Estimation des dyspériodicités vocales dans la parole connectée dysphonique

A. Kacha [(1)], F. Grenez [(1)], J. Schoentgen [(1, 2)]

[(1)] Service Ondes et Signaux, Université Libre de Bruxelles, Bruxelles, Belgique
[(2)] Fond National de la Recherche Scientifique, Belgique
E-mail: akacha@ulb.ac.be

ABSTRACT

Acoustic analysis of connected speech is carried out by means of a generalized variogram to extract vocal dysperiocities. A segmental signal-to-dysperiodicity ratio is used to summarize the perceived degree of hoarseness. The corpora comprise four French sentences as well as vowels [a] produced by 22 male and female normophonic and dysphonic speakers. It is shown that the segmental signal-to-dysperiodicity ratio correlates better with perceptual scores of hoarseness than the global signal-to-dysperiodicity ratio. The perceptual scores are based on comparative judgments by six listeners of pairs of speech tokens.

1. INTRODUCTION

La présentation concerne la classification des voix dysphoniques. Les méthodes de classification basées sur les indices acoustiques sont populaires pour leur nature non invasive et permettent aux cliniciens de suivre l'évolution des patients et de quantifier le degré d'enrouement de la voix.

Plusieurs indices acoustiques sont utilisés pour caractériser la parole des locuteurs dysphoniques. Un nombre de ces indices reflète la déviation du signal de parole voisée par rapport à la périodicité parfaite. Les causes de ces dyspériodicités sont diverses : vibrations non modales des cordes vocales, bruit de modulation comprenant les variations cycle à cycle de la durée de cycle (jitter) et de l'amplitude (shimmer) dues aux perturbations externes, bruit additif dû à une turbulence excessive [7].

La plupart des indices acoustiques sont habituellement obtenus à partir de fragments stables extraits de voyelles soutenues. En effet, les voyelles soutenues sont faciles à analyser lorsque les attaques et les déclins sont exclus parce que les hypothèses de stationnarité et de cyclicité sont alors valables pour beaucoup de locuteurs. L'utilisation d'indices acoustiques obtenus à partir de voyelles soutenues est en fait justifiée par la faisabilité technique plutôt que par la pertinence clinique.

L'évaluation de la qualité de la voix est souvent basée sur la perception de la parole connectée. Par conséquent, on prévoit que les indices acoustiques obtenus à partir de la parole connectée soient mieux corrélés avec l'évaluation perceptuelle de la qualité de la voix. Plusieurs auteurs ont proposé d'extraire les indices acoustiques à partir de la parole connectée. En effet, la parole connectée contient les caractéristiques dynamiques de la source et du conduit vocal tels que les attaques et les déclins et les variations de la fréquence fondamentale et de l'amplitude [5] ce qui la rend plus informative que les voyelles soutenues.

Le nombre d'études sur la parole connectée est relativement faible par rapport aux études traitant les voyelles soutenues. Une revue des travaux publiés sur l'analyse de la parole connectée est donnée dans [1]. La plupart des méthodes d'estimation des dyspériodicities vocales, développées dans le cadre des voyelles soutenues, manquent de robustesse et de précision lorsqu'elles sont appliquées à la parole connectée ou à des voyelles avec attaque et déclin. Le manque de robustesse est une conséquence de l'hypothèse de stationnarité locale qui cesse d'être valable pour la parole connectée produite par des locuteurs enroués.

Dans [6] une approche basée sur un modèle prédictif a été proposée pour l'analyse des dyspériodicités vocales dans la parole connectée. Le modèle comprend un premier étage de prédiction à court terme conventionnel appliqué au signal de parole et un deuxième étage composé d'un prédicteur à long terme appliqué au résidu obtenu à la sortie du premier étage. Plus récemment, dans [1], un modèle de prédiction à long terme bilatérale appliqué directement au signal de parole a été proposé comme alternative. Le minimum des erreurs de prédiction à long terme dans les directions gauche et droite est retenu comme mesure des dyspériodicités, ce qui évite de comparer des fragments de signal à travers les limites phonétiques.

Les méthodes proposées dans [1] et [6] ne garantissent pas que les coefficients de prédiction à long terme soient toujours positifs, ce qui n'est pas cohérent avec la définition mathématique de la périodicité. Pour éviter ce problème, une méthode d'analyse basée sur le variogramme généralisé a été proposée [4].

L'indice acoustique conventionnel utilisé pour quantifier les dyspériodicités vocales dans le signal de parole est le rapport signal à dyspériodicité (RSD) global. La valeur numérique de l'indice global est principalement déterminée par les segments vocaliques dans la parole connectée [1]. Dans cette présentation, on se propose d'utiliser le rapport signal à dyspériodicité segmental (RSDSEG) comme indice acoustique. Il est obtenu par une reformulation locale de l'indice global. Les résultats montrent que le RSD segmental est plus performant que

le RSD global en termes de corrélation avec le degré d'enrouement perçu. Les scores de l'enrouement sont obtenus par une procédure basée sur la comparaison de paires de sons [3].

Le reste de la présentation est organisé comme suit. Dans la Section 2, le corpus utilisé dans l'expérience et les méthodes d'analyse sont présentés. Les résultats expérimentaux sont présentés et discutés dans la Section 3. La conclusion est donnée dans la Section 4.

2. METHODES

2.1. Corpus

Les données comprennent la voyelle [a], incluant attaque et déclin ainsi que quatre phrases produites par 22 locuteurs normophoniques ou dysphoniques (10 hommes et 12 femmes). Le corpus comprend 20 adultes (de 20 ans à 79 ans), un garçon âgé de 14 ans et une fille âgée de 10 ans. Cinq locuteurs sont normophoniques et les autres sont dysphoniques. Les phrases sont les suivantes : "Le garde a endigué l'abbé", "Bob m'avait guidé vers les digues", "Une poule a picoré ton cake" et "Ta tante a appâté une carpe". Ces phrases seront désignées par S1, S2, S3 et S4, respectivement. Elles ont la même structure grammaticale et le même nombre de syllabes. Les phrases S1 et S2 sont voisées par défaut alors que S3 et S4 comprennent des segments phonétiques voisés et non voisés.

Les signaux ont été enregistrés à une fréquence d'échantillonnage de 48 kHz, dans une cabine isolée, au moyen d'un enregistreur audio numérique (Sony TCD D8) et d'un microphone (AKG C41WL) au Département de Laryngologie d'un Hôpital Universitaire à Bruxelles, Belgique. Les enregistrements ont été transférés, par la suite, sur un disque dur d'ordinateur. Les intervalles de silence au début et à la fin des enregistrements ont été supprimés au moyen d'une segmentation manuelle.

2.2. Evaluation auditive

Une évaluation auditive basée sur des jugements comparatifs de paires de stimuli a été utilisée pour déterminer le degré d'enrouement de chaque échantillon (ou stimulus) du corpus composé de la voyelle [a] et des phrases S1 à S4 [3]. Il a été demandé aux auditeurs (ou juges) de comparer deux stimuli en terme du degré d'enrouement. L'objectif est d'hiérarchiser les stimuli, du moins enroué au plus enroué, au moyen de jugements comparatifs de toutes les paires de stimuli extraites de chaque ensemble homogène. La procédure d'évaluation est résumée comme suit :

1. La liste de toutes les paires de stimuli est formée sur la base de l'ensemble des enregistrements. Si N_e représente le nombre d'enregistrements (22 dans notre cas), le nombre de paires de stimuli est $N_e(N_e-1)/2$.

2. Tous les scores des stimuli sont initialisés à zéro.

3. Une paire de stimuli choisie aléatoirement dans la liste est présentée à un auditeur qui doit désigner le stimulus le plus anormal de la paire. L'auditeur a aussi la possibilité de désigner les deux stimuli comme perceptivement identiques.

4. Le score total du stimulus désigné comme étant le plus anormal est augmenté d'une unité. Si les deux stimuli de la paire sont jugés égaux en termes du degré d'enrouement, leurs scores sont augmentés de 0.5 chacun.

5. Les étapes 3 et 4 sont répétées jusqu'à ce que toutes les paires de stimuli appartenant à une même session soient présentées.

6. Les stimuli sont alors caractérisés par leurs scores totaux. Le stimulus qui a été le plus souvent désigné comme étant le plus enroué aura le score le plus élevé tandis que le stimulus le moins enroué recevra le plus faible score.

La procédure est appliquée successivement à la voyelles [a] et aux phrases S1 à S4.

Les stimuli sont présentés via une interface audio numérique-analogique (Digidesign Mbox) et des écouteurs stéréo dynamiques (Sony MDR-7506). L'amplitude des sons est fixée par les auditeurs à un niveau confortable.

Le groupe de juges est composé de 6 auditeurs (une femme, cinq hommes) ayant tous une audition normale. Leurs ages varient de 24 ans à 57 ans. La tâche des auditeurs consiste à classer les stimuli sur la base du degré total de déviance de la voix. Chaque session d'audition est consacrée à un ensemble de 22 stimuli. Le nombre total de sessions est donc égal à 6 x 5 = 30. La même expérience a été répétée par cinq auditeurs après une période d'un jour au moins pour vérifier l'agrément intra-juges.

La moyenne des scores assignés par les six juges a ensuite été choisie comme une mesure subjective du degré d'enrouement perçu.

2.3. Variogramme généralisé

Pour un signal $x(n)$ périodique de période T_0, on peut écrire

$$x(n) = x(n - kT_0),\ k = \cdots, -2, -1, 0, 1, 2, \cdots \qquad (1)$$

Une mesure de l'écart par rapport à la périodicité sur un intervalle de longueur N fournit une indication sur le degré d'irrégularité du signal. Pour les signaux stationnaires, l'énergie des dyspériodicités peut être estimée par

$$\hat{\gamma} = \min_T \left[\sum_{n=0}^{N-1} \big(x(n) - x(n-T)\big)^2 \right], \qquad (2)$$

avec $-T_{\max} \le T \le -T_{\min}$ et $T_{\min} \le T \le T_{\max}$.

L'expression entre crochets dans (2) est connue sous le nom de variogramme et est formellement équivalente à la

différence entre la trame d'analyse courante et une trame décalée de même longueur *N*. L'index temporel *n* positionne les échantillons du signal de parole à l'intérieur de la trame d'analyse. Les bornes T_{min} et T_{max} sont, en nombre d'échantillons, les cycles glottiques acceptables les plus courts et les plus longs. Ils son fixés à 2.5 ms et 20 ms, respectivement (50 Hz ≤ F0 ≤ 400 Hz). Pour les sons voisés, le délai *T* est interprété comme un multiple de la longueur du cycle glottique. Pour les sons non voisés, l'expression (2) demeure valide mais le délai *T* n'est pas interprétable en termes de longueur de cycle glottique.

L'amplitude du signal évolue d'une trame à la suivante à cause des attaques et des déclins, de l'intensité des segments et de l'accentuation. En introduisant un facteur de pondération *a* pour tenir compte de ces variations lentes de l'amplitude du signal, la définition (1) devient

$$x(n) = a\, x(n - kT_0),\ k = \cdots, -2, -1, 0, 1, 2, \cdots \qquad (3)$$

Selon cette définition, le variogramme généralisé prend alors la forme suivante

$$\hat{\gamma} = \min_{T} \left[\sum_{n=0}^{N\text{-}1} \left(x(n) - a\, x(n-T) \right)^2 \right]. \qquad (4)$$

Le gain *a* doit être positif. Il est défini de manière à garantir des énergies identiques dans la fenêtre d'analyse courante et la fenêtre décalée

$$a = \sqrt{\frac{E}{E_T}}, \qquad (5)$$

où *E* et E_T sont les énergies des trames d'analyse courante et décalée,

$$E = \sum_{n=0}^{N-1} x^2(n),\ E_T = \sum_{n=0}^{N-1} x^2(n-T).$$

La longueur de la trame d'analyse *N* et la longueur du décalage sont fixées à 2.5 ms. Ce choix permet de garantir que chaque fragment du signal soit inclus exactement une seule fois dans l'analyse. La valeur instantanée de la dyspériodicité est estimée comme suit :

$$e(n) = x(n) - a\, x(n - T_{opt}),\quad 0 \le n \le N-1 \qquad (6)$$

où T_{opt} est le délai qui minimise le variogramme généralisé (4) pour la position courante de la trame d'analyse. L'analyse est effectuée dans les directions gauche et droite et, par conséquent, le délai T_{opt} peut prendre des valeurs positives ou négatives.

L'approximation du délai optimal par un nombre entier de périodes d'échantillonnage introduit un bruit de quantification. L'effet de ce bruit de quantification est réduit en suréchantillonnant le signal d'un facteur 8.

2.4. Indices acoustiques global et segmental

L'indice acoustique conventionnel utilisé pour quantifier les dyspériodicités vocales dans le signal de parole est le rapport signal à dyspériodicité global exprimé par [1]

$$RSD = 10 \log \left[\frac{\sum_{n=0}^{L-1} x^2(n)}{\sum_{n=0}^{L-1} e^2(n)} \right], \qquad (7)$$

où $e(n)$ est la dyspériodicité instantannée estimée selon (6) et *L* est le nombre d'échantillons dans l'intervalle total d'analyse

Le rapport signal à dyspériodicité segmental (RSDSEG) est connu comme un bon estimateur de la qualité de la parole dans le contexte du codage [2]. Le RSDSEG est obtenu par une reformulation locale du RSD global en calculant le RSD sur des segments courts de l'intervalle d'analyse et en prenant la moyenne de tous les RSD. On prévoit que le RSDSEG d'une production sera mieux corréle avec le degré d'enrouement que le RSD global. En effet, la valeur du RSDSEG est obtenue en appliquant la fonction logarihmique avant de moyenner sur l'ensembles des mesures locales, ce qui permet de donner une plus forte pondération aux segments bruités de faibles niveaux qui sont peu pondérés dans le calcul du RSD global. Par conséquent, les segments de grande amplitude et peu bruités ne masquent pas numériquement la contribution des segments bruités de faible amplitude.

Pour une production donnée, l'intervalle d'analyse est divisé en *K* segments de longueurs *M* et le RSDSEG est calculé comme suit :

$$RSDSEG = \frac{10}{K} \sum_{k=0}^{K-1} \log \frac{\sum_{n=Mk}^{Mk+M-1} x^2(n)}{\sum_{n=Mk}^{Mk+M-1} e^2(n)}. \qquad (8)$$

3. RESULTATS ET DISCUSSION

Les scores de l'enrouement perçu ont été déterminés par six auditeurs. Les scores moyens dépendent légèrement du type de production. Ils varient entre 2.2 et 7.5 pour les locuteurs normophoniques et entre 3.2 et 20.4 pour les locuteurs dysphoniques sur une échelle allant de 0 à 21.

Le RSDSEG a été calculé pour différentes valeurs de la longueur *M* des segments. Les résultat de l'analyse de corrélation entre le RSDSEG et les scores moyens de l'enrouement perçu sont donnés dans le tableau 1 pour les différentes productions (voyelles [a] et phrases S1 à S4). La corrélation dépend légèrement de la longueur du segment et se stabilise à 5 ms. Par la suite, la longueur des segments dans le calcul du RSDSEG a été fixée à cette valeur. Les valeurs du RSDSEG varient de 17.8 dB à 23.8

dB pour les locuteurs normophoniques et de 5.5 dB à 23.8 dB pour les locuteurs dysphoniques.

Le coefficient de corrélation de Pearson des scores moyens de l'enrouement avec les rapports signal à dyspériodicité global et segmental a été calculé et les résultats pour la voyelle [a] et les phrases S1 et S4 sont donnés dans le tableau 2. L'hypothèse nulle ($\rho_P = 0$) a été rejetée pour toutes les entrées du tableau (test unidirectionnel, $\rho_{crit} = 0.36$, $p < 0.05$). L'inspection des résultats du tableau 2 montre que le RSD segmental est plus fortement corrélé que le RSD global pour les phrases S1 à S3. Ceci s'explique par le fait que dans le jugement des sons, les auditeurs sont influencés par les segments bruités quoiqu'ils soient courts et par les fragments vocaliques caractérisés par un grand rapport signal à dyspériocité. Le RSD segmental en fonction du degré d'enrouement correspondant pour la phrase S1 est représenté sur la figure 1.

Le fait que, pour la voyelle, la performance du RSDSEG ne soit pas améliorée, en termes de corrélation avec le degré d'enrouement, par rapport à celle du RSD global est attendu. En effet, les dyspériodicités sont également distribuées dans les sons de parole soutenue. On observe aussi que la performance du RSD n'est pas améliorée pour la phrase S4. Une explication de cette constatation serait que l'évaluation perceptive est moins fiable pour la phrase S4. Les auditeurs ont en effet rapporté que la phrase S4 était relativement difficile à évaluer parce que les intervalles voisés apparaissaient très courts.

Tableau 1 : Coefficients de corrélation de Pearson entre les valeurs du RSD segmental et les scores moyens de l'enrouement pour les voyelle [a] et les phrases S1 à S4. La longueur des segments est indiquée dans la colonne de gauche.

	[a]	S1	S2	S3	S4
30 ms	– 0.71	– 0.84	– 0.80	– 0.78	– 0.65
20 ms	– 0.71	– 0.84	– 0.80	– 0.79	– 0.67
10 ms	– 0.71	– 0.85	– 0.81	– 0.81	– 0.68
5 ms	– 0.70	– 0.86	– 0.81	– 0.81	– 0.70
2.5 ms	– 0.70	– 0.86	– 0.81	– 0.82	– 0.70

Tableau 2: Coefficients de corrélation de Pearson des rapports signal à dyspériodicité global et segmental avec les scores d'enrouement. La longueur des segments est fixée à 5 ms pour le calcul du RSD segmental.

	RSD global	RSD segmental
[a]	– 0.73	– 0.70
S1	– 0.72	– 0.86
S2	– 0.72	– 0.81
S3	– 0.70	– 0.81
S4	– 0.69	– 0.70

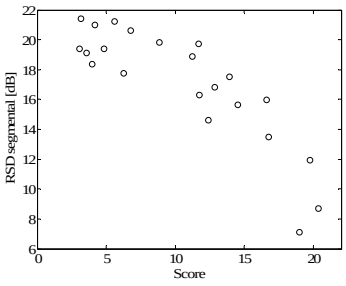

Figure 1 : RSDSEG en fonction des scores d'enrouement associés à la phrase S1 (22 locuteurs) pour une longueur de segment de 5 ms.

4. CONCLUSION

Dans cette présentation, une méthode d'estimation des dyspériodicités vocales dans la parole connectée, basée sur le variogramme généralisé, a été proposée. Le rapport signal à dyspériodicité segmental a été utilisé pour quantifier les dyspériodicités vocales. La performance en terme de corrélation avec le degré d'enrouement perçu a été comparée avec celle du rapport signal à dyspériodicité global conventionnellement utilisé dans l'analyse de la parole dysphonique. Les résultats montrent que l'indice segmental est mieux corrélé avec le degré d'enrouement que l'indice global.

BIBLIOGRAPHIE

[1] F. Bettens, F. Grenez and J. Schoentgen. Estimation of vocal dysperiodicities in connected speech by means of distant-sample bi-directional linear predictive analysis. *J. Acoust. Soc. Am.*, 117: 328-337, 2005.

[2] N.S. Jayant and P. Noll. Digital coding of waveforms :principles and applications to speech and video, Prentice-Hall, Englewood Cliffs, 1984.

[3] A. Kacha, F. Grenez and J. Schoentgen. Voice quality assessment by means of comparative judgments of speech tokens. In *Proc. Intl. Conf. on Spoken Language Processing*, volume 1, pages 1733-1736, 2005.

[4] A. Kacha, F. Grenez, J. Schoentgen and K. Benmahammed. Dysphonic speech analysis using generalized variogram. In *Proc. Intl. Conf. on Acoustic Speech and Signal Processing*, volume 1, pages 917-920, 2005.

[5] F. Klingholtz. Acoustic recognition of voice disorders: A comparative study of running speech versus sustained vowels. *J. Acoust. Soc. Am.*, 87: 2218-2224, 1990.

[6] Y. Qi, R.E. Hillman and C. Milstein. The estimation of signal-to-noise ratio in continuous speech of disordered voices. *J. Acoust. Soc. Am.*, 105: 2532-2535, 1999.

[7] J. Schoentgen. Spectral models of additive and modulation noise in speech and phonatory excitation signals. *J. Acoust. Soc. Am.*, 113: 553-562, 2003.