Identification automatique des parlers arabes par la prosodie

Jean-Luc Rouas[1], *Melissa Barkat-Defradas*[2], *François Pellegrino*[3], *Rym Hamdi-Sultan*[3]

[1] Laboratoire Électronique Ondes et Signaux pour les Transports (LEOST) - INRETS
[2] Laboratoire ICAR-Praxiling UMR CNRS 5191 - Université Montpellier 3
[3] Laboratoire Dynamique Du Langage UMR CNRS 5696 - Université Lyon 2
jean-luc.rouas@inrets.fr, melissa.barkat@univ-montp3.fr, francois.pellegrino@univ-lyon2.fr, rim.hamdi@univ-lyon2.fr

ABSTRACT

This paper presents a study of automatic identification of Arabic dialectal areas based on a prosodic automatic modelling. Inspired from Fujisaki's works, this modelling dissociates long term prosodic variations from short term micro-variations and exploits n-multigrams models. Experiments, achieved on semi-spontaneous recordings from 40 speakers, show that the system reaches 98% of correct identification of the three dialectal areas - Maghreb, Middle-East, and an intermediate area (Tunisia-Egypt) - with test excerpts of 7.6 seconds in average.

1. Introduction

La prosodie concerne l'ensemble des éléments dynamiques de la chaîne parlée - tels que les variations de hauteur, d'intensité (ou d'énergie) et de durée - qui déterminent la mélodie, les tons, les pauses, les accents, le rythme, le débit...etc. Il s'agit d'un phénomène complexe, relativement difficile à étudier dans la mesure où ses manifestations (i.e. ; patrons accentuels et contours intonatifs) sont sujettes à de nombreuses variations. Les schémas prosodiques diffèrent ainsi selon les langues, les dialectes, les registres linguistiques, la structure du discours, la syntaxe des énoncés, les mots constituant ces énoncés et la structure phonétique des unités lexicales, mais aussi selon le genre, l'âge, l'origine sociale voire l'état émotionnel du locuteur. La complexité de ce phénomène est d'autant plus importante que l'ensemble des ces facteurs interagissent. Une conséquence de cette complexité est que les déterminants de cette variabilité ont été relativement peu étudiés jusqu'à récemment, faute d'outils adaptés. Ainsi, la grande majorité des études cherchaient à neutraliser ces variations pour faire émerger une norme ou en tout cas des patrons moyennés. Or, des études expérimentales récentes montrent que la variation prosodique est un élément de discrimination linguistique et/ou dialectal d'importance. Par exemple, plusieurs études portant sur le rythme ont montré que les formes dialectales d'une même langue pouvaient être discriminées au niveau prosodique sur la base de leur structure rythmique. Des différences inter-dialectales pertinentes ont ainsi été rapportées pour l'anglais [10] et l'arabe [12]. A l'inverse, encore peu d'études ont à ce jour tenté d'évaluer la pertinence des variations intonatives pour identifier des dialectes d'une même langue [11]. Dans cette étude nous nous intéressons aux variations de la fréquence fondamentale (F_0) et de l'énergie dans des énoncés en arabe dialectal afin d'évaluer la pertinence et la robustesse de ces paramètres en identification automatique des parlers arabes par la prosodie.

2. État de l'art des études sur la prosodie en arabe

Alors que l'étude du domaine de l'accent de mot en arabe standard et/ou dialectal est relativement bien développée, les recherches expérimentales sur l'intonation et l'organisation prosodique des énoncés sont relativement peu fréquentes. Les travaux qui abordent l'analyse de la courbe mélodique des phrases concernent le plus souvent l'arabe standard, et ont pour principale application le développement d'outils de synthèse de la parole [15, 23]. La question des patterns prosodiques dans les dialectes arabes modernes a été abordée dans quelques rares études portant plus particulièrement sur les patrons mélodiques associés aux structures syntaxiques et énonciatives les plus courantes (i.e. ; questions totales, assertions) [6, 8, 18]. Les différents parlers arabes n'ont pour autant pas fait l'objet de la même attention. Si les contours intonatifs de l'arabe marocain sont quantitativement bien traités [4, 22], les études s'attachant à la description de ces phénomènes dans les dialectes algériens et/ou tunisiens sont, en revanche, très peu fréquentes [3]. Pour l'heure, nous savons :

- que les parlers arabes présentent des différences significatives au plan des structures syllabiques préférentielles [13], et que la position de l'accent varie en fonction de la structure syllabique [5],
- qu'en arabe dialectal la place de l'accent s'étend de la dernière syllabe à la pré-antépénultième selon le parler considéré [14],
- qu'en arabe marocain au moins, les pics de F_0 qui se concentrent autour des syllabes accentuées connaissent une certaine mobilité du fait de l'influence de certains facteurs (position, type et durée de la syllabe, focus) ce qui correspond aux tendances universelles [22].

Toutefois, compte tenu des résultats prometteurs en identification automatique des parlers arabes par la prosodie présentés dans les sections suivantes, il serait fort utile de s'intéresser à la constitution d'une typologie des contours intonatifs des différents parlers arabe dans une perspective comparative, et ce, afin de mieux cerner la nature des éléments prosodiques de discrimination inter dialectale mis en évidence dans ce travail.

3. Identification automatique des parlers arabes par la prosodie

En s'inspirant du travail d'Adami [1], le système d'identification automatique proposé s'appuie sur un codage des trajectoires de fréquence fondamentale et d'énergie. Cependant, la modélisation se fait ici à plusieurs niveaux, en séparant la fréquence fondamentale en deux composantes : la ligne de base et le résidu, et en proposant des modélisations à des échelles différentes pour ces deux contributions : la ligne de base est modélisée à partir d'unités pseudo-syllabiques, et la modélisation du résidu est fondée sur des unités infra-phonémiques. La fréquence fondamentale et l'énergie sont extraites du signal grâce à la bibliothèque Snack [20]. Le système d'identification automatique est décrit sur la figure 1.

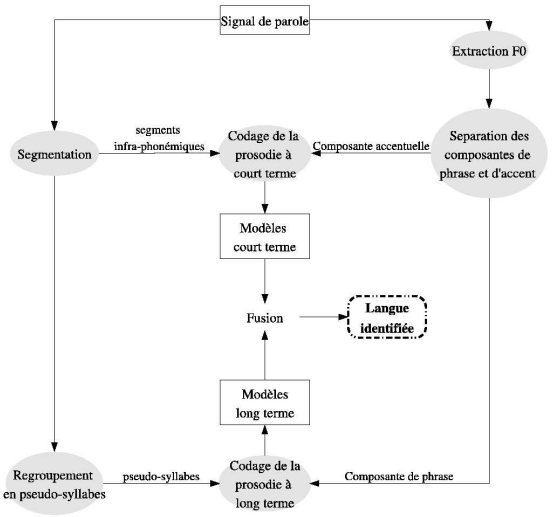

Fig. 1: Description du système

3.1. Segmentation automatique en unités infra-phonémiques et regroupement en unités pseudo-syllabiques

Trois traitements de base conduisent à la localisation de frontières quant à la notion de consonne/voyelle :
- segmentation automatique de la parole en segments quasi-stationnaires [2],
- détection d'activité vocale
- localisation des voyelles [16].

Les segments vocaliques sont étiquetés "V", les segments de non-activité "#", et les autres segments "C". Ces segments, de taille infra-phonémiques, seront utilisés pour l'étiquetage des variations locales de fréquence fondamentale et d'énergie.

Ces segments sont regroupés en unités pseudo-syllabiques : la syllabe est en effet une unité privilégiée pour la modélisation du rythme. Néanmoins, la segmentation automatique en syllabes (en particulier en ce qui concerne la détection des frontières) est une opération délicate et spécifique à chaque langue [17]. Pour cette raison, nous utilisons la notion de pseudo-syllabe [19]. Le signal de parole est segmenté en motifs correspondant à la structure [CC...CV]. Un exemple de segmentation automatique en pseudo-syllabes est donné dans la figure 2.

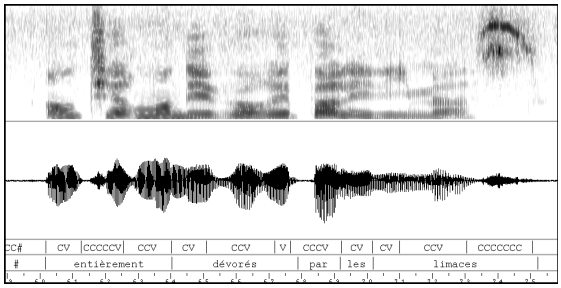

Fig. 2: Segmentation en pseudo-syllabes sur la phrase : "Les choux avaient été **entièrement dévorés par les limaces**". Les segments étiquetés "C" sont regroupés jusqu'à ce qu'un segment "V" soit rencontré.

3.2. Traitement de F_0

En s'inspirant des travaux de Fujisaki [9], le traitement de la fréquence fondamentale est divisé en deux phases permettant de représenter l'accentuation de phrase et l'accentuation locale.

Accentuation de phrase Il est supposé que la ligne de base passe par les minima locaux de F_0, de telle sorte qu'aucun point ne se situe au-dessous d'elle [21]. Pour chaque phrase, le même traitement est appliqué :
- les valeurs de la fréquence fondamentale en Hertz sont converties en demi-tons. Cette quantification permet de se reporter sur une échelle logarithmique, proche de l'échelle de la perception humaine, et de lisser la courbe mélodique.
- la droite de régression linéaire est estimée sur l'ensemble des parties voisées de chaque phrase. Le minimum est alors repéré sur chaque partie voisée située en dessous de cette droite.
- l'accentuation de phrase ou ligne de base est la droite qui rejoint les minima de la phrase.

La pente de la régression est utilisée pour étiqueter la ligne de base. Une étiquette est employée pour chaque pseudo-syllabe : "U" pour une pente positive et "D" pour une pente négative. Les pseudo-syllabes considérées comme peu ou pas voisées (dont le pourcentage en durée de voisement n'excède pas 70%) sont étiquetées "#".

Un exemple d'extraction de la ligne de base est représenté figure 3. Sur cet exemple, la séquence d'étiquettes correspondant à la phrase est : `U.U.U.U.U.-D.D.D.D.D.#`

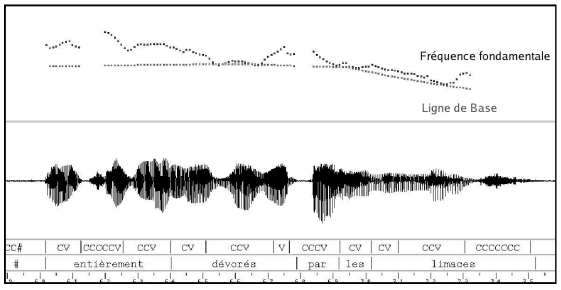

Fig. 3: Extraction de la ligne de base, sur la phrase : « Les choux avaient été **entièrement dévorés par les limaces** ».

Accentuation locale La ligne de base est soustraite du contour original de F_0. La courbe résultante est appelée résidu (figure 4). Le résidu est approximé sur chaque segment infra-phonémique par une régression linéaire. La pente de l'approximation linéaire est utilisée pour étiqueter les mouvements de F_0 avec trois étiquettes possibles ("U", "D" et "#"). La phrase donnée en exemple donne l'étiquetage suivant : `D.D.#.#.#.#.#.D.D.U.D.D.D.D.U.U.D.U.#.#.D.-D.U.D.U.D.D.U.D.#.#.#.#.#.#`

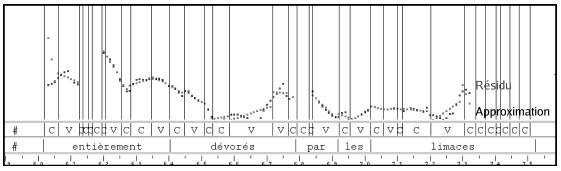

Fig. 4: Approximation du résidu, toujours sur le même exemple. Une approximation linéaire est effectuée sur chaque segment.

3.3. Codage de l'énergie

De la même manière, une régression linéaire est effectuée sur la courbe d'énergie, pour chaque segment infra-phonémique. Pour chacun de ces segments, une étiquette ("U", "D" ou "#") est apposée à la courbe d'énergie.

3.4. Codage de la durée

Les étiquettes de durée ne sont calculées que pour les unités infra-phonémiques. Les étiquettes sont assignées en considérant la durée moyenne selon le type de segment (vocalique, non vocalique ou pause). Par exemple, si un segment vocalique possède une durée supérieure à la durée moyenne des segments vocaliques des données d'apprentissage, il est étiqueté "l" (long). Sinon, il est étiqueté "s". Les étiquettes de durée générées sur la phrase d'exemple sont : `s.l.s.s.s.s.s.s.s.s.l.s.l.s.s.l.s.s.s.-s.l.s.l.s.s.s.s.l.s.s.s.s.s.s.s`

3.5. Modélisation

Les enchaînements des étiquettes sur les phrases de l'ensemble d'apprentissage sont modélisés par des modèles multigrammes. Les séquences les plus fréquentes sont alors identifiées pour chaque langue et des probabilités leur sont associées. Un modèle de langage multigrammes est un modèle statistique qui probabilise chaque motif de suites d'unités [7]. Pour ce système, cette modélisation est utilisée à deux niveaux :

- en employant les étiquettes de variation de F_0 pour la ligne de base, à l'échelle de la pseudo-syllabe. Les modèles multigrammes permettent de travailler sur les enchaînements de plusieurs pseudo-syllabes, ce qui correspondrait à l'échelle du mot.
- en utilisant les étiquettes de variation de F_0 sur le résidu, conjointement avec les étiquettes de durée et d'énergie. Ici, les multigrammes permettent de modéliser les enchaînements de plusieurs segments infra-phonémiques, se rapprochant de l'échelle syllabique.

4. Expériences d'identification automatique

Les données utilisées ont été collectées entre 1995 et 2005 au laboratoire Dynamique Du Langage (corpus Araber). Il s'agit d'enregistrements semi-spontanés (commentaire d'une histoire en images, ...) de 40 locuteurs issus du Maghreb, du Moyen-Orient et de la zone dite intermédiaire (Tunisie, Egypte). Vu la relativement faible quantité de données, tous les enregistrements disponibles pour les locuteurs (soit 5 min. par locuteur, réparties en 40 fichiers de 7,6 secondes de durée en moyenne) ont été utilisés, à la fois en apprentissage et en test, via une procédure de validation croisée (procédure de test d'un seul locuteur en utilisant tous les autres pour apprendre les modèles, itérée pour chaque locuteur, soit 39 fois). Ces expériences sont toutefois à considérer avec précaution, puisqu'elles ne permettent pas d'évaluer la dépendance des modèles vis à vis du texte (les mêmes types de texte apparaissant en apprentissage et en test).

4.1. Résultats de la modélisation à long terme

En ne considérant que les variations de F_0 de la ligne de base, le système permet d'obtenir en moyenne 54 % d'identifications correctes (tableau 1) en utilisant des 5-multigrammes, c'est-à-dire en modélisant les enchaînements jusqu'à 5 pseudo-syllabes consécutives.

Tab. 1: Nombre de fichiers correctement identifiés/testés = 859/1592 (en %)

	Zone occident.	Zone interméd.	Zone orientale.
Zone occident.	49.3	29.0	21.7
Zone interméd.	20.2	59.0	20.8
Zone orientale	18.2	27.1	54.7

Le taux d'identification correcte est significativement supérieur au hasard, mais il est décevant. Les dialectes du groupe « intermédiaire »(i.e., tunisien & égyptien) sont les mieux identifiés mais plus de 40 % des fichiers demeurent mal classés. La courbe de base de l'intonation ainsi modélisée sur une période relativement longue (5 pseudo-syllabes correspondent à une durée d'environ 500 ms) semble donc peu discriminante.

4.2. Résultats de la modélisation à court terme

En considérant les étiquettes de F_0 pour le résidu, l'énergie et la durée des segments, le système permet d'obtenir en moyenne 98 % d'identifications correctes (tableau 2) en utilisant des 5-multigrammes, c'est-à-dire en modélisant les enchaînements jusqu'à 5 segments consécutifs.

Cette modélisation à court terme permet d'obtenir de très bonnes performances quelle que soit la zone dialectale. Le fait que les dialectes de l'est et de l'ouest

Tab. 2: Nombre de fichiers correctement identifiés/testés = 1563/1592 (en %)

	Zone occident.	Zone interméd.	Zone orientale
Zone occident.	99.5	-	0.5
Zone interméd.	1.8	96.0	2.2
Zone orientale.	1.0	0.3	98.7

sont légèrement mieux identifiés que ceux de la zone intermédiaire est dû dans une très large mesure à un unique locuteur, responsable à lui seul de 3/4 des erreurs.

5. Conclusion

Cette étude visait à évaluer la pertinence d'une modélisation prosodique automatique pour l'identification de zones dialectales de l'arabe. Cette modélisation prend en compte des variations de Fo à relativement long terme ainsi que des micro-variations de Fo et d'énergie à court terme. Les expériences, menées avec 40 locuteurs permettent d'atteindre un taux d'identification correcte de 98 % en considérant 3 zones dialectales, à savoir le Maghreb, le Moyen-Orient, ainsi qu'une zone intermédiaire (Tunisie-Egypte). Si les modèles à court terme obtiennent de très bons résultats probablement liés à la prise en compte d'interactions entre accentuation et structure syllabique, les modèles à plus long terme se révèlent peu discriminants. Il est probable que l'échelle temporelle utilisée dans le modèle n-multigramme (supérieur à la syllabe mais inférieur à la proposition) ne soit pas la plus adaptée à des phénomènes intonatifs de plus grand empan temporel.

Références

[1] A. Adami, R. Mihaescu, D.A. Reynolds, and J. Godfrey. Modeling prosodic dynamics for speaker recognition. In *ICASSP*, volume 4, pages 788–791, Hong Kong, China, 2003.

[2] R. André-Obrecht. A new statistical approach for automatic speech segmentation. *IEEE Transanctions on Acoustics, Speech and Signal Processing*, 36(1) :29–40, 1988.

[3] I. Benali. Le rôle de la prosodie dans l'identification de deux parlers algériens : l'algérois et l'oranais. In *Workshop MIDL*, pages 128–132, 2004.

[4] T. Benkirane. *Intonation Systems : a Survey of Twenty Languages*, chapter Intonation in Western Arabic (Moroccan). D. Hirst & A. Di Cristo, Eds, Cambridge University Press, 1996.

[5] R. Bouziri, H. Nejmi, and M. Taki. L'accent de l'arabe parlé à casablanca et à tunis : étude phonétique et phonologique. In *ICPhS*, pages 134–137, Aix-en-Provence, 1991.

[6] D. Chahal. A preliminary analysis of lebanese arabic intonation. In *Conference of the Australian Linguistic Society*, pages 1–17, 1999.

[7] S. Deligne and F. Bimbot. Language modeling by variable length sequences : theorical formulation and evaluation of multigrams. In *ICASSP*, 1995.

[8] Sh. El Hassan. The intonation of questions in english and arabic. *Papers and Studies in Contrastive Linguistics*, pages 97–108, 1998.

[9] H. Fujisaki. Prosody, information and modeling - with emphasis on tonal features of speech. In *ISCA Workshop on Spoken Language Processing*, Mumbai, India, January 2003.

[10] E. Grabe and E.L. Low. Durational variability in speech and the rhythm class hypothesis. *Papers in Laboratory Phonology 7*, 2002.

[11] M. Grice, M. D'Imperio, M. Savino, , and C. Avesani. *Prosodic Typology : The Phonology of Intonation and Phrasing*, chapter A strategy for intonation labelling varieties of Italian, pages 362–389. Oxford : OUP, 2005.

[12] R. Hamdi, M. Barkat-Defradas, and F. Pellegrino. De la caractérisation linguistique à l'identification automatique des dialectes arabes. In *Workshop MIDL*, 29-30 novembre 2004.

[13] R. Hamdi, S. Ghazali, and M. Barkat-Defradas. Syllable structure in spoken arabic : a comparative investigation. In *Eurospeech*, Lisboa, 2005.

[14] D.E. Khouloughli and G. Bohas. *Analyse et Théories*, volume 1, chapter Processus accentuels en arabe (parlers du Caire, de Damas & arabe classique), pages 1–59. 1981.

[15] M. Mahwoub. Prosodie et ordre des constituants dans l'énoncé en arabe standard moderne. In *JEP-TALN*, 2004.

[16] F. Pellegrino and R. André-Obrecht. Vocalic system modeling : A vq approach. In *IEEE Digital Signal Processing*, pages 427–430, Santorini, July 1997.

[17] H.R. Pfitzinger, S. Burger, and S. Heid. Syllable detection in read and spontaneous speech. In *ICSLP*, volume 2, pages 1261–1264, Philadelphia, October 1996.

[18] J. Rosenhouse. Features of intonation in bedouin arabic narratives of the galilée (northern israel). In *Dialectologia Arabica, a Collection of articles in honor of Professor Heikki Palva*, volume 75, pages 193–215. Studia Oientalia, 1995.

[19] J.-L. Rouas, J. Farinas, F. Pellegrino, and R. André-Obrecht. Rhythmic unit extraction and modelling for automatic language identification. *Speech Communication*, 47(4) :436–456, 2005.

[20] K. Sjölander. The snack sound toolkit. *http ://www.speech.kth.se/snack/*.

[21] J. Vaissière. Language independent prosodic features. In *Prosody : models and measurements*, Springer series in language and communication, 14, pages 53–66. Cutler, A. and Ladd, D.R. (eds.), Berlin, 1983.

[22] M. Yéou. Effects of focus, position and syllable structure on f0 alignment patterns in arabic. In *JEP*, 2004.

[23] A. Zaki, A. Rajouani, and Z. Najim. Contours intonatifs de la phrase interrogative en arabe. In *JEP*, pages 249–257, Aussois, Suisse, 2000.