SYNTHESE DE LA PAROLE PAR POINTS-CLES : PREMIERS RESULTATS

AGNES MANTOY

Laboratoire "Image et Parole" et Laboratoire de Phonétique (D.R.L.)
Université Paris 7 - Paris

RESUME

Compte tenu de la redondance inhérente au signal de parole et de la pertinence de certains événements, il est possible de reconstituer un signal de qualité acceptable a partir d'un jeu de paramètres attaches à certains "points-clés" du signal d'origine. Ce principe est mis en oeuvre ici sur des phrases simples de français standard : les points-clés sont recherchés sur les représentations temporelle et spectrale du signal. Entre ces points, les coefficients de réflexion nécessaires à la synthèse par prédiction lineaire sont ensuite calcules par interpolation.

INTRODUCTION

La parole est dotee d'une redondance importante a quelque niveau que se situe l'analyse et en particulier au niveau acoustico-phonétique : la présence d'un phone à un instant donné influe sur la realisation acoustique des phonemes environnants.

Ainsi le signal de parole est constitué de segments stables ou quasi-stables et d'autres, transitoires, reflétant un changement plus ou moins important et plus ou moins rapide de la source sonore et/ou des articulateurs. On peut y repérer, quelquefois non sans difficultés, des discontinuites majeures comme début et fin de voisement, début et fin vocalique, début et fin de friction etc ... (cf Abry & al. [1]).

Les synthétiseurs existants exploitent la redondance du signal pour réduire le débit d'information très élevé du signal d'origine en maintenant son intelligibilité avec une bonne qualité, mais il est encore possible de réduire ce débit d'information en utilisant les propriétés acoustico-phonétiques du signal, c'est-à-dire en tenant compte de ces événements entre lesquels le signal évolue

Certains auteurs ont déjà travaillé dans ce sens, en particulier Olive & Spickenagel [4], et l'on se propose de reprendre ce travail sur des courtes phrases de français standard, de façon plus systématique et en partant de considérations plus phonétiques que techniques.

CORPUS

Il est constitue d'une quarantaine de mots de type CVCV insérés dans la phrase porteuse "C'est ---- ça". La deuxième voyelle est toujours /a/, la première étant /i/, /a/, /u/ ou /ã/. Les consonnes employees sont les suivantes : /t/, /k/, /b/, /d/, /n/, /s/, /ʃ/, /v/, /z/, /l/ ou le glide /j/. Ce corpus a été enregistré en chambre sourde et dans un ordre aléatoire par un locuteur masculin.

Les phrases ont été ensuite numérisées à 16 kHz sur 12 bits. L'analyse LPC, effectuée par tranches de 16 ms, fournit un jeu de 14 coefficients de réflexion auxquels il faut ajouter le gain et le pitch. Après lissage du pitch et du gain, on procède à une synthèse LPC qui restitue un signal "de base" reconstruit directement à partir des coefficients de réflexion d'origine

METHODE

Le traitement consiste tout d'abord a rechercher sur les représentations temporelle et spectrale du signal de base des "points-clés", c'est-à-dire des points indispensables à la reconstitution d'un signal de qualité. A ces points-clés sont associes les coefficients de réflexion de la tranche d'analyse correspondante. Entre ces tranches, les paramètres sont calculés par interpolation linéaire de l'arc sinus des coefficients de réflexion. La synthèse LPC effectuée sur ces jeux de paramètres fournit alors de nouveaux signaux "interpolés".

La qualité de la synthèse dépend bien sûr du nombre de points-clés retenus, de leur emplacement mais aussi du type d'interpolation effectuee tant sur les coefficients de réflexion que sur les paramètres prosodiques. Dans un premier temps cependant, nous avons choisi de limiter notre étude

à l'évolution des paramètres relatifs au conduit vocal indépendamment de ceux associés à la source sonore : nous avons donc conservé, pour chaque signal "interpolé", le gain et le pitch du signal de base. Par ailleurs, les travaux de Nordstrand & Öhman [3] ont montré que l'interpolation linéaire en arc sinus des coefficients de réflexion donne de meilleurs résultats que d'autres méthodes (interpolation linéaire des fonctions d'aire ou des coefficients "LAR" (Log Area Ratio), par exemple). Il va sans dire que lorsque le problème de la recherche des points-clés sans prise en compte de l'évolution des paramètres prosodiques sera résolu, ceux-ci devront être réintroduits et l' on sera amené à revoir le type d'interpolation à effectuer.

SEGMENTATION

Elle est effectuée non pas sur le signal original mais sur le signal de base reconstruit après analyse-synthèse LPC puisque, du fait de l'analyse sur des tranches de 16 ms, il peut y avoir un décalage entre la localisation d'un événement acoustique donné sur l'original et sur le signal synthétisé. Par ailleurs, on utilise le spectrogramme et l'édition du pitch et de l'intensité comme aide à la segmentation.

La segmentation consiste ici à marquer les frontières des zones stables, aucune décision n'étant prise pour les segments transitoires quant à leur appartenance à l'un ou l'autre des phones qui les entourent. Les phones discontinus tels que les occlusives sont subdivisés en deux segments : le premier (silence ou voisement) correspondant à la tenue de la consonne, le second, bruité, à son relâchement.

RECHERCHE DES POINTS-CLES

Si l'on suppose que deux points-clés par phone en moyenne sont nécessaires à la reconstitution du signal (cf Heller [2]), alors plusieurs stratégies sont possibles parmi lesquelles les deux suivantes :

(i) définir les points-clés comme les milieux des parties stables et des transitions : ce marquage indique bien les cibles à atteindre mais ne rend pas compte de la durée respective de ces parties stables et de ces parties transitoires. De plus, la décision concernant la localisation du point situé dans la transition n'est pas toujours facile à prendre.

(ii) définir les points-clés comme les extrémités des zones stables : les cibles à atteindre ainsi que la durée pendant laquelle elles sont tenues sont bien prises en compte. En outre, ces points, entre lesquels les coefficients de réflexion sont interpolés, présentent l'avantage d'être relativement sûrs. Ceci suppose que les formants des zones de transition sont des courbes continues et monotones d'une cible à l'autre, hypothèse qui devra être affinée

C'est cette deuxième méthode qui a été retenue Cependant, le résultat, à l'audition du signal et sur sa représentation graphique, n'est pas toujours satisfaisant. En effet, les courbes des formants dans les parties transitoires n'ont pas une pente constante. De plus, il semble que ces transitions doivent être interprétées plus comme des ajustements des articulateurs en jeu, avec les erreurs que cela comporte, que comme un déplacement monotone de ceux-ci d'une cible à l'autre ("overshoot"). Si tel est le cas, il faut donc déterminer les variations pertinentes dans le mouvement des articulateurs, c'est-à-dire pour nous, les variations pertinentes des coefficients de réflexion dans les zones de transition.

Pour tenter de résoudre ce problème, nous avons également sélectionné un point-clé à l'intérieur de la transition. Le nombre moyen de points-clés par phone est alors de trois et non plus de deux, ce qui rend la synthèse de bien meilleure qualité mais aussi plus coûteuse en stockage de données. Cette méthode conduit à

(i) se limiter, sur les voyelles, uniquement au segment ou la structure formantique est quasiment constante,

(ii) marquer dans la transition le début (dans le cas C-V) ou la fin (dans le cas V-C) de la structure vocalique (établissement ou relâchement de la voyelle).

Sont soumis au même traitement que les voyelles tous les sons vocaliques (présentant une structure formantique) tels que les consonnes nasales ou latérales et les glides ainsi que la consonne approximante /v/.

En fait,le nombre réel de points-clés dépend de la composition de la séquence sonore : une portion de signal ne comportant que des sons vocaliques par exemple, a toujours une structure formantique, variant dans le temps, mais ininterrompue. Il n'y a donc à retenir pour chaque phone que les deux points-clés situés aux extrémités de la partie stable. Certains de ces sons vocaliques ont une partie stable très brève voire inexistante (le glide /j/ ou la latérale /l/ par exemple) et un seul point suffit dans ce cas.

COEFFICIENTS DE REFLEXION ET POINTS-CLES

L'originalité de notre méthode consiste à nous appuyer aussi sur l'évolution temporelle des deux premiers coefficients de réflexion. Le rapport entre représentation temporelle et spectrale du signal et représentation temporelle des coefficients de réflexion nécessite une étude approfondie qui sera menée ultérieurement. Cette relation n'est pas évidente, par exemple, dans les syllabes non accentuées et entièrement voisées qui sont difficiles à segmenter et dont les coefficients de réflexion présentent des fluctuations assez déconcertantes. Toutefois, on constate que la visualisation des coefficients permet le plus souvent, en cas de doute sur deux tranches adjacentes, de sélectionner l'une d'entre elles comme point-clé

Par ailleurs, du fait du principe même de l'analyse LPC, notre méthode se heurte à un problème de résolution temporelle : deux événements successifs choisis sur la représentation temporelle du signal peuvent éventuellement se situer dans la même tranche d'analyse ou dans deux tranches adjacentes : Cela se produit à la frontière d'une consonne occlusive suivie d'une voyelle, où nous sommes toujours amenés à retenir des point-clés dans plusieurs tranches contiguës, l'un marquant la fin de la tenue de l'occlusive, l'autre son relâchement, le troisième la fin du relâchement de la consonne et le début de l'établissement de la voyelle. Il est possible aussi que deux évenements distincts soient situés dans la même tranche d'analyse et contribuent alors, ensemble, aux valeurs que prennent les coefficients. On devra, dans une étape ultérieure (segmentation automatique par exemple), tenir compte de cette difficulté.

APPLICATIONS

Les figures 1 et 2 représentent l'oscillogramme et le spectrogramme des phrases "c'est douta ça" et "c'est vida ça" Elles permettent de comparer les signaux de base synthétisés à partir des coefficients de réflexion originaux (66 jeux de paramètres) avec ceux reconstruits a partir des coefficients interpolés.

Ces signaux "interpolés" ont été resynthétises a partir respectivement de 23 et 19 points-clés Le nombre de points retenus reste donc relativement important, mais en contre-partie la qualite de ces signaux est très bonne et il est difficile de les distinguer à l'oreille des signaux de base.

CONCLUSION

Bien que le type de synthese étudie dans cette communication résulte en une compression du debit d'information, elle apparaît surtout comme un outil bien adapté à la mise en évidence de l'importance perceptuelle d'événements acoustiques dans la compréhension de la parole. On peut penser que les événements qui doivent être retenus sont le reflet de changements articulatoires essentiels lors de la production. Leur détermination nécessite une etude systématique des cas où apparaîssent les phénomènes de coarticulation, etude qui devra être menée sur la production de plusieurs locuteurs. Une analyse plus fine du comportement des coefficients de réflexion sera alors possible et permettra sans doute de distinguer les variations pertinentes de celles qui ne le sont pas et de les rapprocher des changements intervenus dans les représentations tant temporelle que spectrale du signal.

REMERCIEMENTS

Tous les traitements numériques ont été effectués sur système GenRad en TSL (Time Series Language). Nous remercions la société GenRad France et en particulier Messieurs Zeguer et Stohl pour le soutien qu'ils nous ont apporté.

REFERENCES

[1] Abry C. & al. - propositions pour la segmentation et l'étiquetage d'une base de données des sons du français - 14 ème JEP GALF 1985, p. 156-163.

[2] Heller J. - Optimized frame selection for variable rate synthesis - IEEE ICASSP 1982, p. 586-588.

[3] Nordstrand L. & Öhman S.E.G. - Computer resynthesis of speech on phonetic principles - Lund Univ. W.P. n° 19, 1980, p. 74-79.

[4] Olive J.P. & Spickenagel N. - Speech resynthesis from phoneme-related parameters - JASA vol. 59, n° 4, 1976, p. 993-996.

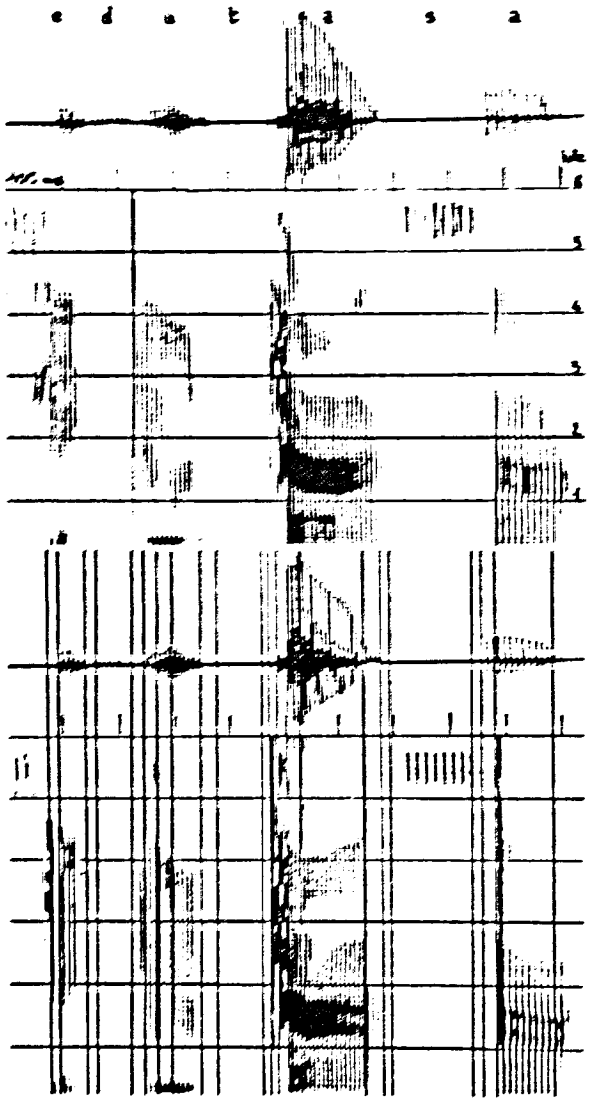

Fig 1 Oscillogramme et spectrogramme de /sedutasa/. En haut : signal de base, en bas : signal "interpolé".

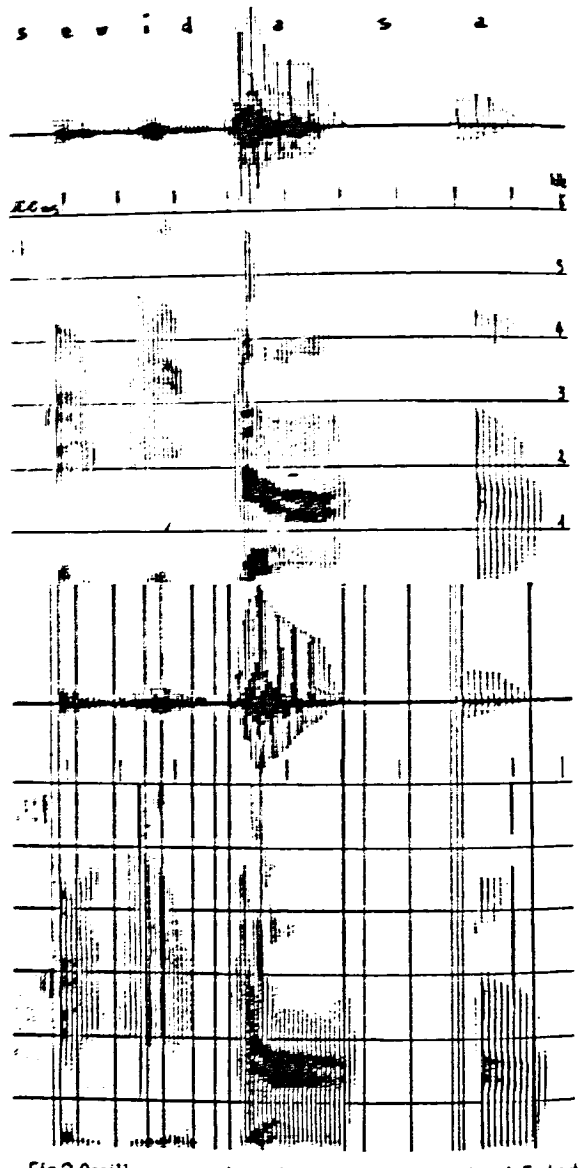

Fig 2 Oscillogramme et spectrogramme de /sevidasa/. En haut : signal de base, en bas : signal "interpolé"