Classification d'Expressions Vocales Passives Versus Actives

Z. HAMMAL[1], B. BOZKURT[2], L. COUVREUR[2], D. UNAY[2], A. CAPLIER[1] et T. DUTOIT[2]

[1]Laboratoire des images et des signaux LIS
46 avenue Félix Viallet, Grenoble, France
[2] Laboratoire de traitement de signal , Faculté Polytechnique de Mons
1 Avenue Copernic, B-7000, Mons , Belgique
[1] hammal,caplier@lis.inpg.fr, zakia_hammal@yahoo.fr
[2]bozkurt,couvreur,unay,dutoit@tcts.fpms.ac.be

Résumé – *Six expressions sont généralement considérées pour caractériser les états émotifs humains : Sourire, Surprise, Colère, Tristesse, dégoût et Neutre. Différentes mesures peuvent être extraites à partir du signal de parole pour caractériser ces expressions, à savoir la fréquence fondamentale, l'énergie, le SPI (rapport des énergies des HF et des BF dans le signal) et le débit de parole. Une classification automatique des cinq expressions basées sur ces caractéristiques présente des conflits entre la Colère, la Surprise et le Sourire d'une part et le Neutre et la Tristesse d'autre part. Ce conflit entre classes d'expressions est également retrouvé chez le classifieur humain. Nous proposons donc de définir deux classes d'expressions: Active regroupant le Sourire, la Surprise et la Colère et Passive regroupant le Neutre et la Tristesse. Une telle classification est également plus réaliste et plus appropriée pour l'intégration d'information de parole dans un système de classification multimodale combinant la parole et la vidéo, ce qui est à long terme le but de notre travail. Dans ce papier, différentes méthodes de classification sont testées: un classifieur Bayésien, une Analyse Discriminante Linéaire (ADL), le classifieur au K plus proches vosins(KNN) et un classifieur à Machine à Vecteur de Support (SVM) avec une fonction de base gaussienne. Pour les deux classes considérées, les meilleurs taux de classification sont obtenus avec le classificateur SVM avec un taux de reconnaissance de 89.74% pour l'état Actif et de 86.54 % pour l'état Passif.*

Abstract *– Six expressions are commonly considered to characterize human emotional states: Happiness, Surprise, Anger, Sadness, disgust and Neutral. Different measures can be extracted from speech signals to characterize these expressions, for example the pitch, the energy, the SPI and the speech rate. Automatic classification of the five expressions based on these features shows a great confusion between Anger, Surprise and Happiness on the one hand and Neutral and Sadness on the other hand. Such a confusion is also observed when humans make the same classification. We propose to define two classes of expression: Active gathering Happiness, Surprise and Anger versus Passive gathering Neutral and Sadness. Such a partition is also better suited for the integration of speech information in a multimodal classification system based on speech and video, which is the long term aim of our work. In this paper, we test several classification methods, namely a Bayesian classifier, a Linear Discriminant Analysis (LDA), the K Nearest Neighbours (KNN) and a Support Vector Machine with Gaussian radial basis function kernel (SVM). For the considered two classes, the best performances are achieved with the SVM classifier with a recognition rate of 89.74% for Active state and of 86.54 % for Passive state.*

1 . Introduction

Les interfaces utilisateurs des systèmes informatiques évoluent vers des interfaces intelligentes et multimodales. Elles prennent en considération aussi bien les gestes de l'utilisateur, sa voix ainsi que ses expressions faciales dans le but de rendre la communication homme-machine aussi proche que possible d'une communication homme-homme[1]. Il existe de nombreux travaux dédiés à l'analyse et la reconnaissance soit des expressions vocales soit des expressions faciales. Dans [2] les auteurs indiquent que l'utilisation de l'une des deux modalités est liée aussi bien aux expressions traitées qu'au contexte de leurs applications. Le travail présenté s'inscrit dans la continuité de nos travaux sur la reconnaissance des expressions faciales [3] basée sur l'analyse d'informations vidéos dans le but de développer un système multimodal de reconnaissance d'expressions.

Plusieurs travaux ont été menés sur l'analyse et la reconnaissance des expressions vocales : analyse de caractéristiques [4,5], classification des expressions vocales [6,7].

Contrairement à ces travaux qui se sont efforcés de déterminer des caractéristiques permettant de reconnaître et dissocier un nombre prédéfini de classes d'expressions, notre objectif est de trouver des classes plus générales et plus réalistes.

Dans un premier temps, une classification en 5 classes d'expressions a été envisagée sur la base d'expressions vocales en langue danoise DES [8]. Cette dernière est composée uniquement de 5 des 6 expressions universelles: *Colère*, *Surprise*, *Sourire*, *Neutre* et *Tristesse*.

Conformément à ce qui a été rapporté dans des travaux précédents [6,7], des confusions entre groupes d'expressions ont été obtenues. Il s'avère que ce sont les mêmes confusions que celles obtenues lors d'une classification par un humain. Ce qui nous amènent à nous intéresser à deux classes d'expressions seulement : la classe des voix *Actives* regroupant la *Joie*, la *Surprise* et la *Colère* et la classe des voix *Passive* regroupant le *Neutre* et la *Tristesse*.

Pour valider le bien-fondé de ces deux nouvelles classes, plusieurs méthodes de classification utilisant un ensemble de caractéristiques statistiques acoustiques sont testées: un classifieur

Bayesien, une classification par Analyse Discriminante Linéaire (ADL), un classifieur aux K plus proches voisins (KNN) et un classifieur à Machine à Vecteur de Support (SVM).
Dans la section 2 est présentée la base d'expressions vocales utilisée. Dans la section 3 est présentée l'analyse des différentes caractéristiques prosodiques utilisées. La section 4 est consacrée aux résultats et à la discussion.

2. Base de données

La base d'expressions vocales DES [8] a été utilisée pour nos expériences. Cette dernière regroupe des expressions vocales de quatre acteurs professionnels (deux hommes et deux femmes). Cinq expressions ont été simulées: *Neutre*, *Surprise*, *Joie*, *Tristesse* et *Colère*. Pour chaque expression, il y a 2 mots simples, 9 phrases et 2 passages de discours continu. La validation des enregistrements a été faite par 40 auditeurs (20 hommes et 20 femmes). Les expressions ont été correctement identifiées avec un taux moyen de 67%. La *Surprise* et le *Sourire* sont souvent confondus ainsi que le *Neutre* et la *Tristesse* (TAB.1). Nous considèrerons que cette matrice de confusion représente la vérité de terrain.

TAB.1 Matrice de confusion à partir de l'évaluation subjective humaine [8]. Les colonnes représentent l'expression vocal à reconnaître et les lignes l'expression sélectionnée par le classifieur humain.

	Neutre	Surprise	Sourire	Tristesse	Colère
Neutre	**60.8**	2.6	0.1	31.7	4.8
Surprise	10.0	**59.1**	28.7	1.0	1.3
Sourire	8.3	29.8	**56.4**	1.7	3.8
Tristesse	12.6	1.8	0.1	**85.2**	0.3
Colère	10.2	8.5	4.5	1.7	**75.1**

3. Extraction et analyse de caractéristiques

En s'appuyant sur de récents travaux en analyse d'expressions vocales [3,4,5,9], plusieurs caractéristiques prosodiques ont été extraites et analysées : la fréquence fondamentale, l'énergie, le SPI et le débit de parole.
La fréquence fondamentale, l'énergie et le SPI sont calculés dans une fenêtre glissante de largeur constante de 30ms avec un pas de 10ms.
Un ensemble de paramètres statistiques normalisés (centrés réduits) ont été calculés et analysés pour chaque caractéristique (TAB.2) .

TAB.2 Paramètres statistiques utilisés pour chaque caractéristique prosodique
('x': utilisé, '-': non utilisé).

	Inter-valle	Média-ne	Ecart-type	Fronts-montants	Fronts-descendants	Maxi-mum
F0	x	x	x	x	x	-
Energie	x	x	x	x	x	-
SPI	-	-	-	-	-	x
Vitesse de parole						

3.1 Fréquence fondamentale et énergie

La fréquence fondamentale (F0) est calculée par un estimateur d'auto corrélation [10]. Tandis que l'énergie est calculée (en décibels) comme la somme des carrés des échantillons du signal discret [10]. On ne prend en considération que l'énergie des zones voisées des signaux de parole.
Pour la fréquence fondamentale F0 et l'énergie du signal de parole sont calculés : le minimum, le maximum, la moyenne, la médiane, l'intervalle et l'écart type ainsi que les médianes des fronts montants et descendants. Les FIG.1.a et FIG.2.a présentent les valeurs de l'écart type, de l'intervalle (maximum-minimum) et de la médiane de F0 (resp. de l'énergie). La valeur de chaque barre correspond à la valeur moyenne pour chaque expression sur l'ensemble des données. Les FIG.1.b, FIG.2.b présentent les médianes des fronts montants et descendants de F0 (resp. de l'énergie) pour chaque expression sur l'ensemble des données. L'analyse de toutes ces statistiques montre que ces dernières sont en moyenne plus importantes pour la *Joie*, la *Surprise* et la *Colère* que pour la *Tristesse* et le *Neutre* aussi bien pour F0 que pour l'énergie. Les FIG.1 et FIG.2 mettent ainsi en évidence deux groupes d'expressions : la *Surprise*, la *Colère* et la *Joie* d'un côté et le *Neutre* et la *Tristesse* de l'autre.

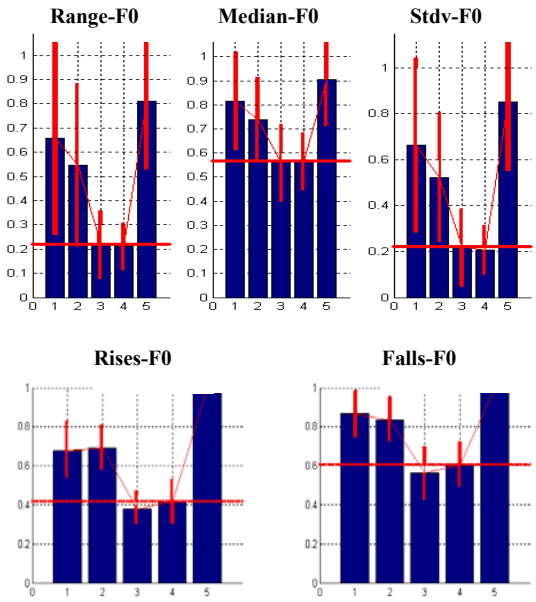

FIG.1 Valeurs des paramètres statistiques de F0 pour toutes les données et toutes les expressions : intervalle (Range), médiane (Median), ecart-type (Stdv), Fronts-montants (rises), Fronts-descendants (falls). Les barres représentent les expressions dans l'ordre suivant : 1) *Colère*, 2) *Joie*, 3) *Neutre*, 4) *Tristesse*, 5) *Surprise*.

3.2 Le SPI

Le SPI est une mesure spectrale du rapport des énergies des basses fréquences (70-1600HZ) sur les hautes fréquences (1600-4500Hz) [11]. L'analyse des caractéristiques statistiques du SPI montrent que la seule caractéristique intéressante pour la classification des expressions vocales est son maximum. Cette caractéristique permet également

de faire une séparation en deux classes (les deux mêmes classes que pour F0 et l'énergie) (FIG.3.b).

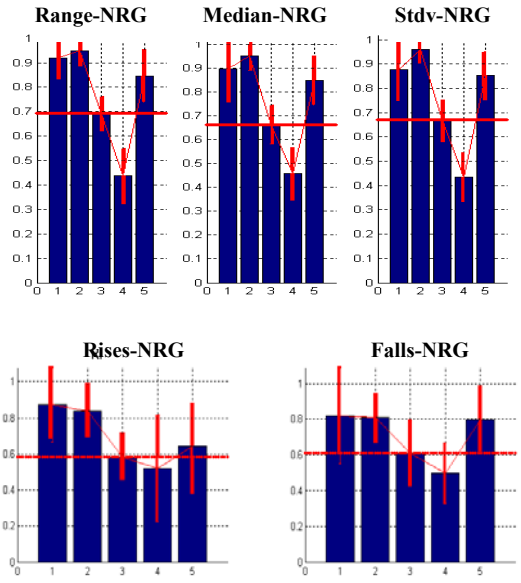

FIG.2 Valeurs des paramètres statistiques de l'énergie pour toutes les données et toutes les expressions : intervalle (Range), médiane (Median), ecart-type (Dtdv), Fronts-montants (Rises), Fronts-descendants (Falls). Les barres représentent les expressions dans l'ordre suivant : 1) *Colère*, 2) *Joie*, 3) *Neutre*, 4) *Tristesse*, 5) *Surprise*.

3.3 Le débit de parole

Le débit de parole est calculé pour chaque enregistrement comme le nombre de phonèmes prononcés dans un intervalle de temps donné. Le nombre de phonèmes de chaque enregistrement est a priori connu pour la base de données utilisée. L'analyse du débit de parole (FIG.3.a) montre que celui-ci est en moyenne plus important pour la *Surprise*, la *Colère* et la *Joie* que pour le *Neutre* et la *Tristesse*. Ceci conduit à la même conclusion que précédemment.

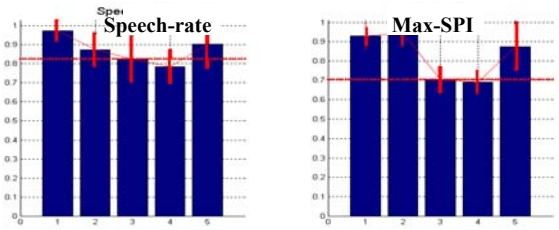

FIG.3 (a) : Valeurs du débit de parole (speech-rate), (b) : valeurs maximales du SPI sur l'ensemble des données (Max). Les barres représentent les expressions dans l'ordre suivant : 1) *Colère*, 2) *Joie*, 3) *Neutre*, 4) *Tristesse*, 5) *Surprise*.

3.4 Conclusion de l'analyse

Les expressions telles que la *Colère*, *Surprise*, *Joie* sont caractérisées par des valeurs de F0, d'énergie et de débit de parole en moyenne plus élevées, ce qui signifie qu'elles sont caractéristiques d'une plus forte activité vocale. A l'inverse, les expressions telles que la *Tristesse* ou le *Neutre* sont associées à des valeurs de F0, d'énergie et de débit de parole moins importantes ce qui signifie qu'elles sont caractéristiques d'une moins forte activité vocale. Ces observations nous amènent à conclure que les paramètres acoustiques les plus fréquemment utilisés ne permettent pas de distinguer les 5 expressions considérées ici. En revanche, ces paramètres ont le même comportement pour deux groupes d'expressions : *Colère*, *Surprise* et *Joie* d'un côté ; *Tristesse* et *Neutre* d'un autre côté.

4. Résultats et Discussion

4.1 Classification en 5 classes

L'analyse de la section 3 montre qu'il y a des similarités prosodiques entre plusieurs expressions vocales. En premier lieu, nous cherchons à confirmer ces résultats et à voir le pouvoir discriminant des paramètres calculés pour la classification des expressions vocales. Les 12 caractéristiques acoustiques décrites dans la section 2 sont donc utilisées avec un classificateur Bayésien pour reconnaître les cinq expressions vocales. Pour palier le manque de données, une technique de Bootstrapping [12] est utilisée sur l'ensemble des données à classer.

TAB.3 présente les résultats de la classification. Dans le but de tester la validité de nos caractéristiques, nous les comparons à ceux de [7] obtenues sur la même base de données. En utilisant plus de caractéristiques (un vecteur de 15 caractéristiques comprenant des statistiques sur la fréquence fondamentale, l'énergie et les 4 premiers formants), le taux moyen de classification obtenu dans [7] est autour de 50% (*Neutre* (51%), *Surprise* (64%), *Joie* (36%), *Tristesse* (70%) et *Colère* (31%)) tandis que le nôtre est autour de 54%. De plus, nos résultats sont plus homogènes : le taux de classification est presque le même pour toutes les expressions ce qui n'est pas le cas dans [7] qui présente un très faible taux de classification pour la *Joie* et la *Colère*.

TAB.3 Matrice de confusion du classifieur Bayesien

	Neutre	Surprise	Joie	Tristesse	Colère
Neutre	**46.76**	23.92	12.26	3.3	13.73
Surprise	20.11	**51.69**	6.5	5	16.61
Joie	7.11	5	**56.61**	24.69	6.5
Tristesse	4.57	3.19	28.76	**61.80**	1.65
Colère	12.5	29.11	4.26	1.84	**52.26**

4.2 Classification en 2 classes

Nos résultats de classification montrent qu'il y a des confusions entre groupes d'expressions. Il s'avère que ce sont les mêmes confusions que celles obtenues lors d'une classification par un humain (TAB.1). Ceci nous amène à dire que si on considère les confusions non pas comme un problème à résoudre mais comme un indice de similitude entre expressions, il est possible de définir deux classes d'expressions seulement : *Active* qui regroupe la *Colère*, la *Joie* et la *Surprise* et *Passive* qui regroupe le *Neutre* et la *Tristesse*.

Pour valider le bien-fondé de ces deux nouvelles classes, quatre classifieurs sont testés : le classifieur Bayésien, le

classifieur par Analyse Discriminante Linéaire (ADL) [13], le classifieur aux K plus prches voisins (KNN) [13] et un classifieur à Machine à Support Vecteur (SVM) [14]. Les taux de classification sont obtenus par validation croisée. Les résultats de la classification (TAB. 4- à 7) montrent que les taux classification du classifieur Bayésien et du classifieur parADL sont inférieurs à ceux du classifieur SVM et du classifieur KNN. Ceci est dû au fait que le classifieur Bayésien suppose une distribution Bayésienne des classes, ce qui n'est pas forcement le cas de nos données.Le classifieur par ADL quant à lui effectue une séparation linéaire alors que nos données peuvent être non-linéairement séparables. Le classifieur KNN obtient de meilleurs résultats que le classifieur par ADL . Toutefois, les SVM donnent les meilleurs taux de classification (TAB. 7). L'ensemble des résultats présentés (TAB. 6 et7) permet de confirmer que les caractéristiques utilisées sont nécessaires et suffisantes pour une classification en deux classes d'expressions vocales.

TAB.4 Résultats de la classification par le Bayesien.

	Active	Passive
Active	**78.84%**	21.15%
Passive	19.23%	**80.76%**

TAB.5 Résultats de la classification par l'ADL.

	Active	Passive
Active	**96.79%**	3.2%
Passive	46.15%	**53.85%**

TAB.6 Résultats de la classification par KNN.

	Active	Passive
Active	**83.33%**	16.67%
Passive	11.54%	**88.46%**

TAB.7 Résultats de la classification par SVM.

	Active	Passive
Active	**89.74%**	10.26%
Passive	13.46%	**86.54%**

5. Conclusion

Afin d'intégrer la modalité parole à notre système de classification d'expressions faciales, nous avons étudié les propriétés acoustiques du signal de parole pour cinq expressions vocales (*Surprise*, *Joie*, *Colère*, *Neutre* et *Tristesse*). Cette analyse a permis de constater que les caractéristiques acoustiques considérées ne sont pas suffisantes pour séparer les cinq expressions vocales. Cependant, les résultats montrent que ces expressions se regroupent en deux classes plus larges: *Joie*, *Colère* et *Surprise* d'un côté et *Neutre* et *Tristesse* de l'autre. Ceci nous a conduit à définir deux classes d'expressions vocales : *Active* et *Passive*. Sur la base de ces classes, les résultats de cette classification sont très satisfaisants.

Remerciements :

Ce travail a été effectué dans le cadre d'une collaboration supportée par le réseau d'excellence Similar.

Bibliographie

[1] http://www.similar.cc
[2] P. Ekman, W.V. Friesen, M. O'Sullivan & K. Scherer, *Relative importance of face, body, and speech in judgments of personality and affect*, Journal of Personality and Social Psychology, vol. 38(2), pp. 270-277, 1980.
[3] Z. Hammal, A. Caplier & M. Rombaut, *Classification d'expressions faciales par la théorie de l'évidence,* Rencontre Francophones sur la Logique Floue et ses Applications, pp. 173-180, Nantes, France, 2004.
[4] K. R. Scherer, *Vocal communication of emotion: A review of research paradigms*, Speech Communication, vol. 40(1-2), pp. 227-256, 2003.
[5] P. N. Juslin & P. Laukka, *Communication of emotions in vocal expression and music performance: Different channels, same code?,* Psychological Bulletin, pp. 770-814, 2003.
[6] V. A. Petrushin, *Emotion recognition in speech signal: experimental study, development, and application*, in Proc. Of International Conference on Spoken Language processing (ICSLP), Beijing, China, 2000.
[7] D. Ververidis & C. Kotropolos, *Automatic speech classification to five emotional states based on gender information*, in Proc. of European Conference on Signal Processing (EUSIPCO), pp. 341-344, Vienna, Austria, 2004.
[8] I. S. Engberg & A. V. Hansen, *Documentation on the Danish Emotional Speech Database (DES)*, Technical Report, Alborg University, Denmark, 1996.
[9] M. Schröder, "Speech and Emotion Research", PhD thesis, university of Saarlandes, 2003.
[10] T. F. Quatieri, "Discrete Time Speech Signal Processing: Principles and Practice", Prentice Hall PTR, 2001.
[11] D. Deliyski,, *Acoustic model and evaluation of pathological voice production*, in Proc. of European Conference on Speech Communication and Technology (EUROSPEECH), pp. 1969-1972, Berlin, Germany, 1993.
[12] R. Kallel, M. Cottrell & V. Vigneron, "Bootstrap for neural model" , Neurocomputing, vol 48, pp.175-183, 2002.
[13] R. O. Duda, P. E. Hart & D. G. Stork, "Pattern Classification" (2nd ed.), John Wiley and Sons, 2001.
[14] J.C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition", Data Mining and Knowledge Discovery, Kluwer Academic Publishers, pp. 121–167, 1998.