DIXIEME COLLOQUE SUR LE TRAITEMENT DU
SIGNAL ET SES APPLICATIONS

NICE du 20 au 24 MAI 1985

DETERMINATION DE LA FREQUENCE FONDAMENTALE A PARTIR DU VOCODEUR DE PHASE

Francis CHARPENTIER Michel STELLA

Centre National d'Etudes des Télécommunications 22300 LANNION FRANCE

RESUME

On présente dans cet article une nouvelle méthode fréquentielle de détermination de la fréquence fondamentale de la parole. Le principe de la méthode est lié au principe des modifications prosodiques dans le vocodeur de phase. Son originalité consiste à traiter l'information contenue dans le spectre de phase, alors que les autres méthodes fréquentielles se limitent à l'analyse du spectre d'amplitude. Le spectre à court terme est calculé par DFT et chaque coefficient DFT est interprété comme la sortie d'un filtre passe-bande présentant un certain recouvrement avec ses voisins. La détection des harmoniques se fait par recherche des groupes de trois filtres adjacents présentant une certaine cohérence de phase. La fréquence de chaque harmonique est alors donnée par la fréquence instantanée de ces filtres. On peut ainsi construire un "harmonogramme" plus facile à interpréter que des représentations analogues à partir de l'amplitude seule. Un algorithme classique de numérotation des harmoniques permet de passer de l'harmonogramme à la fréquence fondamentale proprement dite. Des résultats préliminaires montrent la validité de la méthode.

SUMMARY

A new frequency domain method for determining the fundamental frequency of speech is presented in this paper. The principle of the method is related to the principle of prosodic modifications by use of the phase vocoder. The novelty of the method consists of using the information contained in the phase spectrum whereas the other frequency domain methods are limited to the analysis of the amplitude spectrum. The short-term spectrum is computed by the DFT and it is interpreted as the output of a bank of band-pass overlapping filters. The harmonic components are detected by searching for the sets of three neighbouring filters involving some phase coherence. The frequency of each harmonic is given by the instantaneous frequency of these coherent filters. A "harmonogram" can thus be displayed, which is easier to interpret than similar representations obtained from the amplitude alone. A conventional harmonic numbering algorithm is then used to convert the harmonogram representation to the fundamental frequency itself. Preliminary results show the validity of the method.

DETERMINATION DE LA FREQUENCE FONDAMENTALE A PARTIR DU VOCODEUR DE PHASE

INTRODUCTION

Le vocodeur de phase est un système d'analyse-synthèse qui permet de modifier le déroulement temporel du signal de parole tout en conservant son timbre naturel /1/. Récemment, Seneff a montré qu'un tel système est aussi capable de modifier la fréquence fondamentale F0 sans en déterminer explicitement la valeur /2/. Mais il ne s'agit alors que de modifications relatives, comme la transposition d'une octave du contour mélodique d'une phrase. Pour certaines applications, il est souhaitable d'obtenir des modifications absolues, c'est-à-dire de réaliser exactement un schéma mélodique spécifié, ce qui nécessite alors une estimation absolue de F0. Ceci peut être obtenu par une méthode différente, par exemple par la méthode du filtre en peigne /3/, voire même par une détermination manuelle du pitch pour certaines applications comme la synthèse par diphones /4/. Cependant il existe aussi une méthode naturellement adaptée au vocodeur de phase parce que correspondant au même type d'analyse. On présente ici cette méthode, dont l'originalité par rapport aux autres méthodes fréquentielles est de traiter le spectre de phase à court terme.

La validité d'une telle approche est d'ailleurs suggérée par des expériences de distorsion du signal par manipulation de la phase ou de l'amplitude /5/. Ainsi, l'importance de la phase est bien connue en traitement d'images. La netteté d'images reconstruites à partir de la phase seule est bien plus grande que pour celles reconstruites à partir de l'amplitude, parce que la phase contient une information plus pertinente sur les transitions brusques (contours et discontinuités). Pour les signaux de parole, l'amplitude spectrale à court terme contient une grande partie de l'information phonétique. Cependant, alors que l'oreille est insensible aux relations de phase dans le domaine fréquentiel, le déroulement temporel de la phase à court terme véhicule une grande partie de l'information prosodique. En effet, la synthèse d'un signal de parole après annulation de la phase produit un son avec un pitch artificiel déterminé par la longueur et le déplacement de la fenêtre d'analyse. Inversement, un aplatissement du spectre d'amplitude est perçu comme un signal bruité dans lequel on discerne bien le contour mélodique de la phrase.

1. VOCODEUR DE PHASE ET MODIFICATIONS RELATIVES DE F0

Le vocodeur de phase représente le signal par son spectre à court terme sur une fenêtre glissante. Lorsque le spectre est calculé par DFT, il peut être interprété comme la sortie d'un banc de filtres uniforme /6/. En effet, le coefficient Xk d'une DFT de N points est donné par la formule:

$$X_k = \sum_{l=0}^{N-1} w^{kl}\, x(n-l)\, h(l)$$

où w désigne la racine complexe Nieme de l'unité:

$$w = \exp\left(j \frac{2\pi}{N} \right)$$

x(n) désigne le signal temporel et h(n) désigne la fenêtre d'analyse. Ce coefficient peut être interprété comme le résultat d'un filtrage de réponse impulsionnelle:

$$h_k(l) = w^{kl}\, h(l)$$

Il s'agit d'un filtre passe-bande autour de la fréquence centrale:

$$f_k = k \frac{F_s}{N}$$

où Fs est la fréquence d'échantillonnage du signal. La largeur de bande de ce filtre dépend de la fenêtre d'analyse h. Dans le cas d'une fenêtre de Hanning, cette largeur de bande couvre trois échantillons fréquentiels adjacents. En d'autre termes, la réponse d'un coefficient/filtre Xk recouvre celles des deux coefficients/filtres voisins.

Chaque coefficient complexe Xk est ensuite représenté par son amplitude et par sa fréquence instantanée f'k, qui est définie comme la dérivée temporelle de sa phase. En pratique, la détermination de f'k nécessite de calculer le spectre à court terme sur deux fenêtres successives, décalées d'un échantillon temporel. Ce décalage de la fenêtre d'analyse entraîne pour chaque Xk une certaine variation de phase et on définit la fréquence instantanée f'k comme la pente de cette variation:

$$f'_k = \frac{\Delta\varphi_k}{2\pi} F_s$$

Lorsqu'une forte composante harmonique passe dans le filtre Xk, la fréquence instantanée f'k est en fait égale à la fréquence de l'harmonique elle-même. Ainsi peut-on modifier la fréquence de l'harmonique en changeant la valeur de f'k avant la synthèse. Cette opération peut être faite indépendamment de l'enveloppe spectrale, ce qui fournit une méthode pour modifier le pitch sans en extraire la valeur /2/.

2. DETECTION DES HARMONIQUES

Le principe de la détection des harmoniques repose sur l'analyse des fréquences instantanées dans le banc des filtres Xk. La figure 1 illustre cette analyse sur une portion de signal de parole. La partie droite de la figure représente la fréquence instantanée f'k de chaque filtre Xk en fonction de sa fréquence centrale fk. On obtient une distribution de points qui s'agglommère autour de la droite f'k=fk. Le gabarit en tireté indique un écart déjà important entre les fréquences centrales et les fréquences instantanées:

$$|f'k - fk| = 4 \frac{F_s}{N}$$

La plupart des points se situent à l'intérieur de ce gabarit. La présence d'harmoniques est indiquée par les flèches et se manifeste par une accumulation des fréquences instantanées autour des fréquences des harmoniques. Ceci s'explique par le recouvrement partiel de filtres adjacents.

Sur la partie gauche de la Fig.1, on a donné en mode "bâtonnets" diverses représentations du spectre d'amplitude pour la même portion de signal. La représentation (a) du haut correspond à l'échelle fréquentielle habituelle. La représentation (b) du milieu est obtenue à partir de (a) en déformant l'axe fréquentiel par la fonction f'k = f'(fk). Là aussi, on observe un resserrement des composantes fréquentielles autour des harmoniques du spectre.

Ce type d'analyse constitue la base de notre méthode de détection des harmoniques. L'algorithme comprend deux étapes étroitement liées:
(1) la sélection des harmoniques en utilisant un critere d'harmonicité;
(2) le calcul des fréquences correspondantes.

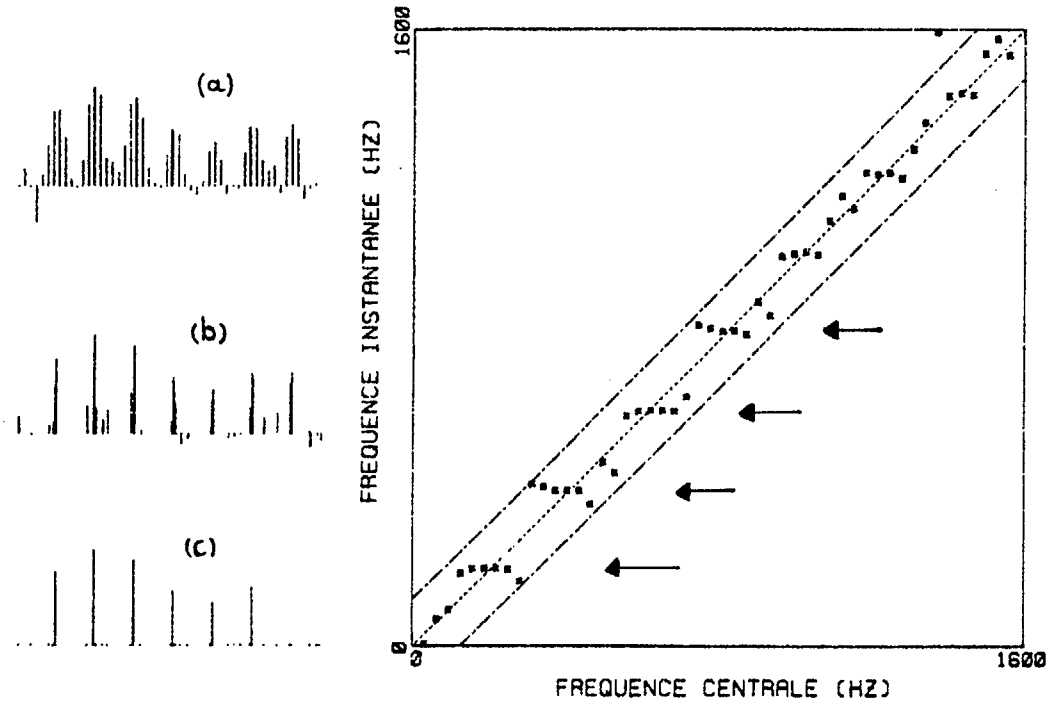

Fig.1 Détection des harmoniques par utilisation des fréquences instantanées des coefficients DFT.
à droite: fonction de déformation f'k = f'(fk)
à gauche: spectre d'amplitude du signal selon une échelle des fréquences:
(a) normale; (b) déformée par f';
(c) déformée par f' et après élimination des composantes non harmoniques.

Le critère de sélection est le suivant: un filtre Xk est susceptible de contenir une harmonique si la fréquence instantanée f'k est égale aux fréquences f'k-1 et f'k+1 des deux filtres voisins à un seuil de détection près. Une valeur typique de ce seuil est la moitié de la résolution fréquentielle Fs/N. Pour obtenir la représentation spectrale (c) du bas de la Fig.1, on a appliqué ce critère de façon à éliminer toutes les composantes fréquentielles non harmoniques. La fréquence de chaque harmonique est alors donnée par la fréquence instantanée f'k du filtre Xk. On obtient ainsi pour chaque fenêtre d'analyse une distribution d'harmoniques non numérotées, ce qui engendre un "spectrogramme d'harmoniques" ou "harmonogramme" lorsque l'analyse est répétée avec un certain pas de progression temporel.

3. ESTIMATION DE LA FREQUENCE FONDAMENTALE

L'estimation de F0 proprement dite se fait par analyse de la distribution des harmoniques détectées à l'étape précédente. De nombreux algorithmes ont été proposés pour réaliser ce passage de l'harmonogramme au contour mélodique /7/. Nous avons utilisé un algorithme simple, analogue aux algorithmes de "tamis" /8/ ou de "peigne" /3/. Une famille de valeurs plausibles pour l'estimation de F0 est constituée à partir des écarts entre harmoniques successives. Pour chacune de ces valeurs, on définit un score par le nombre d'harmoniques qu'elle permet d'expliquer, moyennant une certaine plage d'erreur. La valeur retenue pour F0 est celle qui obtient un score maximal. A l'issue de cet algorithme, les harmoniques reçoivent une numérotation appropriée et celles qui n'ont pas pu être numérotées sont éliminées. L'estimation finale de F0 est calculée par moyenne pondérée des harmoniques significatives.

Les périodes de silence sont détectées au moyen d'un seuil sur l'énergie du signal. Le critère de voisement est défini à partir de l'harmonogramme comme la somme de l'énergie des harmoniques significatives. Des erreurs grossières peuvent se produire à cause de la raréfaction des harmoniques significatives. Elles sont corrigées par un algorithme de lissage temporel, ce qui peut s'interpréter comme un traitement d'image appliqué à l'harmonogramme. Pour effectuer ce lissage, on définit une "fenêtre de voisement" qui correspond à une durée minimale de voisement. Sur cette fenêtre on évalue la valeur la plus probable de F0 à partir de ses estimations brutes et l'on corrige la valeur de l'échantillon central si l'on observe un décalage trop grand.

4. REDUCTION DE LA CHARGE DE CALCUL

Le calcul des fréquences instantanées f'k nécessite celui des incréments de phase $\Delta\varphi_k$ correspondant au décalage de la fenêtre d'un échantillon. En absence de toute optimisation de l'algorithme, il faut calculer la phase de deux spectres successifs, c'est à dire réaliser deux fois la séquence de calcul suivante: une multiplication par une fenêtre de Hanning, une FFT, et un passage en coordonnées polaires.

Heureusement, la charge de calcul peut être réduite si l'on utilise les relations de dépendance entre les deux spectres successifs. On calcule d'abord la DFT sur une fenêtre rectangulaire:

$$Y_k = \sum_{l=0}^{N-1} w^{kl} x(n-l)$$

La DFT pour la fenêtre décalée s'en déduit simplement par la formule:

$$\tilde{Y}_k = w^k \, (Y_k + \delta)$$

où δ représente la différence entre les échantillons entrant et sortant de la fenêtre initiale. D'autre part le coefficient Xk peut être obtenu simplement à partir des Yk si on réalise la fenêtre de Hanning dans le domaine fréquentiel:

$$X_k = Y_k - \frac{1}{2} (Y_{k-1} + Y_{k+1})$$

La variation de phase pour le filtre Xk est alors donnée par la formule:

$$\Delta\varphi_k = 2\pi \frac{k}{N} + \mathrm{Arg}\left[\frac{Y_k - \frac{1}{2}(wY_{k+1} + w^{-1}Y_{k-1}) + \delta(1-\cos\frac{2\pi}{N})}{Y_k - \frac{1}{2}(Y_{k+1}+Y_{k-1})} \right]$$

La charge de calcul pour l'analyse fréquentielle se répartit ainsi en trois étapes:
(1) une DFT;
(2) un ensemble d'opérations sur les nombres complexes: 6N additions, 2N multiplications, N divisions;
(3) N extractions d'arctangentes.
La charge due aux étapes (2) et (3) peut encore être réduite si on limite l'analyse aux basses fréquences (typiquement à la bande 0-1.6Khz) ou bien si l'information d'amplitude est utilisée en parallèle de façon à n'explorer que les zones autour des maxima.

5. RESULTATS ET DISCUSSION

L'algorithme a été testé sur de la parole échantillonnée à 8KHz. La taille de la FFT adaptée à cette fréquence d'échantillonnage est N=256, ce qui correspond à une fenêtre d'analyse de 32 ms. La résolution fréquentielle d'environ 31Hz permet alors de bien séparer deux harmoniques de pitch.

La figure 2 compare un harmonogramme obtenu par notre méthode (Fig.2b) à une représentation analogue obtenue à partir du spectre d'amplitude (Fig.2a). L'harmonogramme de la Fig.2a représente les maxima du spectre d'amplitude en estimant la fréquence des harmoniques par interpolation parabolique sur trois points. L'évolution temporelle des harmoniques subit ainsi un certain lissage. Ce type d'interpolation est souvent utilisé par les méthodes fréquentielles de détermination de F0. Les périodes voisées se distinguent par une plus grande régularité des harmoniques, mais les périodes non voisées contiennent une densité de pseudo-harmoniques aussi importante. Sur l'harmonogramme obtenu par notre méthode (Fig.2b), le suivi temporel des harmoniques est comparable à celui de la Fig.2a, mais le critère d'harmonicité est bien plus sélectif, puisqu'il élimine une grande partie des harmoniques dans les hautes fréquences ou pour les périodes non voisées. Cette sélectivité accrue permet une identification plus facile des périodes voisées, et une meilleure interprétation des harmoniques.

Du fait de sa plus grande sélectivité, le critère de phase peut être combiné au critère de maximum d'amplitude sans entraîner une grosse perte d'information. Comme il est mentionné au paragraphe précédent, cela permet de réduire d'un facteur important le volume des calculs trigonométriques. Le critère de phase peut alors être interprété comme un critère de qualité sur les maxima d'amplitudes.

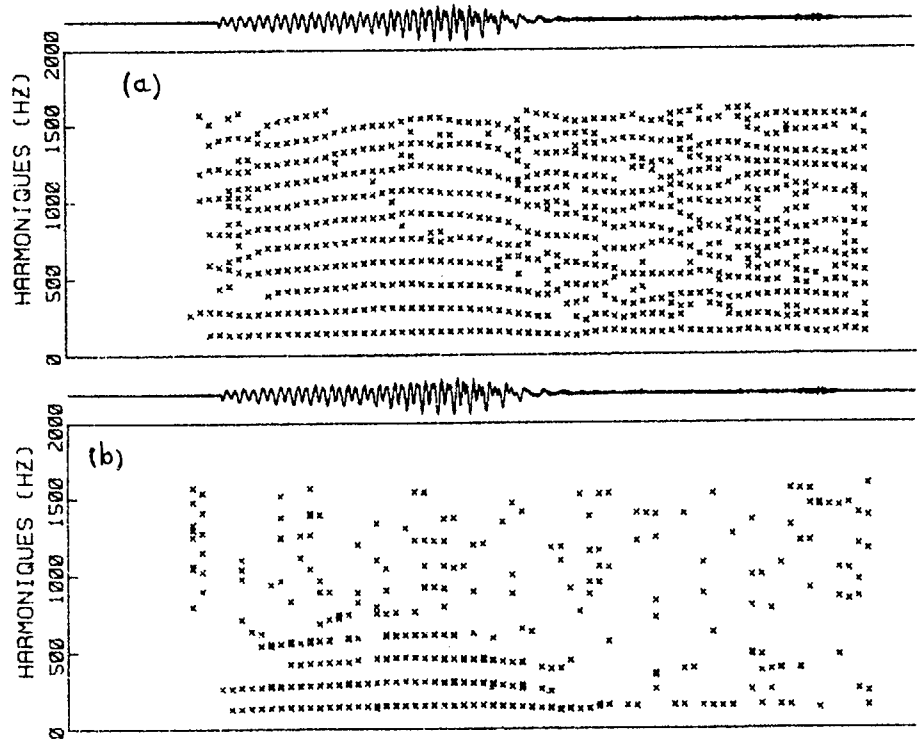

Fig.2 Deux types d'harmonogrammes du signal de parole.
Les critères d'harmonicité sont respectivement:
(a) les maxima du spectre d'amplitude; (b) la cohérence de phase des coefficients DFT
Les fréquences des harmoniques sont estimées
(a) par interpolation parabolique; (b) à partir de la fréquence instantanée

De tels critères ont déjà été proposés à partir de l'information d'amplitude /8-10/. Ces critères ont pour buts d'éliminer certains pics d'amplitude non significatifs. Terhardt a proposé un critère psycho-acoustique pour tenir compte des phénomènes de masquage entre harmoniques voisines /9/. De même, Duifhuis et Willems ont proposé un critère consistant à vérifier que les pics présentent une forme correcte suivant un modèle parabolique /8/. Dans notre cas, le critère consiste à utiliser la variation temporelle de la phase pour vérifier que le pic contient une véritable harmonique.

La figure 3 présente le résultat de notre algorithme sur la phrase: "Le sculpteur taille la pierre". L'harmonogramme est représenté entre 0 et 1.6 KHz. Les harmoniques significatives sont indiquées par des croix, et les points représentent les harmoniques éliminées au cours de la phase de numérotation. La détection de pitch brute est donnée au dessous de l'harmonogramme, et l'on peut distinguer encore quelques erreurs grossières, dont quelques erreurs doubles sur la fin de la phrase. Le critère de voisement est représenté en-dessous sur une échelle relative. Le seuil de voisement, fixé à 0 Db, est tracé en pointillés. Il suit à peu près une distribution bimodale, ce qui permet de distinguer facilement les parties voisées et non voisées entre elles. Enfin la figure présente en comparaison l'estimation de pitch après lissage par notre algorithme et une estimation obtenue par l'algorithme du filtre en peigne /3/. La correspondance des deux méthodes est satisfaisante.

CONCLUSION

La méthode de détection de pitch présentée dans cet article se rattache à la famille des méthodes fréquentielles /7/. Les méthodes fréquentielles se limitent en général à l'estimation des maxima d'amplitude. Elles sont alors contraintes de travailler aux fréquences fixes de la DFT, et elles utilisent souvent des schémas d'interpolation parabolique. L'avantage de notre méthode vient de son utilisation du spectre de phase. Cela permet d'une part une détection plus sélective des harmoniques et d'autre part une estimation directe de leurs fréquences. La charge de calcul comprend principalement une FFT et un calcul de la phase par fenêtre d'analyse. Une combinaison des critères de phase et d'amplitude permet une économie de calcul supplémentaire. Des résultats préliminaires indiquent la validité de la méthode pour des signaux de parole non dégradés. Il serait nécessaire d'effectuer des tests de comparaison afin d'évaluer la méthode finement par rapport aux autres méthodes. Par ailleurs, cette méthode correspond au même type d'analyse que celui effectué par le vocodeur de phase. Il semble donc possible de l'intégrer dans un système d'analyse-synthèse autonome dérivé du vocodeur de phase. Ce système permettrait de manipuler le pitch de façon aussi souple qu'avec le vocodeur à prédiction linéaire, mais avec une meilleure qualité de parole.

REFERENCES

/1/ J.L. Flanagan, R. Golden, "Phase vocoder", Bell Syst. Tech. J., 45, 1494-1509, 1966

/2/ S.S. Seneff, "System to independently modify excitation and/or spectrum of speech waveform without explicit pitch extraction", IEEE Trans. ASSP, 30(4), 566-578, 1982

/3/ P. Martin, "Comparison of pitch detection by cepstrum and spectral comb analysis", Int. Conf. ASSP, Paris, 1982

/4/ M.G. Stella, F.J. Charpentier, "Diphone synthesis using multipulse linear predictive coding and a phase vocoder", Int. Conf. ASSP, Tampa, 1985

/5/ A.V. Oppenheim, J.S. Lim, "The importance of phase in signals", Proc. IEEE, 69(5), 529-541, 1981

/6/ M.R. Portnoff, "Implementation of the digital phase vocoder using the fast Fourier transform", IEEE Trans. ASSP, 24(3), 243-248, 1976

/7/ W. Hesse, "Pitch determination of speech signals - Algorithms and devices", Springer Verlag, 1983

/8/ H. Duifhuis, L.F. Willems, "Measurement of pitch in speech: An implementation of Goldstein theory of pitch perception", J. Acoust. Soc. Am., 71(6), 1568-1580, 1982

/9/ E. Terhardt, "Calculating virtual pitch", Hearing Research, 1, 155-182, 1979

/10/ S. Seneff, "Real-time harmonic pitch detector", IEEE Trans. ASSP, 26(4), 358-365, 1978

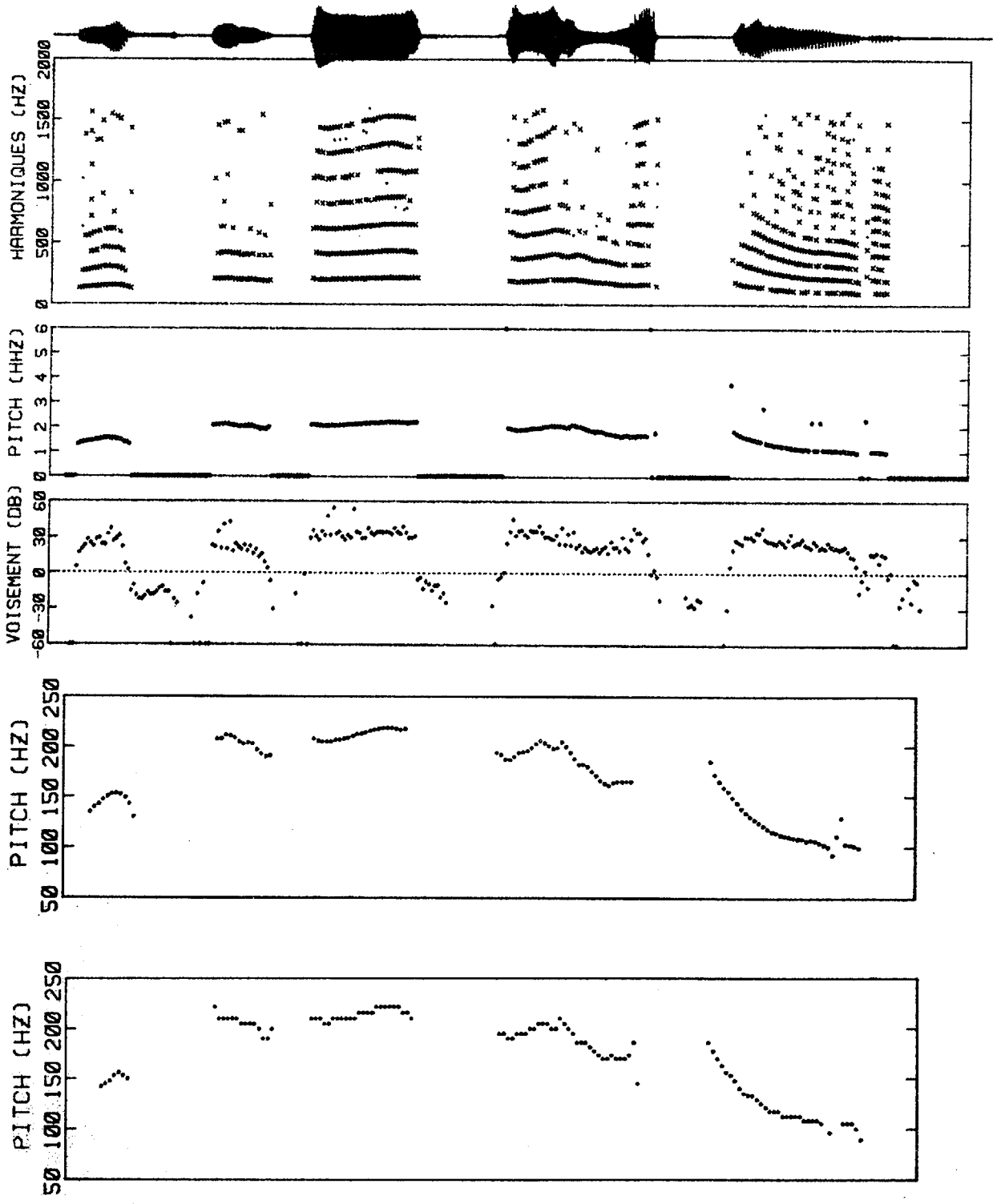

Fig.3 Détection de pitch pour la phrase: "Le sculpteur taille la pierre".
On a représenté respectivement de haut en bas:
-le signal temporel
-l'harmonogramme
-la détection de pitch brute (sans lissage)
-le critère de voisement
-la détection de pitch après lissage
-à titre de comparaison, une estimation du pitch par l'algorithme du peigne (sans interpolation parabolique)