Identification de l'origine des locuteurs non natifs en utilisant des paramètres prosodiques

M. Piat, D. Fohr, I. Illina

Équipe Parole, LORIA-CNRS & INRIA, "http ://parole.loria.fr/"
BP 239, 54600 Vandoeuvre-lès-Nancy, France
{ piat,fohr,illina }@loria.fr

ABSTRACT

In this paper we propose an automatic approach to foreign accent identification. Knowing the speaker origin could allow to adapt the acoustic models for non-native speech recognition. In this study, we use a statistical approach based on prosodic parameters. This approach relies on the fact that prosody is different between languages, and so between accents. This work is done in the framework of the HIWIRE (*Human Input that Works In Real Environment*) European project. The corpus is composed of English sentences pronounced by French, Italian, Greek and Spanish speakers. Results obtained with duration and energy are promising for foreign accent identification : 68.6% with energy and 67.1% with duration. These two parameters combined with MFCC achieve a 87.1% correct foreign accent identification rate.
Keywords: non native speech, foreign accent identification, prosodic parameters

1. Introduction

Les performances d'un système de Reconnaissance Automatique de la Parole (RAP) se dégradent fortement face à des locuteurs non natifs. Pour améliorer la RAP, une solution est d'utiliser des modèles adaptés à l'origine du locuteur. Pour cela, il faut d'abord identifier l'origine du locuteur. Plusieurs études tentent de résoudre ce problème d'identification, souvent par des méthodes statistiques basées sur les MFCC. Chen et al. [3] obtiennent entre 77,5% et 98,5% de bonne identification parmi 4 accents chinois grâce à des GMM : un GMM est appris par langue et par genre en utilisant les MFCC. Kumpf et W.King [8] créent un modèle HMM par phonème et par accent ; 73,8% des locuteurs sont bien classés. Avec le calcul des probabilités bigrammes de phonèmes pour chaque accent, le taux d'identification passe à 76,6%. Arslan et Hansen [1] identifient l'accent turc, allemand, chinois et anglais en langue américaine avec un modèle HMM par mot et par accent, 74,5% des locuteurs sont bien identifiés.
D'autres travaux proposent d'utiliser la confusion phonétique et les contraintes graphémiques comme dans l'article [2] où un taux d'erreur par mot de 6,3% est atteint parmi 4 accents (français, italien, grec et espagnol) en parole anglaise.

Tous ces articles utilisent les paramètres cepstraux, mais il est possible d'utiliser d'autres paramètres. Plusieurs travaux étudient les différences prosodiques entre les langues et entre les accents étrangers. Horgues [6] passe en revue les différences prosodiques entre français et anglais et expose leur importance dans la perception de l'accent anglais et français. Gendrot [4] montre que même altérée, l'auditeur peut reconnaître la langue parlée parmi 8 tant que la f0 et les durées sont conservées. Ramus [10] utilise une technique de délexicalisation donnant lieu au même résultat. Les tests de perception humaine menés par Kumpf et King [7] indiquent que l'auditeur classe en partie les accents selon des caractéristiques prosodiques tels que l'évolution du pitch, le rythme et les pauses.

Certains articles prennent en considération l'aspect prosodique dans la reconnaissance de l'accent. Hansen et Arslan [5] identifient l'accent de locuteurs non natifs parlant américain, parmi 4, avec des modèles HMM de monophones basés principalement sur la fréquence fondamentale, l'énergie et les 4 premiers formants, ils obtiennent jusqu'à 92% de taux de reconnaissance. L'article [12] distingue, parmi 6, l'origine de 25% à 77% des locuteurs lisant un texte français grâce aux fréquences formantiques F1 et F2. L'étude de l'accentuation de la phrase et l'accentuation locale [11] permet d'identifer entre 54% et 98% des parlers arabes.

Notre approche est orientée vers la modélisation de paramètres acoustiques prosodiques (durée, énergie et fréquence fondamentale). La section 2 présente le contexte de notre travail. Le choix de la prosodie au niveau syllabique motivant notre approche est expliquée et détaillée dans la section 3. La section 4 présente la méthodologie utilisée pour réaliser cette approche. Les résultats obtenus sont présentés section 5. Enfin la conclusion est donnée dans la dernière section.

2. Contexte

2.1. Projet HIWIRE

Le projet HIWIRE est un projet européen regroupant 5 laboratoires de recherche en Espagne, Italie, France et Grèce, ainsi que 2 entreprises *Loquendo* en Italie et *Thalès* en France. Un des buts de ce projet est de reconnaître la parole de pilotes d'avion dans un cockpit. Ces pilotes doivent impérativement parler en anglais, mais ne sont pas forcément natifs de cette langue.

2.2. Corpus HIWIRE

Le corpus enregistré dans le cadre d'HIWIRE est composé de phrases correspondant à des tâches de l'aéronautique prononcées par 70 européens : 20 français, 20 grecs, 20 italiens et 10 espagnols. Chaque locuteur prononce 100 phrases de quelques mots en anglais. Par exemple : *Next. Range fourty. Cannot accept nine four four.*
Le lexique est composé de 134 mots. La fréquence d'échantillonnage est de 16 KHz.

3. Motivations

En présence de locuteurs non natifs, l'utilisation de modèles acoustiques de RAP adaptés à l'origine du locuteur permettrait d'améliorer la reconnaissance de la parole. Notre but est donc d'identifier automatiquement l'origine du locuteur. Pour cela nous nous sommes intéressé à la prosodie.
La prosodie se définit principalement par l'étude de l'évolution de la hauteur, de l'intensité et du débit dans le temps. La prosodie au niveau de la phrase est appelée *accent de focus*, tandis que l'*accent lexical* se situe au niveau du mot.

Dans notre étude, nous avons choisi d'étudier l'accent lexical. L'étude au niveau de la phrase aurait été moins pertinente car dans notre cas il s'agit de phrases de l'aéronautique comme par exemple "*position india bravo*". La prosodie de ces phrases courtes ne semble pas véritablement être influencée par l'accent mais par la lecture de mots dissociés.

L'accent lexical est porté sur une syllabe d'un mot particulier et ne tombe pas toujours sur la même syllabe selon le mot et selon les langues. C'est pourquoi nous avons choisi de modéliser les syllabes de chaque mot. D'autre part l'accent lexical marqué sur une syllabe est souvent caractérisé par un allongement de la durée de la syllabe, une augmentation de l'énergie et un pic de f0 (plus ou moins amplifié selon le locuteur). Selon l'origine du locuteur, l'accent lexical sera plus ou moins marqué et pourra être porté par une syllabe différente. Certains locuteurs non natifs tendent à conserver leurs caractéristiques natives. Dans notre application, lorsqu'un non natif parle anglais, il peut être amené à faire des erreurs de prononciations liées à sa langue d'origine. Nous ne cherchons pas ici à connaitre la place de l'accent dans le mot mais à l'utiliser de façon automatique. Pour cela, nous avons étudié ces trois paramètres prosodiques au niveau de la syllabe.
A titre de comparaison, nous avons également étudié les MFCC qui sont considérés comme des paramètres robustes en RAP.

4. Modélisation

Nous avons défini un modèle par syllabe et par mot prononcé avec un accent étranger. Par exemple le HMM `F_z-ih_zero` modélise la syllabe *française* "*z-ih*" du mot "*zero*". Les systèmes sont basés sur la durée, l'énergie et la f0.

4.1. Durée

L'allongement de la durée des syllabes dans un mot peut permettre de détecter l'accent lexical. Grâce aux différences de durées syllabiques, on pourrait identifier l'origine du locuteur. La durée absolue de chaque syllabe n'est pas intéressante car certains locuteurs parlent plus ou moins vite sans que cela ne traduise leur origine. C'est pourquoi le rapport de la durée de la syllabe sur la durée totale du mot est utilisé dans notre approche. La durée moyenne de chaque syllabe par mot et par accent est modélisée par un HMM à 1 état.

4.2. Énergie

L'énergie peut également varier d'une syllabe à l'autre selon la manière dont la syllabe est prononcée. La syllabe portant l'accent lexical sera souvent dotée d'une forte énergie. Un locuteur qui ne s'exprime pas dans sa langue d'origine ne prononce pas forcément correctement le mot. Il est influencé par les habitudes liées à son origine. Ainsi la distribution de l'énergie dans un même mot pourra varier selon l'origine du locuteur. L'énergie est également variable d'un locuteur à l'autre (selon l'intensité de la voix du locuteur). C'est pourquoi il est nécessaire de la normaliser. Pour cela on soustrait la valeur maximum de l'énergie sur la phrase.
L'énergie est extraite pour chaque trame à partir du signal, puis normalisée. Nous utilisons également sa première et seconde dérivées. L'énergie est modélisée de deux façons :
- Dans la première méthode, l'énergie de chaque syllabe est modélisée par un HMM à 3 états pour représenter son évolution (début, milieu, fin) durant la syllabe.
- Une deuxième méthode consiste à calculer l'énergie moyenne de la syllabe afin de distinguer la syllabe accentuée (au sens de l'accent lexical) dans le mot. L'énergie moyenne de chaque syllabe est modélisée par un HMM à 1 état.

4.3. Fréquence fondamentale

La fréquence fondamentale marque l'intonation et peut par exemple en anglais permettre de distinguer le sens de deux mots identiques (comme le nom "*permit*" du verbe "*permit*"). Chaque langue possède ses propres habitudes en terme d'intonation. La façon de parler du locuteur non natif est influencée par son origine. Ces erreurs "typiques" peuvent permettre de identifier l'origine du locuteur. La f0 varie également d'un locuteur à l'autre il faut donc normaliser ces valeurs. La f0 est extraite pour chaque trame à partir du signal. Plusieurs méthodes ont été expérimentées pour utiliser la f0. Seules la méthode de base et celle donnant les meilleurs résultats sont présentées ici :
- Une première méthode consiste à utiliser les valeurs normalisées par le coefficient d'élongation propre à chaque locuteur (VTLN : *Vocal Tract Length Normalization* [9]) et ainsi étudier l'évolution des valeurs non nulles de la f0 durant la syllabe (HMM à 3 états).
- La deuxième méthode consiste à calculer la pente générale de la syllabe (HMM à 1 état) afin de modé-

liser l'accentuation globale de la syllabe par rapport aux autres syllabes du mot. La pente est calculée sur 3 valeurs par la méthode des moindres carrés. (Cette méthode sera notée "F0 pente" dans la suite de cet article.)

4.4. MFCC

Nous comparons nos résultats à l'approche utilisant les MFCC. Une fenêtre de Hamming de 25 ms est appliquée, suivi d'un banc de 24 filtres permettant d'extraire les 12 premiers coefficients cepstraux et leurs dérivées première et seconde. L'énergie n'est pas incluse dans les MFCC. A partir de ces données, des modèles HMM à 3 états pour chaque syllabe, par mot et par accent sont appris.

4.5. Système d'identification

Le corpus HIWIRE étant de petite taille, les expériences sont réalisées par validation croisée. Un groupe de 7 locuteurs est choisi comme groupe de test, tandis que le reste sert à l'apprentissage. Ce processus est réitéré jusqu'à ce que tous les locuteurs soient testés.
Pour chaque paramètre étudié, des modèles HMM de syllabes par mot et par accent sont appris. Le locuteur de test prononce quelques mots dont la transcription est connue. Une segmentation automatique est ensuite effectuée sur le signal afin de connaître les débuts et fins de mots et syllabes. Ensuite un alignement de modèles HMM de syllabes est effectué avec le critère du maximum de vraisemblance. Pour chaque locuteur on obtient une suite de syllabes identifiées d'origines o. Par exemple, le locuteur prononce "*zero seven*", la séquence de modèles reconnue pourra être : `F_zih_zero I_r-ow_zero I_s-eh_seven I_v-ah-n_seven`. L'origine $\hat{o}$ affectée au locuteur est celle recueillant le plus grand nombre de syllabes identifiées d'origine o. Dans notre exemple, le locuteur serait identifié en tant qu'italien. Pour chaque HMM, 16 gaussiennes par état sont calculées.

5. Expériences et résultats

Le vocabulaire a été réduit aux mots polysyllabiques et non acronymes (exemple d'acronyme : *VHF*). De plus, étant donné la petite taille du corpus, seuls les mots les plus fréquents sont retenus pour apprendre les modèles et tester l'identité du locuteur.
Mots retenus : *zero*, *request*, *seven*, *select*, *mayday*, *weather*, *accept*, *level*, *performance*.

Le tableau 1 donne le pourcentage de locuteurs bien identifiés pour chaque paramètre testé. Lorsque les valeurs normalisées (non nulles) de f0 de la syllabe sont modélisées, le pourcentage de locuteurs identifiés est faible (28%). La meilleure modélisation de la fréquence fondamentale est celle où la pente globale de la syllabe est calculée (54,3%) (sans prendre en compte les valeurs nulles). La durée semble un bon indicateur puisqu'elle permet de reconnaître 67,1% des locuteurs. L'énergie moyenne calculée par syllabe permet d'identifier 60% des locuteurs, tandis que l'énergie normalisée identifie 68% des locuteurs. Enfin, les MFCC restent les meilleurs paramètres pour identifier l'origine avec 82,9% de locuteurs bien identifiés. Notons cependant qu'avec les MFCC, les modèles sont élaborés avec 36 dimensions, alors que les paramètres prosodiques n'ont que 1 ou 3 dimensions. L'énergie normalisée et la durée sont donc des paramètres relativement intéressants.

Tab. 1: Pourcentage de locuteurs bien identifiés

Caractéristique	Dimension du vecteur de paramètres	Taux d'identification %
F0 normalisée	1	28,6
F0 pente	1	54,3
Énergie moyenne	1	60,0
Durée	1	67,1
Énergie norm.	3	68,6
MFCC	36	82,9

Tab. 2: Matrice de confusion en pourcentage de locuteurs identifiés avec les **MFCC**

	reconnus			
	F	I	G	E
F	**100**	0	0	0
I	5	**90**	5	0
G	0	5	**95**	0
E	20	30	40	**10**

Tab. 3: Matrice de confusion en pourcentage de locuteurs identifiés avec l'**énergie normalisée**

	reconnus			
	F	I	G	E
F	**60**	5	10	25
I	10	**75**	10	5
G	20	5	**70**	5
E	0	30	10	**60**

Les tableaux 2 et 3 présentent les matrices de confusion en pourcentage de locuteurs dont l'origine est bien identifiée en utilisant comme paramètres les MFCC et l'énergie. Dans le tableau 2, les français, les italiens et les grecs sont identifiés de 90% à 100%, tandis que les espagnols sont très mal identifiés (10%). Ces différences peuvent s'expliquer par le fait que les français sont spécifiques dans leur façon de parler. Tandis que les espagnols sont, dans notre corpus, moins nombreux et plus divers car ils proviennent de différentes régions d'Espagne et peuvent donc avoir des accents espagnols différents.
L'énergie (tableau 3) est intéressante car elle discrimine les 4 différentes origines de façon plus homogène. 60 à 75% des locuteurs sont correctement identifiés dans chaque origine.

5.1. Combinaison

Dans les expériences précédentes, nous avons étudié chaque paramètre indépendamment. Maintenant, nous souhaitons voir si leur combinaison peut ap-

porter une meilleure identification de l'origine du locuteur. Les paramètres combinés sont la durée, l'énergie normalisée, la f0 (calcul de la pente de chaque syllabe) et les MFCC. Plusieurs approches ont été testées, nous ne détaillerons ici que les deux plus intéressantes en terme de taux d'identification du locuteur.

– La première approche (i) consiste à utiliser les MFCC dans un vecteur (premier flux) et à rassembler les paramètrers durée et/ou énergie et/ou fréquence fondamentale dans un autre vecteur (deuxième flux). Les probabilités d'émission $b_j(o_t)$ des deux vecteurs d'observations o_t pour chaque trame t sont calculés selon l'équation suivante :

$$b_j(o_t) = \prod_{s=1}^{S} \left(\sum_{m=1}^{M_{js}} c_{jsm} N(o_{st}; \mu_{jsm}; \Sigma_{jsm}) \right)^{\gamma_s} \quad (1)$$

M_{js} est le nombre de gaussiennes à l'état j et pour le flux s
c_{jsm} est le poids de la m^{ime} gaussienne
$N(o; \mu; \Sigma)$ représente la loi gaussienne du vecteur d'observation o de moyenne μ et de covariance Σ
γ_s est le poids affecté au flux, par défaut 1

– Pour la deuxième approche (ii), chaque paramètre étudié indépendamment fournit pour chaque locuteur testé le nombre de syllabes identifiées de chaque origine. Par exemple, pour le locuteur l testé en utilisant le paramètre $i = durée$, x syllabes prononcées par ce locuteurs sont identifiées françaises, y syllabes sont identifiées italiennes, etc. Pour chaque locuteur et chaque paramètre étudié indépendamment, il s'agit ensuite de sommer le nombre de syllabes δ_o^i identifiées d'origine o par le paramètre i et choisir l'origine $\hat{o}$ du locuteur selon : $\hat{o} = \arg\max_o \sum \delta_o^i$

Tab. 4: Pourcentage de locuteurs bien identifiés

Caractéristique	Taux d'ident.
MFCC + Durée (i)	85,7
MFCC + Durée + Én. norm. (ii)	87,1

La combinaison des MFCC avec la durée dans des flux séparés (i) (tableau 4) donne un taux d'identification de 85,7% contre 82,9% pour les MFCC seuls. On remarque également, méthode (ii), que la F0 ne dégrade pas les résultats mais n'apporte aucune information supplémentaire par rapport à la combinaison MFCC + durée + énergie. Ce sont ces 3 paramètres combinés qui apportent le meilleur taux de reconnaissance de l'origine. Dans les deux cas on distingue une reconnaissance de l'accent supérieure ou égale à 85% pour les français, les italiens et les grecs, alors que les espagnols restent beaucoup moins bien identifiés.

6. Conclusion

Dans cet article, nous avons proposé une approche automatique d'identification de l'origine des locuteurs non natifs. Cette approche est basée sur l'utilisation de paramètres acoustiques prosodiques (durée, énergie et fréquence fondamentale des syllabes). Les expériences sont menées sur une base de données de parole anglaise prononcée par des français, des italiens, des grecs et des espagnols.
Cette approche nécessite d'avoir la transcription phonétique des phrases prononcées par les locuteurs (de test et d'apprentissage).
Selon notre étude, la durée et l'énergie apparaissent comme des paramètres prometteurs donnant des résultats encourageants. De plus, l'ajout de ces paramètres aux MFCC permet de réduire le taux d'erreur relatif de 24% par rapport à l'utilisation des MFCC seuls.

7. Remerciements

Ce travail a été en partie financé par le projet européen *HIWIRE* (*Human Input that Works In Real Environments*), contrat numéro 507943, *sixth framework program, information society technologies.*

Références

[1] L.M. Arslan and J.H.L. Hansen. Language Accent Classification in American English. *Speech Communications*, 18(4) :353–367, 1996.

[2] G. Bouselmi, D. Fohr, I. Illina, and J.-P. Haton. Détection de la langue maternelle de locuteurs non natifs fondée sur l'extraction de séquences discriminantes de phonèmes. In *TAIMA*, pages 397–402, 2007.

[3] T. Chen, C. Huang, E. Chang, and J. Wang. Automatic Accent Identification Using Gaussian Mixture Models. In *IEEE ASRU Workshop*, pages 343–346, 2001.

[4] C. Gendrot. Indices prosodiques et phonotactiques pour l'identification de sa langue natale. In *Modélisations pour l'Identification des langues et des variétés dialectales par les humains et par les machines*, pages 97–100, 2004.

[5] J.H.L. Hansen and L.M. Arslan. Foreign Accent Classification Using Source Generator Based Prosodic Features. In *IEEE ICASSP*, pages 836–839, 1995.

[6] C. Horgues. Contribution à l'étude de l'accent français en anglais. Quelques caractéristiques prosodiques de l'anglais parlé par des apprenants francophones et leur évaluation perceptive par des juges natifs. In *Actes des 8e RJC ED268 Langage et langues*, pages 79–83, 2005.

[7] K. Kumpf and RW. King. Foreign Speaker Accent Classification using Phoneme-Dependent Accent Discrimination Models and Comparisons with Human Pperception Benchmarks. In *Eurospeech 97*, pages 2323–2326, 1997.

[8] K. Kumpf and R. W.King. Automatic Accent Classification of Foreign Accented Australian English Speech. In *ICSLP*, pages 1740–1743, 1996.

[9] Li Lee and Richard Rose. A Frequency Warping Approach to Speaker Normalization. *IEEE Transaction on Speech and Audio Processing*, 6(1) :49–60, 1998.

[10] F. Ramus. La discrimination des langues par la prosodie : Modélisation linguistique et études comportementales. In *Actes de la 1ère journée d'étude sur l'identification automatique des langues*, pages 186–201, 1999.

[11] J.-L. Rouas, M. Barkat-Defradas, F. Pellegrino, and R. Hamdi-Sultan. Identification automatique des parlers arabes par la prosodie. In *16e JEP*, pages 192–196, 2006.

[12] B. Vieru-Dimulescu and P. Boula de Mareüil. Identification perceptive d'accents étrangers en français. In *26e JEP*, pages 163–166, 2006.