HAL
open science

Evaluation d'une nouvelle méthode de suivi de formants sur un corpus Arabe

Imen Jemaa, Oussama Rekhis, Kais Ouni, Yves Laprie

To cite this version:

Imen Jemaa, Oussama Rekhis, Kais Ouni, Yves Laprie. Evaluation d'une nouvelle méthode de suivi de formants sur un corpus Arabe. XXVIIIèmes Journées d'Etude sur la Parole - JEP'10, May 2010, Mons, Belgique. inria-00544361

HAL Id: inria-00544361

https://inria.hal.science/inria-00544361v1

Submitted on 7 Dec 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Evaluation d'une nouvelle méthode de suivi de formants sur un corpus Arabe

Imen JEMAA[1,2], Oussama REKHIS[1] ,Kais OUNI[1], Yves LAPRIE[2]

[1] Unité de Recherche Traitement du Signal, Traitement de l'Image et Reconnaissance de Formes (99/UR/1119)
Ecole Nationale d'Ingénieurs de Tunis, BP.37, Le Belvédère 1002, Tunis, Tunisie
Imen.jemaa@loria.fr , oussamarekhis@gmail.com , kais.ouni@enit.rnu.tn
[2]Equipe Parole, LORIA-CNRS – BP 239 – 54506 Vandœuvre-lès-Nancy, France
Yves.Laprie@loria.fr

ABSTRACT

This paper develops a formant tracking technique based on Fourier ridges detection. In this method we have introduced a constraint of tracking based on the computation of centre of gravity for a set of frequency formant candidates which leads to connect a frame of speech to its neighbours and thus to improve the robustness of tracking. The formant trajectories obtained by the algorithm proposed are compared to those of a hand edited formant Arabic database, created especially for this work, and those given by Praat with LPC data.

Keywords: Arabic database, speech processing, formant labelling, formant tracking.

1. INTRODUCTION

Vu que les formants soient porteurs d'une information phonétique essentielle, les trajectoires formantiques sont très utiles pour l'identification des sons de la parole, en particulier celle des voyelles [1] et des autres sons vocaliques [2], le pilotage des synthétiseurs à formants, la reconnaissance [1] et le codage de la parole. De nombreux travaux ont donc été consacrés à l'élaboration de méthodes automatiques de suivi de formants dont la plupart sont basées sur la détection des racines de LPC [3] comme estimation initiale. Les résultats de plusieurs de ces méthodes ont été utilisés dans des applications de traitement de la parole. Cependant, il y a un manque manifeste de bases de données nécessaires pour évaluer quantitativement ces méthodes, en particulier pour la langue arabe. C'est pourquoi nous avons décidé d'enregistrer et d'étiqueter en termes de formants un corpus en langue arabe standard.

Nous présentons dans ce papier notre nouvel algorithme de suivi automatique de formants basé sur la détection des maxima de la transformée de Fourier, c'est-à-dire les maxima du spectrogramme. Cet algorithme utilise la fréquence instantanée et une contrainte de suivi en calculant le centre de gravité d'un ensemble des fréquences formantiques candidates. Ensuite, nous avons évalué notre algorithme en utilisant la base de données étiquetée que nous avons construite comme référence, et en le comparant à la méthode de suivi automatique de formants utilisant le codage par prédiction linéaire mise en œuvre dans le logiciel Praat.

Ce papier est présenté comme suit. Dans la section 2, nous présentons l'algorithme de suivi de formants proposé, dans la section 3 nous décrivons le corpus arabe, dans la section 4, les différentes étapes du processus d'étiquetage manuel des formants, dans la section 5, les résultats de l'étude et l'évaluation de la méthode de suivi proposée. Enfin, nous donnons quelques perspectives dans la dernière section.

2. ALGORITHME DE SUIVI DE FORMANTS BASÉ SUR LA FRÉQUENCE INSTANTANÉE

Le diagramme de la Figure.1 décrit les principales étapes de l'algorithme proposé. Chaque étape du diagramme est décrite brièvement ci-dessous.

2.1. Prétraitement

Puisque nous nous sommes intéressés aux trois premiers formants, le signal est ré-échantillonné à 8 kHz afin de ne pas prendre en compte de formants candidats au-dessus de 4 kHz, et de nous permettre d'utiliser une analyse d'ordre inférieur plus rapide. Ensuite, le signal de parole est préaccentué par un filtre du premier ordre pour accentuer les hautes fréquences.

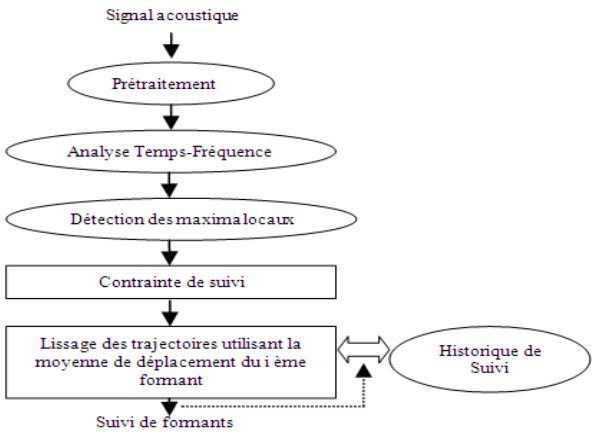

Figure 1 : Diagramme de l'algorithme de Suivi de Formants

2.2. Analyse Temps-Fréquence

Nous utilisons un spectrogramme à large bande pour analyser le signal de parole afin d'obtenir un lissage de

l'enveloppe spectrale qui fasse apparaître les formants et permettre de suivre leur évolution dans le temps. Chaque spectre est le carré de la transformée de Fourier appliquée à une fenêtre de signal.

2.3. Détection des maxima locaux

L'idée du suivi est de trouver les formants à l'aide de la fréquence instantanée. En effet, on peut montrer que les maxima spectraux correspondent à des points pour lesquels la phase est stationnaire. Cela signifie qu'au voisinage d'un pic spectral, les fréquences instantanées des coefficients de la transformée de Fourier sont égales à celle du pic.

L'accumulation de fréquences instantanées proches permet donc de trouver les formants. Cette propriété a notamment été utilisée par Charpentier, en appliquant une fenêtre de signal suffisamment longue, pour trouver les harmoniques de la fréquence fondamentale [11]. Ici nous l'utilisons pour trouver les maxima spectraux correspondant aux formants.

Étant donné que les fréquences des formants varient lentement en fonction du temps, il est possible de les considérer constants durant la fenêtre de signal (de l'ordre de 4 ms) à laquelle on applique la transformée de Fourier.

D'un point de vue théorique il a été démontré dans [9] que la fréquence instantanée est liée à la transformée de Fourier fenêtrée $Sf(u,\xi)$ du signal *f* par la relation suivante (voir Eq.1) telle que $\xi \geq 0$

$$Sf(t,\xi) = \frac{\sqrt{s}}{2} a(t) \exp^{i(\phi(t) - \xi(t))} \times \left[\hat{g}(s(\xi - \phi'(t)) + \varepsilon(t,\xi)\right] \quad (1)$$

où *s* est une échelle appliquée sur la fenêtre, par exemple la fenêtre de Hamming, $\hat{g}$ est la transformée de Fourier de g et $\varepsilon(u,\xi)$ est le terme correctif. Comme $|\hat{g}(\omega)|$ est maximum en $\omega = 0$, l'équation (1) montre que pour chaque *u*, le spectrogramme $|Sf(u,\xi)|^2$ est maximum en sa fréquence centrale $\xi(u) = \phi'(u)$, ce sont ainsi les maxima locaux qui sont validés en tant que fréquences instantanées. Mallat [9] montre aussi qu'il est possible de discerner et calculer la fréquence instantanée de plusieurs signaux composant le signal à analyser pourvu que ces signaux soient raisonnablement écartés en fréquence.

Pratiquement, on détecte donc tous les maxima locaux du spectrogramme aux points $(u,\xi(u))$. On obtient ainsi pour chaque formant la combinaison des fréquences candidates.

2.4. Utilisation de contrainte de suivi

La stationnarité de la fréquence instantanée au voisinage d'un pic spectral signifie que l'accumulation de valeurs de fréquence instantanée dans un même voisinage de fréquence confirme la détection d'un maximum spectral pertinent. Par ailleurs, le calcul de la fréquence instantanée est d'autant plus fiable qu'il correspond à un échantillon spectral d'énergie élevée.

Nous avons utilisé ces propriétés pour contraindre le suivi. D'une part nous avons défini un intervalle de fréquence pour chaque formant de façon à éliminer les valeurs aberrantes [10]. Ensuite, à l'intérieur de chaque intervalle nous calculons la valeur du formant en prenant le centre de gravité des fréquences instantanées pondérées par l'énergie spectrale :

$$\bar{f} = \frac{\sum_{i=1}^{n} p_i f_i}{\sum_{i=1}^{n} p_i} \quad (2)$$

Avec f_i fréquence du i^{ème} candidat, p_i son énergie spectrale et *n* le nombre de valeurs de fréquences instantanées considérées.

3. DESCRIPTION DU CORPUS

Pour élaborer notre corpus, nous avons utilisé une liste de phrases arabes phonétiquement équilibrées proposée par Boodraa et al. [4]. Ce corpus couvre l'ensemble des réalisations phonétiques et phonologiques de la langue arabe standard. La plupart des phrases de cette base est extraite du Coran et de l'Hadith. Elle est constituée de 20 listes chacune formée de 10 phrases courtes. Chaque liste compte 104 CV (C: consonne et V: voyelle), c'est-à-dire 208 phonèmes [4].

Nous avons enregistré ce corpus dans une chambre sourde pour dix locuteurs tunisiens (cinq hommes et cinq femmes) âgés entre 22 et 30 ans. Le signal est numérisé à une fréquence de 16 kHz. Ce corpus contient 2000 phrases (200 phrases prononcées par chaque locuteur) affirmatives ou interrogatives. De cette façon, la base de données présente une sélection équilibrée de locuteurs et de phonèmes. Toutes les phrases du corpus sont riches en contextes phonétiques et présentent donc une grande variété de trajectoires formantiques [5].

Une fois la base enregistrée nous l'avons étiquetée phonétiquement en utilisant le logiciel Winsnoori [6] que nous avons adapté à l'arabe. Une capture d'écran de cet outil est présentée ci-dessous dans la Figure 2 montrant l'annotation de la phrase « أَتُؤْذِيهَا بِآلامِهِمْ؟ » (« 3atu3Di:ha: bi3a:la: mihim») signifiant (tu souhaites la blesser avec leurs chagrins ?) prononcée par une locutrice. Une fois les annotations terminées, le corpus a été examiné par des phonéticiens pour corriger les erreurs d'étiquetage. Cette étape est très importante pour obtenir une bonne base de référence.

4. ETIQUETAGE FORMANTIQUE MANUEL

Pour faciliter le processus d'étiquetage formantique de notre corpus, nous avons d'abord obtenu un ensemble de fréquences formantiques candidates fournies par les racines de LPC [3] à l'aide de logiciel Winsnoori [6] et des outils destinés au pilotage de la synthèse à formants. Sur la base de ces valeurs candidates estimées, nous avons édité les trajectoires des formants à la main à l'aide de la souris. La Figure 2 montre un exemple de phrase prononcée par une locutrice illustrant le processus

d'étiquetage des formants et les résultats. Nous avons suivi et enregistré les trois premiers formants (F1, F2 et F3) toutes les 4 ms pour chaque phrase de la base de données. L'ordre de prédiction utilisé en LPC est 18 et la fenêtre d'analyse utilisée est de 16 ms. La durée de la fenêtre d'analyse spectrale est 4 ms pour avoir un spectrogramme à large bande qui montre mieux l'évolution des trajectoires des formants. Les difficultés se manifestent généralement dans le cas où l'énergie spectrale est faible ou lorsque les renforcements spectraux ne coïncident pas avec les valeurs de référence attendues, particulièrement pour les segments consonantiques. Pour surmonter ces difficultés, nous avons prévu des valeurs nominales spécifiques aux voyelles ainsi que les consonnes [7][8]. Enfin, afin de s'assurer de l'exactitude des trajectoires des formants de chaque phrase, nous avons synthétisé le son avec les trois premiers formants utilisant le synthétiseur Klatt mis en œuvre dans Winsnoori [6] pour vérifier que le signal synthétisé corresponde bien à l'original, c'est-à-dire la phrase enregistrée par les locuteurs. Dans un premier temps l'évaluation a été subjective puisque nous avons évalué la qualité du suivi en jugeant la qualité du signal synthétisé à l'aide des trajectoires formantiques. Ensuite, nous sommes passés à une évaluation plus objective en faisant examiner les trajectoires formantiques étiquetées manuellement par des experts en phonétique de l'arabe.

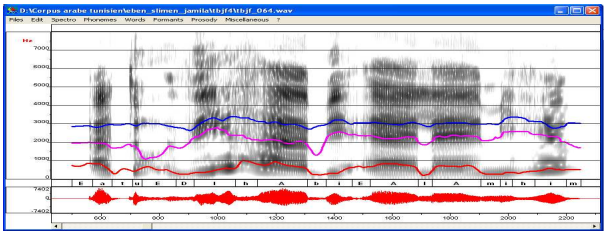

Figure 2 : Etiquetage phonétique et formantique de la phrase «أَتُوْذِيهَا"بِالاَمِهِمْ» (« 3atu3Di:ha: bi3a:la:mihim ») prononcé par une locutrice.

5. ETUDE ET ÉVALUATIONS

Nous avons utilisé ce corpus étiqueté manuellement comme référence pour évaluer notre nouvel algorithme de suivi. Les Figures 2 et 3 montrent d'une part le suivi manuel de référence pour F1, F2 et F3 et d'autre part les résultats du suivi automatique. On peut noter que pour la plupart des segments vocaliques le suivi automatique est correct.

Pour évaluer quantitativement la méthode de suivi automatique, nous l'avons comparé les résultats avec la méthode LPC automatique mise en œuvre dans le logiciel Praat [12]. Nous avons utilisé notre corpus étiqueté comme référence en calculant la différence absolue moyenne (Eq. 3) et l'écart-type normalisé par rapport aux valeurs de références (Eq.4) pour chaque trajectoire formantique (F1, F2 et F3). Nous avons donc étudié les résultats obtenus pour une voyelle brève au sein de la syllabe CV. La Table 1 montre les résultats obtenus pour la voyelle /a/ précédée d'une consonne appartenant à chacune des grandes classes de consonnes et pour les trois formants (F1, F2 et F3). Les CV ont été extraits des trois phrases suivantes prononcées par cinq locuteurs masculins : «عَرَفَ وَالِيًا وَقَائِد» (« εarafa wa:liyan wa qa:3idan ») qui signifie ("Il connaissait un gouverneur et un commandant"), « هِيَ هُنَا لَقَدْ آبَتْ» (« hiya Huna: laqad 3a:bat ») qui signifie ("Elle est ici et elle était pieuse") et « لَقَدْ كَانَ مُسَالِمًا وَقُتِلَ. » (« Laqad ka:na musa:liman wa qutila ») qui signifie ("Malgré qu'il était pacifiste, il a été tué.").

$$Diff = \frac{1}{N} \times \sum_{p=1}^{N} \left| F_r(p) - F_c(p) \right| \text{ Hz} \qquad (3)$$

Avec F_r est la fréquence de référence, F_c la fréquence calculée et N le nombre total de valeurs considérées pour chaque suivi de formants.

$$\sigma = \sqrt{\frac{1}{N} \sum_{p=1}^{N} \left(\frac{\left| F_r(p) - F_c(p) \right|}{F_r} \right)^2} \qquad (4)$$

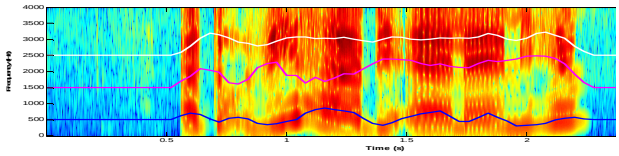

Figure 3 : Trajectoires formantiques (F1, F2 et F3) obtenues à l'aide de notre algorithme pour la phrase (« 3atu3Di:ha: bi3a:la:mihim ») prononcée par une locutrice.

Table 1 : Moyenne des erreurs de suivi de formants (pour la voyelle brève /a/) mesurée pour cinq locuteurs masculins.

		LPC	Fourier	LPC	Fourier	LPC	Fourier
		F1		F2		F3	
Occlusive voisée : /d/	Diff	118	52	110	82	197	125
	σ	46	14	38	19	63	32
Occlusive non voisée: /q/	Diff	67	62	84	109	113	66
	σ	18	15	27	28	31	15
Fricative voisée: /H/	Diff	52	39	87	100	92	87
	σ	9	6	18	16	18	18
Fricative non voisée: /f/	Diff	38	24	77	39	123	50
	σ	11	6	26	9	40	11
Nasale: /m/	Diff	70	90	60	172	192	110
	σ	17	21	16	49	52	27
Latérale: /l/	Diff	66	50	44	64	87	74
	σ	18	11	10	12	22	16
Trille: /r/	Diff	57	37	86	73	173	162
	σ	13	9	23	16	40	36
Semi-voyelle: /w/	Diff	75	85	85	46	154	105
	σ	19	19	32	10	42	25
Total	Diff	68	55	79	86	141	97
	σ	19	13	24	20	39	23

L'examen des valeurs figurant dans la Table 1 montre que l'algorithme proposé atteint des résultats proches des valeurs de référence et meilleurs que le suivi de formants du logiciel Praat en particulier lorsque la voyelle /a/ est précédée par une occlusive voisée ou une fricative non voisée (sauf pour F1 et F2 lorsque /a/ est précédée par une nasale). Pour les autres cas les deux méthodes donnent de bons résultats. Enfin nous constatons que, globalement, il n'y a pas de grande différence en termes d'erreurs entre les deux méthodes de suivi mais dans la plupart des cas notre algorithme présente une différence plus marquée par rapport à la référence pour F3. Il y a plusieurs origines possibles à cet écart. D'une part, le type

d'analyse spectrale n'est pas le même entre l'étiquetage de référence (LPC) et l'algorithme (fréquence instantanée de la transformée de Fourier). D'autre part, il est possible que le calcul du centre de gravité soit légèrement affecté par quelques valeurs parasites.

Les Tables 2 et 3 présentent la moyenne des erreurs mesurées par la différence moyenne et l'écart type normalisé par rapport aux valeurs de référence. Les différentes voyelles, c'est-à-dire les voyelles brèves (/a/, /i/ et /u/) et les voyelles longues (/A/, /I / et /U /) ont été extraites de quatre phrases prononcées par cinq locuteurs et cinq locutrices. Les phrases de test sont : « ” أَتُوْذِيهَا عَرَفَ وَالِيًا وَقَائِدً,» ”بِآلامِهِمْ؟ ”, “هِيَ هُنَا لَقَدْ آبَتْ”, citées ci-dessus et la dernière phrase est « أَسَرُونَا بِمُنْعَطَف ” » (« ʒasaru:na: bimunɛatafin ») qui signifie (Ils nous ont capturés au niveau d’un virage.).

Table 2 : Moyenne des erreurs de suivi de formants pour cinq locuteurs pour les trois voyelles brèves et longues.

		LPC	Fourier	LPC	Fourier	LPC	Fourier
		F1		F2		F3	
a	Diff	38	23	77	40	123	50
	σ	11	6	26	9	40	11
A	Diff	52	63	76	90	82	68
	σ	11	13	21	20	19	14
i	Diff	34	30	53	120	94	128
	σ	12	10	20	39	38	44
I	Diff	49	58	66	198	112	130
	σ	19	18	27	63	47	46
u	Diff	60	75	155	140	314	276
	σ	24	23	67	38	97	76
U	Diff	103	77	257	101	368	175
	σ	38	25	120	32	121	60
Total	Diff	56	54	114	115	182	138
	σ	19	16	47	33	60	42

Table 3 : Moyenne des erreurs de suivi de formants pour cinq locutrices et les trois voyelles brèves et longues

		LPC	Fourier	LPC	Fourier	LPC	Fourier
		F1		F2		F3	
a	Diff	50	30	81	75	70	38
	σ	10	6	18	15	15	7
A	Diff	58	146	81	118	116	84
	σ	9	18	13	18	18	13
i	Diff	59	117	71	92	109	131
	σ	18	36	24	30	40	43
I	Diff	97	41	362	403	226	262
	σ	43	10	104	106	70	71
u	Diff	78	121	179	136	300	183
	σ	25	37	76	33	86	50
U	Diff	133	107	156	146	198	179
	σ	24	20	34	29	47	45
Total	Diff	79	94	155	162	170	146
	σ	21	21	45	39	46	38

Pour les locuteurs, la Table 2 montre que les résultats de la méthode de suivi automatique des fréquences instantanées de la transformée de Fourier sont bons en particulier pour les voyelles /a/, /A/ et /i/ et meilleurs que pour les voyelles /I/, /u/ et /U/ probablement en raison de leur faible énergie. Pour les locutrices, la Table 3 montre que les résultats sont bons pour les voyelles /a/ et /i/ par rapport aux autres voyelles pour les deux méthodes de suivi, mais plus mitigés pour les autres voyelles.

6. Conclusion

Dans ce papier, nous avons présenté le développement d'une base de données arabe étiquetée phonétiquement et en termes de formants. En outre, nous rapportons dans cet article l'utilisation de cette base de données pour évaluer quantitativement un nouvel algorithme de suivi automatique de formants basée sur la détection des maxima spectraux à l'aide de la fréquence instantanée. Cet algorithme fournit un suivi précis des formants pour F1 et F2 et des résultats parfois moins bons pour F3. Nos travaux futurs viseront l'amélioration de cette méthode en particulier pour les formants de haute fréquence.

7. Remercîments

Ce travail est soutenu par le CMCU : Comité Mixte franco-tunisien de Coopération Universitaire (Projet de Recherche CMCU, code 07G 1112).

Bibliographie

[1] F. Thibault. Trajectory Detection using Hidden Markov Models. In *Proc. Of Sound Processing and Control Lab*, Montreal, Canada, 2003.

[2] J.A.M. Ali, J.V.D. Spiegel and P. Mueller. Robust Auditory-based Processing using the Average Localized Synchrony Detection. In *Proc. Speech and Audio of IEEE Trans*, 2002.

[3] S. McCandless. An algorithm for automatic formant extraction using linear prediction spectra. In *IEEE Trans*, 22:135-141, 1974.

[4] M. Boudraa, B. Boudraa and B. Guerin. Twenty Lists of Ten Arabic Sentences for Assessment. In *Act of Communication ACUSTICA*, 86:870-882, 2000.

[5] L. Deng. A Database of Vocal Tract Resonance Trajectories for Research in Speech Processing. In *Proc. of ICASSP*, 2006.

[6] http://www.loria.fr/~laprie/WinSnoori/.

[7] S. Ghazeli. Back consonants and backing coarticulation in Arabic. *PhD In University of Texas*, Austin, 1977.

[8] A. Braham. An Acoustic study of temporal organization in Arabic specific to Tunisian speakers. *PhD In University of Manouba*, Tunis, 1997.

[9] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 1999.

[10] S. Châari, K. Ouni and N. Ellouze. Wavelet Ridge Track Interpretation in Terms of Formants. In *Proc. of INTERSPEECH-ICSLP*, 1017-1020. Pittsburgh, Pennsylvania, USA, 2006.

[11] F. Charpentier. Pitch detection using the short-term phase spectrum. In *Proc. of ICASSP*, Tokyo, 1986.

[12] http://www.praat.org/