme de degré zéro. Nous avons fait le choix de ce degré puisque c'est celui qui va lisser au maximum la courbe.

Nous avons donc appliqué ce lissage sur les prédictions du meilleur système obtenu jusque là, les résultats sont disponibles dans le tableau 5.11.

Descripteurs	Dev		Test	
	Non lissé	Lissé	Non lissé	Lissé
Mfcc-lib (<i>l-rmse</i>)	.698	.719	.513	.570

TABLE 5.11 – Scores CCC avec et sans lissage calculés sur les ensembles de développement et de test d'AlloSat. Score issu du meilleur modèle : système biLSTM-4, descripteurs Mfcc-lib et fonction de coût RMSE.

Nous pouvons observer qu'il s'agit du meilleur score jamais obtenu sur les prédictions de l'axe de satisfaction. De même, la figure 5.4 permet de se rendre compte du profit que nous tirons du lissage. On l'observe sur la figure de droite, tracée en bleu.

Grâce à ce lissage, nous obtenons de très bons scores : 0.719 pour l'ensemble de développement et 0.570 pour le test. Nous utilisons ces scores comme référence sur le corpus AlloSat dans la suite de nos travaux.

5.6 Conclusion

Dans ce chapitre, nous avons décrit les différents choix que nous avons considéré pour construire un modèle de reconnaissance des émotions. Nous avons d'abord présenté la reconnaissance d'émotions discrètes puis continues. Nous avons également validé nos expérimentations et la pertinence de notre corpus en le comparant au corpus de l'état de l'art SEWA.

Nous pouvons conclure que le meilleur système de reconnaissance des émotions continues utilise un réseau de neurones récurrents à 4 couches cachées, un ensemble de descripteurs à base de MFCCs, sur lequel nous appliquons un post-traitement de lissage.

Cependant comme nous l'avons dit précédemment, la modalité acoustique n'est pas la seule permettant de retrouver les états émotionnels. Dans le prochain chapitre, nous

allons mesurer l'apport de la modalité linguistique ainsi que l'apport de descripteurs pré-entraînés de type BERT.

RECONNAISSANCE CONTINUE D'ÉMOTION À PARTIR DE REPRÉSENTATIONS ACOUSTIQUES ET LINGUISTIQUES PRÉ-ENTRAINÉES

6.1 Motivation

Dans ce dernier chapitre de contribution, nous allons détailler les nouvelles méthodes que nous avons appliqués à la reconnaissance d'émotions continues. Comme nous l'avons précédemment indiqué, le corpus AlloSat, quoique contenant 37 heures d'audio, n'est pas considéré comme un grand corpus, surtout si on parle uniquement de la partie utilisée pour l'apprentissage. En effet, les corpus utilisés pour l'apprentissage de système de reconnaissance sont plutôt de l'ordre de grandeur de la centaine d'heures. Cependant les architectures neuronales sont connues pour mieux fonctionner et mieux généraliser si elles sont apprises sur une grande quantité de données. Pour pallier ce problème, nous avons investi deux solutions différentes mais qui se sont révélées être complémentaires.

Tout d'abord, nous avons voulu bénéficier du contenu linguistique des conversations téléphoniques en complément du contenu acoustique. En effet, ces informations sont déjà présentes implicitement dans les données d'entrée des différents systèmes, mais d'une part, elles ont été transformées, et d'autre part, il est possible que les modèles, n'ayant pas assez de données, n'aient pas réussi à dissocier les informations linguistiques des informations acoustiques. Nous avons donc fait le choix de traiter la modalité linguistique à part et de trouver la meilleure fusion permettant de profiter des caractéristiques présentes dans les deux modalités.

De plus, le pré-apprentissage permettant l'extraction de caractéristiques est une approche de plus en plus étudiée pour obtenir de meilleures représentations continues du

contenu audio et textuel. Ces solutions ont pour but de compenser le manque de données sur une tâche courante en utilisant des données sur lesquelles sont appris des systèmes pour des tâches voisines. Ainsi, le travail d’apprentissage est facilité, puisqu’il ne part pas de zéro.

Enfin, nous avons combiné ces deux solutions, afin d’avoir des performances suffisantes permettant d’implémenter une nouvelle fonctionnalité de reconnaissance des émotions au sein des solutions logiciels de la société Allo-Média.

Nous avons mis en place un score de confiance du score CCC. En effet, il est difficile d’analyser si un modèle est plus performant qu’un autre sans avoir mis en place un intervalle de confiance dans nos mesures. La mise en oeuvre de cet intervalle est disponible en annexe. Les différentes expérimentations et leurs résultats sont détaillées dans les sections suivantes.

6.2 Représentation linguistique

Comme nous l’avons dit précédemment, la représentation linguistique peut apporter des informations qui diffèrent de celles retenues par les systèmes de la modalité acoustique. Pour mettre en place un système appris sur cette modalité, nous avons utilisé les transcriptions automatiques des conversations téléphoniques.

6.2.1 Transcriptions automatiques

Ces transcriptions sont obtenues à partir d’un système de reconnaissance de la parole utilisé dans l’entreprise Allo-Media issu des travaux de Rousseau et al. [Rou+14]. Le modèle acoustique est appris en utilisant le framework Kaldi [Pov+11] sur un volume d’environ mille heures de transcriptions manuelles de conversations issues de différents centre d’appels. Ainsi ce modèle est appris sur une grande quantité de données qualitatives, puisque manuelles mais aussi appropriées au domaine, puisque provenant de centre d’appels.

Le modèle linguistique appartient également à l’entreprise et il est mis à jour continuellement en utilisant les conversations transcrites automatiquement par l’entreprise. Chaque segment transcrit est associé à un score de vraisemblance, et seuls les segments ayant de hauts scores sont utilisés pour remettre à jour le modèle linguistique.

Enfin le vocabulaire est également mis en place par l’entreprise, qui est partie d’un

dictionnaire de français qu’un linguiste de l’équipe a phonétisé. Le vocabulaire a ensuite été manuellement vérifié pour retirer toute occurrence de mots n’ayant pas sa place dans des conversations de centre d’appels, afin de réduire les erreurs de reconnaissance dues à des mots à phonétique voisine ou aux homonymes. Ce travail est effectué de façon itérative en fonction des remontées des modérateurs. De plus, le vocabulaire est régulièrement mis à jour avec des noms de famille ou des noms de produits ou de marques.

6.2.2 Synchroniser le linguistique et l’acoustique

Comme notre objectif est d’utiliser conjointement l’information linguistique et l’acoustique, nous avons fait le choix de nous aligner sur les annotations émotionnelles dans les deux cas. Pour l’acoustique, nous extrayons un vecteur descripteur du signal toutes les 250 ms, soit un vecteur par pas d’annotation. Nous faisons la même chose pour le linguistique.

Pour cela, nous partons des CTM (*time-marked conversation*) issus de la transcription automatique, comme par exemple ceux illustrés dans le tableau 6.1. Ce format se compose d’autant de lignes que de mots transcrits avec le code temps de l’émission du mot ainsi que la durée du mot. De plus, on peut retrouver à quel segment de parole sont associés tous les mots.

Nom fichier	Canal	Début	Durée	Mot	Score Confiance	id segment
ConvA	1	5.23	0.09	oui	0.81	ConvA-0000496-0001075
ConvA	1	5.32	0.27	bonjour	1.00	ConvA-0000496-0001075
ConvA	1	5.59	0.36	monsieur	1.00	ConvA-0000496-0001075
ConvA	1	6.19	0.27	voilà	1.00	ConvA-0000496-0001075
ConvA	1	6.46	0.15	j’ai	1.00	ConvA-0000496-0001075
ConvA	1	6.61	0.06	un	1.00	ConvA-0000496-0001075
ConvA	1	6.67	0.39	contrat	1.00	ConvA-0000496-0001075
ConvA	1	7.06	0.24	chez	1.00	ConvA-0000496-0001075
ConvA	1	7.30	0.18	vous	1.00	ConvA-0000496-0001075

TABLE 6.1 – Exemple de la transcription d’un début de conversation au format ctm mise sous la forme d’un tableau pour expliquer les différentes colonnes. Conversation issue du corpus AlloSat.

Nous définissons le protocole suivant pour aligner la transcription et l’annotation :

- Tous les mots sont conservés, même les *stop words*. Nous avons également fait des

tests en supprimant ces mots tels que définis dans le toolkit NLTK¹. Comme les performances finales de reconnaissance ne s’en retrouvaient pas améliorées, nous avons décidé de conserver tous les mots.

- Si un mot est prononcé dans plusieurs trames consécutives, le mot est dupliqué sur toutes les trames concernées.
- Si plusieurs mots sont prononcés dans la même trame, on conserve tous les mots prononcés. Pour les représenter, on fera la moyenne des représentations de chacun de ces mots.

Cette synchronisation nous permet de mettre l’acoustique et le linguistique sur le même plan, ce qui nous permettra par la suite d’utiliser conjointement les deux modalités pour la reconnaissance continue de la satisfaction et de la frustration.

6.2.3 Exploration des word2vec

Les plongements de mots, ou *word embeddings* sont actuellement l’une des représentations continues les plus populaires de données textuelles. Ces plongements sont des représentations vectorielles appliquées à chacun des mots retrouvés dans le texte. Parmi eux, word2vec [Mik+13] est l’un des plus utilisés dans les tâches d’analyse des sentiments telles que la polarité ou la classification des états émotionnels [DM18].

Afin de paramétrer des word2vec cohérents avec nos données, nous avons choisi d’entraîner nos propres représentations. Pour cela, nous avons entraîné un système avec le toolkit GENSIM [ŘS10], en utilisant des données privées détenues par Allo-Media. Ces données sont composées de transcriptions manuelles d’appels reçues par les centres d’appels, totalisant plus de 4,5 millions de mots.

Deux algorithmes peuvent être utilisés pour apprendre ces représentations : *CBOW* et *skip-gram*, illustré dans la figure 6.1. Pour *CBOW*, le modèle prend en entrée les mots du contexte et prédit le mot cible en sortie. Pour *skip-gram*, le modèle prend en entrée le mot cible et apprend à prédire les mots du contexte.

Nous avons utilisé l’algorithme *skip-gram* avec un contexte de 5 mots (soit le mot cible et 4 mots de contexte, 2 à gauche et 2 à droite). Le vocabulaire a été réalisé en incluant tous les mots présents dans le corpus. Tous les mots inconnus ont été remplacés lors de la transformation en une balise $\langle UNK \rangle$ et ont donc tous la même représentation.

Afin de comparer les performances réalisées selon les différentes tailles de représenta-

1. <https://www.nltk.org/api/nltk.html>

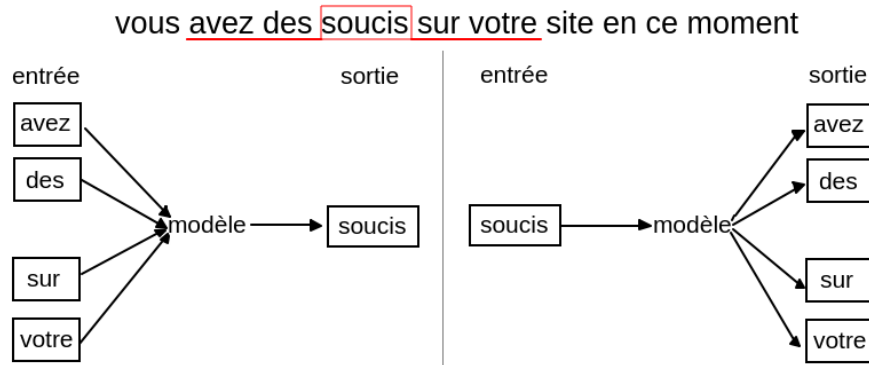


FIGURE 6.1 – Représentation de l’apprentissage des word2vec selon deux algorithmes : le *CBOW* (gauche) et le *skip-gram* (droite). Ici on utilise un contexte de 5 mots, soit le mot cible entouré en rouge et deux mots de chaque côté, soulignés en rouge.

tion utilisées, nous avons créé quatre ensembles de descripteurs, en faisant varier la taille des vecteurs :

- **Word2vec-40** : La taille de la représentation a été fixée à 40 en raison des travaux préliminaires sur la modalité acoustique où nos représentations étaient de cet ordre de grandeur. Nous avons fait ce choix pour minimiser l’impact de la taille des descripteurs sur les performances du système,
- **Word2vec-100** : La taille des plongements est fixé à 100. Il s’agit d’une taille standard, beaucoup utilisée dans des tâches de NLP notamment.
- **Word2vec-150** : La taille des plongements est fixé à 150.
- **Word2vec-200** : La taille des plongements est fixé à 200.

Nous avons utilisé l’architecture biLSTM-4 définie dans le chapitre 5 section 5.3 pour tester la modalité linguistique représentée par des Word2vec. Les résultats sont disponibles dans le tableau 6.2. Comme nous pouvons le remarquer, les scores sont supérieurs à 0.5 pour les quatre ensembles. Nous remarquons que les meilleurs scores sont atteints avec une taille de descripteurs de 100. Les performances se dégradent lorsque l’on a des tailles de descripteurs plus grandes.

Si nous comparons avec le maximum atteint par la modalité acoustique (cf section 5.5.2), soit 0.698 sur le dev et 0.513 sur le test sans post-traitement, nous voyons une amélioration de ce score maximum avec trois de nos ensembles. Avec un score CCC de 0.860 sur l’ensemble de développement et de 0.759 sur l’ensemble de test, la modalité linguistique représentée par des descripteurs Word2vec de taille 100 donne un résultat bien plus performant que tous les expérimentations effectuées sur la modalité acoustique seule.

Descripteurs	Dev	Test
Word2vec-40	0.805	0.569
Word2vec-100	0.860	0.759
Word2vec-150	0.592	0.553
Word2vec-200	0.668	0.414

TABLE 6.2 – Scores CCC des systèmes de reconnaissance des émotions d’une architecture neuronale bilstm à quatre couches en fonction des différents descripteurs d’entrée linguistiques.

Plusieurs justifications peuvent expliquer ce résultat. Tout d’abord, les mots sont porteurs d’informations émotionnelles qui sont moins souvent ambiguës que celles convoyées par la voix. On peut notamment mentionner les injures ou les mots contenus dans des champs lexicaux de polarité négative (par exemple “inadmissible”, “honteux” ou encore “arnaque”).

De plus, nous simplifions le travail du système de reconnaissance des émotions, puisque nous extrayons la modalité linguistique du signal audio et celle-ci est de plus haut niveau. Lorsque l’on traite la modalité acoustique, les informations linguistiques sont toujours implicitement présentes et le système peut avoir des difficultés à les modéliser, notamment à cause du peu de données dont nous disposons.

Ces résultats nous permettent de valider l’utilisation de la modalité linguistique seule dans la détection de la satisfaction et de la frustration.

Nous avons donc cherché à tirer parti de ces deux modalités afin d’améliorer la reconnaissance de la satisfaction et de la frustration.

6.3 Fusion des modalités acoustiques et linguistiques

Comme nous l’avons décrit dans le chapitre 3, il existe plusieurs méthodes pour fusionner des modalités différentes. Chaque fusion a ses avantages et ses inconvénients et peut être plus ou moins adaptée aux descripteurs et aux modèles utilisés. Nous avons donc fait le choix de comparer les trois fusions existantes, à savoir la fusion des descripteurs, la fusion des modèles et la fusion des décisions.

Ces trois fusions sont opérées sur les modalités acoustiques et linguistiques en utilisant le système biLSTM-4 précédemment introduit à la section 5.3. La figure 6.2 résume les quatre expérimentations dont nous rapportons les résultats dans le tableau 6.3.

Les quatre architectures ont les caractéristiques suivantes :

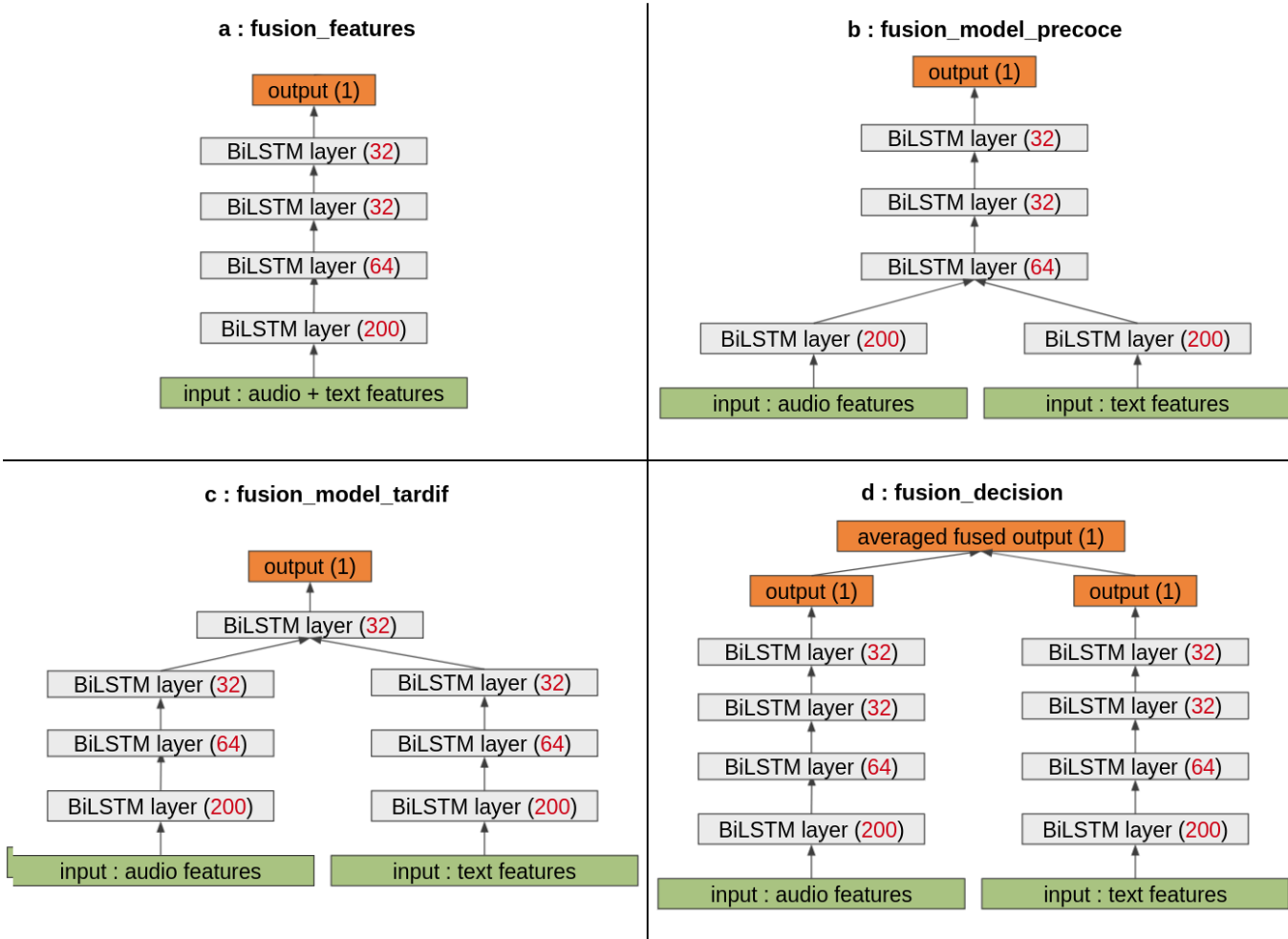


FIGURE 6.2 – Représentation des quatre fusions utilisées pour la reconnaissance des émotions à partir des modalités acoustiques et linguistique. Les nombres entre parenthèses correspondent au nombre de neurones dans chaque couche.

- **Fusion_features** : La fusion s'effectue avant l'apprentissage en concaténant les descripteurs audio et texte. Une passe de normalisation est ensuite effectuée, pour ramener les descripteurs sur une même échelle. Il s'agit de la fusion la moins coûteuse, puisque nous n'avons qu'un seul modèle à entraîner, bien qu'il y ait plus d'entrées. Cette fusion correspond à la sous-figure a.
- **Fusion_model_precoce** : La fusion s'effectue au niveau de la deuxième couche du réseau de neurones. Chaque modalité est d'abord traitée séparément dans la première couche, puis les sorties de ces deux couches sont concaténées avant d'être introduites dans la deuxième couche. La fusion précoce permet également de projeter les modalités dans des dimensions de même taille, sans pour autant que toutes

les informations utiles soient déjà extraites avant d’être fusionnées. Cette fusion est modérément coûteuse, le nombre de paramètres lorsque les tailles des vecteurs d’entrée pour chaque modalité sont inférieures ou égales à 100, est du même ordre de grandeur que précédemment. Cette fusion correspond à la sous-figure b.

- **Fusion_model_tardif** : A l’inverse, la fusion va s’effectuer plus tard, au niveau de la quatrième couche du réseau. La sortie des troisièmes couches sont concaténées avant d’être données à la quatrième couche. Cette fusion tardive va prendre en compte des descripteurs très modifiés, desquels les informations importantes sont normalement déjà extraites. Cette fusion correspond à la sous-figure c.
- **Fusion_decision** : Cette fusion s’effectue sur les sorties des deux modèles. On entraîne séparément deux modèles pour les deux modalités, puis on met en concours les deux vecteurs de sorties des deux modèles en utilisant une moyenne pondérée ou non. Cette fusion est la plus coûteuse puisque nous apprenons deux modèles distincts et que nous appliquons un post-traitement sur les sorties de ces modèles. Cette fusion correspond à la sous-figure d.

Descripteurs	Dev	Test
fusion features		
MFCC \oplus word2vec-40	0.895	0.833
MFCC \oplus word2vec-100	0.836	0.788
fusion modèle précoce		
MFCC-lib \oplus word2vec-40	0.904	0.807
MFCC-lib \oplus word2vec-100	0.829	0.797
fusion modèle tardive		
MFCC-lib \oplus word2vec-40	0.917	0.815
MFCC-lib \oplus word2vec-100	0.881	0.717
fusion décision		
.66 Word2vec-40 + .34 MFCC-lib	0.897	0.840
.72 Word2vec-100 + .28 MFCC-lib	0.844	0.789

TABLE 6.3 – Comparaison des scores CCC des systèmes de reconnaissance des émotions du biLSTM-4 en utilisant différents protocoles de fusion et différents descripteurs pour la modalité linguistique.

Les résultats de ces différentes fusions sont disponibles dans le tableau 6.3. Nous observons que les scores de fusion, quels qu’ils soient, permettent d’égaliser ou d’améliorer les résultats unimodaux. Nous remarquons que dans tous les types de fusion, c’est l’utilisation des word2vec-40 qui donne de meilleurs résultats. Cela peut être dû à la similarité entre les

dimensions acoustiques et linguistiques : les MFCC-lib sont constitués de 48 descripteurs. Nous pouvons également noter que le meilleur score atteint sur l'ensemble de dev (0.917) est attribué à la fusion tardive de modèle. Cependant, nous voyons que ce n'est pas la fusion de modèle qui donne le meilleur résultat sur l'ensemble de test. Le meilleur score (0.840) est atteint avec la fusion de décision. De plus l'écart entre le score de dev et de test est moins important pour la fusion de décision. Nous avons donc décidé de conserver la fusion de décision comme fusion de référence dans la suite de nos expérimentations.

Comme nous l'avons déjà présenté dans le chapitre 3, les word2vec sont statiques : un mot a toujours la même représentation vectorielle, quel que soit le vrai sens du mot et le contexte de son apparition. Cela peut être problématique pour les mots polysémiques, par exemple fréquents en français [Pus+96]. C'est pour cette raison que nous avons souhaité mettre en place des représentations permettant de prendre en compte le contexte, que ce soit sur l'acoustique ou sur le linguistique.

6.4 Descripteurs pré-entraînés

L'apprentissage par transfert, *transfer-learning* est largement utilisé dans l'analyse des sentiments, où de grandes bases de données non spécifiques sont utilisées pour former des caractéristiques génériques qui sont introduites dans le processus d'apprentissage. Ce procédé permet de donner aux systèmes de meilleures capacités de généralisation compte tenu des données d'apprentissage fortement limitées [DM18].

Une autre méthode permettant d'exploiter des modèles qui ont déjà vu une très grande quantité de données est d'utiliser des représentations auto-apprises. L'apprentissage auto-supervisé des représentations de la parole ou du langage a été proposé ces dernières années, par exemple avec le système BERT [Dev+19], utilisé pour la représentation textuelle que nous avons présenté au chapitre 2.

De telles représentations, calculées par des modèles neuronaux entraînés sur d'énormes quantités de données non étiquetées, ont montré leur efficacité, par exemple pour la vision par ordinateur [NGB17] et des tâches de traitement du langage naturel (NLP) telles que de classification, de similarité de texte, ou de classement par pertinence [Liu+19a; You+18; Yan+19]. Elles ont également prouvé leur efficacité dans les domaines de l'ASR [Kah+20; Liu+20] et de la traduction [Ngu+20].

Néanmoins, ces représentations n'avaient jamais été utilisées pour de la reconnaissance d'émotions continues à notre connaissance. Nous avons donc proposé de tester cette

approche pour réaliser notre tâche de reconnaissance des émotions. Il n’est, en effet, pas évident qu’elle soit pertinente pour le domaine de Speech Emotion Recognition. Par exemple, au niveau acoustique, l’ASR a tendance à se concentrer sur des durées d’environ 30 ms, tandis que les émotions sont généralement prises en charge sur environ 1 seconde de parole.

Nous pensons que l’utilisation de ces représentations peut limiter l’impact du manque de données lorsque seules de petites bases de données sont disponibles pour former un réseau de neurones pour une tâche spécifique.

6.4.1 Représentation linguistique

Pour modéliser la partie linguistique, nous sommes partis sur des systèmes dérivés de BERT. Comme nous l’avons défini dans le chapitre 2, BERT permet de donner une représentation contextuelle du texte. Il est cependant appris en majorité avec des occurrences anglaises. Comme nous voulons modéliser du français, nous nous sommes tournés vers CamemBERT [Mar+20] et FlauBERT [Le+20] que nous avons présenté également au chapitre 2.

Chronologiquement, nous avons découvert CamemBERT en premier, et une fois que nous avons expérimenté avec ce modèle, nous avons comparé les résultats obtenus avec le modèle FlauBERT.

Pour l’extraction des descripteurs, nous avons utilisé le toolkit Fairseq [Ott+19]. Nous utilisons le modèle pré-entraîné, appelé *camemBERT-base*, entraîné sur la partie française du corpus OSCAR [OSR19] constitué d’un ensemble de corpus monolingues extraits du *Common Crawl snapshot* et totalisant 138Go de texte brut et 32,7 milliards de tokens après la tokenisation en sous-mots [Wu+16]. Les sous-mots sont des ensembles de caractères qui ne sont pas des mots à proprement parler et qui sont déduits par le système de tokenisation, ici SentencePiece [KR18].

Les caractéristiques ont été extraites sur Allosat à l’aide de ce modèle pré-entraîné, et nous avons résumé les résultats au niveau du segment émotionnel (250 ms) en faisant la moyenne des représentations continues des sous-mots apparaissant dans le segment courant. Au total, nous utilisons un vecteur de caractéristiques à 768 dimensions.

Nous voulions également utiliser la variance de ces représentations, dans une volonté de comparer nos représentations aux précédentes qui utilisaient à la fois la moyenne et la variance des descripteurs. Cependant doubler le nombre de caractéristiques faisait exploser la complexité du réseau et donc le temps d’apprentissage.

Les résultats sont présentés dans le tableau 6.4. Comme nous pouvons l’observer, les modèles sont très performants pour la reconnaissance des émotions continues. Les descripteurs issus de CamemBERT donnent de meilleurs résultats que ceux issus de FlauBERT même s’ils restent du même ordre de grandeur. Si on se compare aux meilleurs résultats des word2vec, à savoir respectivement 0.860 et 0.759 sur le dev et le test, on obtient une amélioration du score CCC de 0.036 sur le dev et de 0.040 sur le test.

Descripteurs	Dev	Test
CamemBERT	0.896	0.799
FlauBERT	0.874	0.733

TABLE 6.4 – Scores CCC des systèmes de reconnaissance des émotions du biLSTM-4 en prenant en entrée des descripteurs linguistiques pré-entraînés.

6.4.2 Représentation acoustique

Dans le domaine du SER, trouver le meilleur ensemble de caractéristiques acoustiques est toujours un sous-domaine de recherche actif [JMC18].

Comme nous l’avons déjà dit (cf section 3.3.1), la plupart des ensembles de caractéristiques *experts* [Eyb+16; Sch+13] visent à décrire la prosodie dans le signal, avec des descripteurs de bas niveau (LLD) capturant l’intensité, l’intonation, le rythme ou la qualité de la voix. Une autre approche consiste à extraire des caractéristiques spectrales : les coefficients cepstraux à fréquence mel (MFCC) sont clairement les plus utilisés car ils sont robustes aux signaux bruités même s’ils n’ont pas été conçus pour la prosodie ou l’émotion.

Récemment, wav2vec [Sch+19] et Audio ALBERT [Chi+20] ont été introduits dans les domaines de la reconnaissance automatique de la parole et dans l’identification du locuteur comme l’une des premières approches pré-entraînées pour extraire des caractéristiques contextuelles des signaux bruts. Comme nous l’avons expliqué au chapitre 2, wav2vec [Sch+19] est un modèle pré-entraîné par auto-supervision : il apprend à prédire les futurs échantillons à partir de l’analyse de la fenêtre courante.

Dans nos expérimentations, nous utilisons deux modèles pré-entraînés. Le premier, appelé *wav2vec-EN*, fourni par Schneider et al., a été entraîné sur le corpus LibriSpeech [Pan+15] qui est composé de 960 heures de livre audio en anglais. Nous formons également notre propre modèle, appelé *wav2vec-FR*, dérivé du premier modèle et fine-tuné

sur les conversations non étiquetées de centres d’appels français correspondant à plus de 500 heures de données privées.

Comme ce modèle est appris avec des audios échantillonnés à 16kHz, nous avons sur-échantillonné nos données pour qu’elles correspondent aux modèles. Nous avons utilisé la fonction de rééchantillonnage FFMpeg [Tom06] avec la fonction d’interpolation *sinc*.

Au final, chaque segment émotionnel est représenté par un vecteur de taille 512 qui se compose des valeurs moyennes des plongements obtenus toutes les 10 ms sur le segment émotionnel, *i.e.* soit une moyenne de 25 valeurs pour AlloSat. Comme pour la représentation linguistique, nous avons uniquement conservé la moyenne et non l’écart type pour ne pas avoir des vecteurs d’entrée trop importants qui auraient fait exploser la complexité et donc le temps d’apprentissage et les ressources nécessaires pour le mener à bien.

Descripteurs	Dev	Test
wav2vec-EN	0.851	0.730
wav2vec-FR	0.865	0.635

TABLE 6.5 – Scores CCC des systèmes de reconnaissance des émotions du biLSTM-4 en prenant en entrée des descripteurs acoustiques pré-entraînés.

Les résultats sont présentés dans le tableau 6.5. Comme nous pouvons l’observer, les modèles sont également très performants pour la reconnaissance des émotions continues. Les descripteurs issus de wav2vec-EN donne de meilleurs résultats que ceux issus de wav2vec-FR même s’ils restent dans la même ordre de grandeur. Cela peut être expliqué par le fine-tuning effectué sur à peine 500 heures de données, ce qui a pu suffisamment dérégler le modèle sans pour autant lui faire intégrer des renseignements pertinents. Si on se compare aux meilleurs résultats des MFCCs, à savoir respectivement 0.698 et 0.513 sur le dev et le test, on peut affirmer que la représentation wav2vec-EN permet au système de mieux reconnaître les états émotionnels, puisqu’on améliore les scores respectivement de 0.153 et 0.217 sur le dev et le test.

6.4.3 Reproduction sur SEWA

Afin d’étendre l’utilisation des descripteurs pré-entraînés, nous avons voulu tester les performances d’un système biLSTM-4 appris avec ce type de descripteurs pour le corpus SEWA. Le corpus étant en allemand et en hongrois, nous avons uniquement considéré le modèle wav2vec-EN.

Descripteurs	Activation	Valence	Liking
wav2vec-EN	0.251	0.215	0.254

TABLE 6.6 – Scores CCC des systèmes de reconnaissance des émotions du biLSTM-4 en prenant en entrée des descripteurs acoustiques pré-entraînés sur l’ensemble de dev du corpus SEWA.

Les résultats sont présentés dans le tableau 6.6. Les scores obtenus sur les dimensions de l’activation et de la valence sont moins élevés que lorsque l’on utilise les ensembles de descripteurs MFCCs ou eGeMAPS. Cependant on peut constater une amélioration sur l’axe du liking, mais qui n’atteint pas un score suffisant pour être considéré comme une réussite.

Dans le cas du corpus SEWA, l’utilisation de descripteurs pré-entraînés ne semble pas permettre d’améliorer les performances du système. Notre hypothèse réside dans la différence de quantité de données entre les deux corpus : le set de train d’AlloSat est composé de 25 heures d’enregistrements audio alors que celui de SEWA a moins de 2 heures de conversations. Nous pensons également que la conception du corpus SEWA peut avoir un impact sur la performance de ces descripteurs. Le fait que SEWA soit composé de conversations dyadiques où deux locuteurs sont présents dans le même document peut complexifier le processus d’extraction de descripteurs, alors qu’AlloSat ne contient que la voix du client.

Nous avons vu précédemment que la fusion des deux modalités, en considérant les descripteurs MFCC et word2vec, permet d’améliorer les performances du système. Nous expérimentons donc la fusion des descripteurs pré-appris.

6.5 Fusion des modalités acoustique et linguistique pré-entraînées

Différents protocoles de fusion ont été explorés : fusion des features (concaténation des vecteurs wav2vec et camemBERT), fusion au niveau du modèle (au niveau de la première et de la dernière couche) et fusion de décision. Les fusions au niveau des features et au niveau du modèle requièrent beaucoup plus de temps d’apprentissage, puisque nous doublons le nombre d’entrées du réseau (768+512). De plus, les résultats que nous avons obtenus sur ces features sont peu intéressants puisqu’au niveau des scores d’unimodalité. Nous ne les présentons pas dans ce document.

Comme jusqu’à présent la modalité linguistique donne de meilleurs résultats que la modalité acoustique et que la dimension des descripteurs camemBERT est grande, nous avons voulu quantifier l’apport de ces descripteurs en les comparant aux descripteurs word2vec. Ainsi en fonction de l’ordre de grandeur de cet apport, nous pourrions statuer sur la pertinence ou non de l’utilisation de plus de ressources.

Descripteurs	nombre features	Dev	Test
.66 word2vec-40 + .34 MFCC-lib	47 + 40	0.897	0.840
.63 wav2vec-EN + .37 word2vec	512 + 40	0.878	0.750
.28 wav2vec-EN + .72 CamemBERT	512 + 768	0.932	0.920

TABLE 6.7 – Comparaison des score CCC des systèmes de reconnaissance des émotions du biLSTM-4 en utilisant la fusion de décision avec des descripteurs pré-entraînés. Nous rappelons les scores obtenus précédemment avec la fusion des descripteurs baseline.

Comme pour les résultats précédents sur la fusion des descripteurs de la baseline (MFCC + word2vec), la fusion de décision permet d’améliorer les résultats même si l’amélioration est moins forte que précédemment. Comme nous le rapportons dans le tableau 6.7, nous atteignons des scores de 0.932 sur le dev et 0.920 sur le test, ce qui correspond à une augmentation significative des scores. On remarque également que la fusion des ensembles wav2vec-EN et word2vec donne de moins bons résultats, et que c’est la modalité acoustique qui prime dans cette fusion. On peut supposer que la trop grosse différence dans le type de descripteur ne va pas permettre une bonne fusion.

Nous atteignons donc des scores très performants pour la reconnaissance des émotions continues sur le corpus AlloSat, permettant à l’entreprise de penser à la commercialisation de cet indicateur.

6.6 Conclusion

Dans ce chapitre, nous avons décrit les différents choix que nous avons considéré pour améliorer nos performances en terme de reconnaissance continue des émotions. Nous avons mis en place une représentation linguistique des conversations que nous avons fusionné avec la représentation acoustique. De plus, nous avons mis en place des nouveaux modèles entraînés avec des descripteurs pré-entraînés.

Nous pouvons conclure que le meilleur système de reconnaissance des émotions continues utilise un réseau de neurones récurrents, une utilisation des deux modalités en les fusionnant et des entrées issues de descripteurs pré-entraînés.

Ces scores à l'état de l'art sont très corrects, mais une question reste en suspens : pourquoi la modalité linguistique permet-elle une telle amélioration de la reconnaissance continue des émotions ? Nous allons tenter de répondre à cette question dans le prochain chapitre.

ANALYSE DES ANNOTATIONS, EXPLICABILITÉ DES MODÈLES AUTOUR DE LA FRUSTRATIONS

7.1 Motivations

Lors de cette thèse, nous avons établi des systèmes de reconnaissance continue de l'émotion qui permettent de détecter la satisfaction et la frustration avec un degré d'erreur acceptable. Dans ce dernier chapitre, nous avons voulu questionner la stratégie qui consiste à fusionner les annotations individuelles de chaque annotateur. Ainsi nous souhaitons proposer une alternative à la conception la plus établie en matière d'annotation en émotion. Dans un second temps, la modalité linguistique étant celle qui induit les meilleures performances, il nous semblait important d'essayer de comprendre la prévalence de cette modalité.

Dans la littérature, on a pour habitude d'utiliser de nombreux annotateurs et de fusionner les annotations pour atténuer le caractère subjectif d'une annotation. En effet, il est difficile de nier que la reconnaissance d'une émotion et de son intensité peut varier en fonction de la personne qui la perçoit et en fonction du moment pour une même personne. Il est donc possible d'avoir autant de versions différentes d'annotations que d'annotateurs. Dans ce cas, ne peut-on pas considérer que chaque annotateur est dans le juste? Nous analysons ce positionnement dans la prochaine section.

Comme nous l'avons établi dans le précédent chapitre, nous pouvons observer que la modalité linguistique permet d'atteindre les meilleurs scores de reconnaissance continue des émotions de satisfaction et de frustration contenues dans AlloSat. Nous aurions pensé dans un premier temps que la modalité acoustique donnerait de meilleurs résultats. En effet, la modalité acoustique, telle que nous la traitons, contient déjà les informations linguistiques. Elles ne sont pas extraites et pré-traitées mais nous pensons que le système

de reconnaissance serait capable d'en retirer les informations pertinentes.

Nous avons donc cherché les marqueurs de l'émotion dans les transcriptions des conversations téléphoniques. Nous avons d'abord conduit des analyses statistiques sur le corpus, puis nous avons travaillé avec un linguiste du LIUM, Pr. Daniel Luzzati, afin de retrouver les clés permettant à l'humain de reconnaître les états émotionnels de satisfaction et de frustration.

7.2 Analyse de l'impact de chaque annotateur

7.2.1 Annotation moyenne ou 3 annotations ?

Comme nous l'avons vu dans le chapitre 4, nous avons choisi de faire annoter le corpus par trois annotateurs. Ce choix a été motivé par l'aspect subjectif non négligeable de toutes les tâches mettant en œuvre des émotions. Nos données contiennent donc des conversations qui ont été annotées trois fois, comme par exemple la conversation illustrée par la Figure 7.1. On peut observer une tendance globale commune entre les annotations avec des différences en intensité et en délais avant de noter un changement d'émotion. Deux positionnements peuvent être pris à partir de ces annotations.

1. Nous pouvons considérer les annotations de chaque annotateur indépendamment. En effet, nous pouvons partir du principe que la perception de l'émotion par chaque individu est légitime et peut être considérée comme une référence pour le système de détection. Ce positionnement va donc considérer que l'individu a plus de valeur que le groupe entier. Nous pouvons donc spécifier nos systèmes de reconnaissance pour reproduire le comportement d'un humain particulier.
2. Nous pouvons également considérer la moyenne des annotations. En effet, faire la moyenne des annotations revient à généraliser ces dernières pour qu'elles soient plus en adéquation avec un ensemble d'individus et atténuer les perceptions qui s'écartent d'une normalité statistique. Ainsi, nous allons considérer que l'annotation de référence pour nos systèmes automatiques est la moyenne de l'avis de différentes personnes. Bien que ce procédé soit plus pertinent lorsque nous avons un nombre élevé d'annotateurs, il reste cohérent pour trois annotateurs et permet de généraliser nos systèmes de reconnaissance pour reproduire non pas le comportement d'un humain particulier mais d'un groupe d'humains.

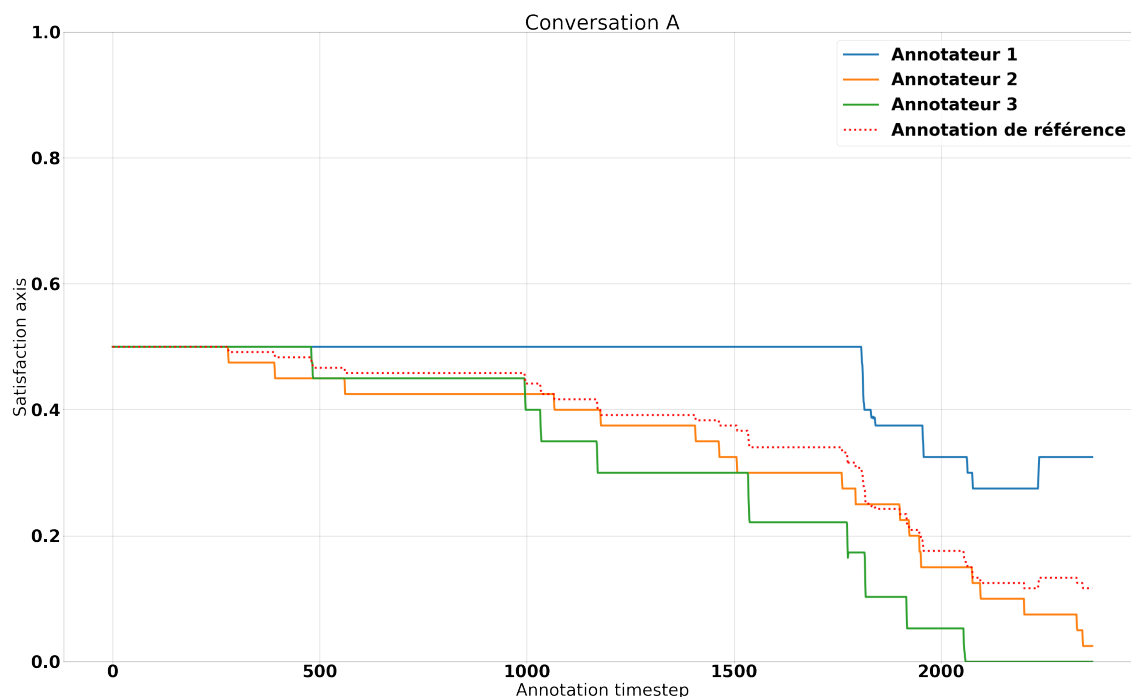


FIGURE 7.1 – Exemple d’annotation d’une conversation selon l’axe de satisfaction déjà présenté chapitre 4. L’annotation de référence correspond à la courbe en pointillée

Au cours de la thèse, nous avons décidé, en adéquation avec les besoins industriels, de nous concentrer sur le deuxième positionnement, comme nous l’avons énoncé dans le chapitre 4. Ce positionnement est également celui de nombreux travaux dans le domaine, notamment les travaux effectués sur les corpus RECOLA [Rin+13] et SEWA [Kos+19].

Cependant pour aller plus loin dans l’analyse du phénomène émotionnel, nous avons voulu explorer la possibilité d’utiliser les annotations de chacun afin de construire des systèmes de reconnaissance.

7.2.2 Un modèle de reconnaissance par annotateur

Si nous considérons que chaque annotateur donne une version légitime de l’émotion, nous pouvons choisir de construire trois modèles de reconnaissance en prenant les annotations des trois annotateurs. Pour ce faire, nous modifions la référence : au lieu d’entraîner un seul modèle sur la valeur moyenne des trois annotateurs, nous entraînons trois modèles différents par annotateur, dans lesquels les références correspondent aux valeurs uniques de cet annotateur. Les prédictions de ces modèles sont évaluées en fonction des annotations individuelles (tableau 7.1) ou de nos anciennes références définies comme la moyenne

des trois annotations individuelles (tableau 7.2).

Les colonnes AVG donnent les performances moyennes sur les trois modèles individuels. Les colonnes CV donnent le coefficient de variation (écart-type sur la moyenne) sur les trois modèles individuels. Diff1 est la différence relative entre linguistique et acoustique pris indépendamment et donne une idée du gain par annotateur.

Annotations individuelles :

Annotateurs	a_1		a_2		a_3		AVG		CV	
	Dev	Test	Dev	Test	Dev	Test	Dev	Test	Dev	Test
Wav2Vec	.834	.734	.731	.785	.841	.597	.802	.705	.077	.138
CamemBERT	.898	.877	.833	.834	.900	.804	.877	.838	.043	.044
Diff1 (%)	7.7	19.5	14.0	6.2	7.0	34.7	-	-	-	-
Features	.884	.870	.815	.753	.883	.834	.861	.819	.046	.073
Modèle précoce	.883	.870	.855	.865	.888	.826	.875	.854	.020	.028
Modèle tardive	.911	.875	.814	.837	.921	.799	.882	.837	.067	.045
Décision	.913	.882	.840	.849	.916	.793	.890	.841	.048	.053

TABLE 7.1 – Résultats des systèmes de fusions pour chaque annotateur. Les modèles sont entraînés et évalués sur des annotations individuelles. AVG : moyenne sur les trois annotateurs. CV : coefficient de variation sur les trois annotateurs. Diff1 correspond à la différence relative entre CamemBERT et Wav2Vec.

Dans le tableau 7.1, nous pouvons remarquer que le coefficient de variation (CV), sans utiliser de fusions, est plus élevé avec les descripteurs acoustiques qu’avec les descripteurs linguistiques en particulier sur le Test. Plus précisément, concernant l’annotateur a_3 , les performances de la modalité acoustique chutent sur le Test pour atteindre un score CCC de 0.597. Ces résultats suggèrent que la prédiction de la satisfaction à partir des descripteurs acoustiques est plus sensible à la subjectivité de la tâche d’annotation qu’à partir des descripteurs linguistiques.

Notre hypothèse est que la variabilité dans l’espace acoustique est très diversifiée, et qu’une même réalisation acoustique peut être perçue avec des niveaux de satisfaction différents par le même annotateur, ce qui produit de moins bonnes performances sur la modalité acoustique.

En ce qui concerne la fusion des modalités, elle améliore les performances dans la plupart des configurations et les meilleures performances en moyenne sont atteintes avec la fusion modèle précoce avec un score de CCC de 0.854 sur l’ensemble de Test. L’améliora-

tion sur le Test est la plus élevée avec l'annotateur a_2 (+3.7% avec la fusion modèle précoce). Cela peut s'expliquer par la très faible différence entre les performances obtenues sur des modalités acoustiques et linguistiques pour cet annotateur (+6.2%), indiquant peut-être que les deux modalités portent des informations différentes pour cet annotateur spécifique.

A partir de ces résultats, nous émettons l'hypothèse qu'au niveau des annotateurs, les modalités acoustiques et linguistiques véhiculent des informations émotionnelles complémentaires. Cependant, si la partie linguistique est bien partagée entre les annotateurs, la perception de la partie acoustique semble assez individuelle. Bien sûr, des expériences supplémentaires avec des annotations croisées sont nécessaires pour confirmer cette hypothèse.

Annotations moyennes :

Annotateurs	a_1		a_2		a_3		AVG		CV	
	Dev	Test	Dev	Test	Dev	Test	Dev	Test	Dev	Test
Wav2Vec	.862	.736	.774	.731	.779	.710	.805	.726	.061	.019
CamemBERT	.916	.878	.755	.793	.851	.833	.841	.835	.096	.051
Diff1 (%)	6.3	19.3	-2.5	8.5	9.2	17.3	-	-	-	-
Features	.896	.845	.741	.688	.868	.861	.835	.798	.099	.120
Modèle précoce	.911	.833	.809	.824	.879	.856	.866	.838	.060	.020
Modèle tardive	.914	.899	.763	.784	.844	.841	.840	.841	.090	.068
Décision	.938	.882	.795	.778	.868	.874	.867	.845	.082	.069

TABLE 7.2 – Résultats des systèmes de fusions pour chaque annotateur. Les modèles sont entraînés sur des annotations individuelles et évalués sur les annotations moyennés de référence. AVG : moyenne sur les trois annotateurs. CV : coefficient de variation sur les trois annotateurs. Diff1 correspond à la différence absolue entre CamemBERT et Wav2Vec. La meilleure fusion est choisie sur le Dev.

En ce qui concerne les modèles individuels évalués avec des annotations moyennes (tableau 7.2), nous remarquons que l'annotateur a_2 a les performances les plus faibles lorsqu'il utilise uniquement des descripteurs linguistiques. Le modèle construit sur cet annotateur atteint les performances les plus basses en utilisant n'importe quel type de fusion sur les ensembles de Développement et de Test. Cette observation confirme ainsi l'importance des performances linguistiques qui ont un poids important dans l'évaluation générale de la satisfaction et la frustration quand on fusionne les modalités. Ce résultat

peut s’expliquer par le fait que parmi les trois annotateurs, nous avons montré que a_2 avait l’accord intra-annotateur le plus faible dans le chapitre 4.

Nous pouvons également confirmer le fait que la fusion permet d’améliorer les performances par annotateur dans tous les cas. Étonnamment, sur le Test, ces modèles fusionnés (CCC=0.884) surpassent même les modèles appris directement sur la référence traditionnelle (meilleur CCC=0.881). Néanmoins comme nous l’avons vu dans le chapitre 6, une différence de 0.003 entre deux scores de CCC n’est pas vraiment significative. La fusion modèle précoce a l’avantage d’avoir des performances moyennes plus élevées que Camembert et d’être le modèle le moins affecté par les annotations individuelles (CV = 0,020 sur l’ensemble de test).

De ces expériences, nous concluons que les approches de fusion semblent être plus robustes à la subjectivité de la tâche d’annotation. Nous avons constaté que la fusion modèle précoce était le meilleur compromis entre performance et robustesse. Ces expérimentations remettent en question le processus d’évaluation largement utilisé qui compare les prédictions à la moyenne des annotations, en effet les valeurs moyennes n’ont pas de réalité perceptive, tandis que les valeurs individuelles en ont une. Il serait intéressant de mettre en place une étude du protocole d’évaluation sur d’autres corpus mais également de mener une étude perceptive sur les résultats de systèmes de reconnaissance ainsi construit.

7.3 Expliquer la frustration dans les conversations

7.3.1 Première écoute humaine : que retire-t-on de l’acoustique ?

Afin de mieux comprendre les données, nous avons arbitrairement choisi d’écouter 57 conversations choisies au hasard dans le corpus. Ces 57 conversations proviennent indépendamment des ensembles d’apprentissage, de développement et de test. Nous les avons classés dans deux catégories : bon ou mauvais, en fonction de leur score de reconnaissance issu de la classification sur la fusion des modalités. On considère comme bon, des scores supérieurs à 0.7. Nous voulions mettre en lumière des facteurs explicatifs de la différence de score entre plusieurs conversations.

Plusieurs phénomènes ont été observés sur ces conversations, qui sont indiqués dans le tableau 7.3, mais nous n’avons pas trouvé un indicateur commun qui en émerge. En effet, ces conversations contiennent ou non du bruit, de la musique, plusieurs locuteurs, des rires, des voix âgées, des silences plutôt marqués (en début, milieu ou fin de conversations) et de

Statistiques	Mauvais	Bon	Total
bruit	29	22	51
musique	7	1	8
conv. autre	11	4	15
silences	22	11	33
locuteur multiple	4	1	5
soupirs	7	2	9
rires	5	0	5
voix âgées	4	0	4
frustration manifeste	5	20	25
femme	26	11	37
homme	9	11	20
assurance	7	12	19
électricité/gaz	4	3	7
santé	2	3	5
voiture	3	3	6
autre	6	13	19
variation annot faible	2	26	28
variation annot forte	20	9	29

TABLE 7.3 – Statistiques sur la présence d’événements retrouvés dans les 57 conversations écoutées.

la frustration manifeste (augmentation du débit de parole, du volume, moins de temps de silence, des injures,...). De plus, nous n’avons rien détecté qui permettrait de relier le sexe du locuteur ou les domaines d’activité dont sont issus les conversations et la variabilité des scores de prédiction.

En se concentrant sur des caractéristiques non linguistiques, nous n’avons pas trouvé de schéma clair et universel de la dimension de satisfaction avec ces observations.

7.3.2 Études statistiques du contenu linguistique des conversations frustrées

Nous avons choisi d’extraire les contenus linguistiques correspondant aux pentes de frustration détectés par le modèle de reconnaissance. Pour cela, nous avons défini les pentes de frustration comme illustré dans la Figure 7.2 par le rectangle bleu. Concrètement, nous nous intéressons au coefficient de variation de la courbe tracée par la prédiction. Si on observe une variation décroissante de la prédiction supérieure à 0.1 point en 2 secondes, on considère que le segment de deux secondes correspond à un pente de frustration. Si

on observe plusieurs pentes qui se chevauchent, on regroupe les segments sous la même annotation de pente.

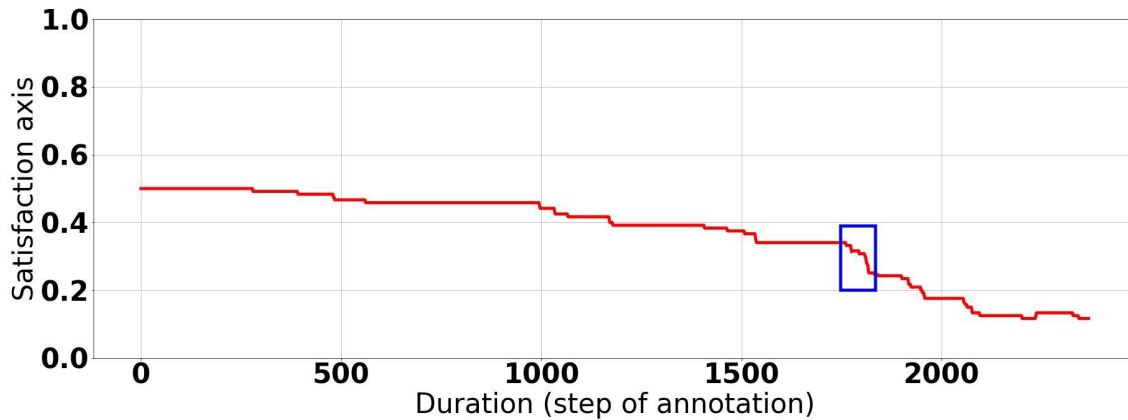


FIGURE 7.2 – Exemple de segment considéré comme une pente de frustration. Le rectangle bleu correspond au segment de pente de frustration.

Une fois cet élément d’analyse mis en place, nous avons extrait les 166 transcriptions correspondant à ces pentes de frustration, en prenant du contexte gauche et droit à hauteur de deux secondes. Des exemples de segments ainsi récupérés sont indiqués dans le tableau 7.4. On peut voir des mots issus de vocabulaire négatif comme *perds mon temps* et *avez fait l’erreur*, ainsi que des constructions spécifiques : des répétitions et une grande insistance sur le sujet avec des *moi je* notamment.

Utilisation de TF-IDF

A partir de ces segments, nous avons conduit une analyse statistique. En utilisant TF-IDF, nous avons analysé les mots, les bi-mots et les tri-mots les plus pertinents. Nous avons ensuite utilisé le résultat de cette fonction pour construire des nuages de mots, illustrés dans la Figure 7.3. Comme nous pouvons le voir, il y a une forte représentation du *je* dans les bi et tri-mots. On observe également beaucoup de tournures négatives (*c’est pas, même pas, je sais pas, je suis pas, c’est pas possible...*) ainsi que des répétitions du *mais*. De plus, on retrouve des mots et des tournures à polarité négatives (*un problème, gros soucis, déposer plainte, j’ai fait opposition, je suis débile, bêtise...*) et des références à des notions temporelles (*dix jours, tous les mois, ce week-end...*). Néanmoins, si nous réalisons les mêmes opérations sur les autres segments, ne correspondant pas aux pentes de frustration, on peut retrouver une grande partie de ces observations.

parce que d'abord je perds du temps alors moi je
 chacun son tour si vous voulez donc moi ce que je vous dis c'est que le véhicule
 ce que je me j'essaie de faire depuis au moins deux semaines
 vais faire appel à mon assistance juridique parce que
 vais tout supprimer et puis c'est tout
 juste pour dire en fait voilà il me faut ça point barre c'est ça
 putain mais visiblement c'est ça
 abonnement pour rien du tout
 c'est vous qui avez fait l'erreur donc je veux ma carte
 alors ça veut dire qu'il faut que je paie une blinde
 j'en ai vraiment besoin je comprends pas

TABLE 7.4 – Exemple de transcription de segments considérés comme pente de frustration.

Nous avons également cherché au niveau de la syntaxe s'il y avait des schémas reconnaissables dans l'expression de la satisfaction. Pour cela, nous avons utilisé l'outil Macaon afin de faire un POS-tagging des segments. À partir du résultat de cette opération, nous pouvons extraire le rôle de chaque mot dans les segments mais nous n'arrivons pas à tirer des conclusions de ces observations.

Comme nous n'arrivons pas à dégager clairement des caractéristiques communes à ces segments, nous avons pensé à utiliser le machine learning pour regrouper les segments en classes, et donc nous permettre de discerner les différences entre les classes. Ainsi nous pourrions peut-être décrire des marqueurs de frustration.

Classification des segments de pente de frustration

Comme nous n'avons pas de classes définies, nous avons fait le choix d'une classification non supervisée. Pour être cohérent avec le nombre limité de segments émotionnels (166), nous avons utilisé un classifieur de type kmeans avec $k=3$. Nous avons classifié les mots, bi-mots et tri-mots suivant ces 3 classes. Nous avons effectué une analyse en composantes principales afin de permettre la visualisation sur les deux premières composantes de nos classes, comme l'illustre la Figure 7.4.

Nous avons ensuite utilisé des TF-IDF, dont nous avons expliqué le fonctionnement au chapitre 3, sur les segments de chaque classe pour dégager des tendances. Les résultats sur les bi-mots et les tri-mots sont relatés dans le tableau 7.5. Nous n'avons pas trouvé de caractéristiques bien saillantes dans ces classes qui permettraient de les dissocier à coup sûr. S'il fallait extrapoler, le premier cluster semble contenir des conversations en



FIGURE 7.3 – Nuages de mots construits à partir des occurrences de bi-mots (gauche) et tri-mots (droite) extraites des transcriptions des segments de pentes de frustration.

rapport avec les mutuelles ; le deuxième cluster semble plutôt concerner des relances ou des clients qui ont déjà appelés ; et le dernier cluster semble être peuplé de litiges de contrat (notamment de l'énergie).

Nous avons donc fait le choix de nous tourner vers un linguiste, afin de collaborer sur l'analyse des transcriptions et retrouver des marqueurs humainement identifiables de la frustration.

7.3.3 Analyses conduites par un linguiste

Dans cette section, nous avons l'intention de fournir des éléments qui pourraient expliquer l'importance de la linguistique pour retrouver la satisfaction ou la frustration. Cette analyse a été faite sur 13 conversations sélectionnées afin de couvrir différentes dynamiques de la dimension satisfaction : globalement plates, occurrences de frustration élevée (annotation de l'axe < 0.4) et occurrences de satisfaction fortement décroissante (pente de frustration). L'analyse a été effectuée à l'aide de la transcription automatique, de l'annotation de l'axe de satisfaction de référence et des balises correspondant à *haute frustration* et *pente de frustration*.

Notre hypothèse est que la parole frustrée comporte principalement des accentuations des phénomènes oraux. Par conséquent, nous avons relevé plus spécifiquement :

cluster 1	cluster 2	cluster 3
au revoir	du tout	que je
affiche bon	en fait	parce que
euh attendez	en ai	je vous
eu lieu	qu on	je ai
la vie	je ai	ce que
monsieur au	du coup	je suis
et on	ai pas	si vous
plus la	je comprends	que vous
quatre fois	comprends pas	pas de
euh quatre	accord euh	et je
la fille	pas accord	ai envoyé
je suis	est pas	moi ai
euh je	on fait	mais mais
il été	ils ont	ai appelé
été livré	je vous	quand même
personne ne	ont fait	je vais
on sait	on est	vous comprenez
ai appelé	fait madame	ai pas
la rubrique	ai fait	suis je
lisez le	je peux	que ça
les choses	peux pas	est pas
mon billet	fait ça	je sais
même allé	ce qu	donc je
procès verbal	fait est	je veux
la commande	ça fait	moi je
faire les	ai ras	est ça
imprimer mon	ras le	vous voulez
deux fois	le bol	que là
ai euh	un petit	que ai
allé la	petit peu	alors que
annuler la	coup juste	je viens
assure votre	juste qu	élégant mais
choses quoi	moi en	souci là
de vous	est ça	mais bon
donc euh	pas du	viens appeler
ils demandent	pour rien	là même
donc votre	rien du	je te

TABLE 7.5 – Classification bi-mots des segments émotionnels correspondant aux pentes de frustration sur laquelle nous avons appliqué un TF-IDF.

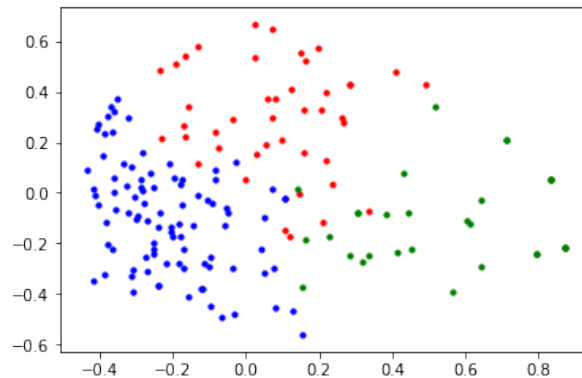


FIGURE 7.4 – Visualisation des données de segments émotionnels projetés sur les deux premières composantes obtenues par une ACP, sur laquelle une classification de type kmeans (k=3) est effectuée.

- Quantité de disfluences,
- Hésitations, répétitions, bégaiements,
- Importance des auto-coupures définies comme *les points où le flux d'énoncé est rompu* [Pal+19],
- Usage des interrogations et des négations,
- Preuves sémantiques de frustration ou au moins d'émotions négatives,
- Quantité de segments significatifs *vs.* segments sémantiquement vides.

Sur la base de ces indices, l'analyse aboutit à différentes observations. Il existe des marqueurs sémantiques de frustration dans les conversations telles que l'usage de la négation (*ça ne m'amuse pas, c'est inadmissible*) : des marqueurs forts (*c'est gonflé, putain de ...*) et des marqueurs faibles (*quand même, franchement*). Il semble également que la quantité de segments significatifs, d'auto-ruptures et de disfluences, soit généralement corrélée à de fortes augmentations de frustration. La structure syntaxique des énoncés interrogatifs semble également corrélée à la frustration.

Dans un second temps, nous comptons aller plus loin dans cette analyse avec l'extraction automatique d'indices de la frustration. Bien entendu, passer d'une extraction manuelle à une extraction automatisée en fonction du temps (avec un pas de 250 ms) implique de faire des choix dans la définition des indices. En essayant de modéliser la quantité de segments significatifs, nous extrayons les balises POS à l'aide de MACAON [Nas+11] directement à partir de transcriptions automatiques et calculons le nombre de verbes et de noms que l'on met en relation avec le temps. Pour capturer les autres indices, nous avons décidé d'extraire automatiquement les sept caractéristiques mentionnées dans le tableau

Caractéristiques	Nombre d'occurrences
Répétition d'un mot (deg1)	26
Répétitions de deux mots (deg2)	4
Pauses dans le discours (<i>euh, bah, hein, eh, etc.</i>)	22
Marqueurs forts (<i>important, inquiet, scandaleux, etc.</i>)	14
Marqueurs faibles (<i>quand même, franchement, etc.</i>)	3
Négations (<i>pas, ne, n'</i>)	30
<i>c'est</i>	44
nombre de mots dans <i>lettre recommandée</i>	1050
nombre de segment de parole dans <i>lettre recommandée</i>	152

TABLE 7.6 – Sept caractéristiques et leur nombre d'occurrences permettent de modéliser les indices supposés être responsables de la frustration dans les conversations. Le nombre total de mots et de segments de parole de la conversation appelée *lettre recommandée*, sont également indiqués.

7.6.

L'idée n'est pas de fournir une analyse exhaustive sur l'ensemble des données mais de fournir quelques indices explicatifs. Nous nous concentrons ici sur l'analyse approfondie d'une seule conversation que nous appelons *lettre recommandée*. Toutes les occurrences des caractéristiques résumées dans Table 7.6 sont synchronisées dans le temps avec la référence de satisfaction annotée.

Les quantités de verbes et de noms ne donnent pas d'information pertinente et ne sont pas représentées ici. L'analyse linguistique dynamique est présentée sur la Fig. 7.5. La conversation *lettre recommandée* a été annotée avec une forte baisse de satisfaction avant 200 sec. La transcription automatique obtenue juste avant cette chute est donnée dans le Tableau 7.7. Juste avant la chute, les occurrences de répétition de mots simples et de *c'est* sont importantes, alors qu'après la chute, le nombre de pauses remplies par des interjections (*euh, bah, hein, eh, etc...*) et de marqueur de négation (*pas*) augmente. On remarque également qu'un marqueur fort (*réclamation*) se produit juste avant la chute, signifiant probablement que ce mot spécifique induit chez l'annotateur la perception d'une frustration notable.

Dans le corpus AlloSat, l'information émotionnelle extraite semble résider davantage dans les mots que dans le contenu prosodique et acoustique. Dans ces données, l'expression de la frustration est principalement liée à l'accentuation des phénomènes oraux : le contenu sémantique et surtout les auto-coupures, les disfluences, les hésitations et les répétitions.

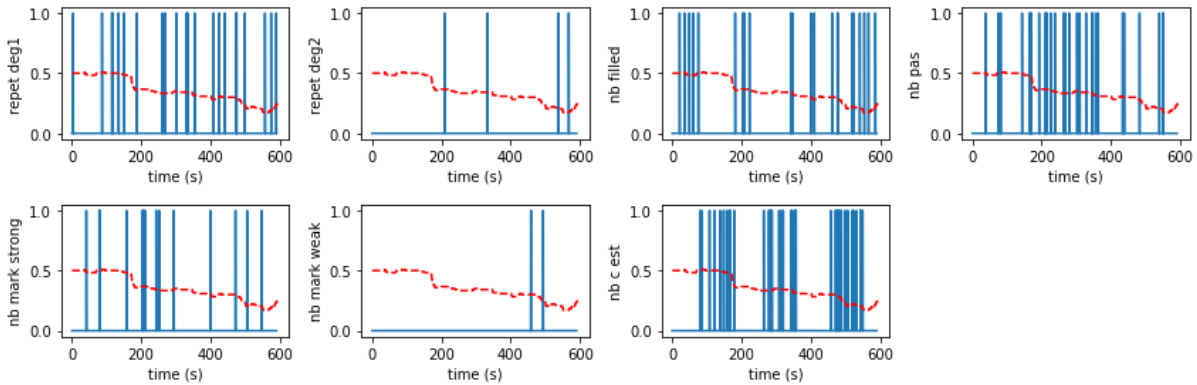


FIGURE 7.5 – Analyse dynamique de la frustration de la conversation appelée *lettre certifiée*. Le nombre d’occurrences des sept caractéristiques linguistiques est tracé par rapport au temps. La référence de l’axe de satisfaction est représentée par la ligne en pointillée rouge.

- *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

TABLE 7.7 – Extrait (137 - 166 sec.) de la conversation *lettre recommandée*. Disfluences : *italic*; Hesitations, bégaiements : underline; Traces Semantiques de la frustration : **bold**; auto-coupures : //

7.4 Considération du genre

Nous avons voulu conduire une analyse genrée des échanges entre les clients et les conseillers. En effet, nous voulons voir si le genre a un impact sur la frustration et son expression dans notre corpus. Pour ce faire, nous avons extrait les paires clients-conseillers dont nous avons les informations, soit 81 conversations sur les 303. En effet, le corpus ne comportant pas de données sur le conseiller, il s’agit de données confidentielles soumises à la RGDP, qui explique que la plupart des métadonnées des conversations ne sont pas conservées par l’entreprise.

Nous avons voulu mettre l’accent sur les conversations portant des pentes de frustration tel que définis dans ce chapitre. Comme nous pouvons le voir dans le tableau 7.8, il n’y a pas assez de données pour donner une quelconque conclusion significative. Dans

Client	Conseiller	Nombre de conversations	Pente de frustration présente
femme	femme	38	9
homme	homme	5	0
femme	homme	17	5
homme	femme	21	7

TABLE 7.8 – Présence de pentes de frustration en fonction du genre de la paire d'interlocuteur.

la Figure 7.6, nous avons l'impression que les hommes sont plus énervés quand ils ont une femme en tant qu'interlocutrice et que la frustration est plus présente sur les paires hétérogènes.

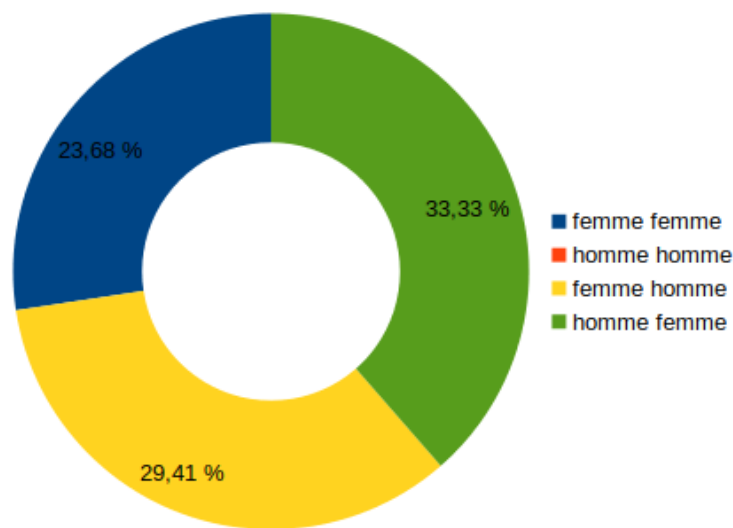


FIGURE 7.6 – Visualisation du pourcentage de conversations contenant des pentes de frustration en fonction du genre de la paire d'interlocuteurs. Le couple homme-homme n'apparaît pas puisqu'il n'y a pas de conversations avec une pente de frustration pour cette paire d'interlocuteurs dans les 81 conversations considérées.

Il serait intéressant de conduire une étude plus large sur l'impact du genre des interlocuteurs sur l'expression de la frustration.

7.5 Conclusion

Dans ce dernier chapitre de contribution, nous avons souhaité prendre du recul sur nos travaux et expliquer nos résultats. Nous avons tout d’abord discuté de l’utilisation d’une annotation moyennée comme référence à nos systèmes de reconnaissance et plus largement de la pertinence de ce protocole qui est très largement adopté au sein de la communauté scientifique pour les tâches ayant une part non négligeable de subjectivité. Nous nous sommes ensuite concentrés sur la modalité linguistique de notre corpus, à savoir les transcriptions automatiques des conversations, et nous avons essayé de mettre en lumière les facteurs expliquant la très bonne performance des systèmes de reconnaissance appris à partir du texte. Enfin nous avons considéré rapidement l’impact du genre sur la frustration.

Nous pouvons conclure qu’il est difficile de statuer sur des marqueurs linguistiques de la frustration. En effet, en tant qu’humain, nous ressentons bien cette frustration dans le texte, sans pour autant pouvoir définir des marqueurs clairs et universels de cette émotion. Nous pouvons néanmoins souligner la présence des disfluences, des auto-coupures et des bégaiements comme des marqueurs pouvant alerter sur une potentielle frustration.

CONCLUSION

Nous concluons ici sur les travaux réalisés pendant ces années de thèse et nous proposons quelques perspectives qui pourront orienter de futurs travaux.

Conclusion

Tout au long de cette thèse nous avons travaillé autour des émotions. Ce travail, inscrit dans un contexte industriel fort, permettra demain à l'entreprise Allo-Média de commercialiser une solution de reconnaissance automatique des émotions de satisfaction et de frustration dans des conversations issues de centre d'appels.

Nous avons tout d'abord défini, en accord avec le besoin industriel, les émotions pertinentes dans la relation clientèle. La satisfaction et la frustration sont des marqueurs forts, vecteurs de gain et de perte de clients par exemple : un client satisfait a plus de chance de continuer avec une entreprise alors qu'un client frustré peut aller voir la concurrence. Ces deux émotions, que nous traitons comme deux faces d'un même ensemble ne sont pas très présentes dans la littérature si on considère nos obligations, à savoir la langue française et le canal téléphonique.

Une première partie de nos travaux a donc porté sur l'élaboration d'un corpus permettant de répondre à ce besoin industriel. En effet, il est difficile d'établir un quelconque système de reconnaissance sans avoir de données sur lesquelles s'appuyer. Nous avons donc construit le corpus AlloSat, qui est annoté en satisfaction et frustration de façon discrète et continue. Ce corpus de 37h d'audio, contient uniquement les voix des clients, puisque c'est sur ces derniers que nous focalisons notre attention. Il est disponible pour toute personne affiliée à un institut public de recherche sur simple demande aux personnes responsables de sa diffusion.

Une fois ce corpus dans nos mains, nous avons pu mettre en place des systèmes de reconnaissance de ces deux émotions. Nous avons tout d'abord travaillé sur les annotations discrètes de ce corpus. Ces études préliminaires nous ont permis de valider des algorithmes d'apprentissage capables de retrouver les émotions de satisfaction et de frustration dans le langage.

Nous avons ensuite mis en place des systèmes de reconnaissance de l'émotion continue.

Ce choix est justifié par les besoins industriels. En effet, nous avons besoin de connaître l'état émotionnel des clients mais aussi de pouvoir situer dans une conversation les instants de frustration. Cette reconnaissance s'appuie sur des réseaux de neurones dont nous avons comparés différentes architectures pour trouver la plus performante dans notre contexte. Nous nous sommes aussi beaucoup attardés sur la représentation des données et donc sur les features que nous donnons en entrée de la reconnaissance d'émotion.

Afin de juger de la pertinence mais surtout de la performance de nos systèmes, nous avons cherché à nous comparer à des expérimentations issues de l'état de l'art du domaine. Pour cela, nous avons étudié le corpus SEWA et nous avons mis en place une série d'expérimentations visant à reproduire nos travaux sur ce corpus. Bien que se ressemblant sur de nombreux points, les corpus restent très différents : la partie de SEWA utilisée est en allemand et hongrois, contient des conversations entre deux interlocuteurs, avec des durées de conversation plutôt homogènes, sur un sujet bien précis, n'utilisant pas le canal téléphonique, annotées en valence, activation et *liking*... Alors qu'AlloSat est en français avec uniquement la voix d'un interlocuteur, dont les durées de conversations varient de quelques secondes à plusieurs dizaines de minutes, concernant plusieurs domaines d'activités, le tout avec un canal téléphonique, annotées en satisfaction/frustration. Autant de justifications possibles qui peuvent expliquer pourquoi ce qui fonctionne sur l'un de ces corpus, ne fonctionne pas forcément sur l'autre.

Pour aller plus loin, nous avons voulu explorer la modalité linguistique de notre corpus. En effet, jusque là, nous utilisons uniquement le signal audio. Grâce à un modèle de reconnaissance automatique de la parole, nous avons accès aux transcriptions de ce signal. Ainsi, il nous est possible de traiter indépendamment la modalité linguistique en se concentrant sur les mots. Cette étude a montré que, en plus d'être pertinente, la modalité linguistique permet d'atteindre de meilleurs scores de reconnaissance.

Comme nous avons deux modalités que nous avons validé comme étant pertinentes, nous les avons fusionnées afin de questionner l'apport de chaque modalité et de savoir si elles sont redondantes ou non. Comme nous améliorons les scores de reconnaissance par ce procédé ce que nous vérifions sur les courbes de prédictions, ces modalités semblent permettre de retrouver des informations complémentaires sur l'état émotionnel des individus.

De plus, nous nous sommes questionné sur des méthodes pour masquer un peu la faible quantité de données dont nous disposons. En effet, même si notre corpus est de taille respectable, il ne contient pas autant de données que des corpus traditionnellement

utilisés pour alimenter des systèmes à base de réseaux de neurones. Notre solution a été d'utiliser des modèles pré-appris pour extraire des embeddings plus spécifiques. Cette technique, nouvelle au début de cette thèse, est aujourd'hui de plus en plus utilisée par de nombreux domaines. Nous avons pu montrer qu'elle est tout à fait pertinente avec notre corpus, nous permettant d'atteindre un score de $CCC = 0.920$ sur le test, ce qui correspond aujourd'hui au meilleur score atteint sur le corpus AlloSat.

Un autre aspect de notre travail a été de vouloir expliquer la performance de la modalité linguistique. En effet, nous ne pensions pas que cette modalité apporterait autant d'informations pertinentes aux systèmes de reconnaissance. Nous avons donc voulu déceler les marqueurs de l'émotion dans les transcriptions. Pour cela, nous avons mis en place des méthodes statistiques et des écoutes humaines, effectuées par des informaticiens et un linguiste. Même si nous avons réussi à dégager quelques indicateurs, de plus amples investigations seraient à envisager.

Perspectives

Il reste de nombreuses pistes que nous aurions voulu emprunter pour approfondir notre travail lors de cette thèse.

Enrichir le corpus AlloSat

Actuellement, le corpus AlloSat est distribué dans son intégralité aux chercheurs qui en font la demande. Nous avons fait ce choix car nous voulons permettre à la communauté de s'en saisir et d'aider à l'avancée dans le domaine, sans aucune restriction. Cependant, il peut être intéressant d'utiliser ce corpus afin de réaliser des campagnes d'évaluation. Ces campagnes permettent de stimuler la communauté en faisant appel à l'esprit de compétition des différents laboratoires. Elles permettent donc de rassembler plusieurs équipes sur une même problématique et donc d'avancer souvent plus rapidement sur cette dernière. Ajouter au corpus AlloSat une partition de test non diffusée permettrait de mettre en place une campagne portant sur la détection de la satisfaction et de la frustration dans les conversations de centres d'appels. Faire annoter de nouvelles données mettrait encore plus en valeur le corpus, et ces données pourraient être utilisées pour entraîner de nouveaux systèmes de reconnaissance.

L'une des problématiques majeures du domaine de la reconnaissance d'émotion reste le manque de données. En effet, même s'il existe une multitude de corpus, ils sont souvent

d'une dimension assez restreinte et non compatibles les uns avec les autres. Les émotions étant subjectives et les protocoles d'annotation différents d'un corpus à l'autre, des études multi-corpus sont assez difficiles à réaliser. Il serait intéressant de mettre en place des outils unifiés, permettant de faciliter ces études multi-corpus. Nous pensons notamment à des outils permettant de régulariser les annotations, voir des systèmes de traduction permettant de décrire toutes les émotions étiquetées par une même représentation. Un premier pas dans ce sens serait de considérer toutes les émotions comme composantes d'un seul et même vecteur de type one-hot par exemple.

De plus, nous pensons que l'aspect multi-domaine du corpus AlloSat pourrait se révéler intéressant sous le prisme de l'étude de la sémantique. En effet, le corpus se compose de conversations traitant de questions énergétiques, d'assurances, du voyage, de la téléphonie et de l'immobilier. Le corpus pourrait servir à construire un module de reconnaissance de concepts sémantiques multi-domaine.

Plus d'explicabilité de la satisfaction et de la frustration

Une grande faiblesse des systèmes de reconnaissance basés sur l'apprentissage automatique est le manque d'explicabilité des résultats. En effet, il est très difficile de justifier les résultats de nos reconnaissances avec des faits. C'est pour cela que nous avons commencé un travail sur la recherche d'indices et de patterns permettant à l'humain de comprendre et d'anticiper la survenue de la frustration. Nous aimerions pousser ces travaux plus loin, en mettant en place une analyse perceptive et participative des conversations par le plus grand nombre possible.

Pourquoi combattre la subjectivité par le nombre ?

Dans la littérature, nous avons l'habitude de palier à la subjectivité d'une tâche en multipliant les annotateurs notamment. Nous partons du principe que plus il y aura d'annotateurs et plus nous serons capables d'approcher le *réel*. Nous avons commencé à questionner ce postulat dans nos travaux et par faute de temps, nous n'avons pas pu aller au bout de ce travail. Il serait intéressant de le poursuivre et, en fonction des résultats, l'étendre à d'autres domaines.

Améliorer les performances de reconnaissance

Depuis la diffusion du système BERT, de nombreux systèmes pre-trained ont vu le jour. Dans cette thèse nous nous sommes attardés sur deux d'entre eux en particulier : wav2vec et camemBERT. Depuis, d'autres systèmes plus performants ou plus spécifiques à notre tâche ou à nos données ont vu le jour. Il serait intéressant de mettre à jour nos systèmes pour bénéficier de l'apport de ces nouveautés. De même, nous pourrions imaginer la construction de ce type de système pour construire des embeddings représentant l'aspect émotionnel des données.

Nous avons réalisé nos travaux en partant du corpus AlloSat. Nous avons également étendu nos travaux au corpus SEWA, afin de confirmer la pertinence de nos choix en terme de modèles et de représentation des données notamment. Il serait intéressant d'étendre encore nos travaux à d'autres corpus du domaine, notamment au corpus SEMAINE, afin de déterminer si nos expérimentations sont valables uniquement sur la modalité de la satisfaction et de la frustration ou si notre approche est compatible avec d'autres émotions, d'autres protocoles d'annotation et d'autres contextes de conversations.

De nouvelles solutions industrielles

Lors de cette thèse, nous n'avons malheureusement pas pu répondre à la problématique de temps réel. En effet, il est très intéressant pour les industriels de détecter la satisfaction et la frustration directement lors de l'appel du client. Ainsi l'entreprise peut répondre instantanément à ces émotions. Un premier pas vers du temps réel a été réalisé après la rédaction de ce manuscrit en remplaçant les architectures bidirectionnelles par de l'unidirectionnelle et par la prise en compte de morceaux de conversation plutôt que de conversations entières (soit des découpages de deux minutes). Ces travaux préliminaires donnant des résultats encourageants, il serait intéressant de les poursuivre.

Enfin, l'entreprise Allo-Média a toujours eu à cœur de participer à des cas d'utilité publique. Mettre en place de la reconnaissance de frustration dans des contextes critiques permettrait de mettre la technologie au profit direct des personnes vulnérables. Nous pouvons par exemple penser à de la modération des priorités dans des centres d'appels d'urgences.

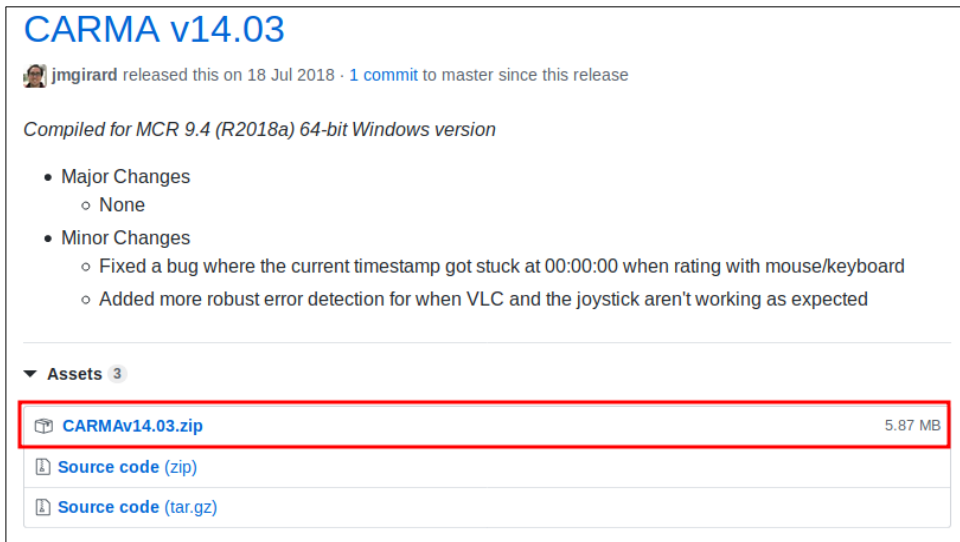
ANNEXES

8.1 Guide d'installation et de configuration de CARMA


Mise en place de l'outil CARMA

1- Télécharger VLC (si non disponible sur les ordinateurs) depuis le lien suivant :
<https://www.videolan.org/vlc/index.fr.html>
Installer le logiciel

2- Télécharger CARMA depuis le lien suivant :
<https://github.com/jmgirard/CARMA/releases>
Prendre le zip CARMAv14.03.zip






CARMA v14.03

 jmgirard released this on 18 Jul 2018 · 1 commit to master since this release

Compiled for MCR 9.4 (R2018a) 64-bit Windows version

- Major Changes
 - None
- Minor Changes
 - Fixed a bug where the current timestamp got stuck at 00:00:00 when rating with mouse/keyboard
 - Added more robust error detection for when VLC and the joystick aren't working as expected

▼ Assets 3

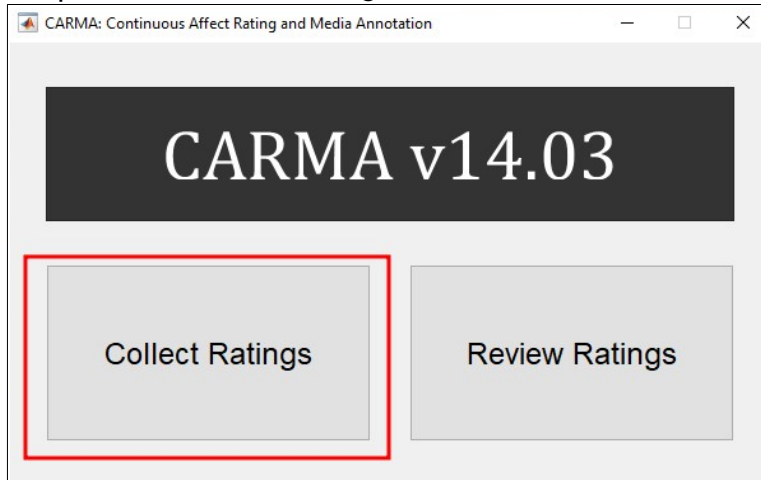
 CARMAv14.03.zip	5.87 MB
 Source code (zip)	
 Source code (tar.gz)	

Dézipper le dossier et lancer l'installer_web.exe
Suivre les étapes de l'installateur
Une fois l'installation terminée, redémarrer l'ordinateur

3- Configuration de CARMA:

Lancer CARMA

Cliquer sur Collect Ratings

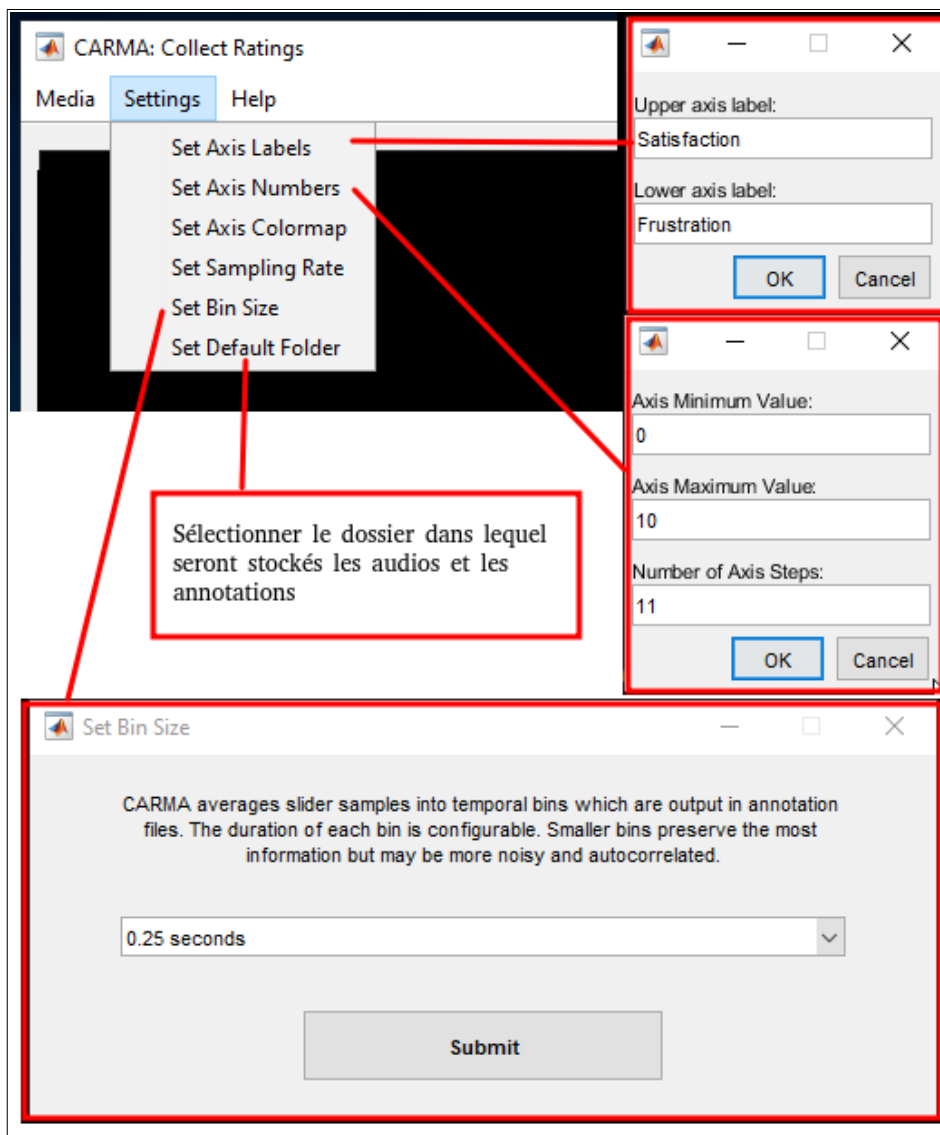


Définir les paramètres comme suivants :

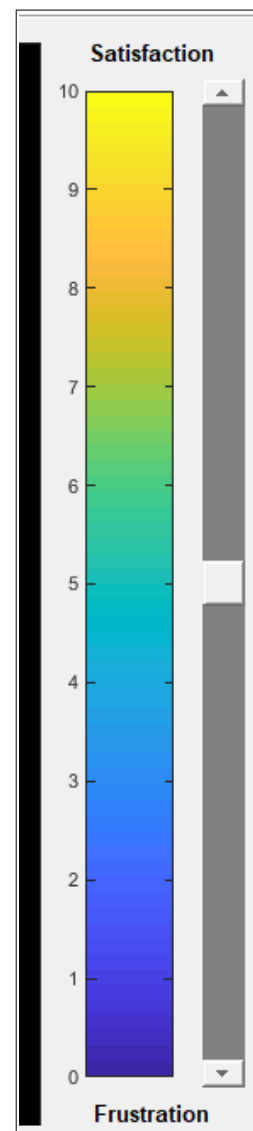
- Axis Label : Upper axis label = **Satisfaction** ; Lower axis label = **Frustration**
- Axis Number : Axis Minimum Value = **0** ; Axis Maximum Value = **10** ; Number of Axis Step = **11**
- Bin Size = Sélectionner **0,25 seconds**
- Default Folder = Sélectionner le dossier dans lequel se trouve les fichiers audio.
Attention, si vous entendez un son de type avertissement Windows, c'est que le dossier n'a pas été pris en compte. Il faudra donc renouveler l'opération.

Vous devez retrouver l'échelle Satisfaction – Frustration indiquée dans l'image suivante.

Les configurations sont sauvegardées pour les prochains lancements du logiciel. Toutefois il peut arriver que celles-ci soient écrasées par la configuration par défaut. Il faudra aller réitérer l'étape 3 avant de commencer une annotation.



The screenshot displays the 'CARMA: Collect Ratings' application. The 'Settings' menu is open, listing options: 'Set Axis Labels', 'Set Axis Numbers', 'Set Axis Colormap', 'Set Sampling Rate', 'Set Bin Size', and 'Set Default Folder'. A red box highlights the 'Set Bin Size' option, with a callout box containing the text: 'Sélectionner le dossier dans lequel seront stockés les audios et les annotations'. Below this, the 'Set Bin Size' dialog box is shown, containing the following text: 'CARMA averages slider samples into temporal bins which are output in annotation files. The duration of each bin is configurable. Smaller bins preserve the most information but may be more noisy and autocorrelated.' A dropdown menu is set to '0.25 seconds' and a 'Submit' button is at the bottom.



8.2 Guide d'annotations

Annotation continue et discrète des émotions: Guideline

Si ce petit guide ne répond pas à toutes vos questions, n'hésitez pas à m'adresser un mail à m.macary@allo-media.fr

Contexte

Dans le cadre de ma thèse, je voudrais construire un système permettant de détecter la frustration et la satisfaction des appelants. Autrement dit, permettre de repérer automatiquement les appels contenant de la frustration et de la satisfaction.

Afin de construire ce système, j'ai besoin de données annotées. Vous allez donc annoter l'émotion des appelants, des clients. Les émotions étant une notion très subjective et très difficile à définir, nous nous concentrons sur deux aspects :

- L'évolution d'une émotion : tous les appels commencent au 'neutre' et évoluent en fonction du temps entre FRUSTRATION et SATISFACTION.
- Une description de l'appel : une fois l'appel terminé, nous voulons avoir un retour sur l'évolution de l'émotion. Pour cela nous aurions besoin de savoir comment était l'appelant au début de la conversation, comment il était à la fin et comment était l'évolution.

Pour essayer de palier l'aspect subjectif, nous avons besoin que plusieurs personnes annotent la même conversation. Ainsi, en faisant la moyenne de vos observations, nous aurons une vue moins subjective sur le sujet.

Outils

CARMA :

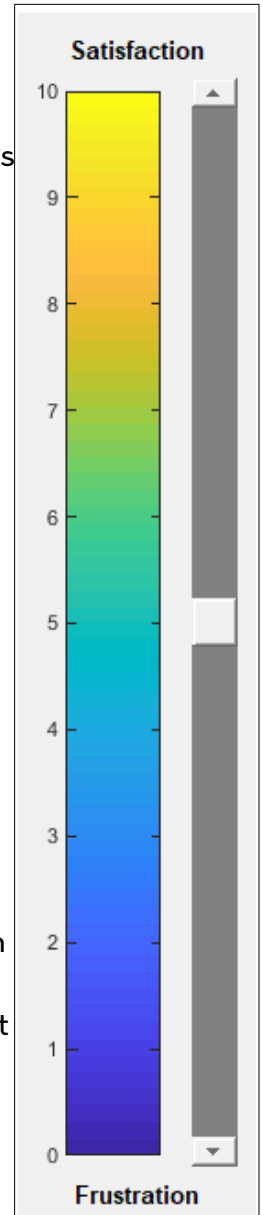
CARMA (Continuous Affect Rating And Media Annotation) est le logiciel d'annotation que nous avons retenu pour cette tâche. Il permet de faire de l'annotation en continu comme le suggère son nom. Nous vous détaillerons comment s'en servir.

Feuille Annotation Excel :

Afin de consigner les résultats des descriptions, vous avez un fichier EXCEL à remplir.

Consignes

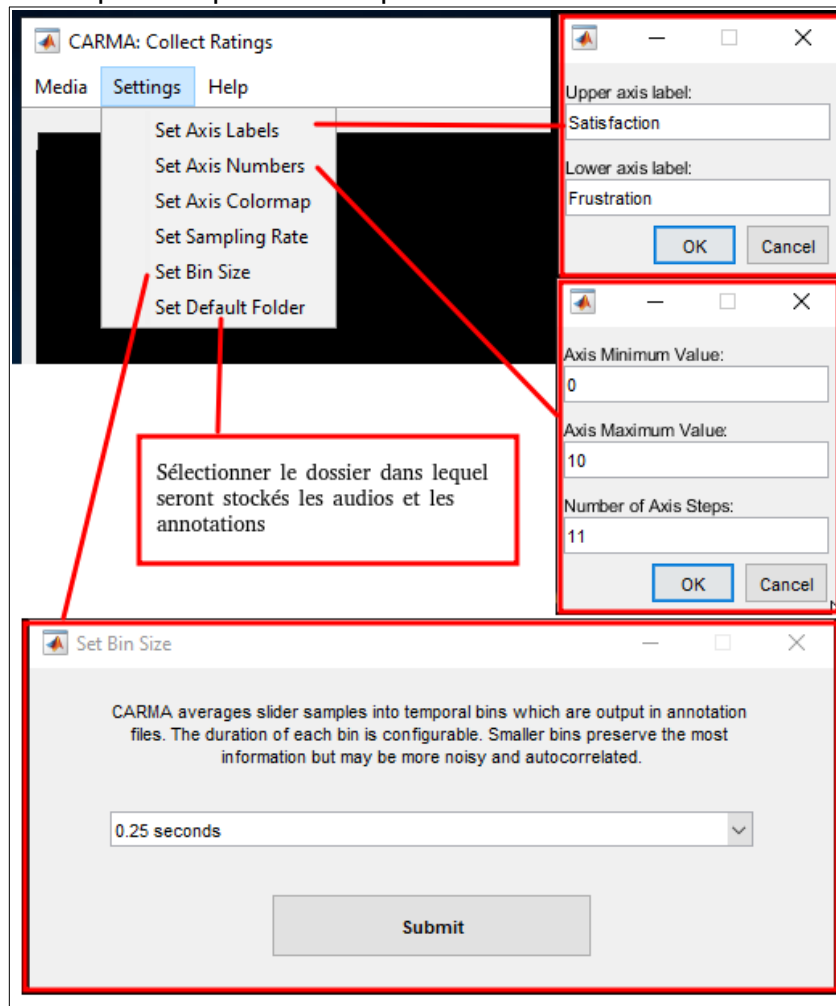
- En début de session d'annotation, vérifier deux paramètres :
 - La barre d'annotation à droite doit avoir les mentions Satisfaction – Frustration et s'étendre de 0 à 10 avec le curseur au milieu au niveau du 5. Si vous constatez des différences, vous trouverez dans la partie "Mise en place" de ce document comment configurer CARMA.
 - Dans Settings > Bin Size : La valeur sélectionnée doit être 0.25 seconds. **Attention** si ce n'est pas le cas, l'annotation ne pourra pas être exploitée.
- Restez le plus neutre possible lors de votre annotation. En effet, vous devez rester le plus objectif possible.
- N'annotez qu'une seule fois un même document. Vous ne pouvez pas revenir en arrière ou avancer l'audio. Si vous voulez une pause, vous pouvez mettre l'annotation en pause et reprendre plus tard. **Attention cependant !** Vous ne pourrez pas enregistrer une annotation tant que le fichier audio n'est pas arrivé à son terme.
- Si vous ne détectez ni Frustration ni Satisfaction, vous pouvez laisser le curseur sur le neutre, tel qu'il est positionné à chaque début de conversation. Il est tout à fait possible qu'il ne se passe rien du tout lors d'une conversation, ce n'est pas la peine de sur-annoter.
- Si vous avez de la mise en attente, du bruit, de la musique sans manifestation de l'appelant, ne touchez pas le curseur, laissez le en place.
- Lorsque vous entendez un petit bruit blanc, comme un grésillement de télévision (joint en copie de ce document appelé silence), cela indique qu'un silence de plus de 2 secondes s'est produit. Ce son est à titre informatif pour vous aider dans l'annotation.
- Vous entendrez également dans certains fichiers un son jazzy qui est utilisé afin d'anonymiser les conversations.
- Lorsque vous avez fini l'annotation continue du fichier audio, remplissez le fichier EXCEL, pour ne pas être biaisé par une autre conversation écoutée entre temps.
- Pensez à sauvegarder le fichier EXCEL souvent, à chaque fois que vous avez fini d'annoter un fichier par exemple.
- Ecoutez les deux fichiers appelés borne_frustration et borne_satisfaction avant les premières annotations pour se faire une idée de l'amplitude possible des émotions.



Mise en place

Comme expliqué dans les consignes, vérifier les informations sur la barre d'annotation. Si celles-ci ne sont pas conformes aller dans "Settings" et modifier :

- Axis Label : Upper axis label = **Satisfaction** ; Lower axis label = **Frustration**
 - Axis Number : Axis Minimum Value = **0** ; Axis Maximum Value = **10** ; Number of Axis Step = **11**
 - Bin Size = Sélectionner **0,25 seconds**
 - Default Folder = Sélectionner le dossier dans lequel se trouve les fichiers audio.
- Attention, si vous entendez un son de type avertissement Windows, c'est que le dossier n'a pas été pris en compte. Il faudra donc renouveler l'opération.



Renommer le fichier EXCEL : annotation_emotion.xls en **annotation_emotion_VosInitiales.xls**. Il y aura donc autant de fichiers EXCEL à livrer qu'il y a d'annotateurs.

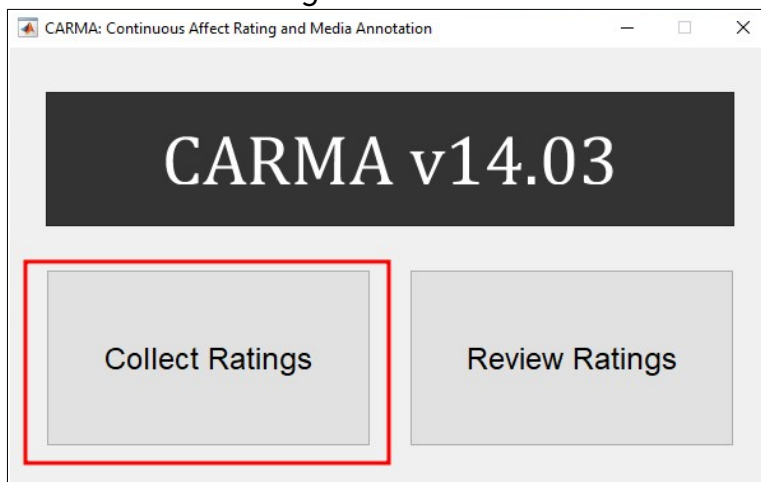
Pas à pas

Voici comment annoter un fichier du début à la fin.

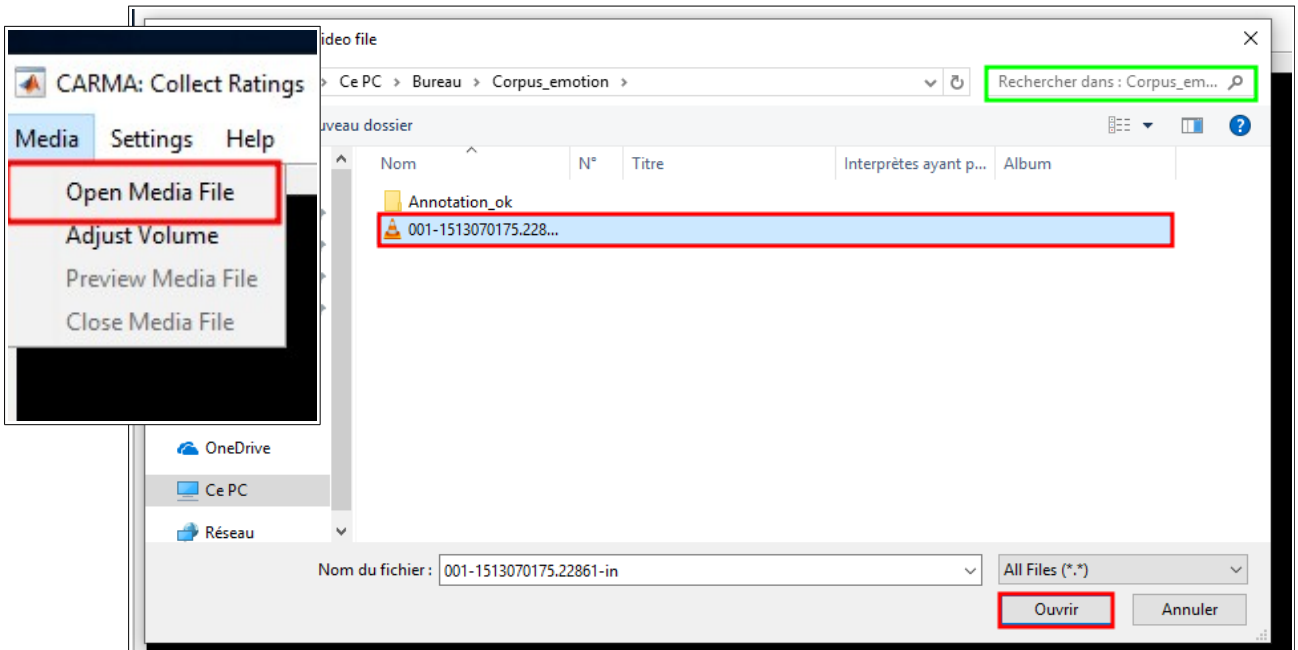
- Ouvrez le fichier annotation_emotion_VosInitiales.xml (par exemple annotation_emotion_MaMa.xml pour moi)
Les noms sont triés dans l'ordre croissant, pour que vous puissiez retrouver le plus facilement possible les documents.
- Reprenez où vous en étiez. Ici : 001-1513070175.22861-in .

NOM	Genre	VALENCE			FRUSTRATION-SATISFACTION					
		début	évolution	fin	début	évolution	fin			
001-1513070175.22861-in									Valence	Très positive
001-1513349340.1020-in										positive
001-1514371642.6441-in										neutre
001-1514554092.8011-in										négative
001-1514974592.9668-in										Très négative
001-1514984921.9908-in									Frustration-Satisfaction	Très satisfait
001-1522239260.20644-in										satisfait
001-1522393254.30278-in										neutre
001-1522405444.32913-in										frustré
001-1522766552.50079-in										très frustré
001-1524047675.171743-in									Évolution	Stagne
001-1524122911.177860-in										Monte
001-1524123022.177885-in										Descend
001-1524229828.190714-in										Fluctue
001-1524812801.236313-in										Fluctue fortement
001-1528288987.68269-in										

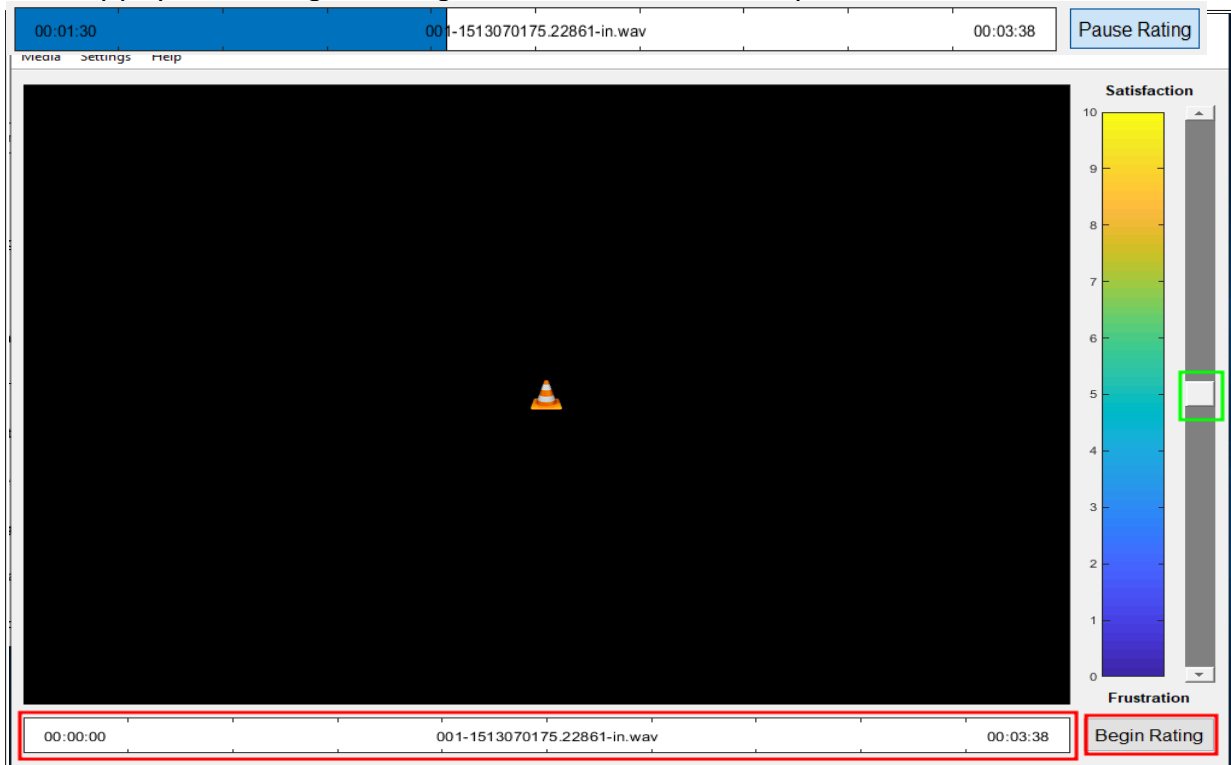
- Ouvrez Carma. Le lancement peut être un peu long, c'est normal. Cliquez sur "Collect Ratings".



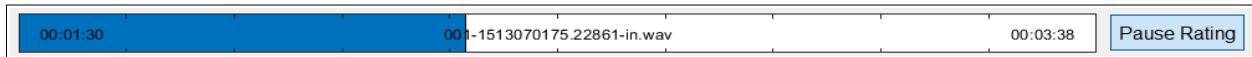
- Vérifiez l'axe d'annotation en Satisfaction – Frustration et de 0 à 10. Vérifiez également dans Settings > Bin Size que la valeur sélectionnée est 0.25 seconds.
- Appuyez sur Media > Open Media File. Une fenêtre va apparaître pour vous demander de sélectionner l'audio. Trouvez l'audio que vous avez identifié à l'étape 2 (Ici 001-1513070175.22861-in) et ouvrez-le. Vous pouvez utiliser la fonction de recherche (encadré vert) pour retrouver plus facilement l'audio.



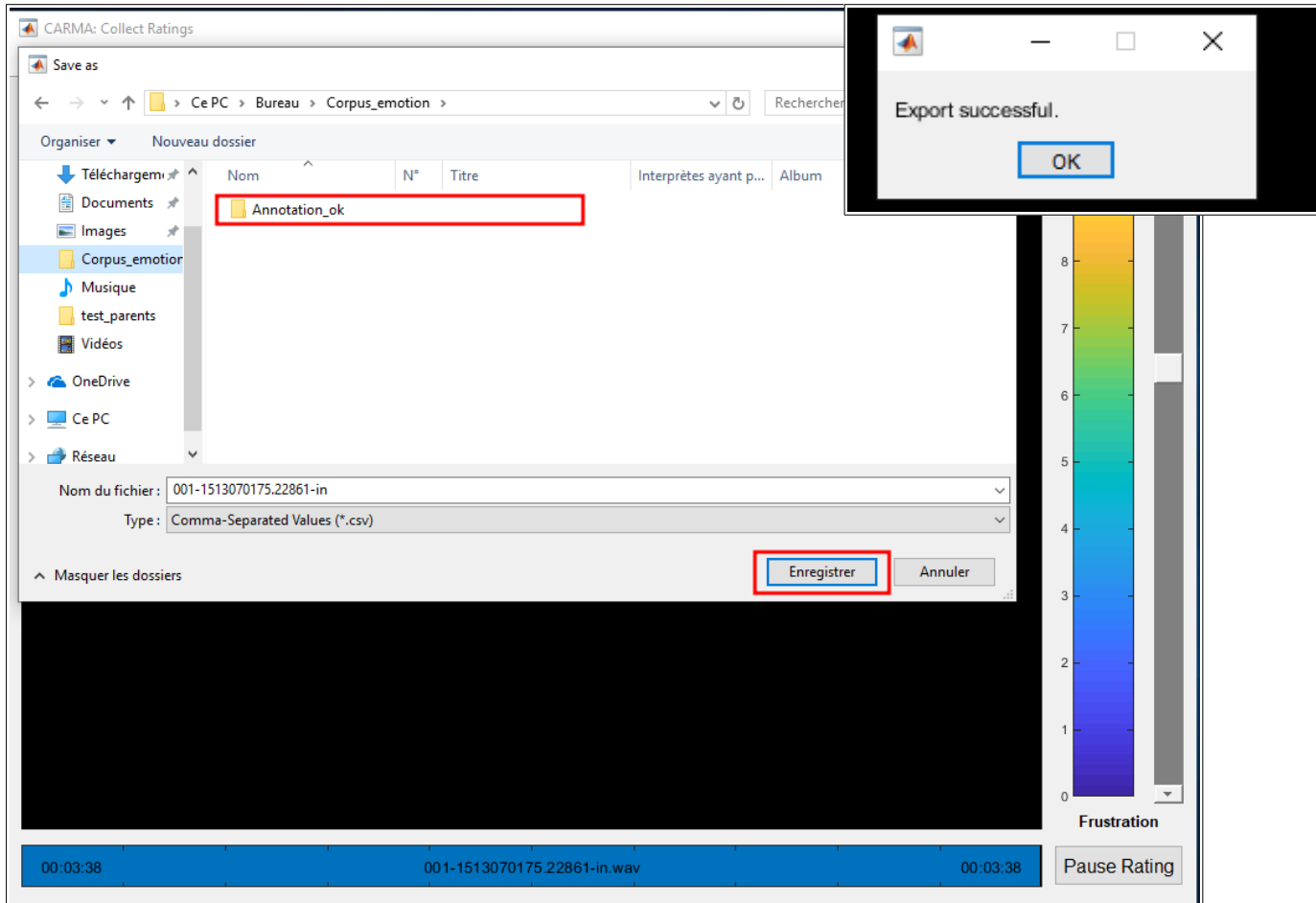
6. Votre fenêtre doit maintenant contenir, en bas, le nom du fichier et la durée. Appuyez sur Begin Rating. Vous verrez un décompte se lancer avant écoute.



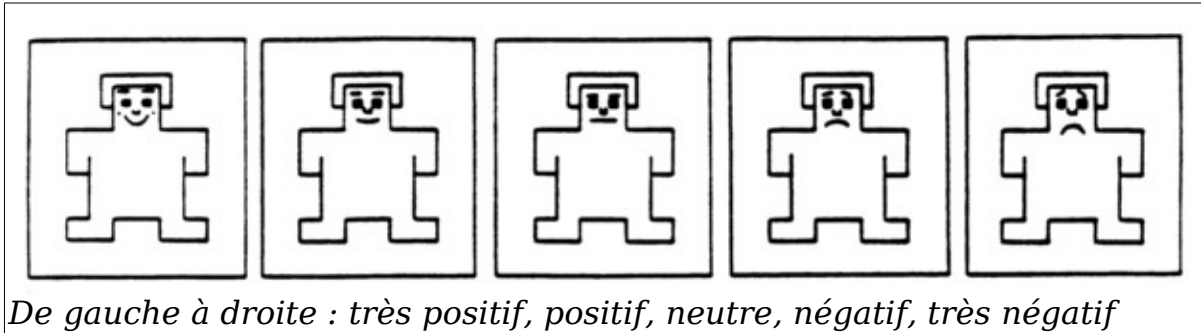
7. Utilisez les flèches de votre clavier pour faire varier le curseur. Vers le haut pour la satisfaction, vers le bas pour la frustration. Vous verrez le curseur bouger sur l'écran (encadré vert sur l'image précédente). Si vous avez besoin de faire une pause, appuyez sur Pause Rating.



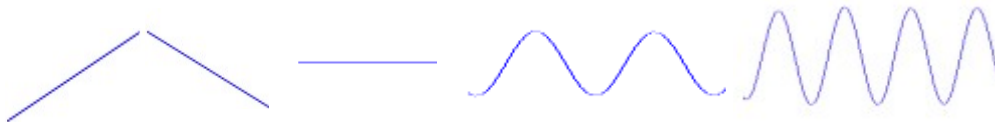
8. A la fin de l'annotation, vous verrez une fenêtre indiquant que l'annotation est terminée et vous demandant de la sauvegarder. Mettez le nouveau fichier dans le dossier Annotation_ok. Enregistrer. Vous verrez apparaître Export Successful qui signifie que tout s'est bien déroulé.



9. Reprenez le fichier EXCEL et remplissez les informations sur le document que vous venez d'annoter. Tout d'abord, le genre de la personne qui parlait : homme ou femme. Vous devez ensuite choisir entre les catégories proposées à droite dans le document. La valence correspond à une émotion positive ou négative, sans la définir. Pour vous aider, voici comment on définit par des pictogrammes la notion de valence :



Si vous trouvez qu'au départ la personne montre une émotion neutre, positive, très positive, négative ou très négative, inscrivez-le ici. Pareil pour la fin. Pour l'évolution, comment ont évolué les sentiments dans la conversation ? Est-ce qu'ils ont fluctué, est-ce qu'ils ont stagné ? Ces évolutions sont représentées par le schéma suivant :



Enfin vous devez annoter les émotions sous la forme de la frustration et de la satisfaction. Comme pour la valence, est-ce que la personne était frustrée, satisfaite, neutre au début et à la fin de la conversation. Les évolutions possibles sont les mêmes que celle de la valence.

Votre résultat devrait ressembler à quelque chose comme ça :

1	2	3	VALENCE			FRUSTRATION-SATISFACTION		
			début	évolution	fin	début	évolution	fin
	NOM	Genre	neutre	stagne	neutre	satisfaite	fluctue	frustrée

10. Vous avez terminé d'annoter un fichier !

Livraison attendue

- Le fichier EXCEL de chaque annotateur
- Le dossier Annotation_ok de chaque annotateur. Il faudra le renommer pour qu'on voie les initiales de l'annotateur apparaître. Par exemple : Annotation_ok_MaMa

8.3 Transcription d'une conversation

001-1529142237.966-in 1 inter-segment-gap 0.000 5.520 <o,f0,>
001-1529142237.966-in 1 001-1529142237.966-in-USAGER 5.520 7.130 <o,f0,unknown>
euh oui bonjour
001-1529142237.966-in 1 001-1529142237.966-in-USAGER 7.130 9.646 <o,f0,unknown>
euh je suis l'agence immobilière
001-1529142237.966-in 1 001-1529142237.966-in-USAGER 9.646 12.400 <o,f0,unknown>
[autre.marque]
001-1529142237.966-in 1 inter-segment-gap 12.400 12.760 <o,f0,>
001-1529142237.966-in 1 001-1529142237.966-in-USAGER 12.760 13.832 <o,f0,unknown>
euh au
001-1529142237.966-in 1 001-1529142237.966-in-USAGER 13.832 18.250 <o,f0,unknown>
[num.tel]
001-1529142237.966-in 1 inter-segment-gap 18.250 18.940 <o,f0,>
001-1529142237.966-in 1 001-1529142237.966-in-USAGER 18.940 30.110 <o,f0,unknown>
et il semblerait qu'il y ait un problème avec euh mes annonces parce que j'ai été avertie
par des clients que mon annonce avait été désactivée une de mes annonces donc j'ai pas
eu le temps de vérifier les autres
001-1529142237.966-in 1 inter-segment-gap 30.110 30.610 <o,f0,>
001-1529142237.966-in 1 001-1529142237.966-in-USAGER 30.610 36.260 <o,f0,unknown>
euh c'est une annonce qui était qui a été publiée euh vendredi je crois
001-1529142237.966-in 1 001-1529142237.966-in-USAGER 36.260 39.142 <o,f0,unknown>
euh un appartement à soixante dans le
001-1529142237.966-in 1 001-1529142237.966-in-USAGER 39.142 41.690 <o,f0,unknown>
[loc.cp]
001-1529142237.966-in 1 001-1529142237.966-in-USAGER 41.690 44.900 <o,f0,unknown>
euh au prix de sept cent quatre-vingt mille euros
001-1529142237.966-in 1 inter-segment-gap 44.900 46.720 <o,f0,>
001-1529142237.966-in 1 001-1529142237.966-in-USAGER 46.720 52.790 <o,f0,unknown>
sept cent quatre-vingt mille et qui aurait disparu euh qui a été désactivée euh voilà
001-1529142237.966-in 1 inter-segment-gap 52.790 53.380 <o,f0,>
001-1529142237.966-in 1 001-1529142237.966-in-USAGER 53.380 56.600 <o,f0,unknown>
effectivement je viens de vérifier elle n'est plus sur le site

001-1529142237.966-in 1 inter-segment-gap 56.600 58.270 <o,f0,>
001-1529142237.966-in 1 001-1529142237.966-in-USAGER 58.270 66.050 <o,f0,unknown>
alors je sais pas si les autres mes autres annonces ont été désactivées aussi euh voilà si
vous pourriez avoir la gentillesse de vérifier s'il vous plaît
001-1529142237.966-in 1 inter-segment-gap 66.050 68.080 <o,f0,>
001-1529142237.966-in 1 001-1529142237.966-in-USAGER 68.080 69.472 <o,f0,unknown>
oui je viens de vous le donner
001-1529142237.966-in 1 001-1529142237.966-in-USAGER 69.472 74.510 <o,f0,unknown>
[num.tel]
001-1529142237.966-in 1 inter-segment-gap 74.510 76.390 <o,f0,>
001-1529142237.966-in 1 001-1529142237.966-in-USAGER 76.390 77.381 <o,f0,unknown>
[pers.nom]
001-1529142237.966-in 1 001-1529142237.966-in-USAGER 77.381 79.580 <o,f0,unknown>
[pers.nom.epl]
001-1529142237.966-in 1 inter-segment-gap 79.580 102.580 <o,f0,>
001-1529142237.966-in 1 001-1529142237.966-in-USAGER 102.580 103.820 <o,f0,unknown>
oui c'est ça
001-1529142237.966-in 1 001-1529142237.966-in-USAGER 103.820 105.320 <o,f0,unknown>
oui oui oui
001-1529142237.966-in 1 inter-segment-gap 105.320 134.440 <o,f0,>
001-1529142237.966-in 1 001-1529142237.966-in-USAGER 134.440 143.230 <o,f0,unknown>
alors et puis je vois que une autre annonce a été supprimée là je suis en train de regarder
ça a été supprimé aussi bon ben je crois que toutes mes annonces ont été supprimées
001-1529142237.966-in 1 inter-segment-gap 143.230 143.500 <o,f0,>
001-1529142237.966-in 1 001-1529142237.966-in-USAGER 143.500 145.420 <o,f0,unknown>
j'en avais déjà pas beaucoup
001-1529142237.966-in 1 inter-segment-gap 145.420 146.650 <o,f0,>
001-1529142237.966-in 1 001-1529142237.966-in-USAGER 146.650 147.430 <o,f0,unknown>
oui
001-1529142237.966-in 1 inter-segment-gap 147.430 149.200 <o,f0,>
001-1529142237.966-in 1 001-1529142237.966-in-USAGER 149.200 151.441 <o,f0,unknown>
alors que sur mon site sur
001-1529142237.966-in 1 001-1529142237.966-in-USAGER 151.441 152.347 <o,f0,unknown>
[autre.produit]

001-1529142237.966-in 1 001-1529142237.966-in-USAGER 152.347 153.610 <o,f0,unknown>
quand je regarde

001-1529142237.966-in 1 001-1529142237.966-in-USAGER 153.610 163.160 <o,f0,unknown>
mon compte mon portefeuille elles y sont elles sont marquées validées on diffusion on euh
voilà comme si de rien n'était alors que moi je peux pas le savoir qu'elles ont été retirées

001-1529142237.966-in 1 001-1529142237.966-in-USAGER 163.160 164.882 <o,f0,unknown>
là je viens de vérifier celle de

001-1529142237.966-in 1 001-1529142237.966-in-USAGER 164.882 166.155 <o,f0,unknown>
[loc.ville]

001-1529142237.966-in 1 001-1529142237.966-in-USAGER 166.155 167.950 <o,f0,unknown>
aussi elle est retirée

001-1529142237.966-in 1 inter-segment-gap 167.950 168.190 <o,f0,>

001-1529142237.966-in 1 001-1529142237.966-in-USAGER 168.190 171.920 <o,f0,unknown>
bon les autres j'ai pas vérifié mais j'imagine que c'est pareil

001-1529142237.966-in 1 inter-segment-gap 171.920 177.250 <o,f0,>

001-1529142237.966-in 1 001-1529142237.966-in-USAGER 177.250 178.390 <o,f0,unknown>
voilà

001-1529142237.966-in 1 001-1529142237.966-in-USAGER 178.390 179.140 <o,f0,unknown>
voilà

001-1529142237.966-in 1 inter-segment-gap 179.140 180.820 <o,f0,>

001-1529142237.966-in 1 001-1529142237.966-in-USAGER 180.820 185.480 <o,f0,unknown>
vous voyez qu'elle a été qu'elle a été désactivée hein vous le voyez ça

001-1529142237.966-in 1 001-1529142237.966-in-USAGER 185.480 188.600 <o,f0,unknown>
donc je sais pas si vous pouvez oui vous devez pouvoir le voir

001-1529142237.966-in 1 inter-segment-gap 188.600 231.400 <o,f0,>

001-1529142237.966-in 1 001-1529142237.966-in-USAGER 231.400 234.247 <o,f0,unknown>
voilà voilà et c'est la même chose pour euh celle de

001-1529142237.966-in 1 001-1529142237.966-in-USAGER 234.247 235.219 <o,f0,unknown>
[loc.ville]

001-1529142237.966-in 1 001-1529142237.966-in-USAGER 235.219 241.300 <o,f0,unknown>
c'est pareil j'ai vérifié elle y est pas donc j'ai pas le temps de vérifier les autres mais j'ai
je pense que tout a été supprimé

001-1529142237.966-in 1 001-1529142237.966-in-USAGER 241.300 245.950 <o,f0,unknown>
est-ce que vous pourriez faire le nécessaire pour qu'on les remette tout de suite s'il vous

plaît

001-1529142237.966-in 1 001-1529142237.966-in-USAGER 245.950 248.330 <o,f0,unknown>

c'est très important parce que

001-1529142237.966-in 1 inter-segment-gap 248.330 248.770 <o,f0,>

001-1529142237.966-in 1 001-1529142237.966-in-USAGER 248.770 258.990 <o,f0,unknown>

nous nous sommes une toute petite agence nous n'avions nous n'avions très peu de d'annonces jusqu'à présent et là on arrive à avoir enfin des des ventes et c'est juste au moment où

001-1529142237.966-in 1 001-1529142237.966-in-USAGER 258.990 259.478 <o,f0,unknown>

[autre.marque]

001-1529142237.966-in 1 001-1529142237.966-in-USAGER 259.478 261.700 <o,f0,unknown>

nous supprime nos annonces on sait pas pourquoi

001-1529142237.966-in 1 001-1529142237.966-in-USAGER 261.700 267.040 <o,f0,unknown>

donc voilà s'il vous plaît euh c'est vraiment important qu'on nous les remette tout de suite

001-1529142237.966-in 1 inter-segment-gap 267.040 267.670 <o,f0,>

001-1529142237.966-in 1 001-1529142237.966-in-USAGER 267.670 270.790 <o,f0,unknown>

est-ce que vous pouvez faire ça ou est-ce que

001-1529142237.966-in 1 inter-segment-gap 270.790 272.470 <o,f0,>

001-1529142237.966-in 1 001-1529142237.966-in-USAGER 272.470 273.700 <o,f0,unknown>

comment ça

001-1529142237.966-in 1 inter-segment-gap 273.700 274.510 <o,f0,>

001-1529142237.966-in 1 001-1529142237.966-in-USAGER 274.510 276.950 <o,f0,unknown>

y'a pas de contrat actif 001-1529142237.966-in 1 inter-segment-gap 276.950 278.320 <o,f0,>

001-1529142237.966-in 1 001-1529142237.966-in-USAGER 278.320 285.770 <o,f0,unknown>

c'est pas possible mais moi je paye euh alors il faut me faire remettre mon contrat immédiatement s'il vous plaît moi je paye pour ça

001-1529142237.966-in 1 inter-segment-gap 285.770 286.000 <o,f0,>

001-1529142237.966-in 1 001-1529142237.966-in-USAGER 286.000 288.400 <o,f0,unknown>

je paye mes factures quand même

001-1529142237.966-in 1 inter-segment-gap 288.400 292.630 <o,f0,>

001-1529142237.966-in 1 001-1529142237.966-in-USAGER 292.630 295.366 <o,f0,unknown>

ah c'est pas vrai 001-1529142237.966-in 1 inter-segment-gap 295.366 297.790 <o,f0,>

001-1529142237.966-in 1 001-1529142237.966-in-USAGER 297.790 299.200 <o,f0,unknown>

oh non

001-1529142237.966-in 1 inter-segment-gap 299.200 304.750 <o,f0,>
001-1529142237.966-in 1 001-1529142237.966-in-USAGER 304.750 314.770 <o,f0,unknown>
oh là là mais non mais là je vais avoir tout le week-end de samedi et dimanche euh mes
annonces vont pas être vues c'est scandaleux moi je suis furieuse là c'est pas normal je
vais
001-1529142237.966-in 1 inter-segment-gap 314.770 315.010 <o,f0,>
001-1529142237.966-in 1 001-1529142237.966-in-USAGER 315.010 317.810 <o,f0,unknown>
c'est c'est insupportable franchement
001-1529142237.966-in 1 inter-segment-gap 317.810 324.670 <o,f0,>
001-1529142237.966-in 1 001-1529142237.966-in-USAGER 324.670 327.280 <o,f0,unknown>
non absolument pas non non non
001-1529142237.966-in 1 inter-segment-gap 327.280 329.530 <o,f0,>
001-1529142237.966-in 1 001-1529142237.966-in-USAGER 329.530 334.790 <o,f0,unknown>
oh là là oh excusez-moi je le prends un peu mal mais là euh c'est normal hein
001-1529142237.966-in 1 inter-segment-gap 334.790 336.610 <o,f0,>
001-1529142237.966-in 1 001-1529142237.966-in-USAGER 336.610 337.340 <o,f0,unknown>
oui
001-1529142237.966-in 1 inter-segment-gap 337.340 339.790 <o,f0,>
001-1529142237.966-in 1 001-1529142237.966-in-USAGER 339.790 345.220 <o,f0,unknown>
on peut on peut joindre personne là chez eux le samedi y'a pas de possibilité de joindre
001-1529142237.966-in 1 inter-segment-gap 345.220 348.460 <o,f0,>
001-1529142237.966-in 1 001-1529142237.966-in-USAGER 348.460 350.270 <o,f0,unknown>
non mais c'est scandaleux
001-1529142237.966-in 1 inter-segment-gap 350.270 352.690 <o,f0,>
001-1529142237.966-in 1 001-1529142237.966-in-USAGER 352.690 360.560 <o,f0,unknown>
alors est-ce que vous pourriez me me rappeler le le nom et le numéro de téléphone de du
commercial qui s'occupe de moi s'il vous plaît
001-1529142237.966-in 1 inter-segment-gap 360.560 367.060 <o,f0,>
001-1529142237.966-in 1 001-1529142237.966-in-USAGER 367.060 370.070 <o,f0,unknown>
ben comment je fais pour le joindre moi lundi
001-1529142237.966-in 1 inter-segment-gap 370.070 378.850 <o,f0,>
001-1529142237.966-in 1 001-1529142237.966-in-USAGER 378.850 381.040 <o,f0,unknown>
alors il s'appelle comment
001-1529142237.966-in 1 inter-segment-gap 381.040 382.930 <o,f0,>

001-1529142237.966-in 1 001-1529142237.966-in-USAGER 382.930 384.076 <o,f0,unknown>
[pers.pre]
001-1529142237.966-in 1 001-1529142237.966-in-USAGER 384.076 384.790 <o,f0,unknown>
comment
001-1529142237.966-in 1 inter-segment-gap 384.790 385.420 <o,f0,>
001-1529142237.966-in 1 001-1529142237.966-in-USAGER 385.420 386.263 <o,f0,unknown>
[pers.nom]
001-1529142237.966-in 1 001-1529142237.966-in-USAGER 386.263 387.830 <o,f0,unknown>
[pers.nom.epl]
001-1529142237.966-in 1 inter-segment-gap 387.830 389.680 <o,f0,>
001-1529142237.966-in 1 001-1529142237.966-in-USAGER 389.680 390.500 <o,f0,unknown>
oui
001-1529142237.966-in 1 inter-segment-gap 390.500 391.930 <o,f0,>
001-1529142237.966-in 1 001-1529142237.966-in-USAGER 391.930 392.810 <o,f0,unknown>
oui
001-1529142237.966-in 1 inter-segment-gap 392.810 393.880 <o,f0,>
001-1529142237.966-in 1 001-1529142237.966-in-USAGER 393.880 394.720 <o,f0,unknown>
oui
001-1529142237.966-in 1 inter-segment-gap 394.720 396.340 <o,f0,>
001-1529142237.966-in 1 001-1529142237.966-in-USAGER 396.340 397.280 <o,f0,unknown>
d'accord
001-1529142237.966-in 1 inter-segment-gap 397.280 399.490 <o,f0,>
001-1529142237.966-in 1 001-1529142237.966-in-USAGER 399.490 405.260 <o,f0,unknown>
très bien donc y'a vraiment aucun aucune possibilité pour vous de de me remettre mes
annonces là
001-1529142237.966-in 1 inter-segment-gap 405.260 405.520 <o,f0,>
001-1529142237.966-in 1 001-1529142237.966-in-USAGER 405.520 408.890 <o,f0,unknown>
ils ont résilié mon compte je n'ai même pas été prévenue
001-1529142237.966-in 1 inter-segment-gap 408.890 410.140 <o,f0,>
001-1529142237.966-in 1 001-1529142237.966-in-USAGER 410.140 415.130 <o,f0,unknown>
vous vous pouvez voir que je paye mes factures là j'ai pas d'impayé j'ai rien du tout
001-1529142237.966-in 1 inter-segment-gap 415.130 427.690 <o,f0,>
001-1529142237.966-in 1 001-1529142237.966-in-USAGER 427.690 428.390 <o,f0,unknown>
bon

001-1529142237.966-in 1 inter-segment-gap 428.390 430.180 <o,f0,>
001-1529142237.966-in 1 001-1529142237.966-in-USAGER 430.180 433.300 <o,f0,unknown>
oui s'il vous plaît c'est gentil merci beaucoup
001-1529142237.966-in 1 001-1529142237.966-in-USAGER 433.300 444.712 <o,f0,unknown>
merci beaucoup madame je vous souhaite une bonne journée merci au revoir

8.4 End User Licence Agreement

AlloSat End User License Agreement

By signing this document, the user, he or she who will make use of the corpus, agrees to the following terms.

1. Disclaimer

Despite all the care taken to anonymize speakers, this resource may include indirectly identifying personal data.

The recorded exchanges are likely to contain data that could infringe the protection of private life or carry an appreciation or value judgment on a natural person named or identifiable, or revealing the behavior of a person under conditions likely to harm it.

As such, this corpus can be used for analytical purposes in the context of historical or scientific research, but in no case may it be intended for other uses giving rise, for example, to public dissemination or to the search for identification of the persons concerned.

2. Commercial use

The user may not use the database for any non-academic purpose. Non-academic purposes include, but are not limited to:

- proving the efficiency of commercial systems
- training or testing of commercial systems
- using audio or annotations from the corpus in advertisements
- selling data from the corpus
- creating military applications
- developing governmental systems used in public spaces

3. Responsibility

This document must be signed by a person with a **permanent position at an academic institute** (the signee). Up to five other researchers affiliated with the same institute for whom the signee is responsible may be named at the end of this document which will allow them to work with this dataset.

4. Distribution

The user may not distribute the audio nor the transcription in **any way**. Annotations and features extracted from the corpus (that cannot be retro-engineered) can be used with the exclusive purpose of clarifying academic publications or presentations. Note that publications will have to comply with the terms stated in article 6.

5. Access

The user may only receive the corpus after this End User License Agreement (EULA) has been signed and returned to the LIUM (Laboratoire d'informatique de l'Université du Mans).

The signed EULA should be returned in digital format by email to one of the authors.

The user may not grant anyone access to the database by giving out their version of the corpus.

6. Publications

Publications include not only papers, but also presentations for conferences or educational purposes. The user may only use annotations and features extracted from the corpus (that cannot be retro-engineered).

All documents and papers that report on research that use any of the AlloSat corpus will acknowledge this as follows:

“(Portions of) the research in this paper uses the AlloSat corpus collected by AlloMedia. The use of the corpus is limited to the agreement of AlloMedia.”

These documents and papers must refer to the original paper by citing

“Manon Macary, Marie Tahon, Yannick Estève, Anthony Rousseau. *AlloSat: A New Call Center French Corpus for Satisfaction and Frustration Analysis*. In Proc of. Language Resources and Evaluation Conference, LREC 2020, May 2020, Marseille, France.”¹.

The user will send a copy of any document or papers that reports on research that uses the AlloSat corpus to one of the authors.

7. Academic research

The user may **only** use the database for academic research.

8. Warranty

The corpus comes without any warranty. The LIUM and AlloMedia can not be held accountable for any damage (physical, financial or otherwise) caused by the use of the corpus.

9. Misuse

If at any point, the administrators of the AlloSat corpus have a reasonable doubt that the user does not act in accordance to this EULA, he/she will be notified and could be prosecuted.

User:

User's Affiliation:

User's address:

User's e-mail:

Additional Researcher 1

Additional Researcher 2

Additional Researcher 3

Additional Researcher 4

Additional Researcher 5

Signature:

Date/Place:

¹ <https://hal.archives-ouvertes.fr/hal-02506086>

8.5 Intervalles de confiance statistiques des scores CCC

Comme nous l'avons expliqué précédemment, le score CCC est devenu un score de référence pour l'évaluation des systèmes de reconnaissance des émotions continues. Cependant, nous n'avons pas trouvé d'expérience en reconnaissance des émotions continues mentionnant l'utilisation d'un intervalle de confiance ou de score de confiance associé à cette mesure.

Nous avons donc cherché à mettre en place un intervalle de confiance statistique. Notre travail se plaçant dans un contexte industriel, il est important de vérifier la pertinence des scores de nos modèles. En effet, comme notre corpus est assez petit, nous avons un nombre limité d'échantillons de test utilisés pour le calcul du score final et une petite variation dans le choix de ces échantillons peut induire une variation importante au niveau du résultat.

À partir des travaux de Liao et al. [LL00] et McBride [B05], nous avons mis en place un intervalle de confiance du CCC. Nous repartons de la définition du CCC, donnée par l'équation 8.1.

$$CCC = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2 + \epsilon} \quad (8.1)$$

Il s'agit d'un coefficient calculé entre deux vecteurs x et y par exemple, x étant la séquence prédite et y la séquence de référence. À partir de ces deux séquences, on peut calculer :

— l'écart type :

$$\sigma_x = \frac{1}{N} \sum_i (x_i - \mu_x)^2 \quad (8.2)$$

— la covariance :

$$\sigma_{xy} = \frac{1}{N} \sum_i (x_i - \mu_x)(y_i - \mu_y) \quad (8.3)$$

— le coefficient de corrélation :

$$\rho = \frac{\sigma_{xy}}{\sigma_x\sigma_y} \quad (8.4)$$

D'après McBride [B05], le CCC ne suit pas une loi normale. En effet, sa distribution est fortement asymétrique, notamment dans les cas où nous avons un coefficient de corrélation proche de ± 1 . Ainsi, on ne peut pas utiliser les intervalles de confiance classiques

Scores CCC	Intervalles de confiance associé
0.7350	[0.7324 ; 0.7375]
0.7587	[0.7563 ; 0.761]
0.8054	[0.8035 ; 0.8074]
0.8524	[0.8509 ; 0.854]
0.8971	[0.8960 ; 0.8982]

TABLE 8.1 – Intervalles de confiance pour les scores CCC obtenus sur le sous-ensemble de Dev. Ils sont calculés sur différentes prédictions d’expérimentations sur l’initialisation des poids du réseau de neurones.

sur le CCC. Afin de pallier ce problème, on utilise la transformation de Fisher. Cette transformation consiste à prendre la tangente hyperbolique inverse du coefficient de corrélation. L’estimateur \hat{Z} du CCC est alors défini par l’équation 8.5 où $\hat{\rho}_c$ représente le CCC.

$$\hat{Z} = \tanh^{-1}(\hat{\rho}_c) = \frac{1}{2} \ln \left(\frac{1 + \hat{\rho}_c}{1 - \hat{\rho}_c} \right) \quad (8.5)$$

L’écart type de cet estimateur est donné par l’équation 8.6.

$$\sigma_{\hat{Z}}^2 = \frac{\frac{(1 - \rho^2)\hat{\rho}_c^2}{(1 - \hat{\rho}_c^2)\rho^2} + \frac{2\hat{\rho}_c^3(1 - \hat{\rho}_c)u^2}{\rho(1 - \hat{\rho}_c^2)^2} - \frac{\hat{\rho}_c^4 u^4}{2\rho^2(1 - \hat{\rho}_c^2)^2}}{N - 2} \quad (8.6)$$

où $u = \frac{\mu_x - \mu_y}{\sigma_x \sigma_y}$.

Ce qui nous permet de déduire un intervalle de confiance à 90% pour le score de CCC selon l’équation 8.7.

$$[\tanh(\hat{Z} - 1.64\sigma_{\hat{Z}}); \tanh(\hat{Z} + 1.64\sigma_{\hat{Z}})] \quad (8.7)$$

Afin de se rendre compte de cet intervalle de confiance, nous avons donné quelques scores et leur intervalle associé dans le tableau 8.1.

Nous pouvons observer que l’intervalle de confiance est relativement faible. En effet, il ne concerne que les centièmes. De même que pour l’intervalle de confiance pour une variable aléatoire normale, on remarque également que plus le score est élevé, plus on a un faible intervalle. Par contre, on peut également constater que cet intervalle n’est pas symétrique.

Dans toutes les expériences suivantes, une différence de performance sera jugée signi-

ficative uniquement si leur intervalle de confiance ne se chevauchent pas.

BIBLIOGRAPHIE

- [Abd+14] O. ABDEL-HAMID et al., « Convolutional neural networks for speech recognition », in : *IEEE ACM Transactions on audio, speech, and language processing* 22.10 (2014), p. 1533-1545.
- [Adr08] R. ADRAIN, « Research concerning the probabilities of the errors which happen in making observations », in : *The Analyst, or Mathematical Museum* 1 (1808), p. 93-109.
- [Ajm+02] J. AJMERA et al., « Unknown-multiple speaker clustering using HMM », in : *Proc. of INTERSPEECH*, Denver, Colorado, USA, 2002, p. 573-576.
- [AM05] Legendre A.M., *Nouvelles méthodes pour la détermination des orbites des comètes*, Firmin Didot, 1805.
- [ANG16] EAS ALZQHOUL, BBT NAIR et BJ GUILLEMIN, « Impact of Background Noise in Mobile Phone Networks on Forensic Voice Comparison », in : *Forensic, Legal & Investigative Sciences* 2.7 (2016), p. 1-9.
- [API07] M. S. AL MASUM, H. PRENDINGER et M. ISHIZUKA, « Emotion Sensitive News Agent : An Approach Towards User Centric Emotion Sensing from the News », in : *Proc. of International Conference on Web Intelligence*, Silicon Valley, California, USA, 2007, p. 614-620.
- [AR14] F. ALAM et G. RICCARDI, « Fusion of acoustic, linguistic and psycholinguistic features for Speaker Personality Traits recognition », in : *Proc. of ICASSP*, Florence, Italy, 2014, p. 955-959.
- [Arn60] M. B. ARNOLD, *Emotion and personality*. Columbia University Press, 1960.
- [Atr+10] P. ATREY et al., « Multimodal fusion for multimedia analysis : A survey », in : *Multimedia Syst.* 16.1 (2010), p. 345-379.
- [B05] McBride G. B., *A Proposal for Strength-of-Agreement Criteria for Lin's Concordance Correlation Coefficient*, rapp. tech., National Institute of Water & Atmospheric Research Ltd, 2005.

-
- [Bae+20] Alexei BAEVSKI et al., « wav2vec 2.0 : A framework for self-supervised learning of speech representations », in : *Advances in Neural Information Processing Systems* 33 (2020), p. 12449-12460.
- [Bah+16] D. BAHDANAU et al., « End-to-end attention-based large vocabulary speech recognition », in : *Proc. of ICASSP*, Shanghai, China, 2016, p. 4945-4949.
- [Bar34] P. BARD, « On emotional expression after decortication with some remarks on certain theoretical views : Part I. », in : *Psychological review* 41.4 (1934), p. 309-424.
- [BBN12] C. BUSSO, M. BULUT et S. NARAYANAN, « Toward effective automatic recognition systems of emotion in speech », in : *Social emotions in nature and artifact : emotions in human and human-computer interaction*, Oxford University Press, 2012, p. 110-127.
- [BDR08] Y. BENAÏBA, M. DIAB et P. ROSSO, « Arabic named entity recognition : An svm-based approach », in : *Proc. of the Arab International Conference on Information Technology (ACIT)*, Hammamet, Tunisia, 2008, p. 16-18.
- [Ben+03] Y. BENGIO et al., « A neural probabilistic language model », in : *The journal of machine learning research* 3 (2003), p. 1137-1155.
- [BGL14] M. BANSAL, K. GIMPEL et K. LIVESCU, « Tailoring Continuous Word Representations for Dependency Parsing », in : *Proc. of ACL*, Baltimore, Maryland, 2014, p. 809-815.
- [BL94] M. M. BRADLEY et P. J. LANG, « Measuring emotion : the self-assessment manikin and the semantic differential », in : *Journal of behavior therapy and experimental psychiatry* 25.1 (1994), p. 49-59.
- [BRS05] Baker B., Vogt R. et S. SRIDHARAN, « Gaussian Mixture Modeling of Broad Phonetic and Syllabic Events For Text Independent Speaker Verification », in : *Proc. of INTERSPEECH*, Lisbon, Portugal, 2005, p. 2429-2432.
- [BSF94] Y. BENGIO, P. SIMARD et P. FRASCONI, « Learning long-term dependencies with gradient descent is difficult », in : *IEEE Transactions on Neural Networks* 5.2 (1994), p. 157-166.
- [Bur+05] F. BURKHARDT et al., « A database of German emotional speech », in : *Proc. of INTERSPEECH*, Lisbon, Portugal, 2005, p. 1517-1520.

-
- [Bus+08] C. BUSO et al., « IEMOCAP : Interactive emotional dyadic motion capture database », in : *Language Resources and Evaluation* 42.335 (2008), p. 335-359.
- [C06] Bishop C., « Neural Networks : Network Training », in : *Pattern Recognition and Machine Learning*, Springer-Verlag New York, 2006, p. 236-241.
- [Cam+13] Erik CAMBRIA et al., « New Avenues in Opinion Mining and Sentiment Analysis », in : *IEEE Intelligent Systems* 28.2 (2013), p. 15-21, DOI : 10.1109/MIS.2013.30.
- [Can15] W. B. CANNON, *Bodily changes in pain, hunger, fear, and rage*, D. Appleton et company, 1915.
- [Can27] W. B. CANNON, « The James-Lange theory of emotions : A critical examination and an alternative theory », in : *The American journal of psychology* 39.1 (1927), p. 106-124.
- [CH67] T. COVER et P. HART, « Nearest neighbor pattern classification », in : *IEEE Transactions on Information Theory* 13.1 (1967), p. 21-27.
- [Cha07] F. R. CHAUMARTIN, « UPAR7 : A knowledge-based system for headline sentiment tagging », in : *Proc. of the International Workshop on Semantic Evaluations (SemEval-2007)*, Prague, Czech Republic, 2007, p. 422-425.
- [Cha95] L. C. CHARLAND, « Emotion as a natural kind : Towards a computational foundation for emotion theory », in : *Philosophical psychology* 8.1 (1995), p. 59-84.
- [Chi+18] C. CHIU et al., « State-of-the-Art Speech Recognition with Sequence-to-Sequence Models », in : *Proc. of ICASSP*, Calgary, Alberta, Canada, 2018, p. 4774-4778.
- [Chi+20] P. CHI et al., « Audio ALBERT : A Lite BERT for Self-supervised Learning of Audio Representation », in : *Pre-print on arXiv/2005.08575*, 2020.
- [Cho+14] K. CHO et al., « On the properties of neural machine translation : Encoder-Decoder approaches », in : *Proc. of Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST)*, Doha, Qatar, 2014, p. 103-111.

-
- [CL08] Z. CALLEJAS et R. LÓPEZ-CÓZAR, « On the Use of Kappa Coefficients to Measure the Reliability of the Annotation of Non-acted Emotions », in : *International Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems*, Irsee, Germany, 2008, p. 221-232.
- [Coc09] T. COCHRANE, « Eight dimensions for the emotions », in : *Social Science Information* 48.3 (2009), p. 379-420.
- [Cos94] J. COSNIER, *Psychologie des émotions et des sentiments*, Fenixx, 1994.
- [Cow+00] R. COWIE et al., « 'FEELTRACE' : An instrument for recording perceived emotion in real time », in : *ISCA tutorial and research workshop (ITRW) on speech and emotion*, Newcastle, Northern Ireland, United Kingdom, 2000, p. 19-24.
- [CR01] A. CHOTIMONGKOL et A. I. RUDNICKY, « N-best Speech Hypotheses Reordering Using Linear Regression », in : *Proc. of INTERSPEECH*, Aalborg, Denmark, 2001, p. 1829-1832.
- [CR33] W. B. CANNON et A. ROSENBLUETH, « Studies on conditions of activity in endocrine organs : XXIX. Sympathin E and Sympathin I », in : *American Journal of Physiology-Legacy Content* 104.3 (1933), p. 557-574.
- [CSR06] W. CAMPBELL, D. STURIM et D. REYNOLDS, « Support vector machines using GMM supervectors for speaker verification », in : *Signal Processing Letters, IEEE* 13.5 (2006), p. 308-311.
- [CV95] Corinna CORTES et Vladimir VAPNIK, « Support-vector networks », in : *Machine learning* 20.3 (1995), p. 273-297.
- [Dam99] Antonio R DAMASIO, *Sentiment même de soi (Le) : Corps, émotions, conscience*, Odile Jacob, 1999.
- [Dan02] R. DANTZER, *Cerveau et émotions*, Presses Universitaires de France, 2002, p. 69-90.
- [Dar72] C. DARWIN, *The expression of the emotions in man and animals*, J. Murray, London, 1872.
- [Dev+19] J. DEVLIN et al., « 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, p. 4171-4186.

-
- [DHS11] J. DUCHI, E. HAZAN et Y. SINGER, « Adaptive Subgradient Methods for Online Learning and Stochastic Optimization », in : *Journal of Machine Learning Research* 12 (2011), p. 2121-2159.
- [DK16] S.F. DODGE et L.J. KARAM, « Understanding how image quality affects deep neural networks », in : *Proc. of International Conference on Quality of Multimedia Experience (QoMEX)*, Lisbon, Portugal, 2016, p. 1-6.
- [DM18] X. DONG et G. de MELO, « A Helping Hand : Transfer Learning for Deep Sentiment Analysis », in : *Proc. of ACL*, Melbourne, Australia, 2018, p. 2524-2534.
- [Doh+04] M. DOHEN et al., « Visual perception of contrastive focus in reiterant French speech », in : *Speech Communication* 44.1 (2004), p. 155-172.
- [Duf34] E. DUFFY, « Emotion : an example of the need for reorientation in psychology. », in : *Psychological Review* 41.2 (1934), p. 184-198.
- [Duf41] E. DUFFY, « An explanation of “emotional” phenomena without the use of the concept “emotion” », in : *The Journal of General Psychology* 25.2 (1941), p. 283-293.
- [EF78] P. EKMAN et W. V. FRIESEN, *Facial Action Coding System*, Consulting Psychologists Press, 1978.
- [Ekm99] P. EKMAN, « Basic Emotions », in : *Handbook of Cognition and Emotion*, Wiley, New-York, 1999, p. 301-320.
- [ELF83] P. EKMAN, R. W. LEVENSON et W. V. FRIESEN, « Autonomic nervous system activity distinguishes among emotions », in : *Science* 221.4616 (1983), p. 1208-1210.
- [Eng+97] I. S. ENGBERG et al., « Design, recording and verification of a danish emotional speech database », in : *EUROSPEECH*, Rhodes, Greece, 1997, p. 1695-1698.
- [Esh+11] Iris ESHKOL-TARAVELLA et al., « Un grand corpus oral “ disponible ” : le corpus d’Orléans 1 1968-2012 », in : *Revue TAL*, Ressources Linguistiques Libres 53.2 (2011), p. 17-46, URL : <https://halshs.archives-ouvertes.fr/halshs-01163053>.

-
- [EWS09] F. EYBEN, M. WOLLMER et B. SCHULLER, « OpenEAR - Introducing the Munich open-source emotion and affect recognition toolkit », in : *Proc. of Affective Computing and Intelligent Interaction and Workshops, (ACII)*, Amsterdam, The Netherlands, 2009, p. 1-6.
- [EWS10] F. EYBEN, M. WÖLLMER et 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, p. 1459-1462.
- [Eyb+16] F. EYBEN et al., « The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing », in : *IEEE transactions on affective computing* 7.2 (2016), p. 190-202.
- [Fel06] L. FELDMAN BARRETT, « Are Emotions Natural Kinds ? », in : *Perspectives on Psychological Science* 1.1 (2006), p. 28-58.
- [Fér+15] R. FÉR et al., « Multilingual bottleneck features for language recognition », in : *Proc. of INTERSPEECH*, Dresden, Germany, 2015, p. 389-393.
- [FH51] E. FIX et J.L. HODGES, *Discriminatory Analysis, Nonparametric Discrimination : Consistency Properties*, rapp. tech., USAF School of Aviation Medicine, Randolph Field., 1951.
- [FR09] M. A. FATTAH et F. REN, « GA, MR, FFNN, PNN and GMM based models for automatic text summarization », in : *Computer Speech & Language* 23.1 (2009), p. 126-144.
- [Fri87] N. H. FRIJDA, « Emotion, cognitive structure, and action tendency », in : *Cognition and Emotion* 1.2 (1987), p. 115-143.
- [Gar+08] M. GARNIER-RIZET et al., « CallSurf : Automatic Transcription, Indexing and Structuration of Call Center Conversational Speech for Knowledge Extraction and Query by Content », in : *Proc. of Language Resources and Evaluation Conference (LREC)*, Marrakech, Morocco, 2008, p. 2623-2628.
- [Gau09] C.F. GAUSS, *Theoria Motus Corporum Coelestium in sectionibus conicis solem ambientium*, Hamburg : Friedrich Perthes et I.H. Besser, 1809.

-
- [GB10] X. GLOROT et Y. BENGIO, « Understanding the difficulty of training deep feedforward neural networks », in : *Proc. of the International Conference on Artificial Intelligence and Statistics*, Chia Laguna Resort, Sardinia, Italy, 2010, p. 249-256.
- [GBC16] I. GOODFELLOW, Y. BENGIO et A. COURVILLE, *Deep learning*, MIT press, 2016.
- [Gen+14] M. GENDRON et al., « Perceptions of emotion from facial expressions are not culturally universal : evidence from a remote culture », in : *Emotion* 14.2 (2014), p. 251-262.
- [GFK05] T. GANCHEV, N. FAKOTAKIS et G. KOKKINAKIS, « Comparative evaluation of various MFCC implementations on the speaker verification task », in : *Proc. of Conference on Speech and Computer (SPECOM)*, Patras, Greece, 2005, p. 191-194.
- [Gha17] S. GHANNAY, « Etude sur les representations continues de mots appliquees à la detection automatique des erreurs de reconnaissance de la parole », thèse de doct., 2017.
- [Gir+13] T. GIRAUD et al., « Multimodal Expressions of Stress during a Public Speaking Task : Collection, Annotation and Global Analyses », in : *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 2013, p. 417-422.
- [Gir14] J. M. GIRARD, « CARMA : Software for continuous affect rating and media annotation », in : *Journal of open research software* 2.1 (2014), p. 1-11.
- [Gor03] P. GORMAN, *Motivation and Emotion*, Routledge, 2003.
- [Gra+06] A. GRAVES et al., « Connectionist Temporal Classification : Labelling Unsegmented Sequence Data with Recurrent Neural Networks », in : *Proc. of International Conference on Machine Learning*, Pittsburgh, Pennsylvania, USA, 2006, p. 369-376.
- [GS13] Hatice GUNES et Björn SCHULLER, « Categorical and dimensional affect analysis in continuous input : Current trends and future directions », in : *Image and Vision Computing* 31.2 (2013), p. 120-136.

-
- [Gu+15] S. GU et al., « Differentiation of primary emotions through neuromodulators : review of literature », in : *International Journal of Neurology Research* 1.2 (2015), p. 43-50.
- [GY08] M. GALES et S. YOUNG, *Application of Hidden Markov Models in Speech Recognition*, Now Foundations et Trends, 2008.
- [Han+18] Jing HAN et al., « Bags in Bag : Generating Context-Aware Bags for Tracking Emotions from Speech », in : *INTERSPEECH*, 2018.
- [HB97] J. HANSEN et S. E. BOU-GHAZALE, « Getting started with SUSAS : a speech under simulated and actual stress database », in : *EUROSPEECH*, Rhodes, Greece, 1997, p. 1-12.
- [HBT96] Jianying HU, M.K. BROWN et W. TURIN, « HMM based online handwriting recognition », in : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 18.10 (1996), p. 1039-1045.
- [Hin+12] G. HINTON et al., « Deep Neural Networks for Acoustic Modeling in Speech Recognition : The Shared Views of Four Research Groups », in : *IEEE Signal Processing Magazine* 29.6 (2012), p. 82-97.
- [HLS07] J. HANCOCK, C. LANDRIGAN et C. SILVER, « Expressing emotion in text-based communication », in : *Proc. of Conference on Human Factors in Computing Systems*, San Jose, California, USA, 2007, p. 929-932.
- [Hoz+02] V. HOZJAN et al., « Interface Databases : Design and Collection of a Multilingual Emotional Speech Database », in : *Proc. of Language Resources and Evaluation Conference (LREC)*, Las Palmas, Canary Islands, Spain, 2002, p. 2024-2028.
- [HS97] S. HOCHREITER et J. SCHMIDHUBER, « Long Short-Term Memory », in : *Neural Computation* 9.8 (1997), p. 1735-1780.
- [Hsu+21] Wei-Ning HSU et al., « Hubert : Self-supervised speech representation learning by masked prediction of hidden units », in : *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), p. 3451-3460.
- [Hu19] D. HU, « An Introductory Survey on Attention Mechanisms in NLP Problems », in : *Intelligent Systems and Applications*, London, United Kingdom, 2019, p. 432-448.

-
- [Hua+15] Z. HUANG et al., « An Investigation of Annotation Delay Compensation and Output-Associative Fusion for Multimodal Continuous Emotion Prediction », in : *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, Brisbane, Australia, 2015, p. 41-48.
- [Hua+17] J. HUANG et al., « Continuous Multimodal Emotion Prediction Based on Long Short Term Memory Recurrent Neural Network », in : *Proc. of the Audio/Visual Emotion Challenge and Workshop*, Mountain View, United States, 2017, p. 11-18.
- [Hua+18] J. HUANG et al., « Multimodal Continuous Emotion Recognition with Data Augmentation Using Recurrent Neural Networks », in : *Proc. of the Audio/Visual Emotion Challenge and Workshop*, Beijing, China, 2018, p. 3-13.
- [HYT14] K. HAN, D. YU et I. TASHEV, « Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine », in : *Proc. of INTER-SPEECH*, Singapore, 2014, p. 223-227.
- [Iza07] C. E. IZARD, « Basic Emotions, Natural Kinds, Emotion Schemas, and a New Paradigm », in : *Perspectives on Psychological Science* 2.3 (2007), p. 260-280.
- [Jac+16] R. E. JACK et al., « Four not six : Revealing culturally common facial expressions of emotion », in : *Journal of Experimental Psychology : General* 145.6 (2016), p. 708-730.
- [Jam84] W. JAMES, « What is emotion ? », in : *Century psychology series*, W. Dennis, 1884, p. 290-303.
- [JMC18] S. JING, X. MAO et L. CHEN, « Prominence features : Effective emotional features for speech emotion recognition », in : *Digital Signal Processing* 72 (2018), p. 216-231.
- [Jor86] Michael I. JORDAN, « Attractor Dynamics and Parallelism in a Connectionist Sequential Machine », in : *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*, Amherst, Massachusetts, USA, 1986, p. 531-546.
- [Kah+20] J. KAHN et al., « Libri-Light : A Benchmark for ASR with Limited or No Supervision », in : *Proc. of ICASSP*, Virtual Conference, 2020, p. 7669-7673.

-
- [KK81] P. R. KLEINGINNA et A. M. KLEINGINNA, « A categorized list of emotion definitions with suggestions for a consensual definition », in : *Motivation and Emotion* 5.3 (1981), p. 45-379.
- [KLG17] H. KAMPER, K. LIVESCU et S. GOLDWATER, *An embedded segmental K-means model for unsupervised segmentation and clustering of speech*, Okinawa, Japan, 2017.
- [Kos+19] J. KOSSAIFI et al., « SEWA DB : A Rich Database for Audio-Visual Emotion and Sentiment Research in the Wild », in : *IEEE transactions on pattern analysis and machine intelligence* (2019), p. 1-1.
- [KR18] Taku KUDO et John RICHARDSON, « SentencePiece : A simple and language independent subword tokenizer and detokenizer for Neural Text Processing », in : *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, Brussels, Belgium, 2018, p. 66-71.
- [Lai74] J. D. LAIRD, « Self-attribution of emotion : The effects of expressive behavior on the quality of emotional experience », in : *Journal of personality and social psychology* 29.4 (1974), p. 475-486.
- [LB19] R. LOTFIAN et C. BUSSO, « Building Naturalistic Emotionally Balanced Speech Corpus by Retrieving Emotional Speech From Existing Podcast Recordings », in : *IEEE Transactions on Affective Computing* 10.4 (oct. 2019), p. 471-483, DOI : 10.1109/TAFFC.2017.2736999.
- [LCC10] Devillers L., Vaudable C. et Chasatgnol C., « Real-life emotion-related states detection in call centers : a cross-corpora study », in : *Proc. of Interspeech*, Makuhari, Chiba, Japan, 2010, p. 2350-2355.
- [Le +89] Y. LE CUN et al., « Handwritten digit recognition : Applications of neural network chips and automatic learning », in : *IEEE Communications Magazine* 27.11 (1989), p. 41-46.
- [Le+20] H. LE et al., « FlauBERT : Unsupervised Language Model Pre-training for French », in : *Proc. of Language Resources and Evaluation Conference (LREC)*, Virtual Conference, 2020, p. 2479-2490.
- [Lec+98] Y. LECUN et al., « Gradient-based learning applied to document recognition », in : *Proceedings of the IEEE* 86.11 (1998), p. 2278-2324.

-
- [Lev03] R. W. LEVENSON, « Blood, sweat, and fears : The autonomic architecture of emotion », in : *Annals of the New York Academy of Sciences* 1000.1 (2003), p. 348-366.
- [Ley10] R. LEYS, « How did fear become a scientific object and what kind of object is it? », in : *Representations* 110.1 (2010), p. 66-104.
- [Lin+13] KA LINDQUIST et al., « The hundred-year emotion war : are emotions natural kinds or psychological constructions? Comment on Lench, Flores, and Bench (2011) », in : *Psychological bulletin* 139.1 (2013), p. 255-263.
- [Lin89] L.I-Kuei LIN, « A Concordance Correlation Coefficient to Evaluate Reproducibility », in : *Biometrics* 45.1 (1989), p. 255-268.
- [Liu+10] Yang LIU et al., « Coherent bag-of audio words model for efficient large-scale video copy detection », in : *Proceedings of the 9th ACM International Conference on Image and Video Retrieval, CIVR 2010*, Xi'an, China, 2010, p. 89-96.
- [Liu+18] Z. LIU et al., « Efficient Low-rank Multimodal Fusion with Modality-Specific Factors », in : *Proc. of ACL*, Melbourne, Australia, 2018, p. 2247-2256.
- [Liu+19a] X. LIU et al., « Multi-Task Deep Neural Networks for Natural Language Understanding », in : *Proc. of ACL*, Florence, Italy, 2019, p. 4487-4496.
- [Liu+19b] Y. LIU et al., « RoBERTa : A Robustly Optimized BERT Pretraining Approach », in : *Pre-print on arXiv/1907.11692*, 2019.
- [Liu+20] A. T. LIU et al., « Mockingjay : Unsupervised Speech Representation Learning with Deep Bidirectional Transformer Encoders », in : *Proc. of ICASSP*, Virtual Conference, 2020, p. 6419-6423.
- [LL00] J. J. Z. LIAO et J. W. LEWIS, « A Note on Concordance Correlation Coefficient », in : *PDA Journal of Pharmaceutical Science and Technology* 54.1 (2000), p. 23-26.
- [LL16] B. LIU et I. LANE, « Attention-Based Recurrent Neural Network Models for Joint Intent Detection and Slot Filling », in : *Proc. of INTERSPEECH*, San Francisco, California, USA, 2016, p. 685-689.
- [LL58] J. I. LACEY et B. C. LACEY, « Verification and Extension of the Principle of Autonomic Response-Stereotypy », in : *The American Journal of Psychology* 71.1 (1958), p. 50-73.

-
- [Llo82] S. LLOYD, « Least squares quantization in PCM », in : *IEEE Transactions on Information Theory* 28.2 (1982), p. 129-137.
- [LLS03] H. LIU, H. LIEBERMAN et T. SELKER, « A Model of Textual Affect Sensing Using Real-World Knowledge », in : *Proc. of the International Conference on Intelligent User Interfaces*, Miami, Florida, USA, 2003, p. 125-132.
- [LPM15] T. LUONG, H. PHAM et C. D. MANNING, « Effective Approaches to Attention-based Neural Machine Translation », in : *Proc. of conference on empirical methods in natural language processing (EMNLP)*, Lisbon, Portugal, 2015, p. 1412-1421.
- [LT15] J. LEE et I. TASHEV, « High-level feature representation using recurrent neural network for speech emotion recognition », in : *Proc. of INTERSPEECH*, Dresden, Germany, 2015, p. 1537-1540.
- [Mac+20a] M. MACARY et al., « AlloSat : A New Call Center French Corpus for Satisfaction and Frustration Analysis », in : *Proc. of Language Resources and Evaluation Conference (LREC)*, Virtual Conference, 2020, p. 1590-1597.
- [Mac+20b] M. MACARY et al., « Multi-corpus experiment on continuous speech emotion recognition : convolution or recurrence ? », in : *Proc. of Conference on Speech and Computer (SPECOM)*, Virtual Conference, 2020.
- [Mar+20] L. MARTIN et al., « CamemBERT : a Tasty French Language Model », in : *Proc. of ACL*, Virtual Conference, 2020, p. 7203-7219.
- [MB15] S. MARIOORYAD et C. BUSSO, « Correcting Time-Continuous Emotional Labels by Modeling the Reaction Lag of Evaluators », in : *IEEE Transactions on Affective Computing* 6.2 (2015), p. 97-108.
- [McK+12] G. MCKEOWN et al., « The SEMAINE Database : Annotated Multimodal Records of Emotionally Colored Conversations between a Person and a Limited Agent », in : *IEEE Transactions on Affective Computing* 3.1 (2012), p. 5-17.
- [Meh20] N. MEHENDALE, « Facial emotion recognition using convolutional neural networks (FERC) », in : *SN Appl. Sci.* 2.446 (2020), p. 1-8.
- [Meh80] Albert MEHRABIAN, *Basic Dimensions for a General Psychological Theory Implications for Personality, Social, Environmental, and Developmental Studies*, Oelgeschlager, Gunn & Hain, 1980.

-
- [Mes+13] G. MESNIL et al., « Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. », in : *Proc. of INTERSPEECH*, Lyon, France, 2013, p. 3771-3775.
- [MF09] Justin MARTINEAU et Tim FININ, « Delta TFIDF : An Improved Feature Space for Sentiment Analysis », in : *Proceedings of the International AAAI Conference on Web and Social Media 3.1* (mars 2009), p. 258-261.
- [Mik+13] T. MIKOLOV et al., « Distributed representations of words and phrases and their compositionality », in : *Advances in Neural Information Processing Systems (NIPS)*, Stateline, Nevada, USA, 2013, p. 3111-3119.
- [ML12] McHugh M.L., « Interrater reliability : the kappa statistic », in : *Biochem Med (Zagreb)* 22.3 (2012), p. 276-282.
- [MM14] P.H. MATTHEWS et P.H. MATTHEWS, *The Concise Oxford Dictionary of Linguistics*, OUP Oxford, 2014.
- [Mor07] K. M. MORRISON, « Natural resources, aid, and democratization : A best-case scenario », in : *Public Choice* 131.3-4 (2007), p. 365-386.
- [MP10] C. MAAOUI et A. PRUSKI, « Emotion recognition through physiological signals for human-machine communication », in : *Cutting Edge Robotics 2010.1* (2010), p. 317-332.
- [MP43] W. S. MCCULLOCH et W. PITTS, « A logical calculus of the ideas immanent in nervous activity », in : *Bulletin of Mathematical Biophysics* 5 (1943), p. 115-133.
- [MPI05] C. MA, H. PRENDINGER et M. ISHIZUKA, « Emotion Estimation and Reasoning Based on Affective Textual Interaction », in : *Proc. of Conference on Affective Computing and Intelligent Interaction (ACII)*, Beijing, China, 2005, p. 622-628.
- [MS14] C. MONNIER et A. SYSSAU, « Affective norms for French words (FAN). », in : *Behavior research methods* 46.4 (2014), p. 1128-1137.
- [Nas+11] A. NASR et al., « MACAON : An NLP Tool Suite for Processing Word Lattices », in : *Proc. of ACL*, Portland, Oregon, USA, 2011, p. 86-91.
- [NGB17] L. NANNI, S. GHIDONI et S. BRAHNAM, « Handcrafted vs. non-handcrafted features for computer vision classification », in : *Pattern Recognition* 71 (2017), p. 158-172.

-
- [Ngu+20] H. NGUYEN et al., « Investigating Self-supervised Pre-training for End-to-end Speech Translation », in : *Proc. of the workshop on Self-supervision in Audio and Speech at the International Conference on Machine Learning (ICML)*, Virtual Conference, 2020.
- [OSR19] P. J ORTIZ SUÁREZ, B. SAGOT et 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, p. 9-16.
- [Ott+19] Myle OTT et al., « fairseq : A Fast, Extensible Toolkit for Sequence Modeling », in : *Proc. of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Minneapolis, Minnesota, USA, 2019, p. 48-53.
- [Owa+12] M. OWAYJAN et al., « The Design and Development of a Lie Detection System using Facial Micro-Expressions », in : *International Conference on Advances in Computational Tools for Engineering Applications, ACTEA*, Beirut, Lebanon, 2012, p. 33-38.
- [PA12] Stephanie PANCOAST et Murat AKBACAK, « Bag-of-Audio-Words Approach for Multimedia Event Classification », in : *INTERSPEECH*, 2012.
- [Pal+19] Berthille PALLAUD et al., « Suspensive and Disfluent Self Interruptions in French Language Interactions », in : *Fluency and Disfluency across Languages and Language Varieties*, sous la dir. de Presses Universitaires de LOUVAIN, Corpora and Language in use 4, 2019.
- [Pan+15] V. PANAYOTOV et al., « Librispeech : An ASR corpus based on public domain audio books », in : *Proc. of ICASSP*, South Brisbane, Queensland, Australia, 2015, p. 5206-5210.
- [Pas+19] R. PASTI et al., « A Sensitivity and Performance Analysis of Word2Vec Applied to Emotion State Classification Using a Deep Neural Architecture », in : *Proc. of the Distributed Computing and Artificial Intelligence Conference (DCAI)*, Ávila, Spain, 2019, p. 199-206.
- [PCB02] P. PHILIPPOT, G. CHAPELLE et S. BLAIRY, « Respiratory feedback in the generation of emotion », in : *Cognition & Emotion* 16.5 (2002), p. 605-627.

-
- [Ped+11] F. PEDREGOSA et al., « Scikit-learn : Machine Learning in Python », in : *Journal of Machine Learning Research* 12.85 (2011), p. 2825-2830.
- [Pes+10] J. PESTIAN et al., « Suicide Note Classification Using Natural Language Processing : A Content Analysis », in : *Biomedical Informatics Insights* 3.1 (2010), p. 19-28.
- [Phi07] P. PHILIPPOT, « Émotion et psychotérapie », in : Mardaga, 2007, p. 11-64.
- [Pic00] R. W. PICARD, *Affective computing*, MIT press, 2000.
- [Pit+17] D. A. PITALOKA et al., *Enhancing CNN with Preprocessing Stage in Automatic Emotion Recognition*, Bali, Indonesia, 2017.
- [Pov+11] D. POVEY et al., *The Kaldi Speech Recognition Toolkit*, Hilton Waikoloa Village, Big Island, Hawaii, US, 2011.
- [PRJ20] A. P. PIMPALKAR, Retna RAJ et R. J., « Influence of Pre-Processing Strategies on the Performance of ML Classifiers Exploiting TF-IDF and BOW Features. », in : *ADCAIJ : Advances in Distributed Computing and Artificial Intelligence Journal* 9.2 (2020).
- [PSM07] R. R. PROVINE, R. J. SPENCER et D. L. MANDELL, « Emotional Expression Online : Emoticons Punctuate Website Text Messages », in : *Journal of Language and Social Psychology* 26.3 (2007), p. 299-307.
- [PSM14] J. PENNINGTON, R. SOCHER et C. D. MANNING, « Glove : Global vectors for word representation », in : *Proc. of conference on empirical methods in natural language processing (EMNLP)*, Doha, Qatar, 2014, p. 1532-1543.
- [Pus+96] A.P.C.S.J. PUSTEJOVSKY et al., *Lexical Semantics : The Problem of Polysemy*, Clarendon paperbacks, Clarendon Press, 1996.
- [PY10] S. J. PAN et Q. YANG, « A Survey on Transfer Learning », in : *IEEE Transactions on Knowledge and Data Engineering* 22.10 (2010), p. 1345-1359.
- [R80] Plutchik R., *Theories of Emotion*, Academic Press, 1980.
- [Ram06] S. RAMÓN Y CAJAL, *The structure and connexions of neurons*, rapp. tech., Lecture of Phisyology or Medecine Nobel Prize, 1906.
- [RH93] L. R. RABINER et Juang B. H., *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, NJ, 1993.

-
- [Rin+13] F. RINGEVAL et al., « Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions », in : *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)* (2013), p. 1-8.
- [Rin+17] F. RINGEVAL et al., « AVEC 2017 - Real-life Depression, and Affect Recognition Workshop and Challenge », in : *Proc. of the Audio/Visual Emotion Challenge and Workshop*, Mountain View, United States, 2017, p. 3-9.
- [Rin+18] F. RINGEVAL et al., « AVEC 2018 Workshop and Challenge : Bipolar Disorder and Cross-Cultural Affect Recognition », in : *Proc. of the Audio/Visual Emotion Challenge and Workshop*, Beijing, China, 2018, p. 57-64.
- [Rin+19] F. RINGEVAL et al., « AVEC 2019 Workshop and Challenge : State-of-Mind, Detecting Depression with AI, and Cross-Cultural Affect Recognition », in : *Proc. of the Audio/Visual Emotion Challenge and Workshop*, Nice, France, 2019, p. 3-12.
- [RJ86] L. RABINER et B. JUANG, « An introduction to hidden Markov models », in : *IEEE ASSP Magazine* 3.1 (1986), p. 4-16.
- [Ros58] F. ROSENBLATT, « The perceptron : a probabilistic model for information storage and organization in the brain. », in : *Psychological review* 65.6 (1958), p. 386-408.
- [Rou+14] A. ROUSSEAU et al., « LIUM and CRIM ASR System Combination for the REPERE Evaluation Campaign », in : *17th International Conference on Text, Speech and Dialogue*, Brno, Czech Republic, 2014, p. 441-448.
- [Roy+18] P. ROY et al., « Effects of degradations on deep neural network architectures », in : *Pre-print on arXiv/1807.10108*, 2018.
- [Roz+12] V. ROZGIC et al., « Ensemble of SVM trees for multimodal emotion recognition », in : *Proc. of the Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, Hollywood, California, 2012, p. 1-4.
- [ŘS10] R. ŘEHŮŘEK et 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, p. 45-50.

-
- [Rus+12] V. RUS et al., « Automated Discovery of Speech Act Categories in Educational Games. », in : *Proc. of International Educational Data Mining Society (EDM)*, Chania, Greece, 2012, p. 25-32.
- [Rus80] J. A. RUSSELL, « A circumplex model of affect. », in : *Journal of personality and social psychology* 39.6 (1980), p. 1161-1178.
- [Sai+18] K. SAILUNAZ et al., « Emotion Detection from Text and Speech - A Survey », in : *Social Network Analysis and Mining (SNAM)*, Springer 8.28 (2018), p. 1-26.
- [SB88] Gerard SALTON et Christopher BUCKLEY, « Term-weighting approaches in automatic text retrieval », in : *Information Processing & Management* 24.5 (1988), p. 513-523.
- [Sch+10] B. SCHULLER et al., « CINEMO-A French spoken language resource for complex emotions : facts and baselines », in : *Proc. of Language Resources and Evaluation Conference (LREC)*, Valletta, Malta, 2010, p. 1643-1647.
- [Sch+11] Björn SCHULLER et al., « The INTERSPEECH 2011 speaker state challenge », in : *Proc. INTERSPEECH*, Florence, Italy, 2011.
- [Sch+13] B. SCHULLER et al., « The INTERSPEECH 2013 computational paralinguistics challenge : Social signals, conflict, emotion, autism », in : *Proc. of INTERSPEECH*, Lyon, France, 2013, p. 148-152.
- [Sch+19] S. SCHNEIDER et al., « wav2vec : Unsupervised Pre-Training for Speech Recognition », in : *Proc. of INTERSPEECH*, Graz, Austria, 2019, p. 3465-3469.
- [Sch05] K. R. SCHERER, « What are emotions? And how can they be measured? », in : *Social science information* 44.4 (2005), p. 695-729.
- [Sch15] M. SCHWARZ-FRIESEL, « Language and emotion », in : *Emotion in Language*. U. Lüdke (ed.). Amsterdam, John Benjamins, 2015, p. 157-173.
- [Sch54] H. SCHLOSBERG, « Three dimensions of emotion. », in : *Psychological review* 61.2 (1954), p. 81-88.
- [Sch59] Stanley SCHACHTER, *The psychology of affiliation : Experimental studies of the sources of gregariousness*. Stanford Univer. Press., 1959.

-
- [Sch84] Klaus R SCHERER, « Emotion as a multicomponent process : A model and some cross-cultural data. », in : *Review of personality & social psychology* 5.1 (1984), p. 37-63.
- [Sch86] K. R. SCHERER, « Vocal affect expression : A review and a model for future research », in : *Psychological Bulletin* 99.2 (1986), p. 143-165.
- [SCS19] M. SCHMITT, N. CUMMINS et B. W. SCHULLER, « Continuous Emotion Recognition in Speech - Do We Need Recurrence? », in : *Proc. of INTERSPEECH*, Graz, Austria, 2019, p. 2808-2812.
- [SD10] B. SCHULLER et L. DEVILLERS, « Incremental acoustic valence recognition : An inter-corpus perspective on features, matching, and performance in a gating paradigm », in : *Proc. of INTERSPEECH*, Makuhari, Chiba, Japan, 2010, p. 801-804.
- [SG64] A. SAVITZKY et M. J. E. GOLAY, « Smoothing and Differentiation of Data by Simplified Least Squares Procedures. », in : *Analytical Chemistry* 36.8 (1964), p. 1627-1639.
- [Sha48] C. E. SHANNON, « A mathematical theory of communication », in : *The Bell system technical journal* 27.3 (1948), p. 379-423.
- [Sil+18] D. SILVER et al., « A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play », in : *Science* 362.6419 (2018), p. 1140-1144.
- [SJA01] T. SUSLOW, K. JUNGHANNS et V. AROLT, « Detection of Facial Expressions of Emotions in Depression », in : *Perceptual and Motor Skills* 92.3 (2001), p. 857-868.
- [SM03] R. J. SRINIVASAN et D. W. MASSARO, « Perceiving Prosody from the Face and Voice : Distinguishing Statements from Echoic Questions in English », in : *Language and Speech* 46.1 (2003), p. 1-22.
- [Sol+07] R. SOLERAURENA et al., « Robust ASR using Support Vector Machines », in : *Speech Communication* 49.4 (2007), p. 253-267.
- [SRS16] M. SCHMITT, F. RINGEVAL et B. SCHULLER, « At the Border of Acoustics and Linguistics : Bag-of-Audio-Words for the Recognition of Emotions in Speech », in : *Proc. of INTERSPEECH*, San Francisco, California, USA, 2016, p. 495-499.

-
- [SS04] A. J. SMOLA et B. SCHOLKOPF, « A tutorial on support vector regression », in : *Statistics and computing* 14.3 (2004), p. 199-222.
- [SS62] S. SCHACHTER et J. SINGER, « Cognitive, social, and physiological determinants of emotional state. », in : *Psychological review* 69.5 (1962), p. 379-399.
- [SS93] S. STEPPER et F. STRACK, « Proprioceptive determinants of emotional and nonemotional feelings. », in : *Journal of personality and social psychology* 64.2 (1993), p. 211-220.
- [SSB09] B SCHULLER, S STEIDL et A BATLINER, « The Interspeech 2009 Emotion Challenge », in : *Proc. Interspeech*, Brighton, UK, 2009, p. 312-315.
- [Ste+02] G. STEMMER et al., « Comparison and Combination of Confidence Measures », in : *Proc. of Text, Speech and Dialogue*, Brno, Czech Republic, 2002, p. 181-188.
- [Ste02] T. STEIMER, « The biology of fear- and anxiety-related behaviors. », in : *Dialogues in clinical neuroscience* 4.3 (2002), p. 231-249.
- [Sto+17] M. N. STOLAR et al., « Real time speech emotion recognition using RGB image classification and transfer learning », in : *International Conference on Signal Processing and Communication Systems (ICSPCS)*, Surfers Paradise, Australia, 2017, p. 1-8.
- [Str96] K. T. STRONGMAN, *The Psychology of Emotion : Theories of Emotion in Perspective*, John Wiley & Sons., 1996.
- [TD16] M. TAHON et L. DEVILLERS, « Towards a Small Set of Robust Acoustic Features for Emotion Recognition : Challenges », in : *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24.1 (2016), p. 16-28.
- [Tho14] M. THOMA, « On-line Recognition of Handwritten Mathematical Symbols », mém. de mast., 2014.
- [TKT18] B. B. TRAORE, B. KAMSU-FOGUEM et F. TANGARA, « Deep convolution neural network for image recognition », in : *Ecological Informatics* 48 (2018), p. 257-268.
- [Tom06] S. TOMAR, « Converting video formats with FFmpeg », in : *Linux Journal* 2006.146 (2006), p. 1-10.

-
- [TP13] C. THURLOW et M. POFF, « 7. Text messaging », in : *Pragmatics of computer-mediated communication*, De Gruyter Mouton, 2013, p. 163-190.
- [TRB10] J. TURIAN, L. RATINOV et Y. BENGIO, « Word Representations : A Simple and General Method for Semi-Supervised Learning », in : *Proc. of ACL*, Uppsala, Sweden, 2010, p. 384-394.
- [Tum11] G. TUMBAT, « Co-constructing the service experience : Exploring the role of customer emotion management », in : *Marketing Theory* 11.2 (2011), p. 187-206.
- [Vas+17] A. VASWANI et al., « Attention is All you Need », in : *Advances in Neural Information Processing Systems (NIPS)*, Long Beach, California, USA, 2017, p. 5998-6008.
- [Wei74] S. WEITZ, *Nonverbal Communication : Readings with Commentary*, Oxford University Press, 1974.
- [WJ02] Wilhelm Max WUNDT et Charles Hubbard JUDD, *Outlines of psychology*, W. Engelmann, 1902.
- [WKW16] K. WEISS, T.M. KHOSHGOFTAAR et D. WANG, « A survey of transfer learning », in : *Journal of Big Data* 3.9 (2016), p. 1-40.
- [WMK00] S. WIENS, E. S. MEZZACAPPA et E. S. KATKIN, « Heartbeat detection and the experience of emotions », in : *Cognition and Emotion* 14.3 (2000), p. 417-427.
- [Wöl+13] M. WÖLLMER et al., « LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework », in : *Image and Vision Computing* 31.2 (2013), p. 153-163.
- [WP16] F. WANG et A. PEREIRA, « Neuromodulation, emotional feelings and affective disorders », in : *Mens sana monographs* 14.1 (2016), p. 5-29.
- [Wu+16] Yonghui WU et al., « Google’s Neural Machine Translation System : Bridging the Gap between Human and Machine Translation », in : *Pre-print on arXiv/1609.08144*, 2016.
- [Yan+19] Z. YANG et al., « XLNet : Generalized Autoregressive Pretraining for Language Understanding », in : *Advances in Neural Information Processing Systems (NIPS)*, Vancouver, Canada, 2019, p. 5753-5763.

-
- [Yao+14] K. YAO et al., « Spoken language understanding using long short-term memory neural networks », in : *Proc. of Spoken Language Technologies Workshop (SLT)*, South Lake Tahoe, California, USA, 2014, p. 189-194.
- [You+18] T. YOUNG et al., « Recent Trends in Deep Learning Based Natural Language Processing [Review Article] », in : *IEEE Computational Intelligence Magazine* 13.3 (2018), p. 55-75.
- [ZH08] S. ZAHORIAN et H. HU, « A spectral/temporal method for robust fundamental frequency tracking », in : *The Journal of the Acoustical Society of America* 123.6 (2008), p. 4559-4571.
- [Zha+19a] F. ZHANG et al., « Construction site accident analysis using text mining and natural language processing techniques », in : *Automation in Construction* 99.1 (2019), p. 238-248.
- [Zha+19b] Z. ZHANG et al., « ERNIE : Enhanced language representation with informative entities », in : *Proc. of ACL*, Florence, Italy, 2019, p. 1441-1451.
- [Zhu+15] Y. ZHU et al., « Aligning Books and Movies : Towards Story-like Visual Explanations by Watching Movies and Reading Books », in : *Pre-print on arXiv/1506.06724*, 2015.
- [ZQR16] B. ZHANG, C. QUAN et F. REN, « Study on CNN in the recognition of emotion in audio and images », in : *Proc. of International Conference on Computer and Information Science (ICIS)*, Okayama, Japan, 2016, p. 1-5.
- [ZW15] R. S. ZHOU et Z. J. WANG, « A Review of a Text Classification Technique : K-Nearest Neighbor », in : *Proc. of the International Conference on Computer Information Systems and Industrial Applications (CISIA)*, Bangkok, Thailand, 2015, p. 453-455.



Titre : Analyse de données massives en temps réel pour l'extraction d'informations sémantiques et émotionnelles de la parole

Mot clés : Reconnaissance de l'émotion continue, Création de Corpus, Satisfaction et Frustration, Embeddings pré-appris

Résumé : Les centres d'appels reçoivent tous les jours des milliers de coups de téléphone permettant de faire le lien entre des clients et des conseillers. Ainsi, de nombreuses informations peuvent être extraites de ces conversations, dont l'aspect émotionnel.

Cette thèse CIFRE a été réalisée en collaboration avec l'entreprise Allo-Media qui est spécialisée dans l'analyse automatique de conversations téléphoniques de centre d'appels. Concrètement, elle met en place des relevés d'information sur différents aspects de la conversation en indexant ces informations pour permettre un traitement automatique des données. L'entreprise cherche à enrichir ses annotations avec une solution innovante permettant de rajouter un aspect émotionnel en adéquation avec le contexte de la relation clientèle afin d'alerter sur les points saillants de la conversation.

Cette thèse tente donc de répondre à plu-

sieurs problématiques : (i) tout d'abord la définition de l'émotion de satisfaction et de frustration dans la parole, (ii) la mise en place d'une reconnaissance automatique de ces émotions de façon continue tout au long de la conversation et (iii) des méthodes d'évaluation de ces systèmes automatiques.

Les contributions de cette thèse sont : (i) la construction d'un corpus à partir de données réelles, annoté de façon continue en satisfaction et frustration, (ii) la mise en place de différentes stratégies pour construire un système de reconnaissance automatique utilisant des réseaux de neurones profonds en nous comparant à l'état de l'art, (iii) l'exploration de la dissociation des aspects acoustique et linguistique des conversations afin d'améliorer nos systèmes de reconnaissance et enfin (iv) la mise en place d'une évaluation nuancée de ces systèmes.

Title: Massive and real-time data analysis in order to extract semantic and emotional information from speech

Keywords: Speech Emotion Recognition, New Corpora, Satisfaction and Frustration, Pre-train embeddings

Abstract: Call centers receive thousands of calls every day in order to connect clients and agents. Thus lots of information can be extracted from these conversations, including the emotional aspect of the speakers.

This CIFRE thesis was carried out in col-

laboration with the Allo-Media company, that is specialized in the automatic analysis of call center conversations. Concretely, they set up information records on different aspects of the conversation by discretizing the information to allow automatic processing of the data. The

company seeks to enrich its annotations with an innovative solution to add an emotional aspect relevant with the context of customer relations in order to alert on the difficult points of the conversation.

This thesis therefore attempts to respond to several issues: (i) first of all the definition of the emotion of satisfaction and frustration in speech, (ii) the establishment of an automatic recognition of these emotions on a continuous basis throughout the conversation and (iii) methods to evaluate these automatic systems.

The contributions of this thesis are: (i) the construction of a corpus from real data, continuously annotated in satisfaction and frustration, (ii) the implementation of different strategies to build an automatic recognition system using deep neural networks by comparing ourselves to the state of the art, (iii) the exploration of the dissociation of the acoustic and linguistic aspects of conversations in order to improve our recognition systems and finally (iv) the implementation of a nuanced assessment of these systems.