COLLOQUE NATIONAL SUR LE TRAITEMENT DU SIGNAL
ET SES APPLICATIONS

NICE du 16 au 21 JUIN 75

LA COMPRESSION D'INFORMATION TELEPHONIQUE

COMPRESSION OF TELEPHONE INFORMATION

MM. BELLEC-BOURGENOT-DECHAUX-DEMAN-PINEL-POTAGE.

THOMSON-CSF - 16 Avenue du Fossé Blanc - 92231 - GENNEVILLIERS -FRANCE

RESUME

L'exposé se propose de présenter les résultats obtenus à la Division Télécommunications de Thomson-CSF sur la transmission de parole sur des canaux de transmission numérique à faible débit (1200 et 600 bits/sec.). Les études poursuivies depuis cinq ans avec le soutien de la DRME ont abouti à la définition d'une maquette temps réel pour le compte du STCAN avec un débit de 1200 bits/sec. pendant que la poursuite de la simulation sur ordinateur conduit à la faisabilité d'un débit de 600 bits/sec.

Le taux de compression et la qualité recherchés a conduit à effectuer une analyse de "formes acoustiques articulées" sur le signal analogique limité aux fréquences 300-3400 Hz.

Les résultats obtenus par simulation de tests de Rimes ont donné une intelligibilité intermédiaire entre le signal d'origine et les Vocoders à canaux à débit double (2400 bits/sec), avec un agrément sensiblement meilleur.

Les caractéristiques les plus originales sont :

- Analyse spectrale à court terme suréchantillonnée avec une structure calquée sur celle de l'oreille,
- Les formes significatives prises en compte dans l'analyse et le codage sont des segments de courbes, y compris les points formant les lignes des crêtes de la surface du spectre à court terme, dans un espace temps-fréquence-amplitude. Le codage à 600 bits/sec utilise l'approximation de ces courbes par des segments de droite de grande longueur,
- Algorithmes d'analyse continue du signal sans discrimination voisé-non voisé,
- Synthétiseur simplifié reconstituant formellement les éléments significatifs.

La validité de l'analyse est démontrée par un résultat original : l'intelligibilité ne dépend pratiquement pas de la présence ou de l'absence de la mélodie, de sa valeur, ni dans une certaine mesure de ses variations.

SUMMARY

The results obtained at Thomson-CSF Telecommunications Division concerning voice transmission on low speed (1200 and 600 bits/sec) digital channels are presented. The outcome of a five years DRME (Direction des Recherches et Moyens d'Essais) supported study has been the design of a real time model for STCAN (Service Technique des Construction et Armes Navales). The rate will be 1200 bits/sec ; however, computer simulations indicate that a 600 bits/sec rate is feasible.

Considering the desired compression ratio and voice quality, an analysis of "articulated acoustical features" has been conducted, making use of the 300-3400 Hz handwidth limited analog signal.

From the resultats obtained by simulating Rimes's tests, the intelligibility is situated between the original signal and 2400 bits/sec channel Vocoders, and the feeling is substantially more pleasant.

The most distinctive characteristics are :

- Oversampled short term spectrum analysis with a structure reproducing the one of the ear.
- The significant features which are taken into account for analysis and encoding are curves segments, including the points which constitute the crest lines of the short term spectrum surface in a time-frequency-amplitude space. For the 600 bits/sec encoding, these curves are approximated by long straight segments.
- Algorithms performing a continuous signal analysis without discriminating between voice and non-voice intervals.
- Simplified synthetizer formally rebuilding the significant features.

The soundness of the analysis is demonstrated by a singular result : intelligibility is nearly independent of the presence or absence of pitch and in the former case of its frequency and to a certain extent of its fluctuations.

Speaker : P. DEMAN

1)- INTRODUCTION AU PROBLEME

La terminologie "compression d'information" est par elle même ambigüe puisqu'il s'agit précisément de conserver toute l'information en s'efforçant de la transmettre à travers un canal de capacité d'information aussi réduite que possible.

Le schéma représentatif du problème posé correspond à la séquence :

signal vocal analogique
traitement de compression
traitement numérique
transmission "comprimée"
traitement de synthèse
signal vocal analogique

avec les contraintes suivantes :

- le signal vocal est supposé déjà acheminé par un réseau analogique et limité par conséquent aux fréquences 300-3400 Hz.
- le retard admissible apporté par les opérations de traitement ne doit pas dépasser 100 à 600 milli/sec.pour autoriser le dialogue.
- la dégradation de la qualité doit rester acceptable pour des taux d'erreurs en transmission de l'ordre de 10^{-3}.
- la qualité "objectif"(intelligibilité, agrément et reconnaissance de locuteur)doit être bonne.
- le traitement retenu doit être universel et pouvoir s'adapter à tout langage articulé humain

L'examen des contraintes du problème montre qu'il est à la fois plus facile et plus vaste que ceux posés dans des systèmes utilisant la reconnaissance ou la synthèse vocale: si la condition de transparence à tout langage articulé élargit le problème, il est possible de laisser sans solution la reconnaissance d'une forme imparfaite déjà dégradée par ailleurs.

2) - LES ORIENTATIONS DE L'ETUDE

Les travaux de recherche ont bien entendu utilisé l'acquit d'études antérieures effectuées sur la parole et plutôt que de revenir sur les éléments connus, nous essaierons de dégager les points originaux qui ont marqué la réalisation et les résultats.

2.1. -La suppression des éléments non significatifs du message devant donner un bénéfice net sur le taux de compression sans accroissement de la vulnérabilité aux erreurs, on a utilisé les connaissances sur la perception auditive pour simplifier un certain nombre de paramètres de traitement ; sans aller jusqu'au bout de la démarche(réalisation d'une oreille artificielle) on retiendra :

- analyse temps fréquence similaire à celle étudiée par Békasy dans la structure de la cochlée,
- exploitation de l'effet de masque des composantes à niveaux forts en négligeant à la synthèse les composantes de faible niveau,
- variation progressive de la précision de la perception avec la durée de stationnarité du signal.

2.2. - Les diverses représentations temps fréquence (spectres à court terme)d'un même signal dépendent de la structure de l'analyseur utilisé (batterie de filtres) ; cette remarque nous a conduit à utiliser une structure voisine de celle de la cochlée et aussi à pratiquer systématiquement un suréchantillonnage des signaux issus de la batterie de filtres. En effet les filtres n'étant pas à bande limitée et le canal de transmission n'étant pas synchrone, il n'est pas possible de définir un système de fonctions orthogonales ; en d'autres termes les réponses temps-fréquences obtenues sont toujours des combinaisons linéaires des composantes du signal analysé. Inversement les réponses obtenues en des points temps fréquences voisins ne sont pas totalement corrélées et seul un sur-échantillonnage permet d'espérer recueillir les paramètres significatifs dans leur intégralité.

A 300 Hz, la largeur de bande à 3 dB des filtres et l'espacement entre filtres sont de l'ordre de la dizaine de Hz, pour une cadence d'échantillonnage de 4 millisecondes.

Utilisant la diminution d'acuïté d'analyse en fréquence de l'oreille pour les fréquences aigües une répartition logarithmique des résonateurs à coefficient de surtension constant a été adopté.

2.3. - On sait que l'intelligibilité dépend peu de la structure voisée ou non, puisque la parole chuchotée reste parfaitement compréhensible. On a donc induit que les règles de traitement de la perception ne devraient pas dépendre de la structure voisée ou non ni de la décomposition en phonèmes, consonnes ou voyelles. L'effet de masque par les composantes acoustiques de fort niveau sur celles de niveau inférieur a conduit à rechercher les éléments significatifs dans les "lignes de crêtes"du spectre à court terme précédemment obtenu et à limiter la synthèse à la représentation formelle d'un nombre fini de ces composantes, de fréquences et d'amplitude variable en fonction du temps ; 3 peuvent être considérées comme suffisantes.

L'expérience a, comme on le verra plus loin, confirmé partiellement ces hypothèses

2.4. - Au fur et à mesure de l'étude, aux hypothèses de travail d'origine "perception" se sont amalgammées des connaissances sur la "phonation" et l'on peut dire que le vocoder à ligne de crête étudié par TH-SCF procède d'une

reconnaissance de "formes acoustiques articulées" :

- acoustiques parce que définies comme discernables à l'audition
- articulées parce que générées par un organe de phonation (poumons, cordes vocales et conduit vocal).

3) - LA METHODOLOGIE UTILISEE

La complexité du problème et l'ampleur des variations paramètriques possibles a conduit à mener l'étude autour d'une simulation sur ordinateur.

Dans une première phase, on a utilisé une cochlée artificielle construite à la Division des Activités Sous Marines et utilisée par M. ALINAT pour ses études de reconnaissances vocales. Les premiers résultats ont été obtenus par traitement de données obtenues à partir d'une vingtaine de phrases de 2 secondes, 5, prétraitées par cette batterie de filtres.

Par la suite la simulation sur ordinateur a été étendue au filtrage, afin de permettre entre autre le test des techniques de filtrage numérique. On a conservé la dimension des phrases pour être compatible à la fois avec les fichiers informatiques intermédiaires et avec les résultats graphiques d'un sonogramme de référence. Si les programmes d'entrées-sorties et de traitement du fichier ont été écrits en assembleur, les programmes de traitement l'ont été en Fortran pour faciliter l'optimisation .

Les phrases successives, qui ont sensiblement correspondu aux contrats successifs de la DRME, ont été les suivantes :

phase 1 - Définition de l'outil de simulation et vérification de la validité de l'analyse.

phase 2 - Intégration du filtrage dans la simulation, traitement de la mélodie, codage à 2400 bits/sec.

phase 3 - Codage à 1200 bits secondes, avec simulation du matériel équivalent nécessaire pour le traitement.

Deux extensions de l'étude sont en cours actuellement :

- Définition et réalisation d'une maquette temps réel (retard $\leqslant$ 100 milli/sec.) pour le compte du STCAN.
- Poursuite de la simulation pour l'obtention d'un débit de 600 bits/sec. avec un retard de traitement allongé à 600 milli/secondes.

L'optimisation des résultats a été dégrossie par un consensus des observateurs et validée par des tests de rimes. Ces derniers entraînant des temps de calcul relativement longs ont exigé une compression des fichiers informatiques et ont été limités à des essais de vérification, sans possibilités pratiques d'application systèmatique à toutes les variantes.

En l'absence des possibilités de tests de longue durée, l'optimisation s'est beaucoup appuyée sur les relations formelles entre les impressions acoustiques et les formes visuelles telles qu'elles pouvaient être déduites de fichiers intermédiaires. Il a été en particulier remarqué qu'il était possible de pratiquer un traitement visuel "à la main" et que s'il était possible de définir un algorithme informatique sensiblement équivalent "à l'oeil" le résultat acoustique se conserverait.

Cet aspect positif a d'ailleurs montré ses limites, car si une amélioration formelle ainsi obtenue a pratiquement toujours fourni un consensus sur l'agrément de l'audition, les tests de rimes ont montré que les algorithmes informatiques pouvaient dégrader la représentation d'autres éléments significatifs.

4) - LE TRAITEMENT DU SIGNAL

La succession des opérations de traitement est la suivante : (fig.1)

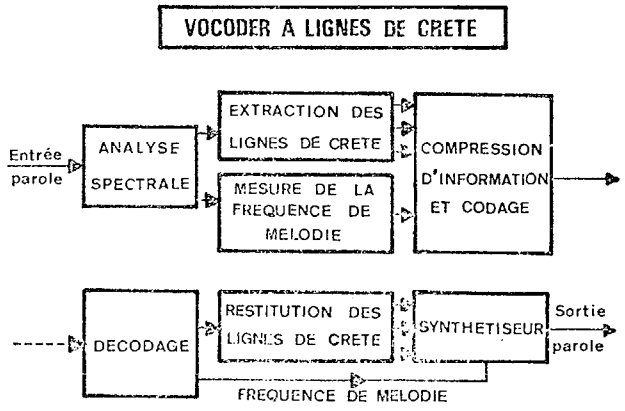
A l'émission :

4.1. - Analyse temps fréquence
4.2. - Détection des maxima instantanés
4.3. - Pistage des lignes de crêtes
4.4. - Extraction de la mélodie
4.5. - Codage au format de transmission

A la réception :

4.6. - Décodage
4.7. - Synthèse

FIGURE 1



4.1. - Analyse temps fréquence (fig.2)

La simulation a montré que cette analyse présentait un optimum peu critique autour des valeurs déduites des modèles bioniques. Ainsi il a été possible de simplifier la structure des filtres (résonateurs simples) et de formaliser la répartition des fréquences et des surtensions.

On a pu également vérifier que la limitation à 32 filtres soit 32 valeurs de définition de paramètre fréquence en répartition logarithmique était compatible avec la qualité cherchée.

4.2. - Détection des maxima instantanés

Des essais variés de recherche de maxima ont été faits à partir des analyses de différentielles premières ou secondes et de combinaison des deux. Parmi les résultats les plus interessants, on a montré le couplage entre cette recherche de maxima et les opérations suivantes en particulier les pistages, couplage lié au choix nécessaire, dus à la limitation des informations transmissibles par le code à débit et retard constant.

FIGURE 2

4.3. - Pistage des lignes de crêtes (fig.3a)

Les paramètres temps fréquence d'une ligne de crête devant être affectée à un des générateurs du synthétiseur en continuité de fonctionnement et le codage devant utiliser les possibilités de transmission de la valeur précise par approximation successive, il était dans le principe du dispositif de reconnaître l'individualité "ligne de crête"que celle-ci soit un élément transitoire de consonne ou qu'elle s'identifie pratiquement avec le "formant" des phonèticiens dans les périodes presque stationnaires des voyelles.

4.4. - Mélodie

Dans les hypothèses de départ, la mélodie pouvait être considérée "vue par une cochlée artificielle"comme une modulation d'amplitude des réponses de sortie des filtres.

Pour des raisons toutes empiriques : structures de la batterie de filtres d'origine et dimensions des fichiers informatiques intermédiaires, la mélodie a été éliminée de la structure des lignes de crête par des filtres passe bas et mesurée globalement à la sortie d'un seul des filtres de la batterie. Une astuce informatique permet l'utilisation en temps partagé de la même batterie de filtres numériques pour extraire le fondamental du signal observé.

Il n'y a pas de détection voisée-non voisée au niveau du traitement, ni au codage, ni au décodage.

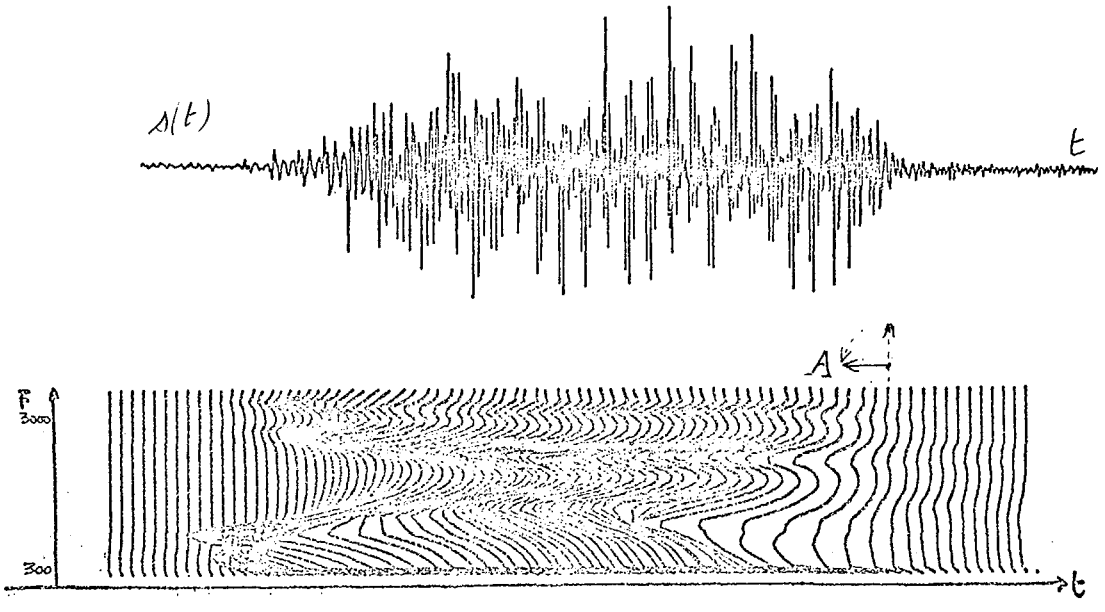
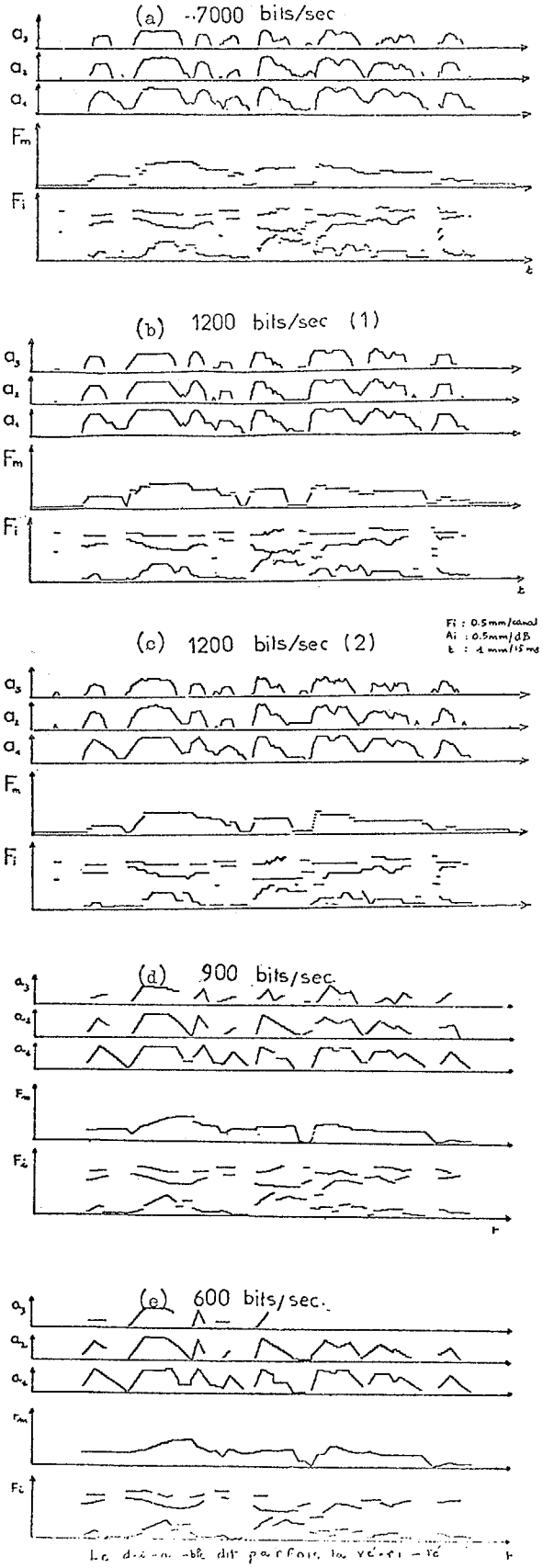


FIGURE 3



4.5. - Codage

Dans cet exposé l'opération "codage" comprend les traitements qui fournissent à partir d'une suite d'échantillons temporels des lignes de crête pistées à la cadence de 4 milli/sec, un message au format binaire synchrone de débit cherché soit 1200 à 600 bits/sec.

Dans la transmission à 1200 bits/sec, on a exploité les structures phonatoires et la redondance liée à la nature "mécanique" du générateur ; la suite des maxima pistés est considérée comme synchrone,le traitement gère les occurences, les priorités et la précision nécessaire pour une représentation correcte. (fig.3 b,c)

Le couplage en retour signalé plus haut est lié à la prise en compte à un instant donné non seulement des évènements présents ou passés de l'évolution de leur existence ou de leur importance dans le futur.

Dans la transmission à 600 bits/sec, les formes retenues comme significatives sont des segments de droite dans l'espace temps amplitude (fig.3 d,e).

Dans ce cas la représentation ne peut plus être considérée comme synchrone et nécessite donc l'adressage temporel des évènements. Il est donc nécessaire d'évaluer la précision avec laquelle doit être restitué le débit"en accordéon" dans le format du message.

En effet, le nombre d'évènements significatifs à décrire varie considérablement entre les silences et les périodes stationnaires du signal (voyelles) d'une part et les transitoires (consonnes) d'autre part.

4.6. - Décodage

Le traitement de décodage est rigoureusement la réciproque du codage précédent ; il s'effectue après la transmission et se propose de restituer au synthétiseur à la même cadence identique de 4 milli/sec.(3,3 milli/sec.dans le projet temps réel en cours) les informations temps fréquence amplitude de chacun des générateurs A1,F1,A2,F2,A3, F3,F0. Suivant les règles inverses du codage, l'organe de traitement procède à toutes les opérations nécessaires d'interpolation ou d'extrapolation.

4.7. - Synthèse

Le schéma du synthétiseur est donné sur la figure 4, il est constitué de 3 générateurs identiques fournissant des éléments de sinusoïdes enchaînés en phase les uns derrière les autres et dont l'amplitude et la fréquence sont commandées numériquement à la vitesse d'échantillonnage et susceptibles de prendre l'une quelconque des 32 fréquences de la batterie d'analyse.

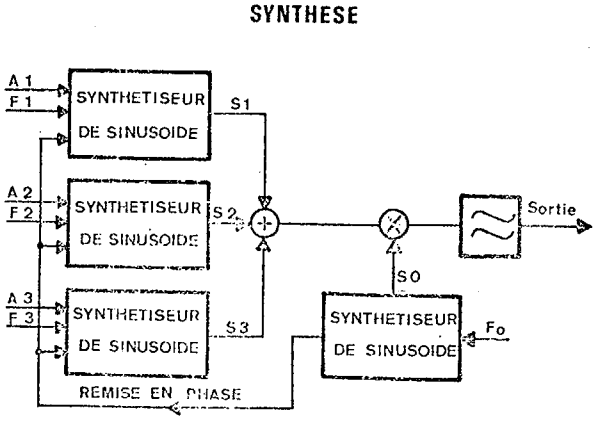
La restitution de la mélodie est obtenue à partir d'un générateur similaire mais dans une gamme de fréquences plus basses.

Ce générateur délivre des impulsions et une loi de modulation en cosinus surélevé de période égale. Les impulsions sont en phase avec le zéro de la loi de modulation et assurent au même instant la remise à une phase donnée des 3 générateurs.

La somme des signaux de sortie de ces trois générateurs est modulée par le signal de modulation de mélodie (So).

La remise en phase au moment du passage par zéro permet l'obtention d'un signal "périodique" dont la largeur de la réponse à la batterie du résonateur est normée ; en terme de phonétique et pour les structures de voyelles, on peut dire que la synthèse est faite à largeur constante des formants. Le prix de cette simplification doit correspondre à une perte sur les éléments significatifs dans l'identification du locuteur.

FIGURE 4



5)- RESULTATS

Les résultats obtenus au stade actuel, qui sont à confirmer par des tests d'intelligibilité plus élaborés sont les suivants :

Les valeurs globales de pourcentage sur test de rimes donnent les valeurs suivantes :

- Vocoder à canaux 93 % (2400 bits/sec)
- Vocoder à lignes de crête 94-95 % (1200 bits/sec)
- Signal vocal 300-3400 Hz 98,5 %

Sur l'agrément, un consensus s'établit avec le même classement. Sur la reconnaissance du locuteur, il n'y a pas de consensus : les essais temps réel devront préciser le résultat.

Sur le plan purement scientifique, on peut noter que le vocoder à ligne de crête semble être le seul dont l'intelligibilité soit indépendante de la détection voisée-non voisée ; celle-ci semble se conserver en l'absence de mélodie, et pour des valeurs constantes ou quelconques. Bien entendu l'intonation et le naturel exigent la restitution correcte de la valeur de la mélodie lorsqu'elle est perceptible c'est-à-dire dans les périodes stationnaires (voyelles voisées).

Les premiers résultats sur les 600 bits/sec sont prometteurs, mais il reste encore à améliorer les critères de lissage et de tri pour ne pas éliminer trop d'éléments significatifs.