Reconnaissance automatique des émotions dans la parole

Jonathan Bonnaud

Janvier 2018

1 Introduction

L'objectif du projet est de reconnaître des émotions discrètes à partir d'un corpus de parole. Pour cela, nous disposons de deux corpus émotionnels en allemand (EMO-DB et AIBO), la partie 2 décrit notamment le corpus AIBO. Les parties 3 et 4 seront consacrées à l'expliquation et au commentaire des approches mises en oeuvre. Deux approches seront abordées : une approche classique et une approche neuronale basées sur les articles (1) et (2) respectivement.

2 Analyse des articles scientifiques

Les corpus utilisés dans les articles sont le corpus AIBO et RECOLA. Le corpus AIBO est un corpus d'émotions spontanées, des enfants réagissent avec le robot Aibo qu'ils pensent contrôler alors qu'il est en fait contrôlé par une personne tiers. Cela permet d'avoir des émotions induites. Les enregistrements ont été fait dans deux écoles différences avec un total de 51 enfants de 10 à 13 ans (21 garçons, 30 filles). Cela représente un total de 9,2 heures de parole. Des annotations sur chaque mot ont aussi été ajoutées par 5 étudiants avancés en langues. Ces annotations d'émotions ont été faites en écoutant le signal dans l'ordre, segmenté en tours de parole (avec un temps de pause de 1 seconde entre chaque tour). Ils ont utilisé 11 étiquettes d'émotions : *joyful*, *surprised*, *emphatic*, *helpless*, *touchy*, i.e. irrité, *angry*, *motherese*, *bored*, *reprimanding*, *rest*, i.e. non-neutre, mais n'appartenant pas aux autres catégories, *neutral*. Pour les tâches de classification, à 5 classes (tâche 1), on peut regrouper les émotions comme suit : **A**nger (angry, touchy, et reprimanding) **E**mphatic, **N**eutral, **P**ositive (motherese et joyful), et **R**est. Et à 2 classes (tâche 2) : **NEG**ative (angry, touchy, reprimanding, et emphatic) and **IDL**e (tous les états non-négatifs).

Le corpus RECOLA est un corpus d'émotions spontanées, composé d'enregistrements de parole de 46 participants francophones de 5 minutes chacun. Les émotions qu'ils cherchent à prédire sont : *Arousal* et *Valence*. Cette approche alternative est introduite grâce à la compréhension psychologique visant à mon-

trer que l'émotion peut être représentée par des emotions primitives telles que *Arousal* et *Valance* .

Dans l'article (1), les descripteurs utilisés sont dit des plus communs, qui sont les plus prometteurs : descripteurs de qualité de prosodie, spectrale et de voix. Ce qui donne un total de 384 descripteurs.

En entrée du réseaux convolutionnel : des séquences de 6 secondes sur le signal brut. (après un prétraitement pour éviter les grosses variations de hauteur de voix entre les locuteurs.)

Dans l'approche classique utilisant le corpus AIBO les auteurs ont utilisé les données d'une école pour l'apprentissage (Ohm) et celles de l'autre école pour le test (Mont) afin de guarantir l'indépendance de locuteur (chaque locuteur est représenté soit dans un ensemble soit dans l'autre). Pour l'approche neuronale l'ensemble des données de RECOLA est utilisé et il est séparé équitablement en 3 partitions : apprentissage (16 personnes), validation (15 personnes) et test (15 personnes). Le genre et l'âge des locuteurs sont bien équilibrés entre les 3 partitions. Ici aussi l'indépendance de locuteur est respectée.

Les métriques d'évaluation utilisées dans l'article (1) sont les moyenne pondérées (WA, i.e. weighted average) et non pondérées (UA, i.e. unweighted average) des rappels (Re) et précisions (Pr) pour chaque classe. Les UA correspondent aux mesures de macro-rappel et macro-précision. Les WA correspondent à la moyenne des Pr (ou Re) pondérées par la répartition des échantillons dans les classes, le tout divisé par le nombre d'échantillons total. L'*accuracy* rapporte seulement le pourcentage de bonnes réponses, cela ne reflète pas le poids qu'a la répartition des échantillons dans les classes. Cela permet aussi de donner des mésures qui permettront la comparaison avec d'autres solutions. Dans l'article (2) la métrique d'évaluation utilisée est une fontion objective : le coefficient de correlation de concordance, qui permet de calculer le taux d'accord entre les prédictions et la référence. Elle est aussi utilisée comme mesure d'optimisation des modèles, tout comme l'erreur quadratique moyenne (MSE) qui est traditionnellement utilisé dans la literrature.

3 Approche état de l'art

Pour l'approche classique, j'ai tout d'abord extrait les descripteurs de chaque fichier audio .wav à l'aide de l'outil OpenSMILE pour générer des fichiers .arff. J'ai bien gardé la séparation des données en 2 (Ohm et Mont). J'ai ensuite lu les fichiers .arff afin d'en extraire les descripteurs et de former mes corpus de train et de test. Chacun des échantillons est donc représenté par un vecteur de dimension 384. J'ai chargé les labels de référence à partir des fichiers `chunk_labels_[2-5]cl_corpus.txt`. Ensuite en utilisant la librairie `sklearn` j'ai utilisé un SVM avec kernel linéaire. Les résultats obtenus sans configuration des paramètres se trouve dans la Table 1. La figure 1 montre les matrices de confusions pour les deux tâches. Les résultats sont assez similaires à ceux de l'article (1), même si mon rappel (UA) est plus faible.

La première expérimentation consiste à utiliser le corpus EMO-DB en plus

	Recall [%]		Precision [%]	
	UA	WA	UA	WA
2-class	66.7	72.0	66.7	72.1
5-class	29.6	58.4	35.8	58.7

TABLE 1 – Résultats de l'approche SVM Linéaire

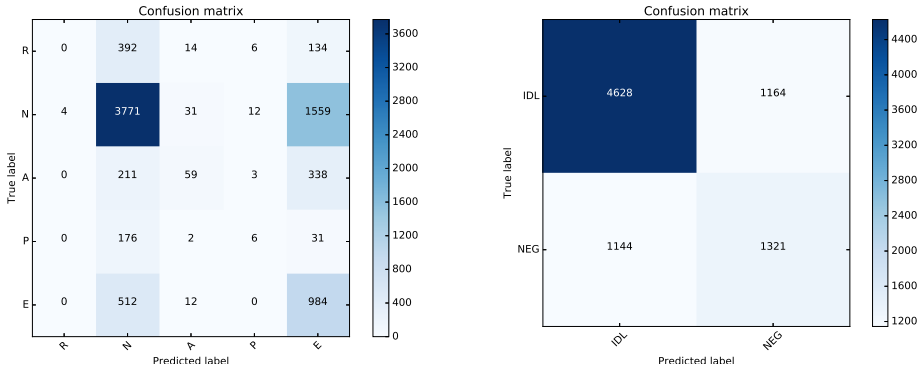

FIGURE 1 – Matrices de confusion pour la tâche 1 (5-class) et 2 (2-class)

d'AIBO. J'ai donc fait l'apprentissage sur tout AIBO (O+M) et le test sur EMO-DB. Le fait d'avoir un plus gros corpus d'entrainement peut peut-être améliorer les performances. Cependant quelques différences entre les corpus peut apporter de la difficulté : les locuteurs ne sont pas de la même tranche d'âge (enfants et adultes) et les émotions sont spontanées et actées, respectivement pour AIBO et EMO-DB. De plus les classes ne sont pas les mêmes. Le corpus EMO-DB contient 535 échantillons de voix en allemand. C'est est corpus d'émotions actéées, les locuteurs sont des acteurs adultes et les enregistrements ont été réalisés dans une chambre anéchoïque à l'Université de Technologie de Berlin. Cette chambre dont les parois absorbent les ondes sonores permet d'avoir des échantillons de voix sans bruit de fond. Étant donné la taille du corpus l'approche la plus logique est de l'utiliser comme test. Pour cela il faut passer par une étape de regroupement des émotions car AIBO et EMO-DB n'ont pas les mêmes classes d'émotions. EMO-DB a la classification suivante : *anger* (W), *boredom* (L), *disgust* (E), *fear* (A), *happiness* (F), *sadness* (T), *neutral* (N). Les configuration expérimentées se trouvent dans les Tableaux 2 et 3. Pour la tâche 1 : Dans la configuration 1 j'ai mis toutes les émotions non représentées par AIBO dans la classe R et les autres dans leurs équivalents. Pour les configurations 1 et 2 j'ai testé le fait qu'on puisse considérer la tristesse (T) et l'ennuie (L) commedes émotions Neutres. La configuration 4 considère la peur (A) comme une émotion *Emphatic*. On peut voir que cette dernière obtient le meilleur rappel (WA) (au détriment de la précision) car dans les autres cas la classes E n'était pas représentée, dans

	Classes de AIBO				
	Anger	**N**eutral	**P**ositive	**R**est	**E**mphatic
Config 1	W	N	F	L, A, T, E	
Config 2	W	N, T	F	L, A, E	
Config 3	W	N, T, L	F	A, E	
Config 4	W	N, T, L	F	E	A

Table 2 – Regroupement des classes d'EMO-DB pour la tâche 1.

	Classes de AIBO	
	IDLe	**NEG**ative
Config 1	F, N, L	W, A, T, E
Config 2	F, N	W, A, T, E, L

Table 3 – Regroupement des classes d'EMO-DB pour la tâche 2.

la configuration 4 elle l'est. Les résultats des expérimentations se trouvent dans le tableau 4. Pour la tâche 2, ma réflexion était la même, peut-on considérer l'ennuie (L) comme une émotions non négative ? Ce changement augmente la précision.

4 Approche end-to-end

Pour une approche end-to-end je pense qu'il est possible d'utiliser les même corpus que pour une approche classique, seulement les pré-traitement seront différents et les features à extraire pourront être différent. Dans l'approche de l'article (2), les données brutes sont prises en entrée, c'est-à-dire l'amplitude en fonction du temps. Les CNNs permettent notamment d'apprendre des charatéristiques acoustiques plus stables (car la convoluation est appliquée sur des fenêtres qui se superposent dans le temps). De plus, des architectures *Deep* ont

	Config	**Recall [%]**		**Precision [%]**	
		UA	WA	UA	WA
5-class	1	**22.3**	14.2	20.7	30.2
	2	19.0	19.8	23.0	29.1
	3	17.9	27.8	**26.4**	**31.5**
	4	19.8	**31.2**	24.1	26.9
2-class	1	55.4	**53.4**	55.8	57.0
	2	**57.5**	49.9	**56.5**	**67.0**

Table 4 – Résultats des différentes expérimentations.

aussi beaucoup de mérite. Elles permettent aux modèles de prendre en comptes plein de facteurs de variabilité dans le signal (qui sont par exemple provoquées lors de la production d'émotions).

5 Conclusion

J'aurais voulu faire plus d'expérimentations comme utiliser EMO-DB pur l'apprentissage et réduire le nombre d'échantillons d'AIBO pour m'en servir comme test. Ainsi que l'implémentation de HMM et d'un réseau convolutionnel.

Références

[1] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Proc. of Interspeech*, (Brighton, U.K.), 2009.

[2] G. T. et al., "Adieu features ? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. of ICASSP*, (Shangai, China), 2016.