Détection d'émotions dans la voix : Synthèse bibliographique[☆]

Alexandre Huat[a,b]

[a] *INSA Rouen Normandie, Département Architectures des Systèmes d'Information*
[b] *Université Rouen Normandie, Master 2 Sciences des Données*

Résumé

Cette synthèse est un état de l'art de la détection d'émotions dans la voix (*Speech Emotion Recognition*, SER). Elle est réalisée dans une optique de développement d'un système basique de SER. En section 1, j'introduirai le sujet en présentant les multiples intérêts et applications de la reconnaissance d'émotions. Puis, je ferai une revue des caractéristiques sonores adaptées à la SER en section 2, des algorithmes existants en section 3 et des performances espérables en section 4. Dans l'intérêt de se comparer à la littérature, des bases de données de références seront aussi nommées en section 4. Enfin, je conclurai sur des perspectives de recherches pour le développement de systèmes évolués.

Mots-clés : reconnaissance d'émotions, émotions acoustiques, caractéristiques acoustiques, apprentissage statistique, intelligence émotionnelle

1. Introduction

Connaître et comprendre les émotions de ses semblables est essentiel pour la communication et la sociabilité chez l'homme. Un déficit dans la reconnaissance des émotions est effectivement associé à de nombreux troubles psychologiques, tel que la phagnosie [1], les psychoses [2] dont la psychopatie [3], ou à une altération de la sociabilité, tel que l'autisme [4] ou le *trolling* [5]. À l'inverse, une intelligence émotionnelle élevée est associée à un plus grand bien être [6] et une meilleure réussite scolaire [7, 8]. Enfin, et de manière non-exhaustive, les émotions participent à divers processus psychologiques individuels, tels que la mémorisation [9], ou des processus hormonaux, comme la ménopause [10].

Ainsi, l'analyse d'émotions présente un intérêt majeur pour l'interaction homme-machine évoluée, la robotique, les psychothérapies [11], ou diverses activités commerciales. Avec les méthodes modernes d'apprentissage automatique et de traitement du signal, il est possible de traiter de grandes quantités de données et d'obtenir des résultats pertinents avec une précision proche voire meilleure que celle de l'humain, suivant des approches multi-modales [12, 13].

Toutefois, selon Perez-Gaspar et al. [14], faisant un riche état de l'art, certaines émotions sont plus facilement reconnaissables par la seule analyse de la voix. Ainsi, dans les sections suivantes, je me concentrerai sur la reconnaissance d'émotions acoustiques.

2. Sélection de caractéristiques acoustiques

Comme tout traitement du signal, la détection d'émotions dans la voix est limitée par le bruit ambiant. C'est pourquoi des étapes de filtrage et de sélection de caractéristiques sont nécessaires pour améliorer la reconnaissance. Les coefficients cepstraux de l'échelle de Mel (MFCC) sont les caractéristiques les plus utilisées pour le traitement du signal acoustique. Notamment, car ils ont des propriétés les rendant facilement utilisables avec des mélanges de modèles gaussiens (GMM) et des chaînes de Markov cachées (HMM) [15].

Mais, depuis les années 2000, leur pertinence est discutée et d'autres alternatives ont été proposées [16]. Dont, des améliorations des filtres de Mel, les splines évolutionnaires de coefficients cepstraux (ESCC) [17], l'analyse des transformées en ondelettes (DWT) [18, 19], les représentations AM-FM, ou des décompositions adaptées aux réseaux d'apprentissage profond.

Sharma et al. [16] propose la décomposition modale empirique (EMD) et ses dérivées comme alternatives. Cette décomposition ne nécessite aucun *a priori* sur le signal, là où les MFCC ou la DWT sont calculés selon des a priori de linéarité. Théoriquement, cela représente un avantage majeur par sa concordance avec la forte non-linéarité de la voix, mais l'EMD manque encore de cadre mathématique.

Tahon and Devillers [20] font un classement de l'importance de la hauteur *(pitch)*, de caractéristiques énergétiques, des formants, de mesures de qualité de la voix et d'autres caractéristiques spectrales. Elles trouvent qu'une réduction de dimensions aux coefficients cepstraux permet au classifieur une meilleure généralisation des performances inter-corpus.

Chenchah and Lachiri [21] comparent les coefficients perceptuels de prédiction linéaire (PLP), les coefficients spectraux de puissance normalisés (PNCC) et les MFCC. Dans leur expériences, en environnement non-bruité, le taux d'erreurs le plus faible est obtenu avec les MFCC. En environnement bruité

☆. Exercice du cours *Interaction Data Analysis*, Master 2 Sciences des Données, d'Alexandre Pauchet.

Adresse email : `alexandre.huat@insa-rouen.fr` (Alexandre Huat)

(0–15 dB, bruits de train, aéroport, voiture et dialogues), les PNCC se rangent en première place juste devant les MFCC, mais la différence n'est pas significative.

Mao et al. [22], après avoir listé un grand nombre de méthodes de réduction de dimensions basées sur des auto-encodeurs, proposent une méthode d'apprentissage de caractéristiques émotives-discriminantes et domaine-invariante (ED-FLM). Cette méthode vise à la fois à considérer les divergences cross-corpus (« domaines ») et isoler les caractéristiques émotives.

Enfin, Hong et al. [23] enrichissent leur base de données de la vitesse d'énonciation, Mansouri et al. [18] tiennent compte des périodes de silence — *zero crossing* — (ZCR), et Kim et al. [24] considèrent le sexe du locuteur et la sincérité de l'émotion (vs. sur-jouée).

3. Modèles et méthodes

Dans cette section, je fais une revue des modèles de classification et des méthodes d'optimisation de SER.

Les GMM et les HMM sont des modèles très répandus en SER depuis des décennies. Chenchah and Lachiri [21] utilisent des HMM. Hong et al. [23] utilisent de même des HMM-GMM, générant un modèle par émotion et par sexe (homme/femme) du locuteur. Dans leur expérience, plus les HMM possèdent d'états, meilleure est la reconnaissance.

Vignolo et al. [17] utilisent un algorithme génétique pour générer les ESCC, puis réalisent la reconnaissance multi-classes d'émotions avec un SVM à noyau polynomial en *one-versus-all*. Tahon and Devillers [20] utilisent également un SVM pour la classification d'émotions et comparent deux méthodes d'optimisation des hyperparamètres. Elles recommandent ainsi une optimisation multi-corpus qui consiste à conserver les valeurs optimales les plus fréquemment obtenues sur l'ensemble des corpus.

Plus récemment, des réseaux de neurones profonds (DNN) ont été conçus pour répondre aux problèmes de SER. Un des avantages mis en avant par cette approche, en comparaison avec les HMM-GMM, est leur capacité à mieux représenter la richesse des caractéristiques spectro-temporelles [15]. Un autre avantage des DNN est l'apprentissage par transfert qui consiste à apprendre les premières couches du réseau — réalisant des opérations communes d'un corpus à un autre, ou d'une émotion à une autre — sur plus de données que les couches plus profondes assignées à une tâche spécifique.

Mao et al. [22] proposent une méthode d'apprentissage de caractéristiques par transfert, puis entraînent des SVM sur les caractéristiques obtenues. Fayek et al. [25] comparent une vingtaine d'architectures de DNN, dont des réseaux totalement connectés, des réseaux convolutionnels (CNN) et des réseaux récurrents à mémoire longue et court terme (LSTM-RNN). Les CNN se montrent les plus performants et les LSTM-RNN les moins performants. Toujours parmi les DNN, Kaya and Karpov [26] utilisent des machines d'apprentissage extrêmes (ELM) à noyau, adaptées à une base de peu d'échantillons pour beaucoup de caractéristiques. Ils proposent également une nouvelle méthode de normalisation des caractéristiques comme alternative à la normalisation standard de centrage-réduction. Kim et al. [24] développent une méthode d'apprentissage profond multi-tâches pour des DNN et des LSTM. Enfin, Mansouri et al. [18] utilisent un perceptron multi-couches (MLP) et Palo and Mohanty [19] un réseau à fonctions de base radiales (RBFNN), qui résiste bien au bruit, est adapté à un set de caractéristiques réduit et est capable d'apprentissage en ligne.

4. Performances

Dans cette section, je restitue les performances les plus représentatives et intéressantes parmi les articles précédemment citées. Celles-ci s'évaluent notamment par le taux de reconnaissance et le rappel moyen non-pondéré [1] (UAR), deux mesures standards de SER. Notons également que les performances intra-corpus (i.e. quand les bases d'apprentissage et de test sont issues du même corpus) sont nécessairement meilleures que les performances cross-corpus (i.e. quand les corpus dont sont issues les bases d'apprentissage et de test sont différents).

Mansouri et al. [18] obtiennent d'excellentes performances avec 90–100 % de taux de reconnaissance intra-corpus sur *Berlin Database of Emotional Speech* (EMO-DB) resp. 77.78–97.83 % sur *Surrey Audio-Visual Expressed Emotion Database* (SAVEE) pour les émotions en colère, joyeux, triste et neutre. Palo and Mohanty [19] atteignent 91.82 % resp. 93.67 % de taux de reconnaissance intra-corpus sur les mêmes bases, avec les émotions en colère, joyeux, dégoûté, neutre et triste. Leurs expériences suggèrent la supériorité prédictive des DWT sur les MFCC.

Hong et al. [23] atteignent un taux de reconnaissance intra-corpus de 82.09 % — sur une base non-renseignée —, pour de la classification des émotions neutre, joyeux, appeuré, triste et en colère.

Vignolo et al. [17] testent leurs systèmes sur *FAU Aibo Emotion Corpus* (FAU AEC) et un corpus Hindi. Ils obtiennent au mieux un UAR de 91.81 % sur le corpus FAU AEC pour les émotions en colère, catégorique, neutre, positif et reposé et 42.50 % sur le corpus Hindi pour les émotions en colère, joyeux, *lombard*, neutre et triste.

Fayek et al. [25] testent plusieurs architectures de DNN et obtiennent au mieux un taux de reconnaissance intra-corpus de 64.78 % et un UAR de 60.89 % sur la base *Interactive Emotional Dyadic Motion Capture* (IEMOCAP) avec des réseaux convolutionnels. Au cours de tests sur six corpus, Kim et al. [24] parviennent à améliorer le taux de reconnaissance cross-corpus d'en moyenne 7.4 % pour les DNN et 5.4 % pour les LSTM d'apprentissage multi-tâches. Des améliorations en reconnaissance intra-corpus ne sont notables que pour les corpus très grands. De manière générale, plus un corpus est grand, meilleures sont les performances.

En détection de valence (positif/négatif), Tahon and Devillers [20] obtiennent jusqu'à 92.7 % d'UAR intra-corpus et 43.4 %

1. *A contrario,* le rappel moyen pondéré (WAR) est la moyenne des rappels pondérés par le nombre d'enregistrements par émotion.

d'UAR cross-corpus, quand Mao et al. [22] obtiennent 65.62 % resp. 61.63 % d'UAR cross-corpus avec comme corpus cible FAU AEC et corpus sources *Airplane Behavior Corpus* (ABC) resp. EMO-DB.

Enfin, notons l'innovation en normalisation apportée par Kaya and Karpov [26], qui testent leurs méthodes sur cinq bases et améliorent l'UAR jusqu'à environ 28 % par rapport à une normalisation classique en détection de l'agitation, et jusqu'à environ 12 % en détection de valence. Ce qui amène à un UAR cross-corpus d'environ 83 % pour l'estimation de l'agitation, et d'environ 68 % pour la mesure de valence. Néanmoins, le taux de reconnaissance n'est pas restitué et nous n'avons donc pas d'indications sur la spécificité [2] de la méthode.

5. Conclusion

Dans cette synthèse, j'ai fait une revue des caractéristiques extraites de la voix et des méthodes d'apprentissage statistique utilisées en SER. Dans la perspective de construire un système de SER multi-classes, sur des enregistrements du même corpus, on peut s'attendre à au moins 80 % de taux de reconnaissance et d'UAR en utilisant des méthodes classiques basées sur les HMM-GMM et diverses caractéristiques du signal, dont les coefficients spectraux et cepstraux, la transformée en ondelettes, la vitesse d'élocution, l'énergie, etc. Des méthodes plus avancées et spécifiques à la tâche à accomplir peuvent cependant être nécessaires pour dépasser les 90 %. Aussi, aucune des systèmes de SER entrevues précédemment n'a atteint 70 % d'UAR en reconnaissance cross-corpus. Des efforts restent à faire, mais les avancées en apprentissage profond et en normalisation de caractéristiques sont prometteuses à ce sujet.

Les méthodes mentionnées dans cette synthèse cherchent à détecter une émotion par traitement direct du son. Toutefois, les processus émotifs sont multi-modaux, et il peut être plus avantageux d'intégrer les tâches de classification des recherches citées précédemment dans des modèles hiérarchiques et plus complexes [27]. Dans cette perspective, je me dirigerais vers les contributions de Gebhard [28] qui intègre la représentation plaisir-agitation-dominance (PDA) dans un modèle tripartite, de Becker-Asano and Wachsmuth [29] qui dissocient émotions primaires et secondaires, ou encore les expérimentations de Mencattini et al. [30] sur le modèle « circomplexe ». Enfin, Albornoz and Milone [31] proposent une méthode ensembliste pour la reconnaissance d'émotions dans des langues inconnues, ce qui ouvre des perspectives de recherche et suggère des progrès en SER cross-corpus.

Références

[1] C. Roswandowitz, S. R. Mathias, F. Hintz, J. Kreitewolf, S. Schelinski, K. von Kriegstein, Two cases of selective developmental voice-recognition impairments, Current Biology 24 (19) (2014) 2348–2353.

[2] A. Mucci, S. Galderisi, Cognitive dysfunctions in the psychoses and their impact on patients' social functioning, European Psychiatry 41 (Supplement) (2017) S48, abstract of the 25th European Congress of Psychiatry.

2. Un système « spécifique » réalise peu de faux positifs.

[3] H. Casey, R. D. Rogers, T. Burns, J. Yiend, Emotion regulation in psychopathy, Biological Psychology 92 (3) (2013) 541–548, specificity, Methodology and Psychopathology of Emotional Attention.

[4] L. J. Moskowitz, T. Rosen, M. D. Lerner, K. Levine, Assessment of Anxiety in Youth With Autism Spectrum Disorder, chap. 5, Academic Press, 79–104, 2017.

[5] N. Sest, E. March, Constructing the cyber-troll : Psychopathy, sadism, and empathy, Personality and Individual Differences 119 (Supplement C) (2017) 69–72.

[6] A. A. Colomeischi, Predictors for wellbeing : emotional factors and expectancy for success, Procedia - Social and Behavioral Sciences 190 (Supplement C) (2015) 48—53, proceedings of 2nd Global Conference on Psychology Researches (GCPR-2014) 28–29 November 2014, University of Barcelona, Barcelona, Spain 28-29 November 2014, University of Barcelona, Barcelona, Spain.

[7] K. Voltmer, M. von Salisch, Three meta-analyses of children's emotion knowledge and their school success, Learning and Individual Differences 59 (Supplement C) (2017) 107–118, ISSN 1041-6080.

[8] P. L. Mihaela, Psychological Factors of Academic Success, Procedia - Social and Behavioral Sciences 180 (Supplement C) (2015) 1632–1637, the 6th International Conference Edu World 2014 "Education Facing Contemporary World Issues", 7th–9th November 2014.

[9] J. S. Leventon, J. S. Stevens, P. J. Bauer, Development in the neurophysiology of emotion processing and memory in school-age children, Developmental Cognitive Neuroscience 10 (Supplement C) (2014) 21–33.

[10] A. Berent-Spillson, C. Marsh, C. Persad, J. Randolph, J.-K. Zubieta, Y. Smith, Metabolic and hormone influences on emotion processing during menopause, Psychoneuroendocrinology 76 (Supplement C) (2017) 218–225.

[11] T. Taguchi, H. Tachikawa, K. Nemoto, M. Suzuki, T. Nagano, R. Tachibana, M. Nishimura, T. Arai, Major depressive disorder discrimination using vocal acoustic features, Journal of Affective Disorders .

[12] B. Felbo, A. Mislove, A. Søgaard, I. Rahwan, S. Lehmann, Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm, in : Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 1616–1626, 2017.

[13] J. H. Janssen, P. Tacken, G.-J. de Vries, E. L. van den Broek, J. H. Westerink, P. Haselager, W. A. IJsselsteijn, Machines outperform laypersons in recognizing emotions elicited by autobiographical recollection, Human–Computer Interaction 28 (6) (2013) 479–517.

[14] L.-A. Perez-Gaspar, S.-O. Caballero-Morales, F. Trujillo-Romero, Multimodal emotion recognition with evolutionary computation for human-robot interaction, Expert Systems with Applications 66 (Supplement C) (2016) 42–61.

[15] A.-R. Mohamed, Deep neural network acoustic models for ASR, Ph.D. thesis, University of Toronto, 2014.

[16] R. Sharma, L. Vignolo, G. Schlotthauer, M. Colominas, H. L. Rufiner, S. Prasanna, Empirical mode decomposition for adaptive AM-FM analysis of speech : A review, Speech Communication 88 (Supplement C) (2017) 39–64.

[17] L. D. Vignolo, S. M. Prasanna, S. Dandapat, H. L. Rufiner, D. H. Milone, Feature optimisation for stress recognition in speech, Pattern Recognition Letters 84 (Supplement C) (2016) 1–7.

[18] B. Z. Mansouri, H. Mirvaziri, F. Sadeghi, Designing and implementing of intelligent emotional speech recognition with wavelet and neural network, International Journal of Advanced Computer Science and Applications (IJACSA) 7 (9) (2016) 26–30.

[19] H. K. Palo, M. N. Mohanty, Wavelet based feature combination for recognition of emotions, Ain Shams Engineering Journal .

[20] M. Tahon, L. Devillers, Towards a small set of robust acoustic features for emotion recognition : Challenges, IEEE/ACM Transactions on Audio, Speech and Language Processing 24 (2016) 16–28.

[21] F. Chenchah, Z. Lachiri, A bio-inspired emotion recognition system under real-life conditions, Applied Acoustics 115 (Supplement C) (2017) 6–14.

[22] Q. Mao, G. Xu, W. Xue, J. Gou, Y. Zhan, Learning emotion-discriminative and domain-invariant features for domain adaptation in speech emotion recognition, Speech Communication 93 (Supplement C) (2017) 1–10.

[23] I. S. Hong, Y. J. Ko, H. S. Shin, Y. J. Kim, Emotion recognition from Korean language using MFCC, HMM, and speech speed, in : The 12th

International Conference on Multimedia Information Technology and Applications (MITA2016), 12–15, 2016.

[24] J. Kim, G. Englebienne, K. P. Truong, V. Evers, Towards speech emotion recognition "in the wild" using aggregated corpora and deep multi-task learning, Computing Research Repository abs/1708.03920.

[25] H. M. Fayek, M. Lech, L. Cavedon, Evaluating deep learning architectures for Speech Emotion Recognition, Neural Networks 92 (Supplement C) (2017) 60–68, advances in Cognitive Engineering Using Neural Networks.

[26] H. Kaya, A. A. Karpov, Efficient and effective strategies for cross-corpus acoustic emotion recognition, Neurocomputing .

[27] M. Tahon, Acoustic analysis of speakers emotional voices during a human-robot interaction, Ph.D. thesis, Université Paris Sud — Paris XI, 2013.

[28] P. Gebhard, ALMA : a layered model of affect, in : 4th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2005), July 25-29, 2005, Utrecht, The Netherlands, 29–36, 2005.

[29] C. Becker-Asano, I. Wachsmuth, Affect Simulation with Primary and Secondary Emotions, Springer Berlin Heidelberg, Berlin, Heidelberg, 15–28, 2008.

[30] A. Mencattini, E. Martinelli, G. Costantini, M. Todisco, B. Basile, M. Bozzali, C. Di Natale, Speech emotion recognition using amplitude modulation parameters and a combined feature selection procedure, Knowledge-Based Systems 63 (2014) 68–81, ISSN 0950-7051.

[31] E. M. Albornoz, D. H. Milone, Emotion recognition in never-seen languages using a novel ensemble method with emotion profiles, IEEE Transactions on Affective Computing 8 (1) (2017) 43–53.