Reconnaissance automatique des émotions dans la parole

La reconnaissance automatique des émotions est une branche très active de l'"affective computing". Alors que la reconnaissance de la parole cherche à identifier les mots qui ont été prononcés, la reconnaissance des émotions cherche à identifier l'état psychologique de la personne qui les a prononcés. Pour cela, les corpus d'apprentissage sont segmentés et annotés suivant des classes d'émotions discrètes ou des dimensions émotionnelles. Un segment pouvant contenir plusieurs mots, on utilisera des descripteurs haut-niveau comme la prosodie. Un segment émotionnel sera représenté dans le modèle par un vecteur de descripteurs acoustiques. Les travaux sur la multi-modalité cherchent à introduire des informations visuelles, physiologiques, linguistiques, phonétiques et contextuels au niveau du vecteur de descripteurs.

Savoir quels descripteurs utiliser et sur quelle temporalité est un problème majeur dans la construction de modèles émotionnels du type de ceux présentés dans l'article [1]. Pour s'en affranchir, des récents travaux ont appliqué des techniques neuronales (également utilisées pour la reconnaissance automatique des mots) qui se passent des vecteurs de descripteurs. C'est le cas de l'article "adieu features" [2]. L'utilisation des modalités linguistique et acoustique pour la reconnaissance des émotions est également une approche très performante aujourd'hui [3]. Enfin, parmi la multitude de représentations possibles pour le signal de parole émotionnelle, on retrouve les x-vecteurs utilisés pour la reconnaissance de locuteur [4].

[1] Björn Schuller and Stefan Steidl and Anton Batliner, "The INTERSPEECH 2009 Emotion Challenge," In proc. of *Interspeech*, Brighton, U.K., 2009.

[2] George Trigeorgis et al., "Adieu Features? End-to-end speech emotion recognition using a Deep Convolutional Recurrent Network," In proc. of *ICASSP*, Shanghai, China, 2016.

[3] Han Feng and Sei Ueno and Tatsuya Kawahara, "End-to-End Speech Emotion Recognition Combined with Acoustic-to-Word ASR Model." In proc. of *Interspeech*, Shanghai, China, 2020.

[4] R. Pappagari et al. 'X-vectors meet emotions: a study on dependencies between emotions and speaker recognition", In proc. of *ICASSP*, Barcelona, Spain, 2020.

Lecture d'articles scientifiques

L'objectif du TD est d'étudier les deux références ci-dessus suivant un certain nombre de critères et d'en dégager des protocoles expérimentaux solides que vous devrez reproduire en projet. En cours de séance, vous devrez remplir la grille de lecture ci-dessous, cette grille pouvant s'appliquer à beaucoup d'autres domaines de l'apprentissage automatique:

	[1]	[2]	[3]	[4]
Contexte				
Date de publication Auteurs / Laboratoire Journal ou conférence Hypothèse de départ				
Corpus utilisés				
Nom du/des corpus Scénario d'enregistrement Multimodalité ? Taille (nb d'heures d'enregistrement, nb de segments) Nb de locuteurs Etiquettes/dimensions émotions (nb et caractéristiques) Répartition des segments train/dev/test				
Entrées des modèles (descripteurs)				
Descripteurs utilisés Dimensions des vecteurs d'entrée Gestion de la dynamique ? (prise en compte du temps) Sélection de descripteurs ?				
Apprentissage des modèles				
Equilibre du corpus d'apprentissage (nb de classes) Type de modèle (justification) Outils utilisés Matériel informatique				
Validation				
Métriques utilisées Référence (baseline utilisée) Gain par rapport à la référence (raisons) Confirmation de l'hypothèse de départ ?				

Table 1: Proposition de grille d'analyse.

Implémentation de l'approche classique (pour aller plus loin)

L'objectif du projet est de reconnaître des émotions discrètes à partir d'un corpus de parole émotionnelle transcrite. Pour cela, vous aurez à disposition deux corpus émotionnels en allemand (AIBO) qui contiennent chacun des étiquettes adaptées au scénario de collecte des données et les signaux audio correspondant. Le corpus AIBO[1] est un corpus d'émotions spontanées (51 enfants, 2 salles différentes, 11 émotions)

Vous reproduirez le protocole qui est proposé dans l'article [1] et vous essayerez d'en reproduire les résultats sur le corpus AIBO. Vous pourrez utiliser la librairie OpenSmile[2] pour extraire les vecteurs acoustiques ou `librosa` déjà abordée dans le cours sur les descripteurs audio, ou encore vos propres descripteurs. Les émotions seront reconnues grâce à un modèle appris sur des données émotionnelles. Les performances des modèles seront évaluées suivant les métriques classiquement utilisées dans le domaine, que vous préciserez.

1. Décrire brièvement la méthodologie que vous allez suivre (vue en TD).

2. L'extraction des descripteurs acoustiques se fait avec l'outil OpenSmile. Installer `openSMILE-2.2rc1`. Il n'est pas nécessaire d'installer l'ensemble de l'outil, seule la commande ci-dessous nous intéresse (cf documentation jointe avec l'outil)

   ```
   ./SMILExtract -C config/IS09_emotion.conf -I test.wav -O test.arff
   ```

   Vérifier que cette commande s'exécute sans problème sur un fichier audio quelconque (ici `test.wav`). Les descripteurs acoustiques en sortie sont écrits directement dans le fichier `test.arff`. Le format est décrit dans la documentation en ligne, on retrouve le même format pour weka. Une ligne correspond à un segment audio, les descripteurs sont séparés par des virgules. Les noms des attributs sont donnés en entête: `@attribute name_attribut numeric`.

3. Proposer un script python permettant de retrouver les résultats de l'article. Les modèles pourront être appris et évalués avec le paquet `sklearn` pour python3. Si vous voyez des différences dans les résultats par rapport à ceux présentés dans les articles, à quoi peuvent-elles être dues ?

4. Il est attendu que vous exploriez plus en avant ce problème. Pour cela vous pourrez formuler de nouvelles hypothèses et proposer une expérimentation qui permet de la valider. Vous pourrez explorez quelques pistes parmi les suivantes (liste non exhaustive):

   - Réduire l'espace des descripteurs et analyser comment se comportent les modèles.
   - Proposer d'autres types de modèles (NN, GMM, Naives Bayes, etc.)

[1] http://www5.cs.fau.de/de/mitarbeiter/steidl-stefan/fau-aibo-emotion-corpus/
[2] http://audeering.com/technology/opensmile/