

Application du formalisme multiéchelles microcanonique pour la segmentation des signaux de parole

Vahid Khanagha¹

Khalid Daoudi¹

Oriol Pont¹

Hussein Yahia¹

¹ INRIA Bordeaux Sud-Ouest (équipe GEOSTAT)

351 Cours de la Libération, Bât. A29, 33405 Talence, France

{vahid.khanagha, khalid.daoudi, oriol.pont, hussein.yahia}@inria.fr

Résumé

Dans la présente communication un cadre d'analyse nouveau, le formalisme multiéchelles microcanonique (FMM), est utilisé pour l'étude des signaux de parole. Le FMM est basé sur le calcul des paramètres géométriques et locaux – les exposants de singularité – qui permettent une analyse non-linéaire de la dynamique complexe et, en particulier, de caractériser la signature intermittente. Également, une mesure cumulative de ces exposants qui a la propriété de produire des changements clairs et distinctifs aux frontières des phonèmes est définie. Des expériences préliminaires sur la base de données TIMIT sont présentées. Elles montrent que les exposants de singularité apportent en effet des informations précises sur la dynamique locale de la parole. Ces expériences montrent également que la mesure proposée a un bon potentiel de fournir une méthode nouvelle et puissante pour la segmentation phonétique indépendante du texte.

Mots clefs

formalisme multiéchelles microcanonique, segmentation de signaux de parole.

1 Introduction

Il est théoriquement et expérimentalement démontré que la turbulence et des phénomènes fortement non-linéaires sont présents dans le processus de production de la parole [8, 2, 9, 10]. Toutefois, l'approche traditionnelle du traitement de la parole est basée sur des techniques linéaires qui s'appuient essentiellement sur le modèle source-filtre. L'approche linéaire ne peut pas prendre suffisamment en compte ou saisir complètement la dynamique complexe de la parole. Pour cette raison, le traitement non-linéaire de la parole a gagné une attention considérable au cours des dernières années.

Dans cet article, la dynamique non-linéaire de la parole en utilisant les concepts et les méthodes du cadre de systèmes turbulents est analysée. Notre approche est basée sur le formalisme multiéchelles microcanonique (FMM), qui est un nouveau cadre pour étudier les propriétés géométriques-statistiques des signaux complexes dans une perspective

multiéchelles [12, 17]. Le FMM s'est avéré être une approche précise pour modéliser et analyser empiriquement des systèmes complexes et turbulents. Cela est particulièrement vrai pour systèmes invariants d'échelle [11].

Le FMM est une extension de son équivalent canonique qui est plus standard [5, 1]. La particularité du FMM est qu'il est basé sur des paramètres géométriques et locaux, plutôt que de s'appuyer sur des moyennes statistiques – telles que les fonctions de structure ou les fonctions de partition – comme c'est le cas dans le cadre canonique [11]. Par conséquent, cela rend possible l'étude de la dynamique locale des signaux complexes.

Nous montrons que les signaux de parole se situent dans le domaine d'applicabilité du FMM. Nous utilisons ensuite les paramètres locaux calculés par le FMM, appelés exposants de singularité [18], et montrons comment ils apportent des informations utiles pour l'identification des frontières de phonèmes.

1.1 Etat de l'art

La segmentation de la parole a des nombreuses applications potentielles dans la technologie de la parole, de la synthèse de la parole à la reconnaissance automatique de la parole (RAP). La segmentation est idéalement la première étape d'un système RAP, mais l'absence d'algorithmes de segmentation précise a conduit à une approche inverse : une large classe de méthodes de segmentation sont des versions adaptées des HMM à base de reconnaissance phonétique [15]. Ces méthodes de segmentation sont connus comme méthodes « dépendantes du texte », car ils reposent sur une base de données externe fournie de vocabulaire cible et ses transcriptions manuelles. D'autre part, il existe une catégorie de méthodes de segmentation, qui sont « indépendantes du texte ». Elles sont basées sur l'identification des variations des distances caractéristiques de la parole [3]. Les méthodes indépendantes du texte ne sont pas limitées à un corpus spécifique et elles s'appuient sur des paramètres acoustiques ou des mesures spectrales de base.

Dans ce papier, les changements temporels de la distribution des exposants de singularité sont exploités avec l'aide d'une mesure cumulative. Nous présentons des expériences

préliminaires qui montrent que cette mesure peut être utilisée facilement pour détecter les frontières de phonèmes. Le document est structuré comme suit : La section 2 introduit des concepts fondamentaux du FMM. D’abord, l’algorithme d’estimation des exposants de singularité est présenté, et ensuite les conditions de validité du FMM sont spécifiées. La section 3 montre que la parole est un signal approprié pour ce formalisme et la section 4 discute l’utilisation des exposants de singularité pour la segmentation des signaux de parole. Enfin, dans la section 5, nous tirons nos conclusions.

2 Le formalisme multiéchelles microcanonique

Dans cette section, un bref aperçu sur les bases du FMM est présenté. Un examen plus approfondi de la théorie et les méthodes peut être trouvé dans [17]. Le FMM est basé sur le calcul des exposants d’échelle locaux d’un signal donné dont leur distribution est la quantité clé définissant sa dynamique intermittente. Ces exposants sont utiles pour l’étude des propriétés géométriques des signaux, et ils ont été utilisés dans une grande variété d’applications allant de la compression de données à l’inférence et la prévision de signaux [16, 13].

Comme avec tous les modèles, avant d’appliquer le FMM à un signal donné, la première étape consiste à étudier sa validité. La validité du FMM pour un signal repose sur l’existence d’une loi de puissance locale à chaque point du domaine du signal [17].

Formellement, pour au moins une fonctionnelle dépendante d’échelle Γ_r , la relation suivante doit être valide pour tout instant t :

$$\Gamma_r(s(t)) = \alpha(t) r^{h(t)} + o\left(r^{h(t)}\right) \quad r \rightarrow 0 \quad (1)$$

où $h(t)$ est l’exposant de singularité du signal $s(t)$ [17]. Le facteur multiplicatif $\alpha(t)$ dépend de la fonctionnelle Γ_r choisie, mais pour certains systèmes invariants d’échelle, l’exposant $h(t)$ est indépendante d’elle. Cependant, il est hors du but de ce document d’étudier si la parole a des propriétés d’invariance d’échelle. Le terme $o\left(r^{h(t)}\right)$ regroupe des perturbations additives qui, pour des échelles petites, sont négligeables par rapport au terme principal. Donc $h(t)$ quantifie la contribution dominante au « degré de régularité » de $s(t)$ à chaque instant du temps.

Si la fonctionnelle est choisie comme l’incrément linéaire, $\Gamma_r(s(t)) = s(t+r) - s(t)$, les exposants qui en résultent sont exposants Hölder et ils caractérisent des corrélations en loi de puissance. Lorsque des données empiriques sont analysées, il est typiquement difficile d’obtenir une bonne estimation des exposants Hölder à partir d’incrément linéaires : De la discrétisation, du bruit et des corrélations à longue portée entravent le calcul pratique de ces exposants avec l’Eq. (1).

Il existe une définition alternative et plus robuste pour la fonctionnelle Γ_r dans l’Eq. (1), qui est définie à partir de

la quantification typique de l’intermittence en turbulence : la mesure « module du gradient ». Cette mesure décrit la dissipation d’énergie à l’échelle r du champ vitesse turbulent. Par conséquent, il est une quantité liée au transfert d’énergie d’une échelle à l’autre. Ainsi, l’exposant associé à la loi de puissance en termes d’échelle caractérise le contenu d’information et les transitions dynamiques du signal [5, 18]. Dans ce cas, la fonctionnelle Γ_r est définie comme la mesure module du gradient à échelle r divisé par le volume de la boule de rayon r :

$$\Gamma_r(s(t)) := \frac{1}{\Lambda(B_r)} \int_{B_r(t)} d\tau |s'(\tau)| \quad (2)$$

où s' est la dérivée de s , B_r est la boule de rayon r et Λ désigne la mesure de Lebesgue sur l’axe réel. La mise en œuvre pratique pour éviter du bruit et des artefacts de discrétisation consiste à utiliser un support d’ondelettes pour la boule $B_r(t)$.

Il existe un ensemble particulier de points qui expriment le plus d’information sur la dynamique non linéaire du signal : la variété la plus singulière. En fait, pour un point donné, la plus petite est la valeur du coefficient de singularité, la plus grande est la prévisibilité future conditionnée à ce point [16]. Les transitions critiques du système surviennent aux points les plus singuliers, et ce fait a été utilisé avec succès dans nombreuses applications telles que la reconstruction de données perdues ou la prévision de valeurs futures [13].

2.1 Estimation des exposants de singularité

La méthode utilisée dans cette étude pour estimer les exposants de singularité est l’équivalent de l’Eq. (2) en termes de transformée en ondelettes continue. Ceci a l’avantage général de faire face aux particularités des données réelles telles que la discrétisation, le bruit d’acquisition et les corrélations à longue portée. En plus, la transformée en ondelettes annule des contributions polynômiales au terme additif $o\left(r^{h(t)}\right)$, qui représentent un obstacle commun pour l’estimation précise des exposants de singularité. En général, nous examinons la loi de puissance à chaque instant :

$$\mathbb{T}_\Psi[|s'|](r, t) \propto r^{h(t)} \quad (3)$$

où $\mathbb{T}_\Psi[x](r, t) := (\Psi_r * x)(t)$ représente la transformée en ondelettes continue, $\Psi_r(t) := r^{-1} \Psi(r/t)$ et Ψ est une fonction appelée ondelette mère.

Il convient de mentionner un autre avantage d’utiliser la transformée en ondelettes continue pour cette estimation : la possibilité de calculer la transformée sur un ensemble d’échelles non entières pour un signal discrétisé. La variable d’échelle r dans l’Eq. (3) peut être assignée des valeurs non entières, fournissant un schéma d’interpolation pour signaux en temps discret.

2.2 Validation du FMM

C’est facile de voir que prendre le logarithme aux deux côtés de l’Eq. (3) révèle une relation linéaire entre le loga-

rithme de la transformée en ondelettes et le logarithme de l'échelle. Il est donc possible d'estimer l'exposant de singularité $h(t)$, à chaque instant t en effectuant une régression log-log de la transformée en ondelettes par rapport à l'échelle. Par conséquent, l'Eq. (3) et donc l'Eq. (1) sont vérifiées pour un signal donné, si nous atteignons des coefficients de corrélation acceptables pour cette régression linéaire. Lorsque cela se produit, le FMM est applicable pour le signal en question.

3 Signaux de parole dans le cadre microcanonique

Dans cette section, nous étudions la validité du FMM pour les signaux de parole. Afin d'estimer les exposants de singularité, nous utilisons une légère modification de l'Eq. (3). En effet, dans notre expérience avec des signaux de parole, nous avons observé qu'une amélioration des coefficients de corrélation est obtenue lorsqu'on prend le logarithme aux deux côtés de l'Eq. (3) et divise par $\log(r)$, de sorte que nous avons la relation linéaire suivante :

$$\frac{\log \mathbb{T}_\Psi [s'](r, t)}{\log r} = \alpha_\Psi(t) \frac{1}{\log r} + h(t) \quad (4)$$

Puis, en effectuant la régression linéaire, l'exposant de singularité $h(t)$ est estimé comme le biais de cette relation linéaire.

Nous utilisons l'ondelette lorentzienne comme ondelette mère. Cette ondelette définit une estimation précise pour les exposants les plus petits, au détriment d'une saturation de tous les exposants ≥ 1 [19]. Ceci est souhaitable car les exposants les plus petits sont les plus informatifs et, dans le cas présenté, les exposants obtenus sont loin de la saturation et donc ce fait n'apparaît pas comme une limitation effective. Pour faire la régression, nous avons choisi 10 échelles qui sont log-uniformément espacés entre 1 et 100 échantillons (qui correspondent à l'intervalle de 62.5 μ s à 6.25 ms).

D'abord, nous vérifions l'existence des lois de puissance, l'Eq. (3), sur les phonèmes, car ils sont les unités acoustiques fondamentales de la parole. Toutes nos expériences sont effectuées sur la base de données TIMIT. Nous utilisons les transcriptions fournies par TIMIT pour construire une base de données d'étude de 3000 phonèmes : 500 instances différentes d'un phonème représentatif de chaque famille de phonèmes (voyelles, fricatives, occlusives, semi-voyelles et liquides, affriquées et nasales). L'estimation des exposants de singularité est effectuée en utilisant l'Eq. (4). La moyenne des coefficients de corrélation est présentée au Tableau 1.

Ensuite, nous avons effectué la même procédure sur des phrases entières. 500 signaux de parole d'une longueur approximative de 3 à 5 secondes ont été utilisés pour cette expérience. Nous avons obtenu un coefficient de corrélation moyen de 0.96. La perte faible par rapport à la moyenne des valeurs dans le Tableau 1 (qui est de 0.98) s'explique

par la présence de longs segments de silence aux phrases entières.

Dans l'ensemble d'étude, cet expérience montre que la procédure d'estimation d'exposants utilisée obtient des coefficients de corrélation d'excellente qualité. Premièrement, ceci suggère que nous attendons une estimation très précise des exposants de singularité. Deuxièmement, ceci suggère aussi que le FMM est applicable aux signaux de parole. Nous pouvons donc procéder maintenant à étudier comment ces exposants apportent des informations utiles pour l'étude de la dynamique de la parole.

4 Application du FMM pour la segmentation du signal de parole

La parole est un signal non stationnaire qui est formé par la concaténation de petites unités acoustiques appelées phonèmes. La détection automatique des frontières entre les phonèmes est une tâche difficile et est encore un problème ouvert qui a de nombreuses applications dans la technologie de la parole. Dans la section 3, nous avons démontré la validité du FMM pour les signaux de parole. Ici, nous présentons nos observations sur les informations apportées par les exposants de singularité sur la dynamique variable des signaux de parole. Puisque des phonèmes différents sont fondamentalement signaux différents, chacun avec ses propriétés dynamiques caractéristiques, nous prévoyons que les exposants de singularité ont un comportement différent à l'intérieur des frontières de chaque phonème.

Afin de démontrer ces changements dynamiques, nous présentons les changements dans la distribution des exposants de singularité conditionnée au temps, $\rho(h|t)$, Figure 1.

Dans la Figure 1–haut, un signal de parole est montré et ses frontières de phonèmes sont représentés par des lignes verticales rouges. Ces limites sont extraites de la transcription manuelle de la base de données TIMIT. La Figure 1–milieu, affiche l'évolution temporelle de la distribution des exposants de singularité. Dans les axes verticaux, nous montrons le rang des exposants de singularité regroupés en classes de 5 centiles de largeur. À chaque instant du temps t , nous prenons une fenêtre de 30 ms, centrée autour de t et nous regroupons les exposants selon son rang global. Lorsque nous voulons représenter la probabilité conditionnelle, chaque colonne a été normalisée par sa norme- ∞ . Il est remarquable qu'il y a des changements dans la position du maximum et la variabilité de la distribution de h . En plus, la distribution alterne entre unimodale et multimodale, avec les cas unimodaux centrés au milieu du domaine et les cas multimodaux généralement avec deux modes : un à chaque extrême de la gamme.

Toutefois, ces changements de comportement de distribution sont visibles à l'œil, mais ils seraient extrêmement difficiles à détecter numériquement et automatiquement. Par conséquent, c'est nécessaire de définir une nouvelle mesure qui exploite ces changements de distribution. Dans ce but, nous remarquons que le principal changement dans les dis-

Type de phonème	Voyelles	Fricatives	Occlusives	Semi-voyelles et liquides	Affriquées	Nasales
Phonème	/aa/	/dh/	/b/	/el/	/ch/	/en/
Moyenne des coef. de corrélation	0.97	0.99	0.99	0.99	0.99	0.99

Tableau 1 – La moyenne des coefficients de corrélation de la régression linéaire l'Eq. (3) pour un phonème représentatif de chaque famille.

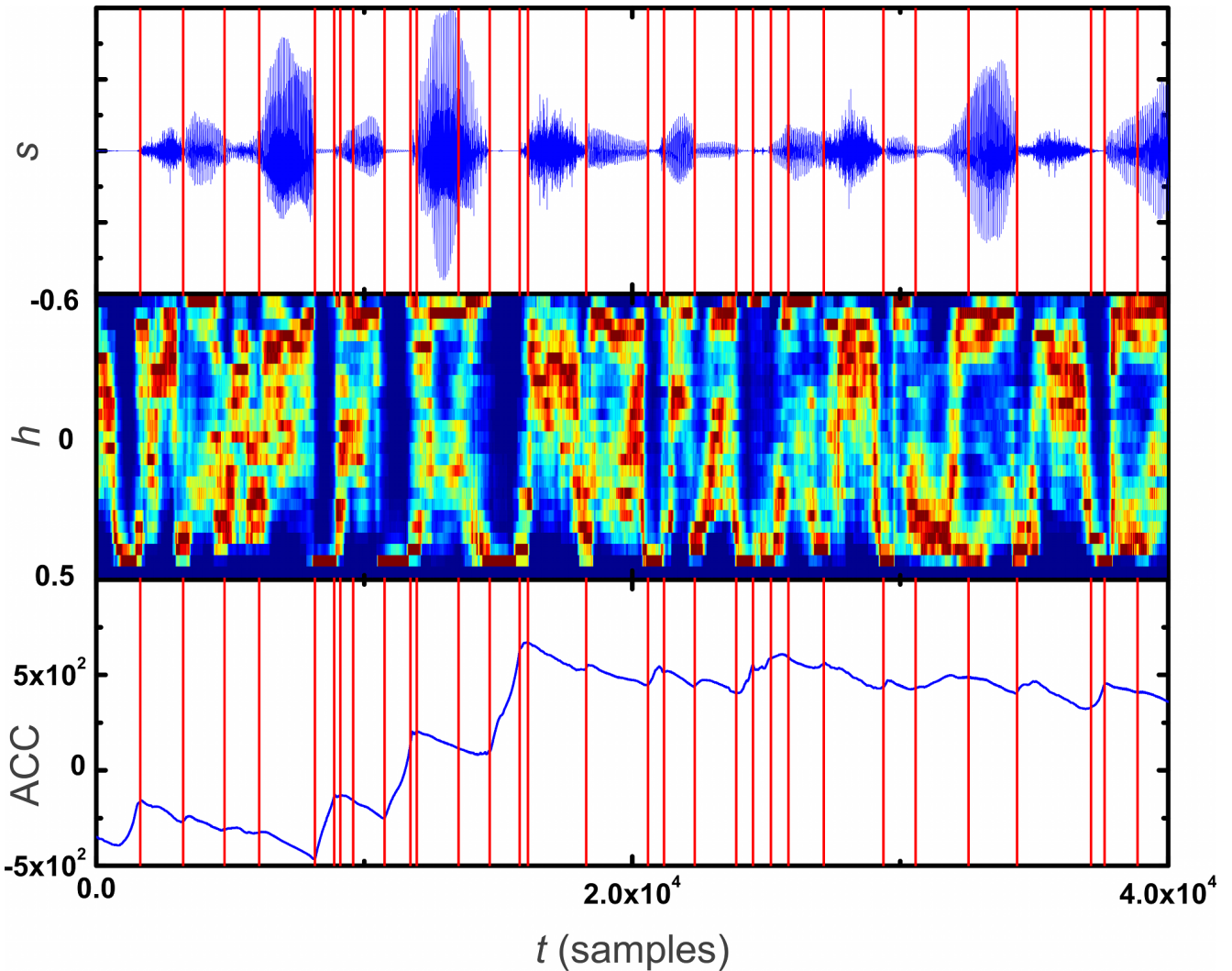


Figure 1 – **HAUT** : Un signal de parole de la base de données TIMIT, consistant en une portion de 40.000-échantillons de la phrase « She had your dark suit in greasy wash water all year ». Le signal est proportionnel à la pression d'air et il est présenté dans une échelle normalisée. Le signal est échantillonné à 16 kHz. Les frontières de phonèmes qui ont été manuellement positionnées, sont marquées par des lignes verticales rouges. **MILIEU** : Histogramme conjoint de la distribution des exposants de singularité (axe vertical) conditionnée à la fenêtre de temps (axe horizontal). Le rouge correspond à la probabilité maximale et le bleu foncé correspond à une probabilité nulle. L'axe horizontal est divisé en fenêtres de 30 ms. L'axe vertical est divisé en classes d'histogramme de largeur 5 centiles, l'échelle de l'axe est proportionnelle au rang des exposants de singularité, et non pas à leur valeur. Cela permet d'éviter les distorsions de probabilité faible. **BAS** : La fonctionnelle proposée pour l'identification des frontières de phonèmes. Il est remarquable que la plupart des frontières de phonèmes concordent avec des forts changements de pente. Pour améliorer la présentation, la tendance globale de la fonctionnelle ACC présentée dans l'Eq. (5) a été soustraite (pour la phrase entière, pas pour la partie présentée).

tributions de la Figure 1–milieu, est le changement dans les moyennes. En autres termes, nous prévoyons que des phonèmes différents ont des moyennes des exposants de singularité différentes par rapport à leurs phonèmes voisins.

Afin de vérifier cela, nous utilisons la primitive de la fonction exposant de singularité comme estimateur de la moyenne instantanée. Formellement, nous définissons la nouvelle fonctionnelle comme :

$$ACC(t) = \int_{t_0}^t d\tau h(\tau) \quad (5)$$

La fonctionnelle obtenue est tracée dans la Figure 1–bas pour le même signal de parole qu’avant. Pour améliorer la présentation des valeurs résultantes dans une fenêtre observable, nous soustrayons la moyenne globale des exposants de singularité de la valeur de l’exposant à chaque instant du temps.

Comme nous attendions, cette nouvelle fonctionnelle révèle les changements dans la distribution d’une manière très précise. En effet, à l’intérieur de chaque phonème la fonctionnelle ACC est presque linéaire (si nous négligeons les fluctuations à petite échelle). En outre, il y a des abruptes changements de pente aux frontières des phonèmes. Ces changements de pente permettent encore d’identifier les frontières entre les phonèmes extrêmement courts, telles que les occlusives. Des observations détaillées sur tout l’ensemble de signaux confirment ce comportement et, donc, la solidité de la fonctionnelle proposée, l’Eq. (5). Ces expériences suggèrent que les exposants de singularité liés au FMM en effet apportent des informations utiles sur les transitions critiques dans les signaux de parole. Elles suggèrent aussi que nous pouvons facilement utiliser ces exposants pour mettre au point une nouvelle méthode robuste pour la segmentation phonétique.

Une évaluation précise de cette mesure nécessite l’application d’un algorithme numérique non-supervisé pour l’identification de points de rupture à courbes bruyants qui sont linéaires par morceaux. C’est le but de nos recherches en cours. Au moment de la rédaction de cet article, nous étudions l’utilisation d’algorithmes Free Knot B-spline [14] pour atteindre cet objectif.

Les méthodes les plus proches de la nôtre que nous avons trouvées dans la littérature sont celles de [7] et [4]. Dans le premier cas, l’analyse de la trajectoire de la variation de la dimension fractale est utilisée pour la segmentation phonétique. Dans la deuxième étude, les auteurs proposent une approche basée sur la fractalité qui utilise les transitions de l’enveloppe de la dimension fractale locale pour déterminer les frontières entre les mots et les phonèmes.

Une comparaison précise entre ces méthodes et la nôtre est au-delà du but du présent document, car ces méthodes – comme la nôtre, pour le moment – seulement distinguent visuellement les phonèmes sans donner une procédure de segmentation automatique. Cependant, nous pouvons dire que la mesure que nous proposons semble beaucoup plus précise et plus facile à incorporer dans un algorithme de

segmentation automatique que les mesures données dans [7] et [4].

5 Conclusions

Dans ce document, nous avons d’abord montré que le FMM est un cadre valable pour l’étude des signaux de parole. Dans cette perspective, nous avons ensuite analysé les propriétés locales des signaux de parole par les exposants de singularité en termes du FMM. Nous avons montré que ces exposants apportent des informations importantes sur la dynamique de la parole. Enfin, nous avons proposé une mesure géométrique de quantification qui produit des changements clairs et distincts aux frontières de phonèmes, et donc elle peut être utilisée pour une segmentation automatique et indépendante du texte. Nous soulignons que l’application complète du FMM pour le signal de parole exige des justifications plus précises pour faire face à toutes ses particularités. Pourtant, l’étude présentée dans le présent document révèle le caractère informatif des exposants de singularité, sans aucune manipulation supplémentaire.

Références

- [1] A. Arneodo, F. Argoul, E. Bacry, J. Elezgaray, and J. F. Muzy. *Ondelettes, multifractales et turbulence*. Diderot Editeur, Paris, France, 1995.
- [2] A. Barney, C. Shadle, and P. Davies. Fluid flow in a dynamical mechanical model of the vocal folds and tract : part 1 & 2. *J. Acoust. Soc. Amer.*, 105(1) :444–466, Nov. 1999.
- [3] A. Esposito and G. Aversano. Text independent methods for speech segmentation. In *Summer School on Neural Networks 2004*, pages 261–290, 2004.
- [4] P. C. Fantinato, R. C. Guido, S.-H. Chen, B. L. S. Santos, L. S. Vieira, S. B. J. L. C. Rodrigues, F. Sanchez, J. Escola, L. M. Souza, C. D. Maciel, P. R. Scalassara, and J. Pereira. A fractal-based approach for speech segmentation. In *Tenth IEEE International Symposium on Multimedia*, pages 551–555, Los Alamitos, CA, USA, 2008. IEEE Computer Society.
- [5] U. Frisch. *Turbulence : The legacy of A.N. Kolmogorov*. Cambridge Univ. Press, Cambridge MA, 1995.
- [6] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren. Darpa timit acoustic-phonetic continuous speech corpus. Technical report, U.S. Dept. of Commerce, NIST, Gaithersburg, MD, 1993.
- [7] W. Kinsner and W. Grieder. Speech segmentation using multifractal measures and amplification of signal features. In *Cognitive Informatics, 7th IEEE International Conference on*, Oct. 2008.
- [8] I. Kokkinos and P. Maragos. Nonlinear speech analysis using models for chaotic systems. *IEEE Transactions on Speech and Audio Processing*, 13(6) :1098–1109, Jan. 2005.

- [9] A. Kumar and S. Mullick. Nonlinear dynamical analysis of speech. *J. Acoust. Soc. Amer.*, 100(1) :615–629, 1996.
- [10] M. Little. *Biomechanically Informed Nonlinear Speech Signal Processing*. PhD thesis, Oxford University, 2007.
- [11] O. Pont, A. Turiel, and C. Perez-Vicente. Empirical evidences of a common multifractal signature in economic, biological and physical systems. *Physica A*, 388(10) :2025–2035, May 2009.
- [12] O. Pont, A. Turiel, and C. J. Pérez-Vicente. Application of the microcanonical multifractal formalism to monofractal systems. *Physical Review E*, 74 :061110–061123, 2006.
- [13] O. Pont, A. Turiel, and C. J. Pérez-Vicente. Description, modeling and forecasting of data with optimal wavelets. *Journal of Economic Interaction and Co-ordination*, 4(1) :39–54, June 2009.
- [14] H. Schwetlick and T. Schutze. Least squares approximation by splines with free knots. *BIT Numerical Mathematics*, 35(3) :361–384, Sep. 1995.
- [15] D. Torre-Toledano, L. Hernandez-Gomez, and L. Villarrubia-Grande. Automatic phonetic segmentation. *IEEE Transactions on Speech and Audio Processing*, 11(6) :617–625, 2003.
- [16] A. Turiel and A. del Pozo. Reconstructing images from their most singular fractal manifold. *IEEE Trans. on Im. Proc.*, 11 :345–350, 2002.
- [17] A. Turiel and C. P.-V. H. Yahia. Microcanonical multifractal formalism : a geometrical approach to multifractal systems. part 1 : singularity analysis. *J. Phys. A, Math. Theor.*, 41 :015501, 2008.
- [18] A. Turiel and N. Parga. The multi-fractal structure of contrast changes innatural images : from sharp edges to textures. *Neural Computation*, 12 :763–793, 2000.
- [19] A. Turiel and C. Pérez-Vicente. Multifractal measures : definition, description, synthesis and analysis. a detailed study. In J.-P. Nadal, A. Turiel, and H. Yahia, editors, *Proceedings of the "Journées d'étude sur les méthodes pour les signaux complexes en traitement d'image"*, pages 41–57, Rocquencourt, 2004. INRIA.