ANALYSE-SYNTHESE DE LA BANDE DE BASE PAR FORMES D'ONDES ELEMENTAIRES

C.d'Alessandro

LIMSI-CNRS: BP 30 F-91406 ORSAY Cedex

ABSTRACT

This paper is a continuation of our work on the representation of speech signal by a set of well-localized time-frequency energy concentrations (elementary waveforms). We present here a method and a system for analysis-synthesis of the speech "baseband" (which will be defined below). The quality problems which were encountered in the lower part of the spectrum during modelling and synthesis are thus circumvented. We use the same kind of method for the processing of both formantic areas and "baseband". After an introduction in section 1, we describe the synthesis formulae in section 2 and the system developed according to these principles in section 3. Some conclusions are presented in section 4.

1 INTRODUCTION

Le traitement automatique de la parole reste tributaire de la représentation préalable du signal qui en est le support acoustique. Parmi les nombreuses méthodes disponibles, l'analyse en formes d'ondes élémentaires se présente comme un moyen neuf, et prometteur dans la mesure où il vise à une représentation permettant une reconstruction parfaite du signal et manipulant des objets pertinents tant du point de vue de la perception que de celui de la production [Liénard 87].

Dans un papier précédent, une méthode de représentation du signal de parole en fonctions d'ondes élémentaires a été développée, en se basant sur une décomposition en parallèle de la fonction de transfert du conduit vocal [d'Alessandro 87]. Les fonctions d'ondes élémentaires apparaissent comme des contributions bien localisées dans le plan spectro-temporel, et permettent ainsi de rendre compte des phénomènes de production (formants, excitations du conduit vocal, explosions ...) de façon explicite, par un ensemble discret d'éléments. L'exploitation de la structure particulière du signal de parole guide la recherche des formes d'ondes élémentaires dans les régions de maximum d'énergie spectrale ("formants", au sens large) et temporelle ("impulsions", au sens large) et permet l'obtention de paramètres perceptivement pertinents [d'Alessandro 88].

Notre système d'analyse-synthèse en fonctions d'ondes élémentaires (s'appuyant sur les fonctions d'ondes formantiques) permettait une bonne représentation du signal de parole, sans perte de qualité du point de vue perceptif, sauf dans la "bande de base" (région spectrale jusqu'au premier formant inclus) où des problèmes de modélisation apparaissaient. Par une démarche semblable à celle adoptée dans les vocodeurs à bande de base, nous présentons ici une nouvelle méthode pour décomposer la bande de base du signal en formes d'ondes élémentaires utilisant un processus d'analyse-synthèse analogue à celui employé précédement et s'appuyant sur une représentation sinusoïdale.

2 REPRESENTATIONS

2.1 représentation formantique

Le modèle linéaire classique de production du signal vocal suppose le filtrage d'une certaine fonction d'excitation par un filtre linéaire évoluant dans le temps.

$$s(t) = e(t) * R(t)$$

- e(t): signal d'excitation.
- R(t): réponse impulsionnelle du filtre.
- s(t): signal résultant.

Dans ce qui suit on suppose le signal stationnaire (sur une tranche de temps assez courte) et donc le filtre de réponse impulsionnelle R invariant. Ce filtre, associé au conduit vocal, peut être décomposé en n sections parallèles, chacune d'elle représentant une résonance (ou formant). Dans le domaine temporel il est ainsi possible d'identifier le signal de parole avec la somme des réponses de chaque section au signal d'excitation. En première aproximation, si celui-ci est constitué d'une série d'impulsions idéales, on peut écrire:

$$s(t) = \sum_{j=1}^{m} \sum_{i=1}^{n} \delta_0(t - t_j) * R_i(t)$$

où R_i représente la réponse impulsionnelle de la $i^{ième}$ section, et $\delta_0(t - t_j)$ une impulsion d'excitation à l'instant t_j.

Si l'on assimile de plus les sections parallèles à des résonateurs du second ordre [Klatt 80], alors:

$$R_i(t) = G_i e^{-\alpha_i t} sin(\omega_i t + \phi_i)$$

où α_i règle la largeur de bande (à -6 dB du sommet), G_i le gain à la résonance, ω_i la fréquence centrale du $i^{ième}$ résonateur et ϕ_i sa phase.

soit:

$$s(t) = \sum_{j=1}^{m} \sum_{i=1}^{n} \delta_0(t - t_j) * (G_i e^{-\alpha_i t} sin(\omega_i t + \phi_i))$$

Pour une représentation en formes d'onde, on peut de plus rendre indépendantes les excitations des différentes sections, ce qui affine le compromis entre précision fréquencielle et précision temporelle en le localisant, et estimer les paramètres pour chaque réponse impulsionnelle; le pavé spectro-temporel où l'on suppose le signal stationnaire est ainsi délimité par la forme d'onde:

$$s(t) = \sum_{i=1}^{n} \sum_{j_i=1}^{m_i} \delta_0(t - t_{j_i}) * (G_{j_i} e^{-\alpha_{j_i} t} sin(\omega_{j_i} t + \phi_{j_i}))$$

Les fonctions d'ondes élémentaires choisies sont donc ici identiques aux fonctions d'ondes formantiques [Rodet 80].

2.2 représentation sinusoïdale.

Pour la partie grave du spectre (en deçà du premier formant), l'utilisation d'un signal d'excitation trop simple pose de sérieux problèmes de qualité. Pour pallier à ce défaut de nombreux modèles d'excitation ont été proposés, en particulier la représentation sinusoïdale [McAulay 86]:

$$s(t) = \sum_{i=1}^{k} A_i sin(\omega_i t + \phi_i)$$

Le nombre k de sinusoïdes ainsi que l'amplitude A_i, la fréquence ω_i et la phase ϕ_i évoluent dans le temps et doivent donc être estimés sur une tranche de temps pendant laquelle le signal est quasi-stationnaire. Une alternative aux différentes méthodes proposées pour cette estimation est l'utilisation de formes d'ondes élémentaires pour représenter chaque segment de sinusoïde, pendant la durée desquelles on suppose le signal stationnaire:

$$s(t) = \sum_{i=1}^{k} \sum_{j_i=1}^{l_i} \delta_0(t - t_{j_i}) * (A_{j_i} env_{j_i}(t) sin(\omega_{j_i} t + \phi_{j_i}))$$

l'enveloppe temporelle choisie $env_{j_i}(t)$ doit permettre la reconstitution de la sinusoïde initiale; il s'agira par exemple de segments sinusoïdaux.

$$env_{j_i}(t) = 1/2(1 + cos(\beta_{j_{i_1}} t))$$

pour $0 \leq t < \pi/2\beta_{j_{i_1}}$

$$env_{j_i}(t) = 1/2(1 + cos(\beta_{j_{i_2}}(t - \pi/2\beta_{j_{i_1}}) + \pi/2))$$

pour $\pi/2\beta_{j_{i_1}} \leq t < \pi/2\beta_{j_{i_2}} + \pi/2\beta_{j_{i_1}}$

Les β_i sont calculés de façon à conserver un nombre de cycles constant dans chaque forme d'onde (qui est alors une ondelette au sens de [Goupillaud 85]).

2.3 représentation par formes d'ondes

La représentation par forme d'onde complète s'appuie sur une segmentation spectrale préalable, qui permet de dégager les régions de maximum d'énergie, auxquelles est appliquée une représentation en formes d'ondes élémentaires de type "formantiques" (au delà du premier maximum) ou "sinusoïdales" (en deçà du premier maximum) (fig. 1).

$$s(t) = (\sum_{i=1}^{n} \sum_{j_i=1}^{m_i} \delta_0(t - t_{j_i}) * (A_{j_i} env_{j_i}(t) sin(\omega_{j_i} t + \phi_{j_i}))) +$$

$$(\sum_{a=1}^{k} \sum_{b_a=1}^{l_a} \delta_0(t - t_{b_a}) * (G_{b_a} e^{-\alpha_{b_a} t} sin(\omega_{b_a} t + \phi_{b_a})))$$

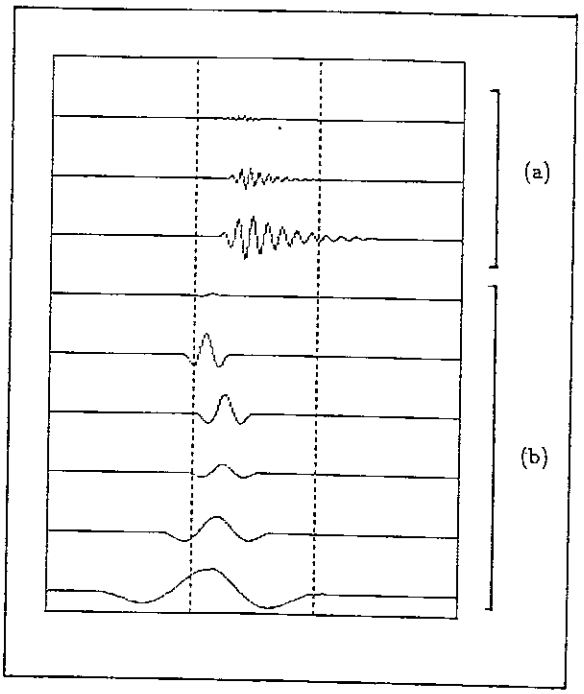

Figure 1: Modèles (a) "formantiques", (b) "sinusoïdaux" de formes d'ondes élémentaires

3 PROCESSUS D'ANALYSE-SYNTHESE

Le processus d'analyse-synthèse en fonctions d'ondes formantiques a déjà été décrit. Rappelons qu'à la suite d'une modélisation de l'enveloppe spectrale, un filtrage à phase nulle permettait d'obtenir des signaux à bande large centrés sur les maxima spectraux. Les formes d'ondes élémentaires etaient ensuite détectées grâce à l'enveloppe temporelle de ces signaux, puis modélisées comme précédemment pour la synthèse.

Pour le traitement de la bande de base, un procédé analogue est mis en œuvre, trame par trame (les trames sont de durée assez courte, 6 ms, pour que l'on puisse supposer le signal quasi-stationnaire):

- Modélisation de l'enveloppe spectrale par prédiction linéaire.
- Détection des maxima spectraux, associés aux formants, et définition de la "bande de base", comme la région spectrale jusqu'au premier maximum inclus.
- Calcul du module de la transformée de Fourier d'une tranche de signal centrée sur la trame.

- Recherche des maxima spectraux, associés aux harmoniques pour de la parole voisée, et segmentation spectrale autour de ces maxima. Tout ce qui suit ne concerne évidemment plus que la bande de base (fig. 2).
- Filtrage à phase nulle dans chacune des bandes ainsi définies, pour obtenir des signaux à bande étroite (fig. 3).
- Dans chaque bande, détection des formes d'onde (qui appartiennent à la trame si leur maximum y appartient) par recherche des maxima du signal.
- Synthèse, par les formules vues précédemment, après estimation des paramètres d'amplitude, de fréquence, de phase et d'enveloppe de chaque forme d'onde.

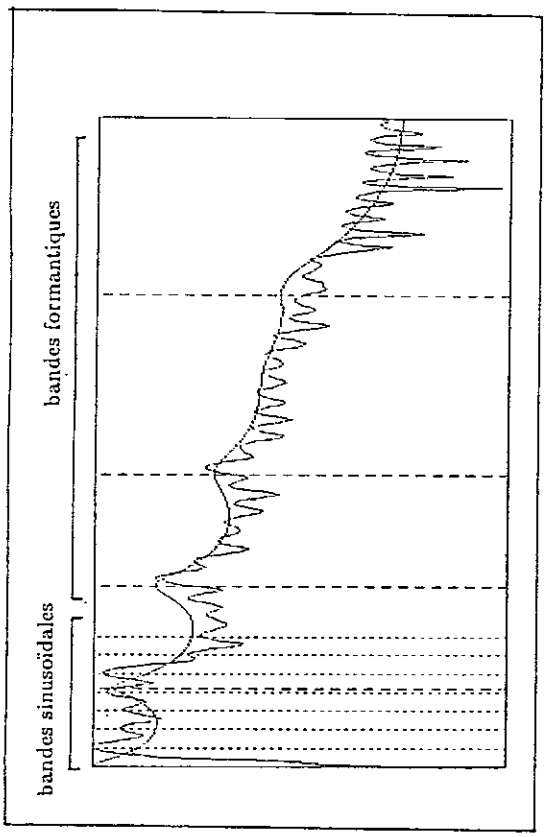

Figure 2: Modélisation et segmentation spectrale.

De par la largeur des bandes spectrales, les signaux obtenus dans chaque bande d'analyse sont des segments de sinusoïdes. Pour la parole voisée, il s'agit bien sur des premiers harmoniques, et l'enveloppe de ces signaux évolue beaucoup plus lentement que celle des signaux issus des bandes formantiques. Par contre, il est de toute première importance d'estimer précisément leur phase et leur fréquence (la vitesse d'évolution de la phase dépend évidemment de la fréquence).

Le choix du modèle de forme d'onde élémentaire adopté (nombre de cycles de la sinusoïde constant quelle que soit la fréquence) est motivé par ce souci de résolution spectro-temporelle, et non par un examen de l'enveloppe temporelle qui perd ici de son importance.

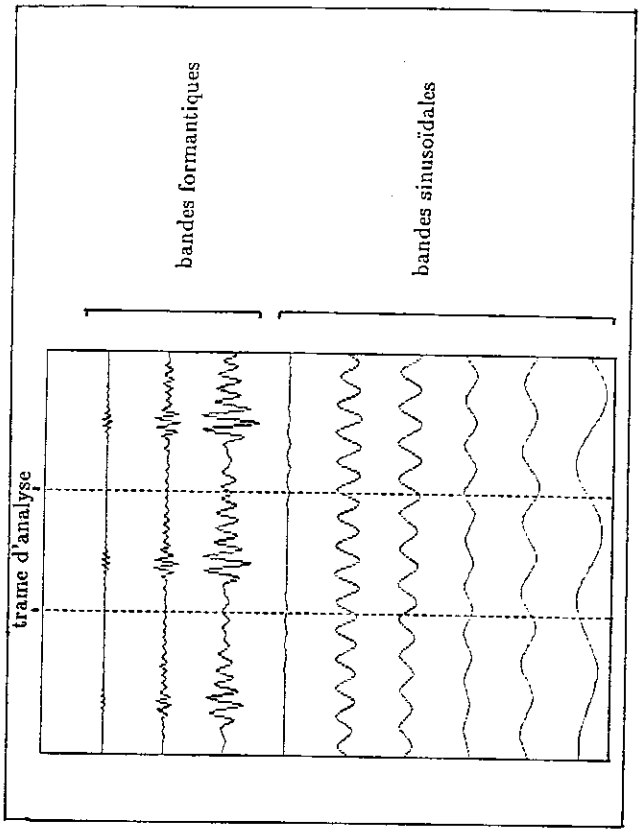

Figure 3: Filtrage dans les bandes précédemment définies.

Ainsi, le critère de segmentation d'une forme d'onde n'est plus basé sur une considération d'amplitude, comme c'est le cas pour les signaux formantiques en bande large, mais plutôt sur une considération de périodicité locale.

De même que dans les bandes d'analyse formantique, Ce processus donne des résultats satisfaisant tant pour de la parole voisée (on détecte alors des segments d'harmoniques) que pour la parole non voisée: le caractère local de la détection des formes d'onde permet en effet de reproduire des signaux transitoires trés brefs ou des signaux aléatoire.

4 RESULTATS

Le système réalisé permet le traitement automatique d'un segment de parole, et a été testé pour diverses voix tant féminines que masculines. La qualité de synthèse est excellente (pas ou très peu de différence avec l'original), mais doit maintenant faire l'objet de tests systématiques.

Le procédé s'appuie sur le modèle linéaire classique de production de la parole (de par la segmentation sur l'enveloppe spectrale) mais il autorise une analyse d'une grande finesse spectro-temporelle tout en ne délivrant qu'un jeu discret d'objets porteurs de toute l'information contenue dans le signal.

L'affichage des formes d'ondes dans le plan temps/fréquence permet de visualiser le résultat obtenu, qui paraît particulièrement intéressant, en ce sens que la plupart des caractères perceptivement pertinents (formants, pitch, voisement, explosions ...) sont représentés par un ensemble réduit de formes d'ondes (fig. 4, fig. 5, fig. 6).

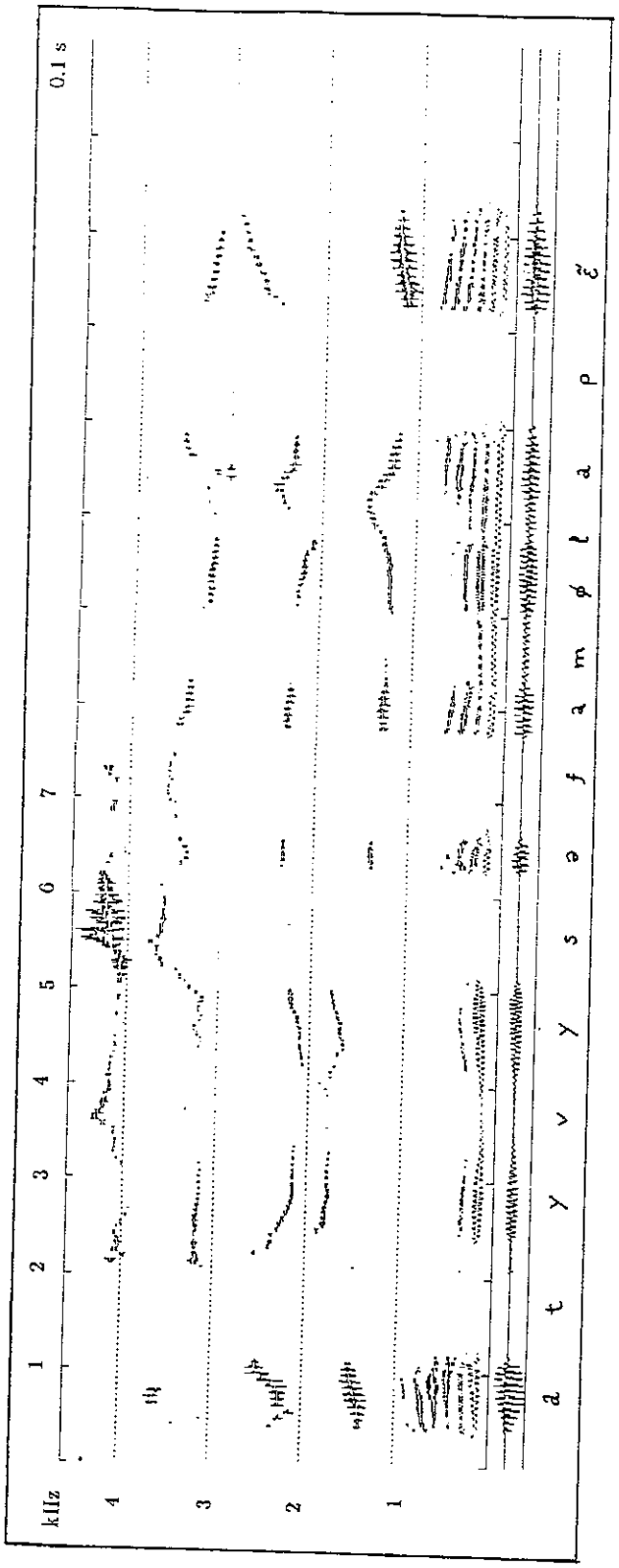

Figure 4: Affichage des formes d'ondes dans le plan temps/fréquence.

5 CONCLUSIONS

Nous avons présenté les fondements et la réalisation d'un système de représentation du signal de parole en formes d'ondes élémentaires. De part le choix opéré sur les formes d'ondes à rechercher, qui dérive du modèle classique de production du signal vocal, un traitement différent doit être appliqué aux différentes régions spectrales: une méthode spécifique semble en effet nécessaire pour rendre compte de la partie grave du spectre, et se trouve ici développée.

Il s'agit désormais, en raprochant cette méthode de modèles de perception, de l'appliquer à l'analyse de la parole. Les paramètres qu'elle fournit semblent également utiles pour une variante de la synthèse à formants en parallèle.Le débit d'information obtenu, qui reste à estimer de façon systématique, paraît offrir de bonnes potentialités en vue du codage: un gain semble en effet possible par rapport au codage sinusoïdal de part l'agglomération de plusieurs harmoniques dans une seule forme d'onde.

Ce travail représente le fruit de nombreuses discussions avec M^{rs} J.S.Liénard & X.Rodet, que l'auteur tient à remercier ici.

REFERENCES

[Liénard 87] Liénard, J.S. "Speech Analysis and Reconstruction Using Short-Time, Elementary Waveforms". IEEE-ICASSP 87, Dallas.

[d'Alessandro 87] d'Alessandro, C. & Rodet, X. "Fonctions d'ondes formantiques: extraction des paramètres et synthèse vocale". $16^{ième}$ JEP, 1987 Hammamet.

[d'Alessandro 88] d'Alessandro, C. & Liénard, J.S. "Decomposition of the Speech Signal into Short-Time Waveforms Using Spectral Segmentation". IEEE-ICASSP 88, New-York.

[Klatt 80] Klatt, D. "Software for o cascade/parallel formant synthesizer". JASA vol. 67(3), Mar. 1980.

[Rodet 80] Rodet, X. "Time Domain Formant-Wave-Function Synthesis". in "Spoken Language Generation and Understanding", J.C. Simon ed., D.Reidel publishing compagny, Dordrecht.

[McAulay 86] McAulay, R. & Quatieri, T. "Speech Analysis/Synthesis Based on a Sinusoidal Representation". IEEE trans. on ASSP, vol. ASSP 34 no. 4 1986.

[Goupillaud 85] Goupillaud, P., Grossmann, A. & Morlet, J. "Cycle-Octave and Related Transforms in Seismic Signal Analysis". Geoexploration 1985.

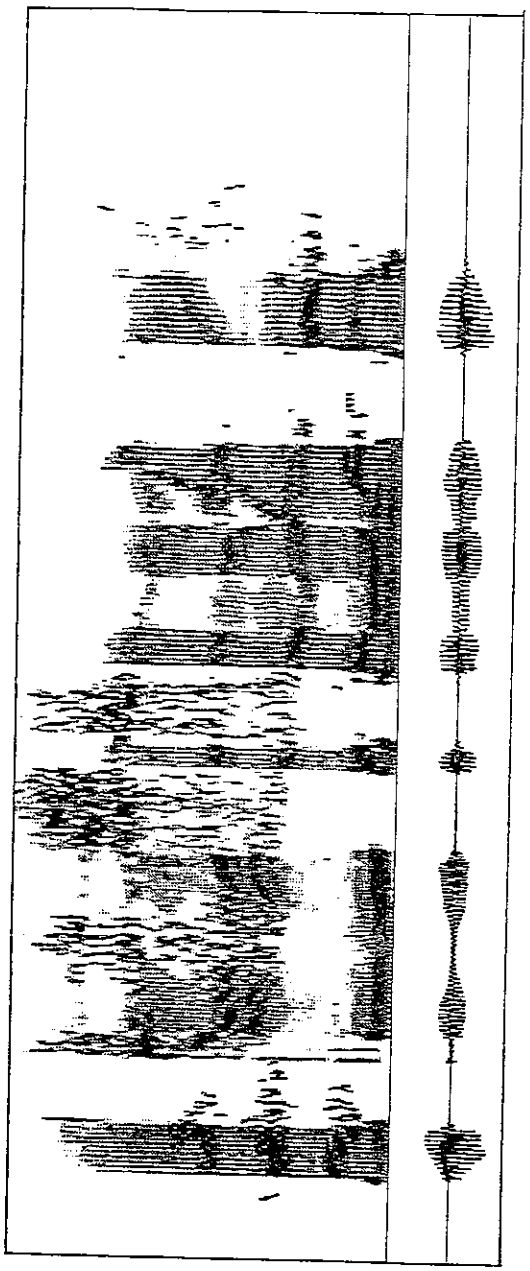

Figure 5: Spectrogramme du signal naturel correspondant à la figure 4.

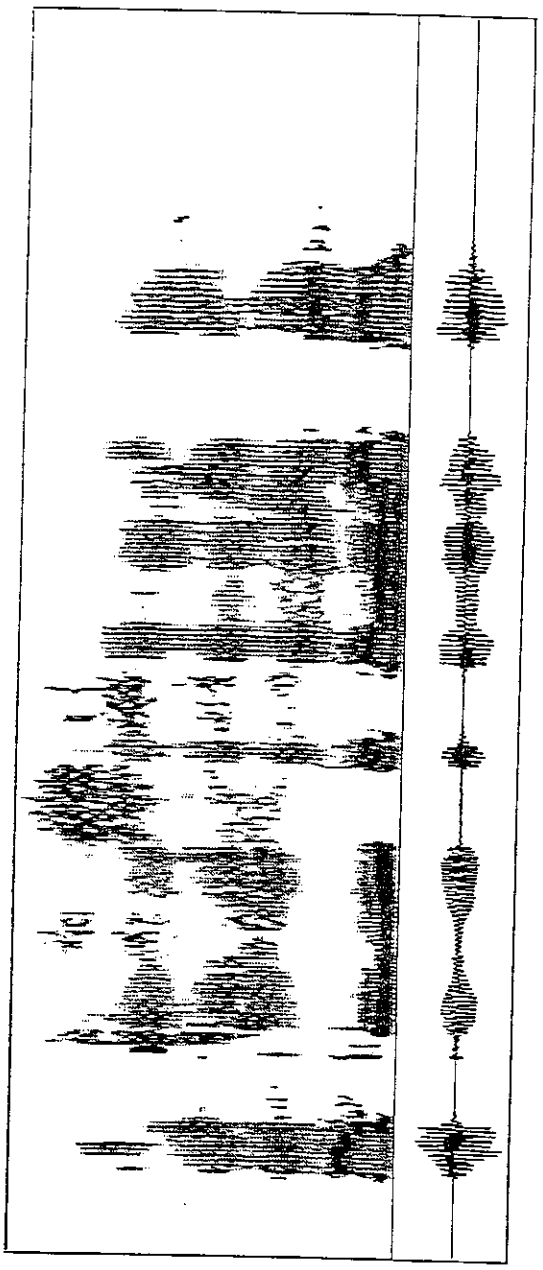

Figure 6: Spectrogramme du signal synthétique correspondant à la figure 4.