QUATORZIEME COLLOQUE GRETSI - JUAN-LES-PINS - DU 13 AU 16 SEPTEMBRE 1993

UNE SIMULATION DE L'INCIDENCE DU PITCH SUR L'ESTIMATION FORMANTIQUE EN PAROLE HYPERBARE

J. CRESTEL, M. GUITTON, L. BARBIER

ENSSAT / LASTI
6 rue de Kérampont, BP 447
22305 Lannion Cedex (France)

RESUME

Trois assertions fixent le cadre du problème: fondamentalement la restauration de l'intelligibilité de la parole hyperbare exige une correction de la fonction de transfert du conduit vocal, le filtre vocal peut être modélisé par un filtre tout pôle, l'identification AR du filtre à partir du signal de parole doit être réaliste. Ajoutons l'observation suivante: le pitch du signal de parole hyperbare est particulièrement élevé. Alors, dans quelle mesure l'identification du filtre vocal peut-elle être réaliste? La question fait l'objet de la présente communication. La réponse est fondée sur une approche théorique procédant en trois étapes: formaliser le lien entre identification AR et pitch, établir des données tests, quantifier l'approximation en termes de fréquence et bande passante des formants. L'étude révèle une estimation formantique relativement plus précise que celle résultant de l'analyse de données "air" similaires.

ABSTRACT

Three assertions state the framework of the problem: basically the restoration of the hyperbaric speech requires a correction of the transfer function of the vocal tract, the vocal tract can be modelized by an all pole filter, and the AR identification of the filter, derived from the hyperbaric speech signal, must be realistic. Let us add the following comment: the hyperbaric speech signal pitch period is particularly low. Then, in which extent can the vocal filter AR identification be realistic? This question is subject to the present lecture. The answer is based on a theoretical approach which proceeds in three steps: formalize the relationship between identification and pitch, state checking data, and appraise the approximation in terms of formant frequencies and bandwiths. The formantic estimation computed from hyperbaric data is proved relatively more accurate than that resulting from similar "air" data.

I. INTRODUCTION

La parole hyperbare fait référence à la parole produite par les plongeurs qui opèrent en saturation, dans la plupart des cas sur des champs pétrolifères sous-marins. Sur les sites profonds (au delà de quelques dizaines de mètres), pour des raisons d'ordre physiologique, les plongeurs doivent respirer un mélange gazeux synthétique léger tel que l'héliox, l'hydrox ou l'hydréliox. De tels mélanges induisent une quasi-inintelligibilité. Cernons l'origine de ce phénomène [1].

Le signal de parole est le résultat du contrôle simultané du mécanisme d'excitation (énergie pulmonaire, vibration des cordes vocales ou bruit) et du résonateur (conduits vocal et nasal). D'un point de vue fondamental, il est naturel d'associer un modèle mécanique et aérodynamique à la glotte, par exemple le modèle à deux masses de Flanagan, et un modèle acoustique -ou son analogie électrique- aux conduits vocal et nasal (Fant, Flanagan). Ces modèles intègrent explicitement les paramètres physiques du mélange respiratoire: densité, coefficient de viscosité, chaleurs spécifiques, célérité du son. Les relations mécaniques s'avèrent prépondérantes dans l'expression du comportement théorique de la glotte, tandis que la fonction de transfert du résonateur s'avère très sensible aux paramètres physiques, en particulier à la célérité et à la densité: les formants subissent un déplacement vers les fréquences hautes dans un rapport au moins égal à celui des célérités dans le mélange respiratoire synthétique et dans l'air [2]. L'étude expérimentale confirme l'altération formantique. Par contre, s'agissant de la production des sons voisés, elle contredit l'invariance de fonctionnement de la glotte: statistiquement la fréquence fondamentale subit un accroissement de l'ordre de 20% quel que soit le locuteur. Néanmoins, la seule correction formantique réalise une restauration de l'intelligibilité qui est jugée "relativement" satisfaisante.

Cette analyse suscite un principe fondamental de restauration de l'intelligibilité: en supposant connues l'excitation et la fonction de transfert hyperbares, le signal corrigé peut être synthétisé à partir de l'excitation inchangée et de la fonction de transfert modifiée en intégrant les paramètres physiques du mélange respiratoire. *Le problème crucial tient alors en l'identification, à partir des observations du signal de parole*

hyperbare, à la fois de l'excitation et de la fonction de transfert hyperbare. Toute solution est nécessairement fondée sur un modèle paramétrique du processus de production du signal de parole. Nous considérons ici le modèle linéaire (fig. 1), dans lequel le filtre résonateur, abstraction faite du filtre nasal, est un filtre tout-pôle 1/A(z). La question abordée est alors la suivante: l'identification AR du filtre vocal peut-elle être réaliste, sachant que le signal de parole hyperbare est caractérisé par un pitch particulièrement élevé?

Le lien entre estimation AR et pitch est développé dans la section II. La simulation de l'incidence du pitch sur l'estimation formantique est nécessairement fondée sur des données tests: ces données sont respectivement présentées et traitées dans les sections III et IV.

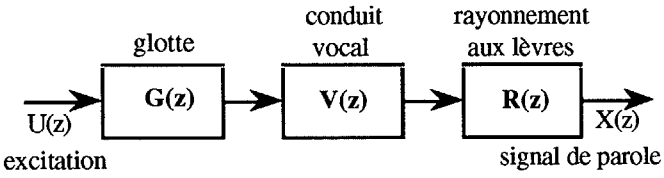

Fig. 1. Modèle linéaire du système vocal

II. RELATION ENTRE ESTIMATION AR ET PITCH

La connaissance a-priori du signal de parole hyperbare $x_{hyp}(.)$ et des filtres G(z) et R(z) donne accès à un signal y(.) assimilable à la sortie du filtre V(z) lorsque l'excitation est U(z):

$$Y(z) = X(z) / [G(z) . R(z)]$$

L'identification du filtre vocal consiste, à partir des observations de y(.), en l'estimation d'un ensemble de M coefficients $\{a_i\}$ telle que

$$V(z) = \frac{1}{A(z)} = \frac{1}{\sum_{i=1}^{M} a_i z^{-i}} \qquad (1)$$

Dans la suite, nous considérons une estimation par la méthode d'autocorrélation [3]. La donnée n'est plus le signal y(.), mais sa fonction d'autocorrélation $R_y(.)$. Explicitons l'incidence de deux excitations idéales: l'impulsion et le peigne de Dirac.

1er cas: $u(t) = \delta(t)$

La réponse impulsionnelle y(.) du filtre vocal, sa fonction d'autocorrélation R(.) et sa d.s.p. $\gamma(\nu)$, ν désignant la fréquence normalisée, vérifient:

$$y(k) = \sum_{i=1}^{M} a_i \, y(k-i) \qquad (2)$$

$$R(l) = \sum_{i=1}^{M} a_i \, R(l-i) \qquad (3)$$

$$R(l) = \int_{-\frac{1}{2}}^{\frac{1}{2}} \gamma(\nu) \; e^{j2\pi\nu l} \, d\nu \qquad (4)$$

$$\gamma(\nu) = \sum_{l=-\infty}^{+\infty} R(l) \; e^{-j2\pi\nu l} \qquad (5)$$

2ème cas: $u(t) = \delta(t - kP)$

Ce peigne de Dirac, où P représente la période de pitch, modélise l'excitation des sons voisés. Dans le domaine de Fourier U(v) est le peigne de Dirac

$$U(\nu) = \frac{1}{P} \sum_{k=-\infty}^{+\infty} \delta(\nu - \frac{k}{P}) \qquad (6)$$

La d.s.p. et la fonction d'autocorrélation du signal y(.) s'écrivent alors respectivement:

$$Y(\nu) = \frac{1}{P} \sum_{k=0}^{P-1} \gamma(\nu) \; \delta(\nu - \frac{k}{P}) \qquad (7)$$

$$R_y(l) = \frac{1}{P} \int_{-\frac{1}{2}}^{\frac{1}{2}} \sum_{k=0}^{P-1} \gamma(\nu) \; \delta(\nu - \frac{k}{P}) e^{j2\pi\nu l} \, d\nu \qquad (8)$$

Il vient:

$$R_y(l) = \frac{1}{P} \sum_{k=0}^{P-1} \gamma(\frac{k}{P}) \, e^{j2\pi \frac{k}{P} l} \qquad (9)$$

soit, en reportant (5) dans (9)

$$R_y(l) = \sum_{n=-\infty}^{\infty} R(l - nP) \qquad (10)$$

Cette relation formalise le lien , en vue de la détermination des M coefficients a_i, entre la donnée effective $R_y(.)$ et la donnée de référence R(.). Deux interprétations peuvent y être associées.

1-Par essence, y(.) et $R_y(.)$ sont respectivement les résultats des convolutions du peigne de Dirac avec la réponse impulsionnelle du filtre et sa fonction d'autocorrélation R(.). Par conséquent, si la longueur de la réponse est supérieure à 1/P, alors $R_y(.)$ est susceptible d'être une version dégradée de R(.). Il s'agit d'une approche "temporelle" qui s'avère pertinente dans [4,5].

2-Plus abstraitement, la relation (10) exprime la conséquence d'une discrétisation de l'enveloppe spectrale en relation avec le pitch (relation 7). Cette interprétation établit la cohérence avec les travaux de Makhoul et El-Jaroudi relatifs à la discrétisation fréquentielle en modélisation AR [6].

L'analyse LPC produit une estimation $\hat{A}(z)$ du filtre A(z) à partir de la donnée $R_y(.)$ calculée par (10). S'agissant du filtre vocal, il est naturel de quantifier l'approximation en termes d'écart relatif sur les fréquences et les bandes passantes des formants.

III. DONNEES TESTS

Les données tests consistent en un ensemble de filtres vocaux tout-pôle 1/A(z) paramétrés par la configuration articulatoire et l'environnement hyperbare, et par les caractéristiques physiques des mélanges respiratoires hyperbares. Les coefficients de prédiction déterminent la connaissance de la réponse impulsionnelle par (2) et de la fonction de transfert par (1) pour $z=e^{j2\pi\nu}$. L'objectif étant d'évaluer la robustesse de la méthode d'autocorrélation, la donnée test appropriée est la d.s.p. de la réponse impulsionnelle, c'est-à-dire:

$$\gamma(\nu) = \frac{1}{|A(e^{j2\pi\nu})|^2} \qquad (11)$$

En effet, elle induit la connaissance de la fonction $R_y(.)$ par (9) en évitant toute opération de fenêtrage direct du signal. En d'autres termes, elle permet de s'affranchir des erreurs systématiques engendrées par la forme et la longueur de la fenêtre ainsi que son placement par rapport au signal [7].

Les filtres tests sont obtenus par la méthode suivante. A un couple configuration articulatoire-mélange respiratoire est associé un modèle n-tubes du conduit vocal dont le fonctionnement est régi par les équations de propagation du son de Webster [8]. La fonction de transfert T(ν) du modèle, calculable pour toute fréquence ν d'un intervalle [0..Fs], est dotée d'une symétrie hermitienne garantissant une transformation de Fourier inverse réelle. La borne Fs, assimilée à une fréquence d'échantillonnage, détermine le lien avec la discrétisation du temps: elle est fixée à 32 KHz, sachant que la bande du signal de parole hyperbare est toujours d'un ordre de grandeur 3 fois supérieur à celle du signal de parole conventionnel. T(k) est obtenue par discrétisation de T(ν) en $N=2^{11}$ points sur l'intervalle de définition de ν. La fonction d'autocorrélation de la réponse impulsionnelle associée à T(k) est alors calculée par:

$$R_T(l) = \sum_{k=0}^{N-1} |T(k)|^2 e^{j2\pi\frac{k}{N}l} \qquad (12)$$

Le choix de N apparaît crucial: il induit une troncature de la fonction de référence $R_T(l)$ à la longueur N/Fs, ce qui revient implicitement à admettre que $R_T(l)$ est négligeable au delà de N/Fs. Des travaux antérieurs consacrés à l'étude de la réponse impulsionnelle du filtre vocal en environnement hyperbare ont démontré la validité des valeurs numériques attribuées à N et à Fs [9]. Le prédicteur $\{a_i\}$ associé à T(k) est calculé à partir des M+1 premiers termes de $R_T(l)$: il caractérise alors le filtre test A(z). Notons que l'égalité (3) est en conséquence complétée par

$$R(l) = R_T(l) \text{ pour } 0 \le l \le M+1$$

Pour tous les filtres tests l'ordre M est fixé à 14: le critère de choix est fondé sur l'interprétation de M en termes de mémoire de filtre [9].

IV. QUANTIFICATION DE L'INCIDENCE DU PITCH

Etant donné une configuration articulatoire et un environnement acoustique, l'incidence du pitch sur l'estimation formantique est quantifiée par les erreurs induites sur les fréquences et les bandes passantes des 3 premiers formants. Une configuration articulatoire simulée produit un graphe, un environnement simulé produit une courbe sur un graphe. Une courbe est déterminée discrètement conformément au schéma algorithmique suivant.

Description des indices de boucle P, k, l (cf. section II):

$40 \le P \le 200$, (pas de 2, soit 0.2ms) dans le cas d'un environnement "air";

$128 \le P \le 640$, (pas de 8, soit 0.25ms) dans le cas d'un environnement "hyperbare";

$0 \le k \le P-1$ (pas de 1);

$0 \le l \le M+1$ (pas de 1);

Procédure:

{ Identification d'un filtre test;
Pour l'ensemble des valeurs de P, faire:
{ Pour l'ensemble des valeurs de k: calculer $\gamma(k/P)$;
Pour l'ensemble des valeurs de l: calculer $R_y(l)$;
Identifier A(z);
Calculer les 6 erreurs relatives induites; }}

Dans cette communication les résultats sont illustrés par 5 graphes correspondant à la voyelle [a] (fig. 2). Chaque graphe permet de comparer les estimations formantiques calculées pour 3 environnements simulés: air, héliox 100m, héliox 500m. En l'occurence les valeurs absolues des erreurs relatives ont été tracées afin d'obtenir un meilleur discernement visuel.

Dans tous les cas les erreurs convergent vers 0 pour les valeurs croîssantes de P. En moyenne ces erreurs sont relativement importantes dans le cas de l'environnement "air", en particulier en ce qui concerne les bandes passantes: cette observation corrobore l'inadéquation de la prédiction linéaire au signal de parole pour certaines applications lorsque le pitch est élevé (pitch de femmes par exemple).

Par contre, lorsque l'environnement est de type "hyperbare", les erreurs sont faibles tant en ce qui concerne les fréquences que les bandes passantes, et ce d'autant plus que la profondeur de plongée simulée est grande: la convergence vers 0 est pratiquement atteinte pour un pitch de 100Hz en "héliox 500m". A titre indicatif, l'incidence d'un pitch de 150Hz est moindre en conditions hyperbares que celle d'un pitch de 75Hz en conditions naturelles.

Dans tous les cas les erreurs présentent une nature oscillatoire. Les fréquences des oscillations sont probablement à corréler avec les fréquences des formants. Pour justifier cette

hypothèse, imaginons une configuration articulatoire qui serait caractérisée par une seule fréquence de résonance f: alors R(.) serait périodique en 1/f, et l'addition de répliques de R(.) décalées de 1/f (cf. relation 10) produirait une fonction $R_y(.)$ peu dégradée par rapport a R(.), et par conséquent une bonne estimation $\hat{A}(z)$ de A(z).

V. CONCLUSION

En parole conventionnelle, un pitch élevé est reconnu préjudiciable à l'application de la modélisation AR. S'agissant de parole hyperbare, précisément caractérisée par un pitch a-priori défavorable, la mesure de la qualité de l'estimation formantique est une information majeure pour la recherche d'algorithmes de restauration de l'intelligibilité. L'approche adoptée exploite des filtres vocaux tests paramétrés par la configuration articulatoire et le mélange respiratoire, et évalue l'identification AR (autocorrélation) par une méthode originale. Ainsi, en parole hyperbare, la dégradation de l'estimation formantique induite par le pitch s'avère particulièrement faible. Ce résultat est confirmé par l'étude d'un ensemble de graphes, dont seulement 5 exemplaires sont présentés dans cette communication. Une analyse plus fine montre que la célérité du son dans le mélange respiratoire est le facteur responsable prépondérant.

BIBLIOGRAPHIE

[1] J. Crestel, Contribution à l'amélioration de l'intelligibilité de la communication orale en plongée hyperbare, Thèse de l'Université de Rennes I, mai 1987.

[2] G. Fant, J. Lindquist, Pressure and gas mixture effects on diver's speech, Royal Inst. of Tech. Stockholm, STL-QPSR 1/68, 1968.

[3] J.D. Markel, A.H. Gray, Linear prediction of speech, Springer-Verlag, New-York 1976.

[4] J. Crestel, M. Guitton, V. Le Calvé, M. Corazza, Sur la quasi-stationnarité du filtre vocal en conditions hyperbares, Colloque GRETSI, Juan-Les-Pins, septembre 1991.

[5] J. Crestel, M. Guitton, On robust estimation of LP coefficients in high pitch speech analysis, Sixth European Signal Processing Conference, Brussels, August 1992.

[6] A. El-Jaroudi, J. Makhoul, Discrete All-Pole modelling, IEEE Trans. On Signal Processing, Vol. 39, n°2, February 1991.

[7] C.H. Lee, On robust linear prediction of speech, IEEE Trans. On ACSSP, 1988.

[8] M. Mrayati, Contribution aux études sur la production de la parole, Thèse de docteur d'Etat, Grenoble, 1976.

[9] M. Guitton, J. Crestel, Characterization of vocal impulse responses for articulatory configurations of vowels in hyperbaric conditions, Proc. ESCA "Speech processing in adverse conditions", Cannes, novembre 1992.

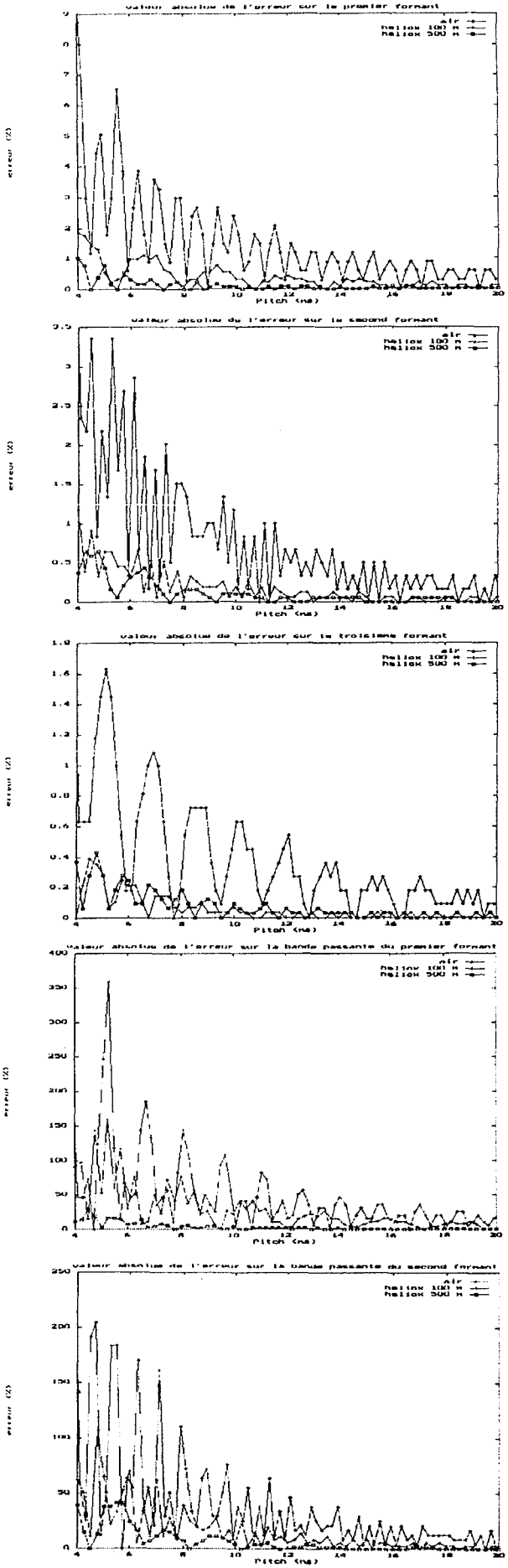

Fig. 2. Incidence du pitch sur l'estimation formantique (voyelle [a], 3 environnements simulés)