COLLOQUE NATIONAL SUR LE TRAITEMENT DU SIGNAL
ET SES APPLICATIONS

NICE du 16 au 21 JUIN 75

DECOMPOSITION ET DESCRIPTION INTEGRALE DU SIGNAL VOCAL

M. Ismaël EL-MALLAWANY

C.N.E.T. Route de Trégastel 22301 LANNION

RESUME

Les caractéristiques d'un son de parole sont déterminées par les propriétés de la fonction d'excitation, de la propagation dans les cavités bucco-nasales et du rayonnement des lèvres au capteur. La description du signal de parole est possible à l'aide d'un modèle paramétré du conduit vocal (C.V.), soit sa fonction de transfert. Cette fonction est continûment variable mais presque fixe dans un intervalle de 10 à 30 ms. L'étude de la propagation dans le C.V., moyennant certaines hypothèses, conduit à un modèle du type filtre numérique. La détermination adaptative des paramètres du modèle est effectuée à l'aide d'un algorithme à base de prédiction linéaire. Ces paramètres sont liés à la fonction d'aire du C.V. Dans le cas des sons à excitation glottale un signal d'erreur fonction du temps est déterminé à partir du signal de parole. Un algorithme, faisant appel à ce signal d'erreur, permet la détermination de l'Intervalle de Fermeture de la Glotte (IFG). Il est ensuite possible de déterminer une période de mélodie. L'identification du modèle, après une adaptation au locuteur et une égalisation auto-adaptative des pertes dans le C.V., peut s'effectuer soit sur l'IFG, soit sur une période de mélodie. Le profil de la fonction d'aire du C.V. est ensuite obtenu à partir du modèle déterminé. La fonction d'excitation glottale est déterminée par déconvolution du signal de parole avec le modèle codé sur l'IFG.

Les applications possibles de cette approche comptent, outre l'analyse, la synthèse et la reconnaissance de la parole, l'identification du locuteur.

SUMMARY

The characteristics of a speech sound are determined by the properties of the excitation function, the transmission through the vocal and nasal tracts and the radiation from the lips to the microphone. The speech signal can be described with reference to a model of the vocal tract (VT), which is its transfer function. This latter function is continually variable but almost stationary over time spans of 10 to 30 ms. Given certain assumptions, the analysis of the propagation within the VT leads to a digital filter model, whose parameters can be adaptively determined using a linear predictive coding algorithm. These parameters are related to the area function of the VT. An error time function is determined from the speech signal for voiced sounds. An algorithm applied to this error signal leads to the determination of the closed glottis interval (CGI). A pitch period can then be delimited. The model can be identified by linear prediction either over the CGI or the pitch period, after an auto-adaptive equalization of the source and radiation contributions has been performed. The shape of the area function of the V.T. can then be derived from the identified model. Finally, the excitation function is obtained from the integrated speech signal and the model determined over the CGI by deconvolution.

The potential applications of this approach include not only analysis and synthesis of speech, but equally speech recognition and speaker identification.

DECOMPOSITION ET DESCRIPTION INTEGRALE DU SIGNAL VOCAL

I. INTRODUCTION.

Le signal de parole est le principal support d'information dans les télécommunications. De ce fait, une connaissance approfondie de la structure et des propriétés de ce signal est un préalable à la réalisation de tout système de traitement soumis à des contraintes d'ordre économique et technique et de qualité (intelligibilité et agrément). L'intérêt porté aux études sur le codage de la parole tient au problème d'écoulement d'un trafic toujours plus intense sur le réseau téléphonique. Le développement des techniques numériques a permis la mise à l'étude de divers services spéciaux tels que le centre de renseignement automatique, le stockage de message, la dénumérotation etc... Ces services relèvent de la communication homme -machine, dont les deux composantes principales sont un organe de reconnaissance de la question posée par l'abonné et un système de synthèse de la réponse à lui fournir.

Quelle que soit l'application en vue, le préalable est le codage de la parole en vue de la réduction des redondances soit l'extraction des seuls paramètres significatifs. Il existe différents systèmes de codage tels que les vocoders [1], et, plus récemment, les méthodes dites de prédiction linéaire [1-5]. L'approche retenue dans ces "codeurs" est du type global (boîte noire) et l'analyse porte sur un intervalle important (10 à 25 ms), que la variation relativement faible du spectre en fonction du temps justifié en partie. Néanmoins, ces méthodes se sont révélées imprécises en raison de la non-stationnarité du signal et du caractère pseudo-périodique de la source d'excitation dans le cas des sons sonores [6]. Par ailleurs, aucune séparation n'est faite entre la source et la réponse du "filtre" : cavités bucco-nasales. De ce fait, les paramètres obtenus sont fortement dépendants du locuteur ce qui rend la reconnaissance de la parole très difficile. Afin de contourner cette difficulté, les chercheurs ont reporté leurs efforts sur l'introduction des contraintes d'autres niveaux (phonétique, lexical, syntaxique et sémantique) pour surmonter les ambiguïtés du niveau acoustique. Les résultats obtenus semblent indiquer qu'une description plus complète au niveau du codage soit souhaitable. Plus précisément, la décomposition du signal en ses deux composantes (source - conduit vocal, C.V.) pourrait réduire la dispersion des valeurs des paramètres significatifs pour un phonème prononcé par divers locuteurs. Par ailleurs, une description plus précise du signal de parole (ex : analyse en synchronisme avec la mélodie : fondamental de la voix) devrait permettre de dégager des règles plus simples en synthèse.

Dans cet article, une approche automatique est proposée au problème de la détermination de l'intervalle de fermeture de la glotte (IFG). L'identification des paramètres d'un modèle du C.V. à partir du signal dans cet intervalle permet la détermination de la fonction d'excitation par déconvolution. De cette fonction les bornes d'une période de la mélodie sont déterminées. Une nouvelle identification du modèle est réalisée sur cette période, après égalisation des influences de la source et du rayonnement (lèvres-capteur) sur le signal de parole. Le profil de la fonction d'aire et le spectre du CV sont obtenues de ce dernier modèle.

II. LE MODELE.

L'analyse de la structure d'un signal est un préalable à toutes définitions d'un modèle. Une telle analyse doit permettre d'identifier les composantes du signal, de déterminer ce qui les caractérise, et enfin d'intégrer ces divers éléments dans des relations susceptibles de reproduire les caractéristiques observées du signal.

Une simplification consiste à ne distinguer que 2 catégories de sons : les sons sonores pour lesquels l'excitation du C.V. par les cordes vocales prend la forme d'une suite de créneaux presque-périodiques, et les sons sourds pour lesquels l'excitation est engendrée par le passage turbulent de l'air à travers des constrictions du C.V. Dans tous les cas les caractéristiques du son émis sont déterminées par les propriétés de la fonction d'excitation (la source), de la transmission à travers les cavités bucco-nasales et du rayonnement des lèvres au capteur du son.

La longueur et l'aire de section variable caractérise chacun des deux conduits vocal et nasal. Des deux le C.V. est le plus important. En première hypothèse, on supposera que le comportement du C.V. est identique à celui d'un tuyau rectiligne dans lequel ne se propagent que des ondes acoustiques planes normales à l'axe. Cette hypothèse qui consiste à ne pas tenir compte de la propagation des modes transversaux est valable en première approximation si les dimensions du tube sont inférieures aux longueurs d'onde étudiées soit pour des fréquences inférieures à 5000 Hz. Une deuxième hypothèse consiste à représenter le C.V. par une succession de N sections cylindriques d'égale longueur, Δ, et d'aire de section An. Si Δ est bien plus petite que la longueur d'onde de la fréquence la plus élevée prise

en compte dans le signal, l'erreur introduite par la quantification de la fonction d'aire (de section) du C.V. sera faible. Ces hypothèses d'analyse sont encore valables si des variations brutales dans le tube provoquant des réflexions internes de l'onde sont exclues. Le coefficient de réflection, γ_n, entre deux sections adjacentes est définie par

$$\gamma_n = (A_n - A_{n+1}) / (A_n + A_{n+1}) \ldots (1)$$

Nous ne traiterons ci-après que de la propagation des ondes sonores à l'intérieur du C.V. La transmission dans une section est décrite en termes de la pression ou de la vitesse volumique d'une onde aller (glotte-lèvres) et d'une onde retour (du fait des réflexions subies au moment des discontinuités entre 2 sections). La description des caractéristiques de cette propagation est facilitée par l'établissement d'une analogie acoustique/électrique (tube élémentaire ⟷ résonateur équivalent ; pression ⟷ tension ; vitesse ⟷ courant). Pour simplifier davantage l'analyse on avance l'hypothèse d'un conduit sans pertes. L'application des conditions de continuité de vitesse volumique et de pression permet d'établir deux relations entre les paramètres de deux sections adjacentes [2,3]. On peut avancer, de plus, l'hypothèse que la source a une vitesse volumique constante à l'entrée de la première section. Le rayonnement aux lèvres peut être considéré comme provenant d'une source sphérique, dans quel cas le schéma équivalent est une résistance en parallèle avec une self (autre approx.impédance aux lèvres résistive). Le modèle recherché qui n'est autre que la fonction de transfert du C.V. est déduit des considérations précédentes [2,3].

Dans la mesure où les variations du conduit vocal peuvent être approximées par une succession de configurations stationnaires, il est possible de définir une fonction de transfert en la variable complexe z pour le conduit. Dans le cas de sons sonores non nasalisés cette fonction ne comprendra que des pôles. Dans le cas contraire (sons nasalisés et sons sourds), il y aura en plus des antirésonances. Ces zéros, situés à l'intérieur du cercle unité, peuvent être remplacés par des pôles (dont le nombre dépendra de la précision requise), ce qui réduit le modèle à un filtre numérique linéaire constitué de pôles exclusivement de la forme

$$G(z) = K_g / \left(1 - \sum_{k=1}^{N} b_k z^{-k}\right) \ldots\ldots (2)$$

La source d'excitation peut être représentée approximativement par 2 pôles et le rayonnement des lèvres au capteur par une dérivation. Le tout peut être réduit à un modèle à 2 pôles [2]. Il vient, le modèle global.

$$H(z) = K_h / \left(1 - \sum_{k=1}^{p} a_k z^{-k}\right) \; ; \; p = N+2 \ldots\ldots (3)$$

Toutes les hypothèses et simplifications énumérées précédemment ne sont pas nécessaires à la dérivation du modèle global, mais principalement pour parvenir à des relations simples entre les paramètres du modèle et ceux du C.V.

La méthode de décomposition du signal, s_t, s'applique au sous-ensemble des sons sonores. Dans la mesure où ces phonèmes ne sont pas nasalisés les coefficients de réflection, γ_n, peuvent être calculés à partir des b_k dans (2) à l'aide d'une équation de récurrence [2-6]. Enfin, le profil d'aire du C.V. est obtenue des γ_n. Par ailleurs, le nombre N de coefficients nécessaire au codage doit être tel que N * T(période d'échantillonnage) soit égal à deux fois le temps de propagation d'une onde de la glotte aux lèvres [2, 6].

Détermination des paramètres.

La détermination des paramètres fait appel à une méthode d'optimisation, et, par conséquent, à un critère d'optimisation. Ces méthodes nécessitent des calculs assez longs et des algorithmes difficiles à exploiter en temps réel dans l'état actuel de la technologie des semiconducteurs. Par conséquent, les traitements sont réalisés par simulation sur calculateur numérique. Nous noterons que la suite d'échantillons s_n du signal de parole dans un intervalle D constitue la réponse impulsionnelle d'un filtre numérique.

Différentes méthodes d'optimisation sont possibles, à savoir le codage prédictif linéaire [1 - 6], le filtre inverse optimal [4,7], et les coefficients d'autocorrélation partielle [4,8], et le filtre de Kalman [4,9]. Nous décrivons brièvement le principe de la première de ces méthodes.

Exception faite du premier échantillon de l'intervalle, les valeurs de s_n sont approximativement obtenues par une somme pondérée des p valeurs précédentes. Il en découle une erreur d'estimation e_n.

Le critère d'optimisation est la minimisation de l'erreur d'estimation au sens des moindres carrés sur l'intervalle D. En posant $\partial \langle e_n^2 \rangle / \partial a_k = 0$; on obtient p équations en p inconnues, d'où on détermine les a_k. La séparation de l'effet de la source est réalisée, ensuite, en éliminant les deux pôles réels ou proches de l'axe des réels de l'expression de H(z). Dans le cas des sons scurds, l'optimisation porte sur tous les échantillons de l'intervalle sans exception. Si les influences de la source et du rayonnement sont égalisés p = N et aucune séparation ne s'impose. L'erreur de prédiction normalisée est donnée par

$$E_p = \prod_{k=1,p} (1-\gamma_k^2) \quad \ldots\ldots\ldots\ldots\ldots (4)$$

III. L'intervalle de fermeture de la glotte.

L'IFG est un laps de temps pendant lequel le CV est en oscillation libre. Par conséquent, le codage sur cet intervalle détermine les caractéristiques propres du CV. La fonction d'excitation peut être obtenue du modèle optimisé et de l'intégrale de s(t) par déconvolution. Le principe de la méthode proposée repose sur le fait qu'à l'instant de fermeture de la glotte le signal contient des hautes fréquences d'énergie élevée. De ce fait, une analyse de la vitesse volumique aux lèvres sur des intervalles de temps relativement courts manifeste une pointe d'erreur de prédiction sur l'IFG. La sélection de cette pointe parmi d'autres se fait en détectant une pente négative sur le signal source au niveau de ce maximum.

Les étapes de l'algorithme sont les suivantes :

1) Une indication de la fréquence de mélodie, f, caractéristique du locuteur permet la détermination approximative du nombre de paramètres, p, du modèle le mieux adapté à l'analyse du signal (ex : p=(12-f/100)/(10000 * T) arrondie à l'entier inférieur).

2) L'intervalle d'analyse élémentaire $\tau = 2(p-1)*T$ secondes.

3) Décision Parole/Silence : test niveau de signal supérieure ou non au niveau de bruit du canal.

4) Cadre d'analyse , $D = 3/f + \tau$ secondes.

5) Intégration de s_t sur D : $v_t = s_t/(1-z^{-1})$

v_t = vitesse volumique aux lèvres.

6) Calcul du signal d'erreur, e(nT), n=1,2..., $(D-\tau) / T$

$e(nT)=E_1(nT)-E_p(nT)$

où $E_i(nT)$=erreur de prédiction normalisée du modèle d'ordre, i, calculé sur v_t dans l'intervalle de temps t=nT à $(nT+\tau/T-T)$ après avoir mis la moyenne de v_t à zéro.

7) Détecter le minimum M=e(kT) de e(nT) dans l'intervalle 0.2*D/T à 0.8*D/T. M représente un point de départ , au début d'un intervalle τ, qui se situe toujours dans l'intervalle d'ouverture de la glotte, et dans lequel les harmoniques élevés ont une très faible énergie.

8) Localiser trois maxima significatifs, dont deux directement avant M (ê(jT) et ê(iT) ; i>j) et le troisième juste après, ê (mT). Le terme significatif implique qu'aucune autre valeur de e(nT) ne dépasse la valeur de la pointe à une distance $\leq (2pT/3)$. Si l'instant kT correspond à un point sur le front montant du créneau glottale, la probabilité est grande que la pointe recherchée soit ê (jT), car il y a généralement une pointe à l'instant d'ouverture de la glotte. Ce minimum, M, peut être très proche du début d'ouverture de la glotte ce qui mènera à retenir ê(iT). Si kT est sur le front descendant la pointe sera ê (mT).

9) La décision finale est prise après : i) la détermination du modèle (2) sur les intervalles correspondant aux trois pointes ; ii) Une déconvolution $F(z)=V(z) * (1- \sum_{k=1}^{p} b_k z^{-k})$ sur l'intervalle (jT) à $(mT+\tau)$ par modèle calculé ; iii) La détection du minimum de chaque F(z) et sa localisation dans un des trois intervalles retenus. L'intervalle dans lequel y figure au moins deux minima est l'IFG.

L'indication de la mélodie caractéristique du locuteur est utile principalement pour les traitements ultérieurs. S'il s'agit de déterminer le profil d'aire du CV il est indispensable que la valeur de p soit correctement choisie. Si ce n'est pas le cas il suffirait d'indiquer que le locuteur est un homme, une femme ou un enfant. Dans certaines situations, il peut être difficile de donner la moindre indication. Dans des cas aussi extrêmes il faudra adopter une valeur de p valable dans tous les cas, soit p = 7/(10000 *T). Néanmoins, pour que l'algorithme demeure fiable pour les voix d'homme il faudra également retenir trois points avant le minimum et deux après.

IV. Extraction des paramètres significatifs.

L'IFG étant déterminée, il est possible de :

1) Déterminer les paramètres b_k à partir de V_t sur l'IFG. Effectuer une déconvolution et obtenir "le signal d'excitation". Ce signal peut être codé à l'aide d'un modèle à 2 ou 3 paramètres.

2) Obtenir les b_k à partir de s_t sur l'IFG et calculer le spectre caractéristique du CV

$$G(jw) = 1./(1. - \sum_{k=1}^{p} b_k \exp(-jwkT))$$

Les maxima du spectre représentent les formants (fréquences de résonances des cavités buccales), ou en d'autres termes si

$$(1. - \sum_{k} b_k z^{-k}) = \prod^{k} (1 + P_k z^{-1})$$

Les formants sont donnés par

$$f_k = \mathrm{Im}(\mathrm{Log}_e P_k) / (2\pi T)$$

Et leur largeur de bande par

$$lb_k = -\mathrm{Re}(\mathrm{Log}_e P_k) / (\pi T)$$

3) Egaliser les influences n'appartenant pas au CV dans s_t sur l'IFG à l'aide d'un pole réel adapté, calculer les b_k sur ce signal égalisé et en déduire le profil d'aire du CV.

Il se pose, néanmoins, quelques difficultés du fait que l'IFG est de durée variable. A titre d'exemple, des "a" de femme ou d'enfant sont caractérisés par des IFG trop faible, ce qui conduit à des profils d'aire peu précis. La solution retenue consiste à localiser l'IFG, déterminer les paramètres du modèle, obtenir la fonction d'excitation, délimiter une période de mélodie à partir de la forme de l'excitation, et appliquer la procédure d'extraction du profil d'aire à cet intervalle.

Cette méthode appliquée au mot OUI a donné les résultats portés sur la figure 1. Les graphiques représentent de haut en bas, le signal à analyser, s_t , la fonction d'erreur, e(nT), le profil de la fonction d'aire, A_i, le spectre du CV, et le signal d'excitation f(nT). De gauche à droite, l'évolution dans le temps est de 6ms, et l'on peut suivre la transition de la fonction d'aire de la forme d'un OU vers celle d'un I.

V . Conclusion.

Cette méthode s'est avérée très fiable pour toutes les voyelles non-nasales et occlusives voisées (b, d, g) du français, pour différents locuteurs (homme, femme, et enfant) et à deux cadences d'échantillonnage (10 KHz et 12 KHz). L'intérêt principal de cette approche ne réside pas dans les résultats obtenus, mais dans les prolongements possibles.

Cette approche présente un intérêt direct dans d'autres domaines tels que :

- La vérification et l'identification de locuteur (signature vocale)
- La mise en évidence des caractéristiques d'élocution dues à des affections d'origine diverse (cancer, aphasie)
- Les sciences du langage (voix mal posées, rééducation, apprentissage des langues).

VI. Bibliographie.

[1]I. El Mallawany, P. Lorand, F. Platet : Méthodes de codage en vue de l'analyse et de la synthèse de la parole, Congrès AFCET, Informatique et Télécommunications, Rennes, Nov.1973; tome 1, pp 61-70.

[2]I. El Mallawany : Détermination de la Fonction d'aire du conduit vocal par codage prédictif ; Journées d'Etudes sur la Parole : GALF - LANNION (1972).

[3]I. El Mallawany :"Fonction de transfert et Fonction d'aire du conduit vocal", Note technique CEI/CSI/43 CNET-LANNION (1er février 1974).

[4]I. El Mallawany : "Fonction de transfert et fonction d'aire du conduit vocal", analyse et synthèse de la parole Vol-I (1972-1973) CNET-LANNION.

[5]J.I.Makhoul,J.J.Wolf : Linear prediction and the spectral analysis of speech, Rept 2304 Bolt Beranek and Newman, Cambridge (Aug.1972).

[6]B.S.Atal, S.L. Hanauer : "Speech Analysis and synthesis by linear prediction of the speech wave". Jour. Acous. Soc. Amer., 50, pp 637-655,1971.

[7]J.D. Markel, A.M. Gray : On autocorrelation Equation as applied to speech analysis". IEEE Trans on Audio and Electroacoustics Vol AU-21, no 2 April 1973.

[8]F. Ittakura, S. Saito : Digital Filtering Techniques for Speech Analysis and Synthesis, Proc. Int. Congr Acoust, 7 th, Budapest (Aug. 1971)

[9]C.Gueguen,G.Carayannis:Analyse de la parole par filtrage optimal de Kalman.Automatisme.Tome 18 n°3.1973

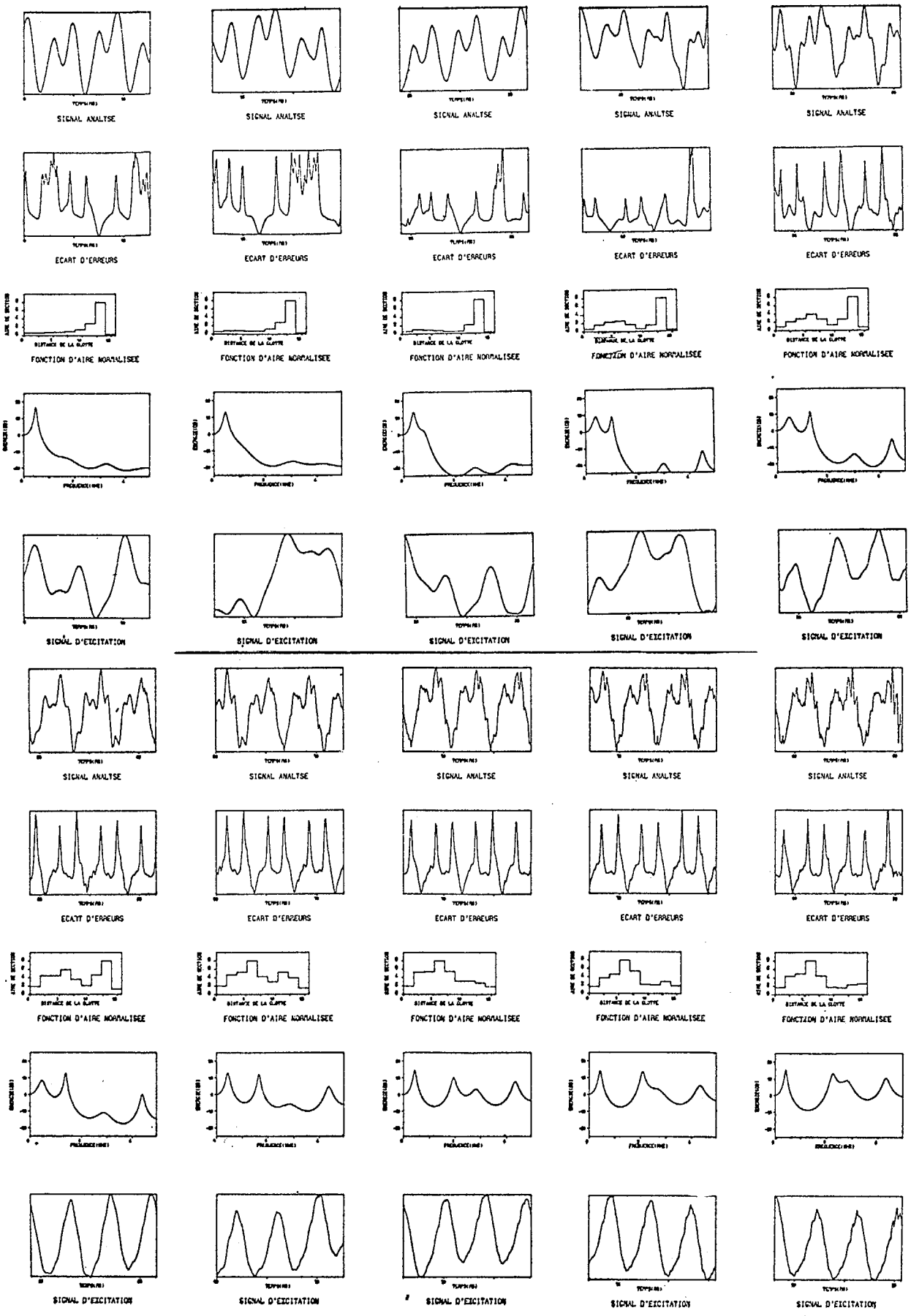

Figure 1 - Diphonème Analysé = OUI ; Voix femme.
T = 1.E-4 ; Filtre Inverse Optimal.