



HAL
open science

APPARIEMENT DES PICS SPECTRAUX ET RÈGLES POUR LA SYNTHÈSE DE LA PAROLE PAR CONCATÉNATION DE DIPHONES

C. Laura, X. Rodet

► **To cite this version:**

C. Laura, X. Rodet. APPARIEMENT DES PICS SPECTRAUX ET RÈGLES POUR LA SYNTHÈSE DE LA PAROLE PAR CONCATÉNATION DE DIPHONES. *Journal de Physique Colloques*, 1990, 51 (C2), pp.C2-531-C2-536. 10.1051/jphyscol:19902125 . jpa-00230419

HAL Id: jpa-00230419

<https://hal.science/jpa-00230419>

Submitted on 4 Feb 2008

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

**APPARIEMENT DES PICS SPECTRAUX ET RÈGLES POUR LA SYNTHÈSE DE LA PAROLE PAR
CONCATÉNATION DE DIPHONES**

C. LAURA et X. RODET

*LAFORIA, CNRS UA 1095, Laboratoire de Traitement de la Parole,
Université Pierre et Marie Curie, 4, Place Jussieu, F-75252 Paris Cedex
05, France*

ABSTRACT:

In the framework of diphone speech synthesis, we present a method for smoothing the discontinuity at the frontier of two successive diphones. Diphones are represented as a succession of vectors of parameters corresponding to successive analysis frames. The parameters are formant or peak parameters computed on the LPC spectral envelope. The smoothing procedure is based on a matching of spectral pics between the superposed frames, each one taken from one of the two diphones to be concatenated.

11 INTRODUCTION:

Les meilleurs systèmes de synthèse par diphones sont construits sur une modélisation spectrale par prédiction linéaire ou par TFCT. Pourtant les paramètres de LPC ou de TFCT se prêtent mal aux modifications nécessaires en synthèse.

Pour mieux rendre compte des phénomènes de coarticulation, nous avons introduit une couche supplémentaire de paramétrisation du signal dans notre système d'analyse-synthèse basé sur la modélisation AR. L'enveloppe spectrale est alors caractérisée par ses maximas. Nous utilisons cette représentation en pics spectraux pour la synthèse par diphones.

La reconstruction d'une phrase, c.à.d d'une évolution spectrale continue par concaténation de diphones nécessite une procédure de lissage aux frontières des diphones. Un lissage produisant des trajets aussi continus que possible paraît souhaitable. Pour cela, il nous est donc nécessaire d'être capable d'apparier les pics d'une trame d'un diphone (censés représenter des formants) avec les pics spectraux d'une trame de l'autre diphone (censés représenter les mêmes formants).

Une technique d'appariement des pics spectraux est proposée, et des règles d'interpolations sont étudiées pour résoudre les problèmes de la discontinuité des trajets à la frontière de deux diphones successifs.

Un dictionnaire de diphones a été constitué sur la base de cette représentation pour expérimenter la méthode de synthèse.

21 PARAMETRISATION DU SIGNAL:

Notre méthode d'analyse-synthèse repose sur le modèle linéaire de production de la parole. La fonction de transfert est estimée par prédiction linéaire. Le signal échantillonné à 16 KHz est analysé toutes les 5 à 10 ms (en synchronie avec la fréquence fondamentale) avec un filtre de prédiction d'ordre 30. La source d'excitation est caractérisée par la fréquence fondamentale, et le degré harmonique par bande de fréquences. Cette modélisation harmonique-bruitée de la source permet une synthèse de meilleure qualité que la méthode classique de synthèse par prédiction linéaire. Pour reconstruire le signal, nous utilisons un filtre linéaire en échelle. On obtient ainsi une parole de bonne qualité /6/.

Dans le cadre de la synthèse par diphones les paramètres d'analyse doivent être interpolés entre les diphones pour éviter les discontinuités aux frontières. Dans ce cas une paramétrisation de type LPC paraît difficile à maîtriser notamment pour contrôler les phénomènes de coarticulation. Une approche plus favorable consiste à

choisir une description des enveloppes spectrales privilégiant les formes caractéristiques au niveau perceptif. Des études ont montré d'une part l'importance de la contribution des maximas d'énergie, d'autre part la faible contribution des vallées spectrales sur la perception /3/.

L'enveloppe spectrale est estimée par prédiction linéaire. Les maximas détectés sur l'enveloppe par une simple procédure de peak-picking fournissent une couche de la paramétrisation du signal en termes de pics définis en fréquence, en amplitude et en largeur de bande. Inversement, nous pouvons calculer l'enveloppe à partir de sa description en pics. Les paramètres du filtre de synthèse sont alors obtenus par estimation tout pôle de cette enveloppe, en augmentant légèrement l'ordre du filtre /7,4/.

3) CONCATENATION DE DIPHONES ET LISSAGE AUX FRONTIÈRES:

La technique de synthèse par concaténation de diphones consiste à reconstruire une phrase par concaténation des diphones codés en paramètres, généralement des coefficients de réflexion (K_i) issus d'une analyse LPC. Mais cette concaténation introduit en général des discontinuités spectrales aux frontières des diphones. Ces discontinuités sont parfois atténuées par un lissage sur les Log Area Ratio (LAR) correspondant aux K_i .

Un lissage sur les trajets de formants semble mieux adapté. De plus il permet alors d'envisager des règles sur les valeurs des paramètres de formants conformément aux connaissances phonétiques et phonologiques qui en général sont connues en termes de valeurs de fréquence ou amplitude des formants. Mais la représentation de la parole en trajets de formants présente de grandes difficultés. On observe souvent des trajets qui se rejoignent ou qui disparaissent, rendant peu utilisable la numérotation traditionnelle des trajets de formants. Les méthodes habituelles d'extraction automatique des trajets /2,8/ donnent les trois premiers formants, mais l'existence même de trois trajets bien définis est sujette à caution en particulier pour certaines consonnes. De plus, un nombre aussi limité de pics ne permet pas de reconstruire des enveloppes suffisamment naturelles pour donner des paramètres corrects au filtre de synthèse.

Notre solution à ces difficultés est d'effectuer un lissage respectant la continuité des "lignes de crête" sans chercher à les simplifier en trajets individuels numérotés. Soit TU et UW deux diphones à concaténer. Pour respecter la continuité des "lignes de crête", il suffit d'être capable de mettre en correspondance les pics d'une trame de la partie droite de TU avec les pics d'une trame de la partie gauche de UW. Nous appellerons "appariement des pics spectraux" cette mise en correspondance.

4) APPARIEMENT DES PICS SPECTRAUX :

Nous utilisons donc, une technique d'appariement introduite dans le cadre de la reconnaissance pour construire une mesure de distorsion spectrale entre deux enveloppes à partir de leur description en pics. Le principe consiste à définir une distance interpic et à appairer les pics les plus proches au sens de cette distance par comparaison dynamique des deux enveloppes /1/.

Nous avons établi une distance interpic $d(i,j)$ qui rend compte des déformations en fréquence, en amplitude et en largeur de bande de deux pics i et j . Les écarts en fréquence sont calculés sur une échelle de type MEL.

$$d(i, j) = a_1 * E_f(f_i, f_j) + a_2 * E_a(a_i, a_j) + a_3 * E_l(l_i, l_j)$$

où a_1, a_2, a_3 sont des coefficients de pondération qui permettent de normaliser la quantification sur l'ensemble des paramètres et de fixer leurs contributions respectives en fonction des seuils différentiels perceptifs, et où E_f, E_a, E_l sont les distances construites pour

f_i, f_j les fréquences centrales,
 a_i, a_j les amplitudes et
 l_i, l_j les largeurs de bande.

Un algorithme de programmation dynamique permet d'apparier les pics d'une trame à l'autre. Soient I et J le nombre de pics des 2 trames, l'algorithme de programmation dynamique utilise une fonction de coût de passage au point i, j de la forme :

$$D(i, j) = \min \begin{cases} D(i-1, j-1) + 2d(i, j) \\ D(i-1, j) + d(i, j) \\ D(i, j-1) + d(i, j) \end{cases}$$

$D(I, J)$ est alors une mesure de dissimilarité entre les deux trames. Cependant, on rencontre des pics qui manifestement n'ont pas de correspondant: c'est le cas lors des apparitions ou des disparitions de trajets. Ceci est détecté automatiquement sur la distance interpic lorsqu'elle dépasse un seuil fixé expérimentalement.

51 APPLICATION A LA RECONSTRUCTION DES ENVELOPPES:

Remarquons que ce type de mesure de dissimilarité nous fournit également un critère objectif pour évaluer le calcul des coefficients du filtre de synthèse à partir des pics. Nous l'avons expérimenté sur des phrases naturelles en effectuant trois types de mesures:

- 1 mesure de la distance, entre les pics de l'enveloppe originale et les pics de l'enveloppe reconstruite.
- 2 mesure de la distance entre deux enveloppes originales consécutives.
- 3 mesure de la distance entre deux enveloppes reconstruites consécutives.

En moyenne (fig 1) on montre que :

a) Pour 76% des enveloppes, la distance dans le cas 1 reste inférieure à 5 et s'élève en moyenne à 3.3. Ces chiffres sont relativement faibles comparés aux mesures relevées dans le cas 2 qui fournissent une distance moyenne de 10.7 et une distance comprise entre 5 et 20 pour 86% des enveloppes. On peut en déduire que nos enveloppes ont été "bien" reconstruites.

b) Une comparaison des calculs effectués pour chacune des trames d'un signal analysé dans le cas 2 et dans le cas 3 nous montre qu'en général sur l'ensemble des trames, les distances sont du même ordre pour les enveloppes originales et pour les enveloppes reconstruites. Ce résultat est confirmé globalement puisque 82% des enveloppes reconstruites fournissent une distance trame à trame comprise entre 5 et 20 avec une moyenne de 11.2.

61 APPLICATION A LA SYNTHÈSE PAR DIPHONES:

Pour construire une base de données de dipphones, nous avons enregistré un corpus de phrases porteuses. Chaque diphone est inclut au milieu d'un logatome pour limiter les effets de coarticulation. Ces logatomes ont été prononcés dans la phrase porteuse

"c'est" logatome "ça"

afin de réduire les effets de liste pour un locuteur, et permettre la localisation automatique des logatomes. L'extraction des logatomes est effectuée par un algorithme qui utilise le taux de passage par zéro et l'énergie du signal. Les logatomes entiers sont analysés et un dictionnaire de dipphones est constitué avec les paramètres de l'analyse /9/.

Un diphone TU est représenté par 4 instants (fig2):

tu_deb : début du diphone dans la partie stable du phonème T
 tu_sup : début de la transition TU
 tu_inf : fin de la transition TU
 tu_fin : fin du diphone dans la partie stable du phonème U

Le calcul des paramètres de synthèse est fondé sur la technique d'appariement des pics afin de lisser les "lignes de crêtes" entre deux diphones TU et UW. Les trames tu_inf et uw_sup délimitent le phonème U et permettent de fixer 3 instants sur la durée de ce phonème: la frontière tf autour de laquelle les diphones TU et UW sont superposés et les limites t1 et t2 du segment à interpoler pour reconstruire le phonème U /7,5/.

Les diphones TU et UW sont superposés autour de la frontière tf. Une fonction de pondération r(t) variant de 1 à 0, permet alors de passer progressivement de TU à UW en fonction des décisions d'appariement; Chaque mise en correspondance de deux pics i1 et i2 pour les trames superposées de TU et de UW à un instant t, détermine un nouveau pic pour la trame résultante. Chaque paramètre de ce pic k est calculé en fonction de la contribution relative des trames superposées. Considérons le paramètre f, la fréquence par exemple, à l'instant t, f_{k1}(t) est la valeur du paramètre dans le diphone TU. La valeur du paramètre du pic k dans la trame résultante est alors :

$$f_k(t) = (1 - r(t)) f_{k1}(t) + r(t) f_{k2}(t)$$

Le contexte du phonème U doit être pris en compte au cours du lissage. Pour cela, les deux trames aux limites du phonème, tu_inf et uw_sup sont appariées. Un segment linéaire f'(t) est calculé de t1 à t2 pour chaque pic en correspondance.

Le trajet final est obtenu en fonction du degré d'articulation dg, retenu pour la transition TUW, comme:

$$f_k(t) = dg \cdot f_k(t) + (1-dg) \cdot f'_k(t)$$

71 CONCLUSION :

Une comparaison (fig3) des trajets calculés et des trajets obtenus par une simple concaténation de diphones, montre que le lissage a permis de résoudre en général le problème des discontinuités apparentes à la frontière des diphones concaténés. En particulier, nous obtenons des trajets continus y compris pour des transitions très coarticulées pour lesquelles le contexte a pu être pris en compte.

Par ailleurs, notre système d'analyse-synthèse étant fondé sur une paramétrisation du signal de type LPC, le signal synthétique est finalement reconstitué à partir des paramètres du filtre de synthèse, déduits des pics ainsi calculés. A l'écoute, la synthèse obtenue par lissage "des lignes de crête" gagne en naturel par rapport à la synthèse par concaténation brute des diphones.

Une étude en cours devrait améliorer encore la synthèse dans les cas où les décisions d'appariements restent insuffisantes pour contrôler la coarticulation sur toute la durée du phonème.

En effet, un apprentissage des règles de synthèse sur des phrases naturelles, permettra d'adapter ces décisions en fonction des contextes et de lever les ambiguïtés qui demeurent.

REFERENCE :

/1/, MJ.CARATY, X.RODET Distance interspectrale à critères perceptifs, 14ème JEP PARIS
 Etude comparative de mesures de distorsion spectrale, 15ème JEP AIX

- /2/, MC.CANDLESS ,An alorithm for automatic formant extraction using linear prediction spectra,IEEE ASSP april 74.
- /3/, R.CARLSON,B.GRANSTROM,The representation of speech in the peripheral auditory system ,Elvesier Biomedical Press,Amsterdam 1982.
- /4/, T.GALAS,calcul des coefficients LPC à partir des pics spectraux,Rapport Interne LAFORIA
- /5/, X.RODET,P.DEPALLE,G.POIROT, Analysis and synthesis methods based on spectral envelopes and voiced/unvoiced functions ,ECST,edinburg U.K ,sept 87
- /6/, X.RODET,P.DEPALLE,G.POIROT, Analyse et synthèse de voix parlées et chantées 16ème JEP HAMMAMET
- /7/, X.RODET,P.DEPALLE,G.POIROT,Diphone sound synthesis based on spectral envelopes and harmonic/noise excitation functions,ICMC 88.
- /8/[,G.KOPEC, A family of formant trackers based on hidden markov models, ICASSP 86 TOKYO.
- /9/, STELLA, synthèse de la parole, écho des recherches 84.

	or / rec	or / or	rec / rec
0,0 à 2,0	43,70%	1,80%	0,90%
2,0 à 5,0	32,30%	9,20%	10,00%
5,0 à 9,0	17,10%	28,70%	25,20%
9,0 à 12,0	3,30%	27,00%	24,30%
12,0 à 20,0	2,70%	30,50%	32,10%
> 20,0	0,90%	2,80%	7,50%
moyenne	3,3	10,7	11,2

Fig. 1. - Répartition et moyenne des mesures des distances dans les 3 cas.

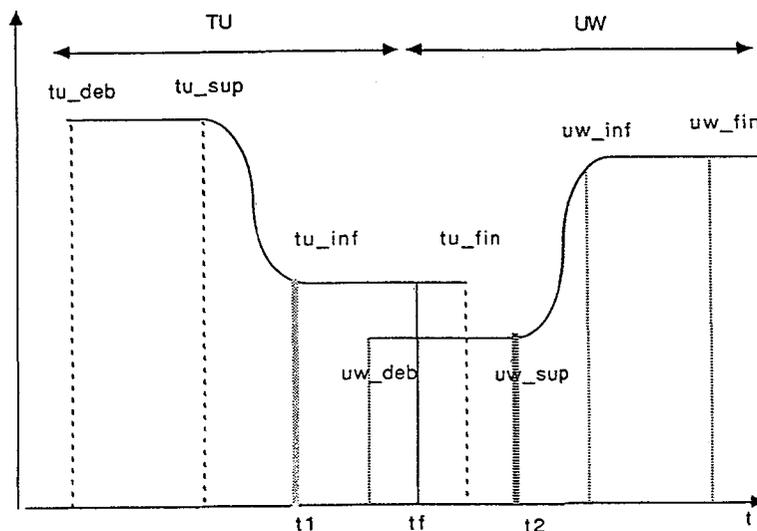


Fig. 2. - Superposition de deux diphones TU et UW du dictionnaire.

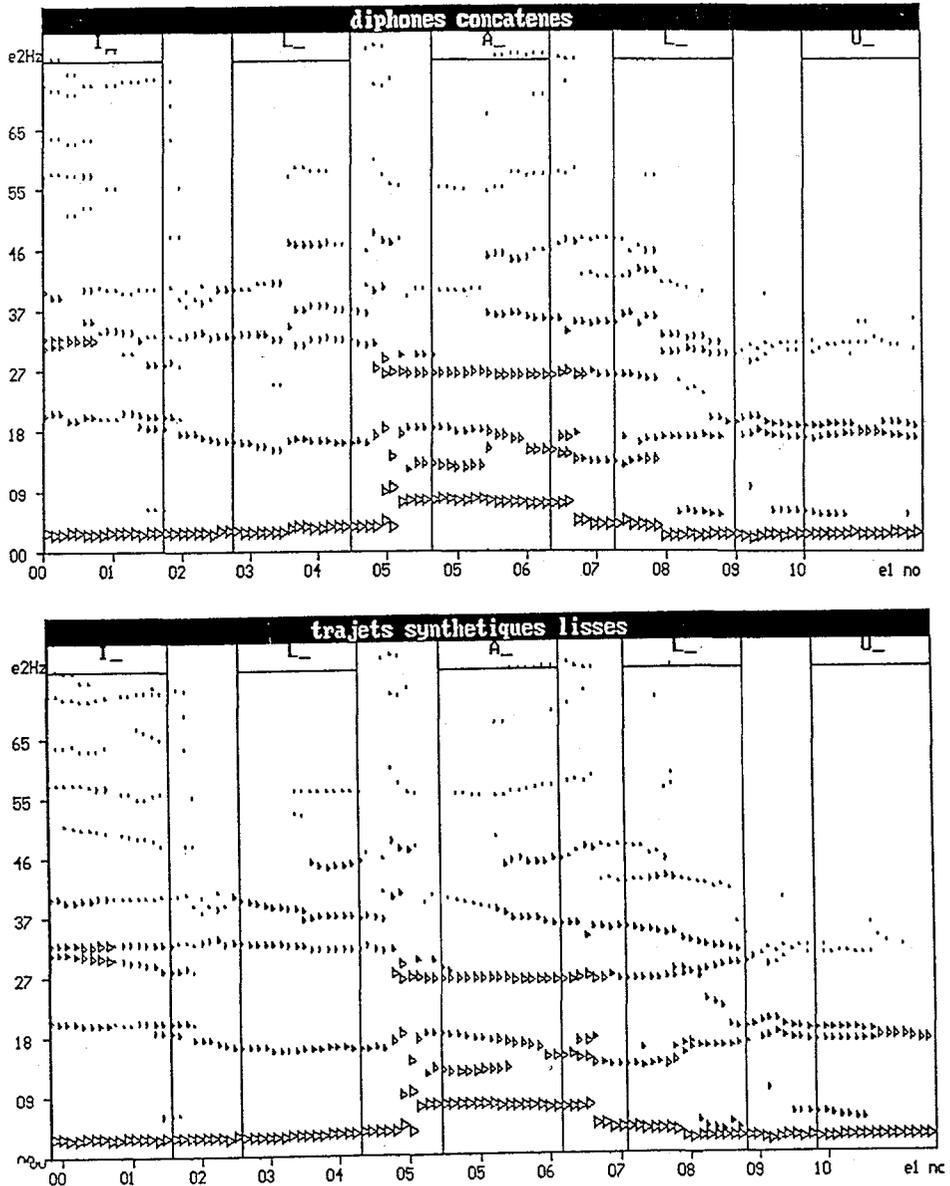


Fig 3 - Représentation des trajets concaténés et des trajets lissés pour la phrase "IL A LU".
Le lissage permet d'éliminer les discontinuités à la frontières des 2 phonèmes "L" aux instants (03,04) et (07,09).