Une Approche Basée Voyelle pour la Reconnaissance d'Émotions Actées

Fabien Ringeval, Mohamed Chetouani

Institut des Systèmes Intelligents et Robotique
3 rue Galilée, 94 200 Ivry sur Seine, France
fabien.ringeval@isir.fr, mohamed.chetouani@upmc.fr

ABSTRACT

This paper is dedicated to the description and the study of a new feature extraction approach for emotion recognition. Our contribution is based on the extraction and the characterization of phonemic units such as vowels and consonants, which are provided by a pseudo-phonetic speech segmentation phase combined with a vowel detector. Concerning the emotion recognition task, we propose to extract both MFCC acoustic and prosodic features from these pseudo-phonetic segments (vowels and consonants), and we compare this approach with traditional voiced and unvoiced segments. The classification is achieved by the well-known k-nn classifier (k nearest neighbors) on the Berlin corpus.

Keywords: Emotion recognition, automatic speech segmentation, vowel detection

1. INTRODUCTION

La manifestation des émotions est un domaine particulièrement complexe de la communication humaine, et concerne des sciences pluridisciplinaires telles que la psychologie, la cognition, la sociologie, la philosophie et les sciences computationnelles orientées émotion. Cette pluridisciplinarité, qui est due à la haute variabilité du comportement humain, influence à la fois la production et la perception de ses émotions. L'affect (les sentiments et leurs changements physiques associés), la cognition, la personnalité, la culture et l'éthique sont les principales propriétés des émotions décrites dans la littérature [5]. Bien que certaines études référencent plus d'une centaine de termes reliés aux émotions [2], six émotions primaires qualifiées de 'full-blown' sont largement acceptées dans la littérature : peur, colère, joie, ennui, tristesse et dégoût. Plutchik [13] postule que les émotions plus complexes seraient des états dérivés ou mixés, et apparaîtraient comme des combinaisons, des mélanges ou composés des émotions primaires.

Les sciences computationnelles orientées émotion visent à la reconnaissance et à la synthèse automatique des émotions dans la parole, les expressions faciales, ou tout autre canal de communication biologique [12]. L'une des difficultés majeures rencontrées dans leur classification réside en la détermination de leurs caractéristiques et des classifieurs [18]. D'autres problèmes apparaissent quant à la définition même des émotions et de leur annotation [7]. Les méthodes d'extraction de caractéristiques usuellement utilisées reposent sur des paramètres acoustique et prosodique, regroupés dans un vecteur de caractéristiques. Les paramètres acoustiques sont dérivés du traitement de la parole (e.g. MFCC, LPCC), tandis que la prosodie est caractérisée par des mesures statistiques de ses principales composantes (pitch, énergie et durée des sons produits) calculés sur les segments voisés. Les algorithmes de classification reposent sur des méthodes d'apprentissage telles que les mesures de distance (k-ppv), les arbres de décisions binaires, les mixtures de modèles gaussiens (GMM), les machines à support vecteur (SVM) [16].

2. UNE APPROCHE BASÉE VOYELLE

Nous proposons dans cet article un nouveau schéma d'extraction de caractéristiques basé sur une approche pseudo-phonétique (figure 1). Le point clé de nos travaux est d'extraire les caractéristiques selon différents segments tels que les voyelles et les consonnes. Ces segments sont identifiés par une segmentation du signal de parole en zones stationnaires (DFB), combinée avec un détecteur de voyelles. Le processus de segmentation est indépendant de la langue et ne vise pas à l'identification exacte des phonèmes comme pourrait le faire un alignement phonétique. Les segments obtenus lors de cette phase sont alors qualifiés d'unités pseudo-phonétiques.

Cette méthode a été introduite pour la reconnaissance automatique des langues [15]. Des unités similaires aux syllabes (pseudo-syllabes) y sont alors définies par le regroupement des consonnes précédent les voyelles détectées (structure de type C^nV). L'étude des pseudo-syllabes (durée et complexité) a permis de caractériser deux groupes de langue décrites dans la littérature : accentuelle (anglais, allemand et mandarin) et syllabique (français et espagnol).

2.1. Description du corpus Berlin

La base de données Berlin [3] contient de la parole émotionnelle (six émotions primaires et une 'neutre'), et a été étudié en reconnaissance automatique d'émotions par de nombreux auteurs [6,17,19]. 10 phrases (5 longues et 5 courtes) issues de discussions de tous les jours ont été émotionnellement colorées par 10 acteurs allemands (cinq hommes et cinq femmes) sur des équipements de haute qualité (chambre anéchoïque). 535 phrases évaluées alors comme naturelles et reconnaissable par 20 personnes (à respectivement 60% et 80% minimum) lors

d'un test de perception ont été conservées, et étiquetées dans une transcription comprenant un alignement temporel en mots et phonèmes. Le corpus Berlin possède un lexique de 59 phonèmes : 24 voyelles et 35 consonnes.

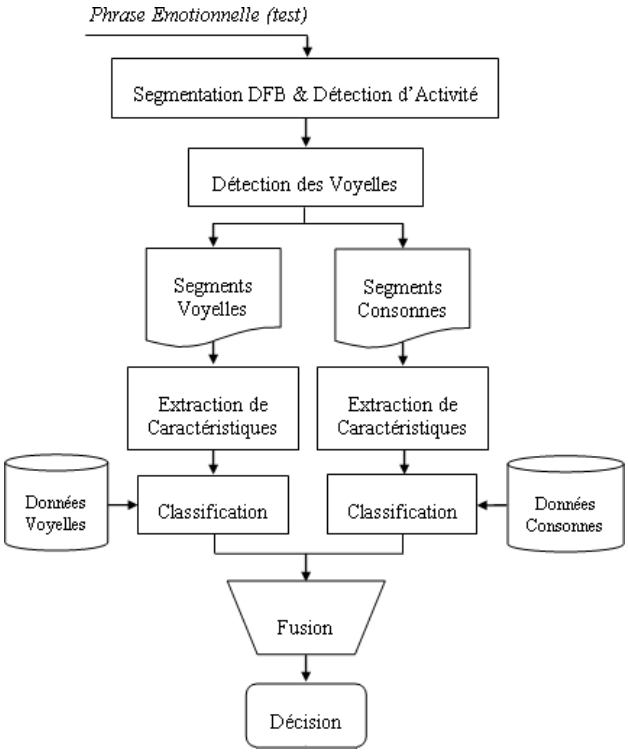

Figure 1 : Approche pseudo-phonétique pour la reconnaissance automatique des émotions.

2.2. Extraction d'unités pseudo-phonétique

La méthode de segmentation pseudo-phonétique repose sur l'algorithme Divergence Forward Backward (DFB) [1]. Trois types de segments y sont alors identifiés : courts, transitoires et quasi-stationnaires. Les segments contenant de la parole sont détectés par un seuillage de leur variance, et les voyelles sont identifiées par la mesure spectrale proposée par F. Pellegrino et al. [11] : 'Reduced Energy Cumulating' (REC) (équation 1). Le signal est pour cela segmenté en trames. N valeurs d'énergie E_i sont alors extraites pour chaque trame k, et celles supérieures à leur valeur moyennes respectives $\bar{E}$ sont cumulées puis pondérées par la proportion d'énergie contenue dans les basses fréquences (rapport d'énergie entre les basses fréquences E_{BF} et le spectre total E_{T}). Les segments de parole issus du DFB sont étiquetés 'voyelle' lorsqu'un pic de la courbe REC y est détecté. Ceux n'étant pas identifiés comme 'voyelle' sont alors étiquetés 'consonne'.

$$\mathrm{Rec}(k) = \frac{E_{LF}(k)}{E_T(k)} \sum_{i=1}^{N} \left(E_i(k) - \bar{E}(k) \right)^+ \qquad (1)$$

Le principe de détection des voyelles décrit ci-dessus a été évalué lors d'une précédente étude [14] sur trois corpus phonétiquement étiquetés. Le premier est Berlin, et contient de la parole émotionnelle (Allemand), tandis que les deux autres contiennent de la parole lue (Anglais américain) pour une qualité laboratoire ou téléphonique (TIMIT [8], NTIMIT [10]). La mesure de performance utilisée pour évaluer le détecteur de voyelles est le 'Vowel Error Rate' (VER) [15]. Cette mesure regroupe les taux de non-détection et d'insertion des voyelles par rapport aux transcriptions phonétiques contenues dans les données. Les résultats obtenus (table 1) sont en accord avec les travaux menés sur la contribution de la qualité vocale selon les émotions. Elles sont en effet les mieux détectées comme les plus confondues. Les performances globales de notre système sont d'autre part très bonnes sur l'ensemble des données puisque le VER augmente de moins de 4% pour une configuration unique du détecteur.

Table 1 : Performances du détecteur de voyelles.

Corpus	**Référence**s (quantité)	**Détection**s (en %)	**Insertion**s (en %)	**VER** (en %)
Berlin	6437	89.96	19.05	29.08
TIMIT	57501	87.56	7.07	19.50
NTIMIT	57493	81.06	5.13	24.07
Tous	121431	84.62	6.79	22.17

3. RECONNAISSANCE ACOUSTIQUE

Deux approches sont étudiées pour la reconnaissance acoustique des émotions : celle issue de l'état de l'art reposant sur des segments voisés, et une utilisant les unités pseudo-phonétiques (voyelles et consonnes). La figure 1 illustre les méthodes employées pour les phases d'extraction de caractéristiques et de classification. Durant l'approche voisée, le signal de parole est segmenté en trames de 32ms avec un recouvrement de moitié. Un détecteur de voisement est ensuite utilisé pour distinguer les trames voisées de celles non voisées. Pour l'approche basée voyelles, la segmentation en trames est réalisée sur les unités pseudo-phonétiques (32ms), selon les voyelles et les consonnes. L'extraction de caractéristiques est obtenue par le calcul de 24 coefficients MFCC (Mel Frequency Cepstrum Coding). Lors de la phase de classification, les vecteurs d'énergie MFCC sont étiquetés par la méthode des k-ppv (k plus proches voisins) pour chaque trame testée. L'estimation des scores repose sur 5 validations statistiques croisées (n-fold cross validation). Différentes répartitions des données en apprentissage et en test y sont réalisées afin de minimiser le risque empirique.

3.1. Fusion voisé et non voisé

La classification des paramètres MFCC (calculés sur les segments voisés et non voisés) produit deux vecteurs d'étiquettes émotionnelles (V et NV). Afin de les fusionner, nous estimons tout d'abord leur probabilités conditionnelles $p(C_i \mid \mathrm{V})$ et $p(C_i \mid \mathrm{NV})$ selon les sept émotions du corpus Berlin (C_1 *à* C_7). Deux approches

sont ensuite employées pour les fusionner : statique et dynamique [4]. La méthode de fusion mise en œuvre est une combinaison linéaire des probabilités conditionnelles. La décision de l'émotion reconnue selon les segments voisés V et non voisés NV est prise par l'équation 2 :

$$E = \text{argmax}\left(\lambda_V * p(C_i \mid \text{V}) + \lambda_{NV} * p(C_i \mid \text{NV})\right) \quad \textbf{(2)}$$

La différenciation entre la fusion statique et dynamique apparaît lors de l'estimation des poids des classifieurs voisé λ_V et non voisé λ_{NV}. Pour la fusion statique (équation 2), les poids sont fixes durant toute la phase de test, et sont estimés par la combinaison aboutissant au meilleur score. Alors que la fusion dynamique utilise des poids différents pour chaque fichier testé. Ces poids sont définis par le taux de voisement *r* correspondant à la proportion de trames voisées contenues dans une phrase. Une fonction puissance est ensuite appliquée pour paramétrer la vitesse de variation des poids de fusion r^{α}. La constante α été optimisée par la base de test (α = 0.2).

$$E = \text{argmax}\left(r^{\alpha} * p(C_i \mid \text{V}) + \left(1 - r^{\alpha}\right) * p(C_i \mid \text{NV})\right) \quad \textbf{(3)}$$

Comme on peut s'y attendre, les performances obtenues en classification des trames voisées 70.10% est bien meilleure que celles non voisées 42.33%, ce qui n'est pas non plus un mauvais score comparé au classifieur naïf 24.27% (retournant en permanence la classe la plus représentée).

3.2. Fusion voyelle et consonne

Deux vecteurs d'étiquettes émotionnelles sont obtenus par la classification des unités pseudo-phonétiques. Les données contenues dans la transcription phonétique ont été exploitées lors du calcul des coefficients MFCC. Le même principe de classification utilisé pour la fusion statique des classifieurs voisés est employé sur les segments 'voyelle' et 'consonne' :

$$E = \text{argmax}\left(\lambda_{Voy} * p(C_i \mid Voy) + \lambda_{Csn} * p(C_i \mid Csn)\right) \quad \textbf{(4)}$$

Comme la segmentation DFB tend à sur-segmenter les consonnes (ratio voyelle/consonne de 2.09 contre 1.69 pour les références), la fusion dynamique n'a pas été étudiée. Les taux de reconnaissance obtenus par les voyelles de référence en fusion statique (λ_{Voy} = 0.8 et λ_{Csn} = 0.5) sont très proches de celles détectées (λ_{Voy} = 0.3 et λ_{Csn} = 0.2) (table 2). Malgré que les unités phonétiques aboutissent aux meilleurs scores, elles apparaissent comme moins complémentaires que les unités pseudo-phonétiques.

Table 2 : Taux de reconnaissance acoustique selon les classifieurs pour les deux approches étudiées.

Classifieur	**Score en %**
Voisé	70.10
Non Voisé	42.33
Fusion Statique	70.10
Fusion Dynamique	70.68
Voyelles (références)	72.13
Consonnes (références)	65.89
Fusion	73.10
Voyelles (détectées)	71.85
Consonnes (détectées)	57.28
Fusion	73.98
Naïf	24.27

Bien que la quantité d'informations disponible sur les segments voisés soit bien plus importante que sur les segments voyelles, les meilleurs scores sont obtenus sur ces derniers. Les scores issus de l'état de l'art sont équivalents à ceux obtenus sur les voyelles détectées. Ces résultats sont très intéressants puisqu'ils mettent en lumière le poids des voyelles dans la perception des émotions.

4. RECONNAISSANCE PROSODIQUE

Les segments voisés et ceux identifiés comme 'voyelle' sont exploités pour l'extraction des caractéristiques prosodiques. Les paramètres mesurés sont communs aux deux approches, et reposent sur le pitch, l'énergie et la durée des sons produits. Les deux premiers paramètres sont caractérisés par divers mesures statistiques (table 3), tandis que les durées sont modélisées par la mesure proposée par Grabe : 'Pairwise Variability Indice' (PVI) [9]. Cette mesure vise à quantifier la variabilité des durées d_k des *N* intervalles vocaliques ou intra vocaliques successifs (équation 5). Elle a été utilisée pour caractériser les dialectes de l'anglais britannique.

$$PVI = \frac{1}{N\text{-}1}\sum_{k=1}^{N-1}\left|d_k - d_{k+1}\right| \quad \textbf{(5)}$$

Le classifieur nous retourne un vecteur de probabilité conditionnelle $p(C_i \mid \text{Seg}_x)$ pour chaque segment testé Seg_x (voisé ou voyelle). Ces vecteurs de probabilité sont ensuite normalisés selon la longueur respective des segments Seg_x. La décision de l'émotion est obtenue par la fonction *argmax* calculée sur la somme des probabilités normalisées. Les résultats obtenus par les modèles prosodiques confirment l'intérêt des voyelles pour la reconnaissance des émotions puisqu'ils sont à nouveau supérieurs à ceux obtenus sur les segments voisés (table 4). A titre de comparaison, Shami et al. [16] obtiennent un score de 59% sur les zones voisées avec un niveau supplémentaire d'extraction de caractéristique.

Table 3 : Mesures prosodiques.

Paramètres	Mesures	Nombre
Pitch	Moyenne, écart-type, amplitude normalisée, max, écart-type Δ, écart-type ΔΔ, 3ème coefficient du polynôme de régression d'ordre 3	7
Energie	moyenne, écart-type, 3ème coefficient du polynôme de régression d'ordre 3	3
Durée	durée, PVI inter	2

Table 4 : Taux de reconnaissance selon les classifieurs.

Classifieur	Score
Voisé	50.68%
Voyelles (références)	49.51%
Voyelles (détectées)	55.73%

5. CONCLUSION

Un nouveau schéma d'extraction de caractéristiques pour la reconnaissance des émotions a été présenté : l'approche basée voyelle. Cette approche exploite des unités pseudo-phonétiques extraites automatiquement dans un signal de parole. Une évaluation du système de détection des voyelles a montré un comportement relativement robuste sur trois corpus différents avec un taux d'erreur moyen (VER) de 22.17%. Ces unités ont ensuite été exploitées dans une tâche de reconnaissance automatique d'émotions actées. Bien que la quantité d'informations portées par les segments 'voyelle' et 'consonne' soit bien inférieure à celles des segments voisés et non voisés, les résultats obtenus par ces unités sont supérieurs sur les plans acoustique et prosodique. Les segments phonétiques (ou pseudo-phonétiques) apparaissent plus appropriés que les segments voisés pour l'extraction d'informations. Nous proposons par conséquent d'intégrer les unités pseudo-phonétiques présentées dans cet article dans les systèmes de reconnaissance des émotions afin d'améliorer leur performances.

BIBLIOGRAPHIE

[1] R. André-Obrecht. A New Statistical Approach for Automatic Speech Segmentation. IEEE Transaction on ASSP, volume 36, numéro 1, pages 29-40, 1988.

[2] Appendix F. Labels describing affective states in five major languages. In Scherer, K (ed.): Facets of emotion: Recent research. Hillsdale, NJ: Erlbaum. [Version revised by the members of the Geneva Emotion Research Group] pages 241-243, 1988.

[3] F. Burkhardt et al. A Database of German Emotional Speech. In *Proc.* of Interspeech, 2005.

[4] C. Clavel, I. Vasilescu, G. Richard et L. Devillers. Voiced and Unvoiced Content of Fear-type Emotions in the SAFE Corpus. In *Proc*. of Speech Prosody, 2006.

[5] R. Cowie. Emotion-Oriented Computing: State of the Art and Key Challenges. Humaine Network of Excellence, 2005.

[6] D. Datcu et L.J.M. Rothkrantz. The Recognition of Emotions from Speech using GentleBoost Classifier. A Comparison Approach. CompSys-Tech, session 5, 2006.

[7] L. Devillers, L. Vidrascu et L. Lamel. Challenges in Real-Life Emotion Annotation and Machine Learning Based Detection. Journal of Neural Networks, volume 18, numéro 4, pages 407-422, 2005.

[8] J.-S. Garofolo et al. DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM. NIST, 1993.

[9] E. Grabe, F. Nolan et K. Farrar. IViE – A Comparative Transcription System for Intonational Variation in English. In *Proc.* of ICSLP, 1998.

[10] C. Jankowski et al. NTIMIT: A Phonetically Balanced, Continuous Speech, Telephone Bandwidth Speech Database. ICASSP, volume 1, pages 109-112, 1990.

[11] F. Pellegrino et R. André-Obrecht. Automatic Language Identification: an Alternative Approach to Phonetic Modelling. Signal Processing, volume 80, pages 1231-1244, 2000.

[12] R. Picard. Affective Computing. The MIT Press, 1997.

[13] R. Plutchik. A General Psychoevolutionary Theory of Emotion. In Plutchik R. & Kellerman H. (eds.): Emotion: theory, research, and experience, New York: Academic, volume 1, pages 3-33, 1980.

[14] F. Ringeval et M. Chetouani. Exploiting a Vowel Based Approach for Acted Emotion Recognition. Springer-Verlag, 2008.

[15] J-L. Rouas, J. Farinas, F. Pellegrino et R. André-Obrecht. Rhythmic Unit Extraction and Modeling for Automatic Language Identification. Speech Communication, volume 47, issue 4, pages 436-456, 2005.

[16] M. Shami et W. Verhelst. An Evaluation of the Robustness of Existing Supervised Machine Learning Approaches to the Classification of Emotions in Speech. Speech Communications, volume 48, issue 9, pages 201-212, 2007.

[17] K. Truong et D. Van Leeuwen. An 'open-set'

Detection Evaluation Methodology for Automatic Emotion Recognition in Speech. Workshop on Paralinguistic Speech - between models and data, pages 5-10, 2007.

[18] D. Ververidis et C. Kotropoulos. Emotional Speech Recognition, Features and Method. Speech Communication, volume 48, issue 9, pages 1162-1181, 2006.

[19] T. Vogt et E. André. Improving Automatic Emotion Recognition from Speech via Gender Differentiation. In *Proc.* of Language Resources and Evaluation Conference, 2006.