Suivi temporel de stimuli dynamiques interférants par marquage du plan temps-fréquence utilisant une statistique de passages par zéro

François Gaillard, Frédéric Berthommier, Jean-Luc Schwartz, Gang Feng

Institut de la Communication Parlée

CNRS URA 5008 – Université Stendhal - INPG

ICP/INPG, 46 avenue Félix Viallet 38031 GRENOBLE cedex 01
Tél : (+33) 04 76 57 47 15 Fax : (+33) 04 76 57 47 10

{gaillard, schwartz, bertho, feng}@icp.inpg.fr

Résumé – Dans un cadre d'Analyse de Scènes Auditives Computationnelle (CASA), ce papier présente un modèle de marquage du plan temps-fréquence par détection d'harmonicité. L'originalité du modèle tient à l'exploitation d'une statistique des passages par zéros du signal temporel pour le marquage, statistique qui fournit une mesure de la fiabilité du marquage par le biais de l'écart-type des longueurs d'intervalles inter-zéros du premier ordre. Après avoir présenté le modèle et son comportement, nous montrons que celui-ci peut-être utilisé pour le suivi de stimuli dynamiques présentant de fortes variations prosodiques.

Abstract – In a CASA context, a harmonicity-based time-frequency labeling method is described. This method exploits a statistics on first-order inter-zeros interval lengths to perform the labeling; moreover, the time-frequency plane is labeled thanks to a reliability criterion, based on the standard deviation of interval lengths, which provides information about the presence (or not) of interference. After a description of the algorithm and its properties, we show that it can be used for tracking applications involving dynamical stimuli.

1. Introduction

Notre système auditif est capable, lorsqu'il est plongé dans une scène auditive, de structurer son environnement sonore : en contexte de sources acoustiques mélangées, il sait *regrouper* les différentes composantes d'une même source pour l'identifier, et *séparer* les différentes sources. L'Analyse de Scènes Auditives (ASA) [1] suggère, pour expliquer cette capacité, l'existence d'un processus faisant intervenir une coopération entre plusieurs traitements d'indices primitifs extraits du signal acoustique lui-même, tels que des indices d'harmonicité ou de différence de marche interaurale par exemple. Dans ce cadre, l'Analyse de Scènes Auditives Computationnelle (CASA) se propose de modéliser ce processus par le biais de l'outil informatique [6]. Classiquement, les modèles CASA construisent d'abord des images de la scène auditive, appelées "représentations intermédiaires", et issues du marquage du plan temps-fréquence selon chaque indice primitif. Ces représentations alimentent ensuite, en coopération, un processus de reconnaissance [2], capable de grouper et d'identifier les composantes de chacune des sources présentes dans la scène auditive.

Dans une stratégie CASA de construction d'une représentation intermédiaire par détection d'harmonicité, nous proposons dans cet article de réhabiliter une méthode ancienne et simple d'extraction de l'indice premier de l'harmonicité des signaux de parole: la fréquence fondamentale (F0) des sons voisés. Cette méthode, basée sur une statistique des longueurs d'intervalles inter-zéros du premier ordre, et connue pour sa grande sensibilité au bruit, a toujours été présentée comme une méthode inutilisable en conditions interférentes. Nos travaux précédents ont montré qu'en l'adaptant aux conditions interférantes, cette méthode permettait de fournir un marquage efficace et rapide du plan temps-fréquence pour la construction de la représentation intermédiaire correspondante [3]. La Section 2 permettra de rappeler le principe et l'architecture de l'algorithme que nous avons développé, ainsi que les grandes lignes de ses propriétés, issues d'une évaluation menée en différents paradigmes d'interférences. Puis, alors que ces résultats témoignent essentiellement d'une analyse fenêtre temporelle par fenêtre temporelle d'une situation de mélange, nous proposerons en Section 3 de généraliser son utilisation au suivi de stimuli dynamiques au cours du temps.

2. Marquage du plan temps-fréquence par extraction de passages par zéro

2.1. Principe

Le principe de la méthode utilisée peut être résumé par les deux observations suivantes [3] (figure 1):

- Les intervalles inter-zéros (en pente fixée, dans notre cas, montante) consécutifs d'un son pur sont tous égaux à la période de ce son pur; dans ce cas, la moyenne (μ) des longueurs d'intervalles contenus dans une fenêtre temporelle est égale à la période du son pur.
- Considérons que la dégradation de ce son pur peut avoir deux causes, que sont la perte de périodicité d'une part (modélisable par un élargissement de bande passante ΔF) et la perte de dominance énergétique d'autre part (modélisable par l'apparition d'un second son pur

interférant, de niveau relatif d'énergie NR variable), ces deux dégradations vont engendrer une dispersion des longueurs d'intervalles inter-zéros, et donc une augmentation de l'écart-type (σ) des longueurs d'intervalles contenus dans une fenêtre temporelle; en regard de cette augmentation, la moyenne (μ) des longueurs d'intervalles ne fournit plus la période du son pur.

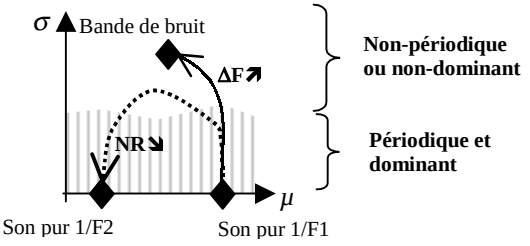

FIG 1. Le principe de l'utilisation des statistiques de passages par zéro. NR=niveau relatif en énergie (cas d'un mélange progressif de deux sons purs); RSB=rapport signal/bruit (cas du bruitage d'un son pur)

Ainsi, en se dotant de deux estimateurs classiques de moyenne et d'écart-type (notés $\hat{\mu}$ et $\hat{\sigma}$) pour lesquels les propriétés statistiques ont été proprement caractérisées, et pour chaque fenêtre temporelle de signal, l'ordre de grandeur de la valeur de $\hat{\sigma}$ va permettre de disposer d'un "détecteur de son pur": une faible valeur de $\hat{\sigma}$ indique la présence d'un son pur dominant en énergie, et dans ce cas, $\hat{\mu}$ fournit sa période fondamentale; en revanche, une forte valeur de $\hat{\sigma}$ indique toute autre situation (situation d'interférence, perte de périodicité), et, dans ce cas, $\hat{\mu}$ est inutilisable.

2.2. Le modèle complet

Le principe décrit ci-dessus ayant été énoncé sur la base d'observation de sons purs et mélanges de sons purs, il est nécessaire, pour traiter des situations d'interférences entre sons complexes, de développer un modèle intégrant la dimension spectrale des sons complexes, et donc travaillant dans le plan temps-fréquence. Nous avons donc développé un modèle de type CASA, dont le rôle est de marquer le plan temps-fréquence selon la présence (ou non) d'une composante périodique et dominante en énergie dans chacune de ses régions. Ce plan temps-fréquence étiqueté constitue alors une représentation intermédiaire contenant les informations d'harmonicité, et pourra alimenter un processus de reconnaissance.

Dans un esprit CASA, le modèle se doit de rester proche des réalités physiologiques; pour cette raison, celui-ci est constitué des étages suivants (figure 2):

- Une pré-accentuation (6dB/octave, par dérivation), qui filtre l'énergie basse fréquence provenant de la source pour rehausser les formants des spectres de signaux de parole ; cette pré-accentuation modélise le filtrage effectué par l'oreille moyenne;
- Modélisant la décomposition fréquentielle effectuée par l'oreille interne, une analyse spectrale par banc de filtres dits "gammatones" (γ) (proposés pour la modélisation du système auditif [4]) à 32 canaux produit la représentation temps-fréquence de chaque portion temporelle du signal de parole. Les filtres gammatones ont des fréquences centrales comprises entre 38Hz et 5kHz, et leurs largeur de bande augmente avec cette fréquence centrale.
- Un étage de démodulation, constitué d'une rectification simple-alternance (RSA) suivie d'un filtrage passe bande 50-200Hz est appliqué aux canaux de hautes (HF) et moyennes (MF) fréquences ; dans ces canaux en effet, la non-résolution des harmoniques due à la largeur de bande des filtres gammatone appelle, pour retrouver F0, l'utilisation d'un étage de démodulation des battements entre harmoniques non-résolus. Dans les bas canaux en revanche (canaux BF), la faible largeur de bande rend les harmoniques résolus; l'étage de démodulation n'est pas requis. Le fonctionnement de cet étage est crucial dans notre modèle ; une étude fine, constituée de résultats théoriques et de simulations, a montré que notre démodulateur remplace proprement un signal harmonique interférant de fréquence fondamentale F0 par un son pur (de même fréquence F0) interférant, dans des conditions de dominance spectrale similaires; il joue donc bien le rôle de transfert (non-linéaire) attendu.
- Un étage d'extraction de passages par zéro (PPZ), qui, à partir des deux estimations $\hat{\mu}$ et $\hat{\sigma}$ effectuées sur une fenêtre temporelle de 80ms, permet de prendre une décision sur la présence (ou non) d'une composante périodique et dominante en énergie dans le pavé temps-fréquence considéré. Cette décision est prise dans un plan ($\hat{\mu}, \hat{\sigma}$), selon la position du point ($\hat{\mu}, \hat{\sigma}$) par rapport à une frontière séparatrice de ce plan. Cette frontière, construite par analyse statistique sur bruit blanc, sépare le plan en deux régions : faibles valeurs de $\hat{\sigma}$ vs. fortes valeurs de $\hat{\sigma}$. Notre postulat de départ est alors le suivant: si le point est localisé sous la frontière, alors il existe, dans le pavé temps-fréquence correspondant, une composante de source périodique et dominante en énergie, et la période de cette source (ou l'un de ses sous-multiples, dans le cas où l'harmonique résolu issu des bas canaux n'est pas le fondamental mais un harmonique d'ordre supérieur) est donnée par $\hat{\mu}$; si le point est localisé au-dessus de la frontière en revanche (région des fortes valeurs de $\hat{\sigma}$), aucune source périodique et dominante n'est présente dans ce pavé temps-fréquence.

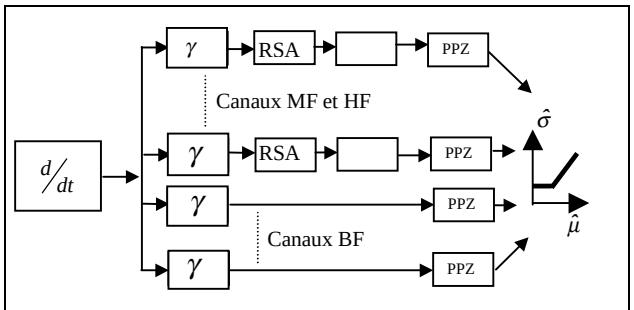

FIG 2. Le modèle complet de marquage du plan temps-fréquence

En ne conservant que les points localisés sous la frontière, on peut alors marquer le plan temps-fréquence selon la période identifiée.

2.3. Evaluation

De nombreuses expériences ont été menées selon différents paradigmes d'interférences, mélangeant voyelles synthétiques ou naturelles, de diverses fréquences fondamentales, en contexte bruité par bruit blanc ou non. Ces expériences ont montré que ce modèle est capable de suivre les zones spectrales les plus énergétiques d'une source harmonique (typiquement, les formants d'une configuration vocalique), même si celle-ci est mélangée à du bruit blanc Gaussien ou à une autre source harmonique. Pour l'évaluation, des indices variés, tels que scores d'identification de F0 ou efficacité en termes de détection de formants par exemple, ont été associés à ces paradigmes de tests.

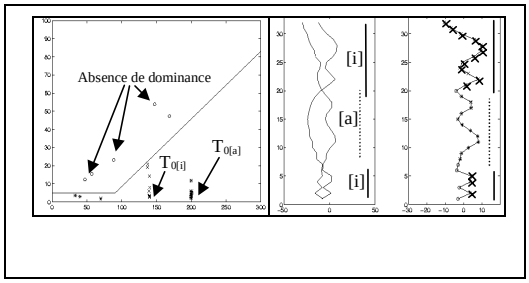

FIG 3. Un exemple de marquage du plan temps-fréquence dans le paradigme des doubles- voyelles : [a], $F0_{[a]}$=100Hz + [i], $F0_{[i]}$=142Hz, de même niveau d'énergie global : plan ($\hat{\mu}$, $\hat{\sigma}$) et spectre marqué ('∗'=[a], 'x'=[i], 'o'=canal perdu) [4].

La figure 3 présente un exemple de marquage du spectre d'un mélange de deux voyelles ($[a]_{F0[a]=100Hz}$ + $[i]_{F0[i]=142Hz}$): le spectre du mélange est marqué à l'aide de trois symboles (deux symboles pour ces deux fréquences fondamentales lorsque le point ($\hat{\mu}$, $\hat{\sigma}$) est localisé en-dessous la frontière, et un symbole pour les points localisés au-dessus de la frontière). En correspondance sont présentées les 32 estimations correspondant aux 32 canaux, dans le plan ($\hat{\mu}$, $\hat{\sigma}$).

Après étude quantitative, constituée, pour chaque paradigme, d'une analyse statistique sur un grand nombre d'estimations, le modèle s'avère être très sélectif: lorsqu'un point ($\hat{\mu}$, $\hat{\sigma}$) se situe sous la frontière, une composante périodique et dominante est effectivement présente dans le pavé du plan temps-fréquence considéré, et sa période (i.e. l'inverse de la fréquence fondamentale du son voisé) est identifiée quasiment sans erreur. Cette tendance se confirme même à fort niveau de bruit dans le cas de bruitage d'une source harmonique, et se confirme également malgré les fluctuations de la parole naturelle. En revanche, la méthode est peu spécifique, dans la mesure où des canaux peuvent être perdus (i.e. des points ($\hat{\mu}$, $\hat{\sigma}$) sont situés au-dessus de la séparatrice malgré la présence d'une composante périodique et dominante en énergie).

2.4. Vers du suivi dynamique…

Le modèle ainsi développé permet de marquer chaque canal spectral selon la présence (ou non) d'une composante de source harmonique et dominante, et ce pour chaque fenêtre temporelle d'analyse. Pour le suivi temporel de stimuli à forte dynamique, on peut alors se demander si le choix d'une fenêtre temporelle de 80ms est un inconvénient majeur : cette analyse à moyen terme ne sera-t-elle pas un obstacle au suivi temporel, dans le cas de variations rapides de F0 dues à la prosodie?
C'est à cette question que nous allons essayer de répondre à présent.

3. Suivi dynamique de signaux à fortes variations prosodiques

3.1. Peignage harmonique

L'étape d'identification de F0 n'était pas requise pour le marquage du plan temps-fréquence, et la représentation intermédiaire binaire (sous la frontière vs. sur la frontière) pourrait alimenter un processus de reconnaissance. Pour le suivi temporel de F0 en revanche, il est nécessaire d'avoir recours à une étape d'identification de F0. Par ailleurs, et du fait de l'application de l'étage de démodulation uniquement sur les canaux MF et HF, certaines estimations $\hat{\mu}$ issues des canaux BF peuvent fournir non pas la période fondamentale, mais un de ses sous-multiples (détection d'harmonique). Il est donc nécessaire, pour suivre l'évolution de F0, de disposer d'un algorithme capable, à partir d'une série harmonique, d'en déterminer la fréquence fondamentale. Nous avons testé l'algorithme en utilisant une technique de repli d'harmoniques proposée par Schroeder [7]. Le principe en est le suivant: pour une estimation sur 32 canaux spectraux ne sont conservés que les canaux correspondant à des points ($\hat{\mu}$, $\hat{\sigma}$) localisés sous la frontière. Les valeurs de $\hat{\mu}$ correspondantes sont consignées dans un tableau, associées à leur pondération (en nombre de canaux). On construit alors un histogramme constitué de toutes ces valeurs $\hat{\mu}$ pondérées par leur poids, ainsi que de tous leurs multiples également pondérées. L'histogramme résultant est alors lissé par un critère d'identification de 5% en valeur relative de F0: ce lissage a pour effet d'intégrer dans une même raie deux raies voisines de moins de 5% (en valeur relative). L'histogramme fait alors apparaître une succession régulière de maxima, et la valeur de la période fondamentale est supposée correspondre au plus petit de ces maxima.

3.2. Protocole

Afin de répondre à la question du choix de la longueur de fenêtre, nous avons voulu tester cet algorithme sur des stimuli dynamiques; les stimulis que nous avons utilisés sont constitués de séquences voisées naturelles à fortes variations prosodiques, les fréquences fondamentales présentant des variations importantes à l'intérieur même de

fenêtre de 80ms [5]. Ces séquences, constituées de six répétitions de la syllabe /ma/, ont été prononcées par un seul locuteur; elles peuvent être classées en six modalités prosodiques (Déclarative, Doute-Incrédulité, Evidence, Exclamative, Question, Ironie de soupçon), pouvant présenter des variation de F0 jusqu'à 0.6Hz/ms.
Le suivi temporel consiste alors à effectuer une série d'estimations sur des fenêtres temporelles de 80ms glissantes par pas de 12.5ms. On dispose, pour chaque modalité prosodique, d'environ 70 fenêtres temporelles, soit 70 estimations ($\hat{\mu}$, $\hat{\sigma}$).

Les estimations ont été menées d'abord sur signaux purs, puis sur signaux bruités par bruit blanc pour différents rapports signal/bruit (RSB, rapport de l'énergie du signal sur l'énergie du bruit, exprimé en dB) RSB$\in\{12; 6; 0\}$ (dB). La Figure 4 présente, pour la modalité "Ironie de Soupçon", le type de résultats obtenus: en regard sont présentés les signaux temporels et les spectrogrammes marqué, ainsi que les patrons mélodiques F0(t) déduits par l'algorithme de peignage harmonique pour les différents niveaux de bruit utilisés.

3.3. Résultats

En termes de marquage du plan temps-fréquence, et selon l'attente, celui-ci se devrait d'être uniformément marqué, les signaux étant toujours voisés. Dans le cas des signaux purs, et pour toutes les modalités, l'harmonicité est, dans le cas général, bien détectée. Si l'on effectue des coupes spectrales des spectrogrammes marqués, on constate que les formants du [a] sont très correctement identifiés. Dans le cas du [m], on retrouve, par coupe spectrale, des caractéristiques typiques du murmure nasal, et notamment un pic d'énergie à 250Hz et un puic vers 1500-2000Hz. En conditions bruitées, on constate que de nombreux canaux sont perdus. En revanche, l'information pertinente est toujours présente; cette représentation intermédiaire pourra donc alimenter un processus de reconnaissance partielle placé en aval.
En termes de suivi de F0, la première observation est que malgré la longueur de notre fenêtre d'analyse, les variations brutales de F0 sont correctement suivies. l'algorithme de peignage harmonique permet de façon satisfaisante de replier les harmoniques issues des canaux BF.
Dans l'exemple de la figure 4, on constate, en absence de bruit, l'émergence des formants du [a] (deux formants localisés en moyennes fréquences), et la possibilité de suivre ces formants. Par addition de bruit, et malgré la perte d'un plus grand nombre de canaux spectraux, les trajectoires formantiques continuent à émerger à des niveaux de bruit importants. L'intérêt de cette méthode, et malgré la croyance commune que les passages par zéro sont très sensibles au bruit, est de continuer à fournir des estimations fiables de F0 malgré l'addition de bruit. La perte de certaines portions de trajectoires formantiques, à fort niveau de bruit, pourrait ensuite être corrigée par lissage temporel de ces trajectoires dans le plan temps-fréquence.

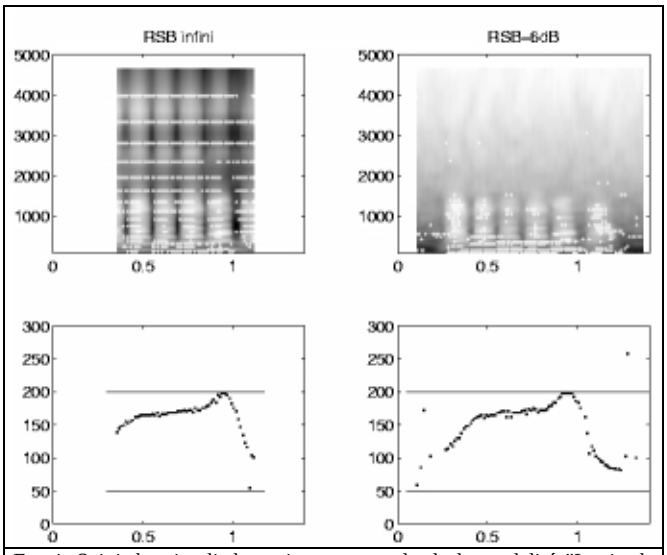

FIG 4. Suivi de stimuli dynamiques: exemple de la modalité "Ironie de Soupçon", sans bruit et avec bruit additif à RSB=6dB. ligne 1: marquage du plan temps-fréquence (une croix indique la détection d'une composante harmonique et dominante de F0 compatible avec le résultat du peignage harmonique dans le pavé du plan temps-fréquence correspondant); ligne 2: patron mélodique F0(t)

4. Conclusion

Ces premières expériences de suivi temporel sont les prémisses de l'exploitation d'un système de marquage du plan temps-fréquence sur stimulis dynamiques. Elles montrent d'une part qu'il est possible de suivre, au cours du temps, les trajectoires formantiques des voyelles en contexte pur ou bruité à l'aide de ce système, et d'autre part que la longueur de la fenêtre d'analyse (80ms), indispensable pour disposer de suffisamment d'intervalles dans chaque fenêtre temporelle, n'est pas un obstacle au suivi, même en cas de variations rapides de F0.
Enfin, ce travail annonce deux grandes voies de poursuites de recherche: d'une part, l'utilisation de la cohérence temporelle des signaux de parole, qui permettra, par lissage temporel de l'image marquée du plan temps-fréquence, de récupérer des portions manquantes de patrons mélodiques, et d'autre part, une validation exhaustive du système par reconnaissance partielle.

Références :

[1] BREGMAN A.S., *"Auditory Scene Analysis"*, MIT Cambridge, MA, 1990.
[2] COOKE M.P. *et al.*, "Recognising occluded speech", *Proc. of Workshop on the Auditory basis of speech perception*, Keele, pp. 297-300, 1996.
[3] GAILLARD F., *"Analyse de Scènes Auditives Computationnelle (CASA) : un outil original de marquage du plan temps-fréquence par harmonicité, exploitant une statistique de passages par zéro"*, Thèse de Doctorat Signal-Image-Parole, INPG, 1999.
[4] HOLDSWORTH J. *et al.*, *"Auditory / connexionist techniques for speech"*, ESPRIT BASIC Research Action 3207, Periodic Progress Report n°2, 1991.
[5] MORLEC Y. *et al.*, "Synthesising attitudes with global rhythmic and intonation contours", *Proc. of Eurospeech'97, Rhodes*, pp. 219-222, 1997.
[6] ROSENTHAL D.F. *et al.*, *"Computational auditory scene analysis"*, Lawrence Erlbaum Associates, publishers. Mahwah, New Jersey, 1998.
[7] SCHROEDER M.R., "Period histogram and product spectrum : new methods for fundamental frequency measurements", *J. Acoust. Soc. Amer.*, vol 43, pp. 829-834, 1968.