Conventions de segmentation pour la construction de diphones

LAIP - Lettres, Université de Lausanne
Version 1.0, 31.8.98
Contribuants:
Sandra Schwab, Eric Keller, Brigitte Zellner, Pierre-Yves Connan, Beat Siebenhaar

Note: Afin d'éviter des problèmes d'impression, ce document utilise l'alphabet phonétique à 7 bits du LAIP (document séparé)

A. Objectif

L'objectif de ces conventions est de spécifier les critères de segmentation des signaux de la parole, afin d'assurer une cohérence optimale pour la génération automatique de diphones à partir de cette segmentation.

B. Conventions générales

Définition d'un diphone. En général, un diphone s'étend du milieu d'un premier segment au milieu d'un second, afin de capter les éléments essentiels d'une transition phonétique. Un exception à ce principe est faite pour les occlusives où nous calculons le début du diphone à partir du *début du "burst"* de plosion. Par conséquent un diphone débutant par une occlusive (p.ex., le diphone [ps]) commence au début du "burst", tandis qu'un diphone terminant sur une occlusive (p.ex., [sp]) achève sur la fin de la période (plus ou moins) silencieuse.

Placement des étiquettes. Nous délimitons les deux segments participant au diphone en plaçant une étiquette *après* chaque segment concerné. Par conséquent, trois étiquettes seront placées pour un diphone donné (Figure 1):

1) Une étiquette au point de transition entre le segment précédant la transition visée et le premier segment de la transition en question.

2) Une étiquette au point de transition entre le premier et le deuxième segment de la transition visée.

3) Une étiquette au point de transition entre le deuxième segment de la transition visée et le segment suivant la transition en question.

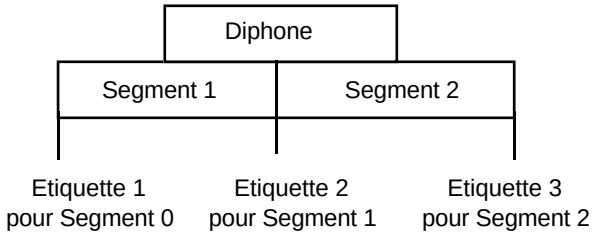


Figure 1. Placement des étiquettes

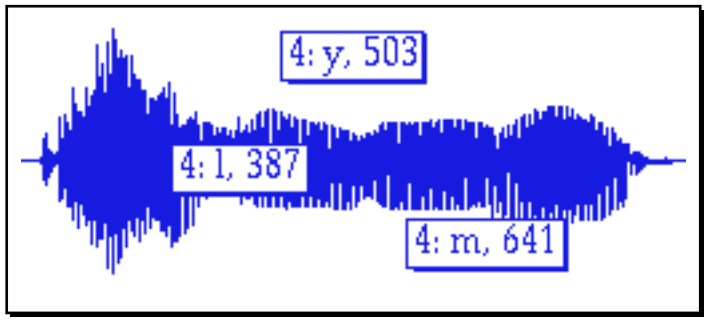


Figure 2. La séquence [y-m] dans « allumette »

Le cas particulier des occlusives. Lorsqu'une occlusive sourde ou sonore fait partie de la transition ou qu'elle précède la transition, on place une étiquette supplémentaire pour marquer la fin du silence des occlusives sourdes [#] et la fin du "silence voisé" (du pré-voisement) des occlusives sonores [V]. Comme indiqué, ceci est nécessaire pour la construction subséquente des diphones (Figure 2).

Il arrive cependant que la barre de prévoisement d'une occlusive sonore [V] soit remplacée par le pré-voisement d'une autre consonne, p., dans "membre" [m@mbR%]. Dans ce cas, l'étiquetage montrera la suite [@] [m] [b], et non pas la suite [@] [m] [V] [b] ou [@] [V] [b] (Figure 3).

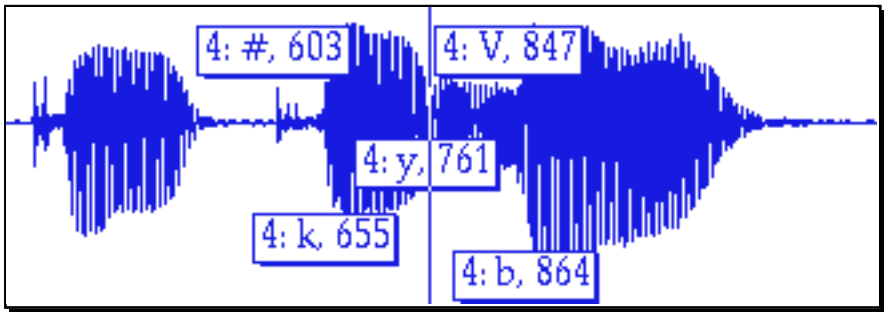


Figure 3. La séquence [kyb] du mot "concubin" est étiquettée [#kyVb], où l'étiquette [#] marque la partie silencieuse de l'occlusive [k] et l'étiquette [V] marque la partie sonore et préplosive de l'occlusive [b].

C. Voyelles

Début et fin de voyelles (Figure 4). Pour l'identification des débuts et fins de voyelles, la segmentation s'effectue généralement en fonction des 2ème et 3ème formants (particulièrement si la voyelle est en fin d'énoncé). Ceci tient compte de la sensibilité maximale de l'oreille entre 1000 et 2000 Hz, ainsi que des diminutions de cette sensibilité de chaque côté de cette plage de fréquences (diminution rapide en dessous de 1000 Hz et diminution moins rapide au-dessus de 2000 Hz).

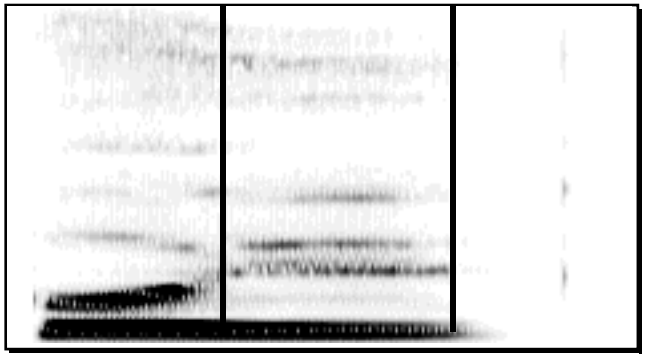
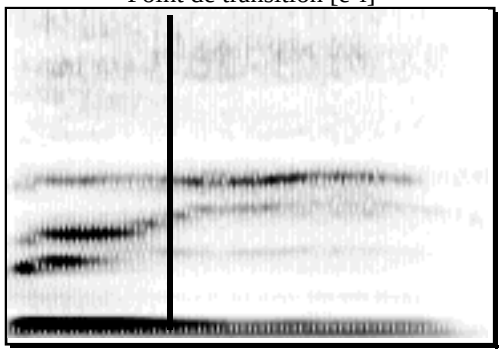


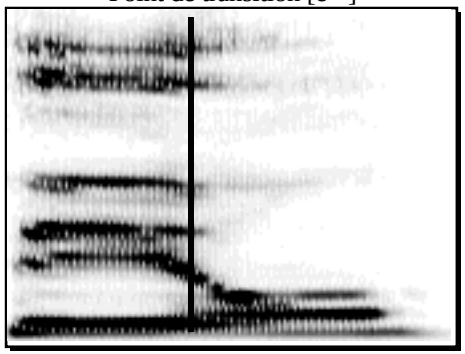
Figure 4. Segmentation de [y] dans « cohue ».

Deux voyelles voisines. Dans le cas de deux voyelles voisines ayant des structures formantiques distinctes, la segmentation s'effectue au *milieu de la transition*, au point de la pente la plus forte (Figure 5).

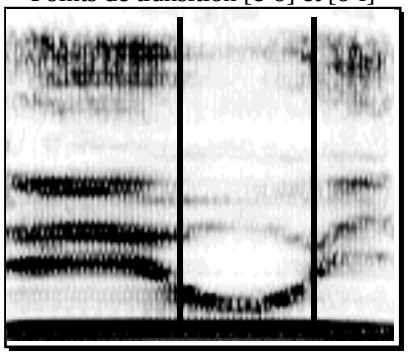
Point de transition [e-i]



Point de transition [e-^]



Points de transition [e-o] et [o-i]



Point de transition [e-@]

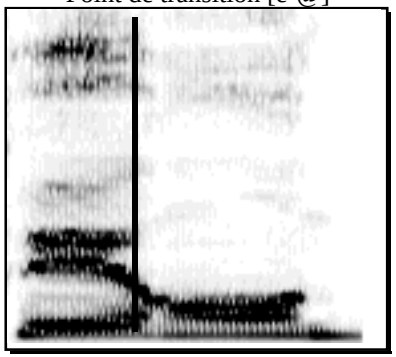


Figure 5. Quelques exemples de segmentation dans les voyelles.

Deux voyelles très semblables ou identiques. Dans le cas de deux voyelles très semblables ou identiques, une différentiation entre les deux voyelles est parfois possible, soit au niveau de la fréquence fondamentale, soit au niveau de l'amplitude globale du signal. Dans ce cas, choisir le point de transition en fonction de ces indicateurs. Dans tous les autres cas, la segmentation d'effectue au milieu arithmétique entre les frontières latérales des deux voyelles (Figure 6).

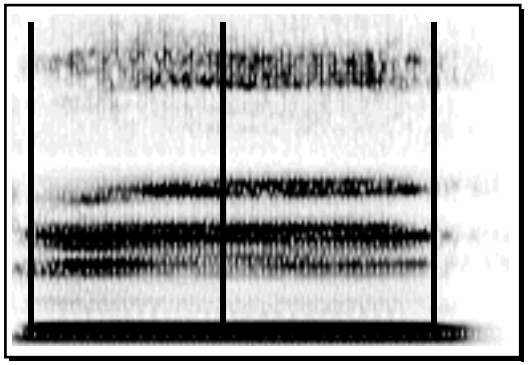


Figure 6. La transition [e-e] dans «créer»

D. Semi-voyelles

Les semi-voyelles peuvent être délimitées généralement entre deux états pseudo-stables. Dans le cas d'une voyelle suivie d'une semi-voyelle, la segmentation se fait *après et avant la partie pseudo-stable* de la voyelle (et non au milieu de la transition entre la voyelle et la semi-voyelle). Ceci assure une saisie satisfaisante de la semi-voyelle (Figure 7).

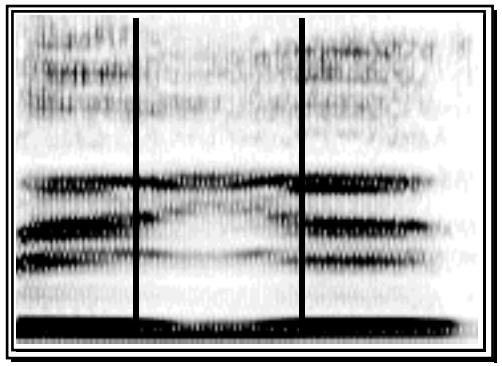


Figure 7. Segmentation de la semi-voyelle [j] dans « impayé ».

E. Consonnes

Les occlusives sourdes, indépendamment de leur emplacement dans l'énoncé, sont décrites en deux étapes, afin de permettre la resynthèse satisfaisante à partir de ces deux étapes. La première étape se compose d'un silence non voisé, représenté par [#]. La deuxième étape se compose de l'effet plosif (du "burst") associé l'articulation de la consonne en question (soit *p*, *t*, *k*) (Figure 8).

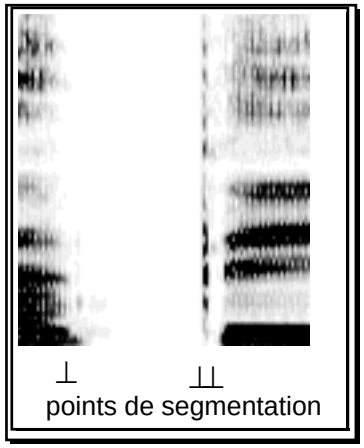


Figure 8. Les trois points de segmentation de l'occlusive sourde [p] dans « impayé ». La transcription montrera donc [5#p].

Les occlusives sonores, indépendamment de leur emplacement dans l'énoncé sont décrites en deux étapes. La première étape se compose d'un "silence voisé" (c.-à-d., d'une barre de pré-voisement), représenté par [V] (Figure 9). La deuxième étape se compose de l'effet acoustique de la plosion associée l'articulation de la consonne en question (soit *b*, *d*, *g)*. Une plosion mineure, ou un "burst" mineur, est souvent visible sous forme d'une barre de plosion.

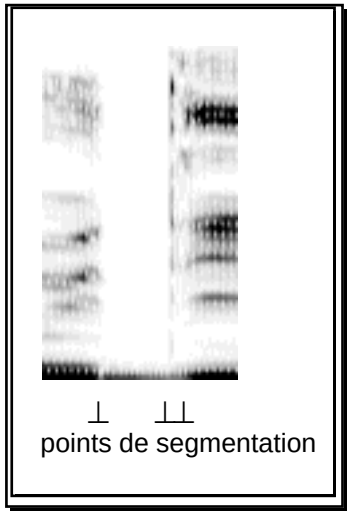


Figure 9. Les trois points de segmentation de l'occlusive sonore [d] dans « érudition ». La transcription montrera donc [yVd].

Remarque: Le pré-voisement peut ne pas être présent, par exemple dans le cas d'une consonne nasale suivie d'une consonne occlusive sonore.

Latérales. En règle générale, les consonnes latérales ne posent pas de problèmes particuliers de segmentation, étant donné leur représentation spectrographique caractéristique (Figure 10).

Segmentation de [R] dans « érudition »

Segmentation de [l] dans « accumulateur »

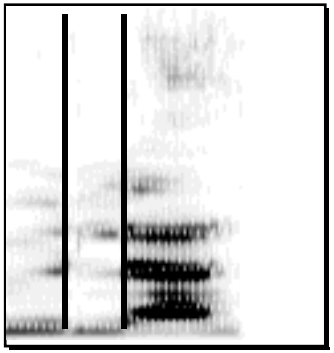


Figure 10. Exemples de segmentation de latérales.

Néanmoins, si un silence apparaît avant ou après la consonne, la segmentation se fait au milieu du silence (Figure 11).

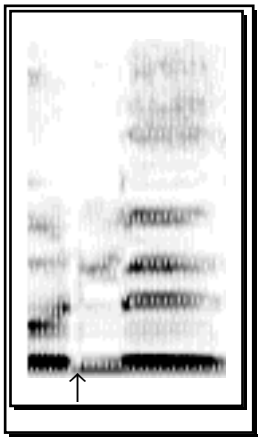


Figure 11. Illustration d'une latérale avec un petit silence séparant la latérale du segment précédent.

Fricatives. En élocution lente, un petit silence apparaît avant et/ou après la consonne; la segmentation s'effectue alors au milieu du silence (Figure 12).

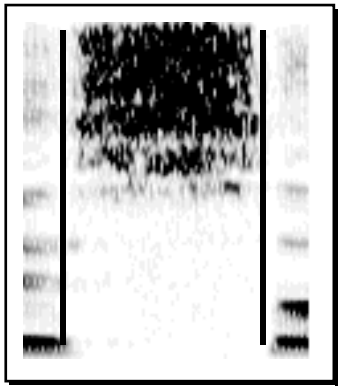


Figure 12. Segmentation de [s] dans « adolescent »

Si aucun silence n'apparaît, la segmentation s'appuit sur les indices de bruit (Figure 13).

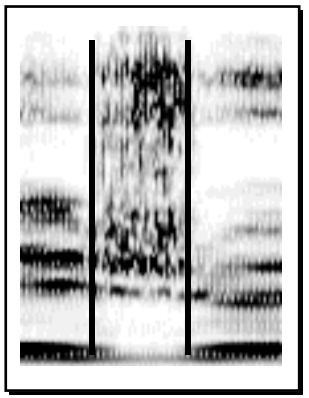


Figure 13. Segmentation de [Z] dans « déjeuner »

Nasales. En règle générale les consonnes nasales ne posent pas de problèmes de segmentation étant donné leur représentation spectrographique caractéristique (Figure 14).

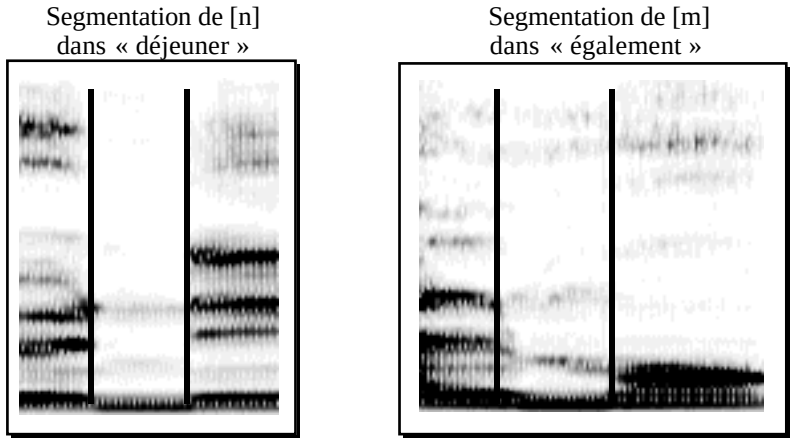


Figure 14. Exemples de segmentation des nasales.

F. Silences et bruits secondaires

• Si la transition donnée concerne une voyelle suivie d'un silence [#] (généralement en fin d'énoncé), l'étiquette du silence se place à **200 ms** de l'étiquette de la voyelle précédente (en y incluant les "bruits" d'aspiration ou déglutition, s'il y en a). Si le signal débute moins que 200 ms avant le début de la parole, on place une étiquette initiale aussi proche du début du signal que possible.

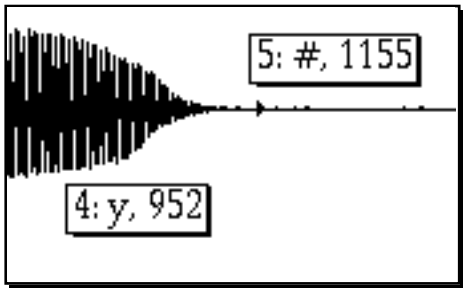


Figure 15. Etiquetage du silence à la fin d'un énoncé.

• Si la transition donnée concerne un silence suivi d'une voyelle (généralement en début d'énoncé), l'étiquette du silence se place à **200 msec** avant le début de l'énoncé (en y incluant les "bruits" d'aspiration ou déglutition, s'il y en a) (Figure 15). Si le signal achève moins que 200 ms après la fin de la parole, on place une étiquette initiale aussi proche de la fin du signal que possible.

• Dans le cas d'un silence "impur", c'est-à-dire un silence comprenant des bruits d'aspiration ou de déglutition, il est considéré comme un silence "normal". *Remarque:* les coups de glotte sont marqués comme segment (symbole [q]) seulement s'ils sont linguistiquement pertinents (p. "hibou" [qibu], où le [q] est linguistiquement pertinent, car il empêche la liaison [leqibu]) (Figure 16).

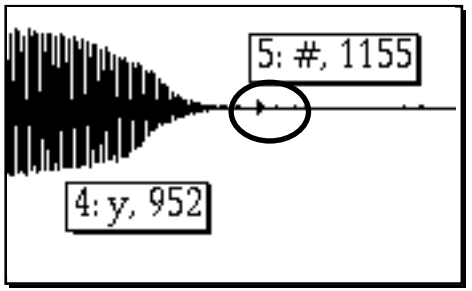


Figure 16. Illustration d'un silence bruité.