Chapitre 1

Traitement et analyse du signal de parole

1.1. Introduction

Alors que Daniel Jones — le phonéticien britannique dont les travaux réalisés pendant la première moitié du 20ème siècle ont exercé une influence majeure — se préparait un jour à partir pour une étude sur le terrain, quelqu'un lui demanda quels instruments il allait emporter avec lui. Jones lui désigna ses oreilles et répondit : « seulement ces instruments-là »[1]. L'oreille, et l'œil, devrait-on rajouter, constituent en effet les outils de base du phonéticien et du phonologue, et c'est par leur truchement que de nombreuses langues ont d'abord été décrites dans leur forme sonore. Si l'on s'accorde à penser qu'un locuteur parle pour être entendu, et qu'il est donc nécessaire pour le phonéticien de s'assurer que les phénomènes qu'il analyse dans le signal de parole sont détectables par le système perceptif, le recours à l'oreille et à l'œil se montre inévitable. Les études expérimentales menées depuis des décennies sur le traitement de la parole chez l'humain peuvent d'ailleurs et d'une certaine manière être considérées comme une façon de pousser cette analyse perceptive jusqu'au bout, dans la mesure où celle-ci est alors pratiquée dans des conditions hautement contrôlées, par un certain nombre d'auditeurs ne connaissant pas les objectifs de l'expérience, et de telle sorte que les résultats obtenus soient généralisables à une population plus large de sujets. Mais l'analyse perceptive présente des limites, liées entre autres choses à son imprécision et au fait qu'elle ne nous offre pas un accès direct aux caractéristiques acoustiques du signal de parole. Elle est ainsi et depuis longtemps utilisée de pair avec différentes méthodes instrumentales, dont l'objectif de ce chapitre sera de présenter un aperçu.

Chapitre rédigé par Christine MEUNIER et Noël NGUYEN.

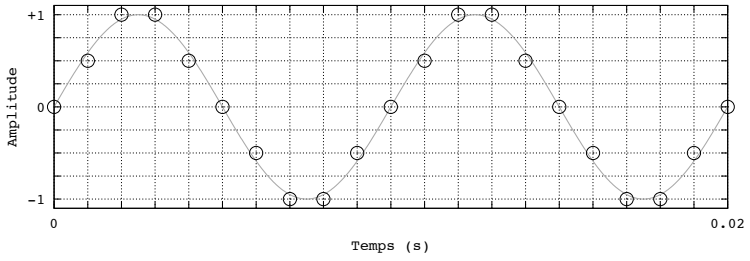
1. Cité dans [LAD 97].

Pour la plupart d'entre elles, les analyses pratiquées sur la parole font intervenir une opération clé que l'on appelle l'étiquetage. Cette opération consiste à établir une correspondance entre le signal de parole, d'une part, et une suite d'unités dont l'inventaire est pré-établi et qui sont relatives à un ou plusieurs niveaux d'analyse, d'autre part. Le plus souvent, on entreprend également de déterminer le domaine temporel associé à chaque unité à l'intérieur du signal de parole, et cette seconde opération est désignée sous le terme de segmentation. Les analyses elles-mêmes peuvent schématiquement se ranger en trois grandes catégories, selon la nature des caractéristiques acoustiques étudiées, qui se rapporteront à la forme spectrale du signal, à la fréquence fondamentale, ou bien encore à la durée et aux paramètres qui lui sont associés. Après quelques rappels sur les procédures d'acquisition du signal de parole, nous aborderons ainsi les questions relatives à la segmentation et à l'étiquetage multi-niveaux.

1.2. Acquisition du signal de parole

Il existe une très volumineuse littérature, destinée au grand public comme aux spécialistes, sur les instruments et les techniques utilisables pour l'enregistrement des sons, et nous n'aborderons donc pas ces aspects-là dans ce chapitre, en renvoyant le lecteur à [LAD 97, TAR 03, TES 01] pour des synthèses centrées sur l'acquisition des données de parole. Les quelques rappels que nous commencerons par faire ne concerneront que le dernier élément de cette chaîne d'acquisition. Pour pouvoir être stocké et être traité sur ordinateur — comme cela est le cas dans toutes les études dont nous traitons ici — le signal de parole doit être soumis à une procédure que l'on appelle la numérisation, ou la digitalisation. Cela consiste à convertir la grandeur continue constituée par ce signal en une suite de valeurs numériques (les échantillons) représentant l'amplitude du signal à intervalles réguliers. Cette procédure est illustrée sur la figure 1.1.

Figure 1.1. *Représentation schématique de la numérisation d'un signal périodique continu.*



Dans cet exemple, le signal d'origine est un son périodique simple se présentant sous la forme d'une sinusoïde et les échantillons sont représentés par des cercles. L'intervalle temporel entre deux échantillons successifs est désigné sous le terme de période d'échantillonnage, et il est ici de 0.001 seconde (soit 1 milliseconde). L'inverse de la période d'échantillonnage, que l'on appelle la fréquence d'échantillonnage, et qui se mesure en Hertz, représente le nombre d'échantillons recueillis par seconde, et sa valeur est ici de 1000 Hz. Chaque échantillon renvoie à une valeur numérique entière encodée sur un certain nombre de bits. Ce nombre détermine ce que l'on appelle le niveau de quantification. Il a été fixé à 3 dans notre illustration. L'un de ces bits symbolise le signe algébrique relatif à chaque valeur (+ ou $-$) et les deux autres bits permettent de représenter l'amplitude du signal sous la forme de $2^2 = 4$ niveaux régulièrement espacés au-dessus et au-dessous de 0.

Tout se passe donc comme si les valeurs associées aux différents échantillons prenaient place sur une grille, représentée en pointillés sur la figure 1.1, et dont l'espacement entre les points dépend *a)* sur l'axe temporel, de la fréquence d'échantillonnage, et *b)* sur l'axe d'amplitude, du niveau de quantification. Les points dont la grille est formée sont en nombre fini et ils permettent de ne saisir qu'imparfaitement les variations présentées par le signal d'origine, comme le montrent les écarts qui s'établissent entre les échantillons et ce signal. La précision avec laquelle le signal sera caractérisé obéit donc directement à deux paramètres, la fréquence d'échantillonnage et le niveau de quantification.

Dans de nombreuses études acoustiques sur le signal de parole, la fréquence d'échantillonnage employée est de 16000 Hz. Cela offre la possibilité d'analyser des sons dont la composante la plus aiguë est de 8000 Hz. Cette fréquence se calcule en divisant la fréquence d'échantillonnage par deux, et elle est désignée sous le terme générique de fréquence de Nyquist. Les consonnes fricatives telles que /s/ et /ʃ/ par exemple, peuvent se caractériser par la présence d'une énergie acoustique importante entre 6000 et 8000 Hz. Les capacités offertes par les ordinateurs d'aujourd'hui en matière de stockage des données nous permettent d'utiliser des fréquences d'échantillonnage plus élevées (44100 Hz par exemple, avec une fréquence de Nyquist qui sera donc portée à 22050 Hz, même s'il est encore une fois estimé que les composantes susceptibles de véhiculer de l'information pour l'auditeur dans le signal de parole, ont une fréquence qui ne dépasse guère 8000 Hz).

Le niveau de quantification le plus couramment utilisé est de 16 bits. Cela signifie que l'amplitude relative à chaque échantillon est représentée par une valeur entière comprise entre -32768 (-2^{15}) et +32768 ($+2^{15}$). Lorsqu'il est mesuré en décibels (dB), l'écart entre la valeur minimale et la valeur maximale, c'est-à-dire la dynamique du signal numérisé, est dans un tel cas de 96 dB (chaque bit représentant +6 dB). Cet écart est supérieur à la dynamique réelle du signal de parole, si l'on estime que celle-ci se situe entre 40 et 60 dB [HAR 99], et un niveau de quantification de 16 bits peut

donc être considéré comme suffisant dans les analyses acoustiques pratiquées sur la parole.

1.3. Segmentation et étiquetage

Comme nous l'avons indiqué en introduction, l'opération d'étiquetage consiste à projeter sur le signal de parole une séquence d'étiquettes ou de labels se rattachant à un ou plusieurs niveaux d'analyse. L'exemple le plus traditionnel est celui d'une séquence de phonèmes que l'on fera correspondre au signal par le biais de différents critères, mais les étiquettes peuvent se situer à des niveaux extrêmement variés, depuis les propriétés acoustiques associées aux différentes phases d'un geste articulatoire par exemple, jusqu'au tour de parole dans un échange conversationnel. Les étiquettes appartiennent à un inventaire pré-établi (tel qu'une liste de phonèmes ou de mots), mais des modifications peuvent leur être apportées dans le décours des analyses, par exemple lorsqu'un élément fait surface dans les données analysées qui ne figure pas dans l'inventaire de départ. En règle générale, l'étiquetage s'accompagne d'une segmentation du signal de parole, dont l'objectif est d'associer à chaque étiquette un intervalle temporel dans ce signal. La segmentation présente deux caractéristiques importantes. En premier lieu, et au sein d'un certain niveau d'analyse, les intervalles relatifs à différentes étiquettes peuvent se chevaucher de manière partielle ou totale sur l'axe temporel. Dans la séquence /ba/ pour prendre l'exemple le plus simple, il est possible de considérer que le début de la voyelle coïncide avec celui du bruit d'explosion de la consonne précédente. En second lieu, il est à notre connaissance toujours considéré qu'un segment, quelle que soit sa taille, débute en un certain point dans le temps et se termine en un autre point, et une hypothèse plus complexe consistant à supposer que le début (ou la fin) d'un segment puisse lui-même s'étaler à l'intérieur d'une certaine fenêtre temporelle ne semble pas avoir été jamais explorée. Les segments correspondent donc chacun à un fragment du signal de parole, et ils sont en ce sens-là des éléments discrets. Cela confère à la segmentation un caractère partiellement arbitraire, dans la mesure où les frontières propres à chaque segment sont souvent intrinsèquement floues, comme peuvent le montrer les séquences de voyelles.

Si la segmentation et l'étiquetage peuvent être accomplis à la main, au moyen d'un éditeur de signal tel que Praat[2] ou Wavesurfer[3], il existe à présent des outils nous permettant de procéder à ces deux opérations de manière automatique ou semi-automatique. Au premier rang de ces outils figurent les aligneurs aujourd'hui disponibles pour le français, l'anglais, et d'autres langues. Un aligneur offre la possibilité de découper un signal de parole en mots, et à l'intérieur de chaque mot en une suite de segments associés chacun à un phonème, à partir d'une transcription orthographique

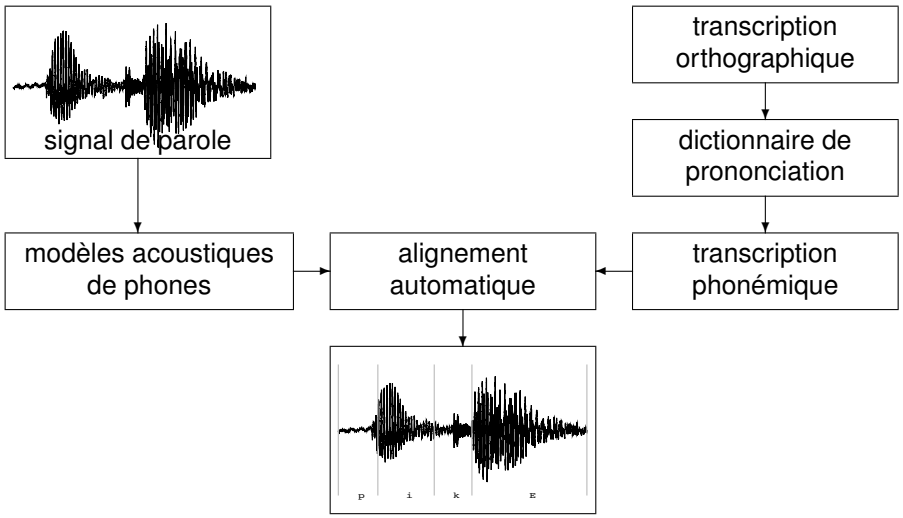
2. `www.praat.org`

3. `www.speech.kth.se/wavesurfer`

antérieurement établie de l'énoncé prononcé. Tous les aligneurs ne font pas appel à la même méthode [MAL 03], mais la figure 1.2[4] représente celle qui est la plus fréquemment utilisée [BÜR 08].

Comme le montre la figure 1.2, la transcription orthographique est d'abord convertie en une transcription phonémique par l'intermédiaire d'un dictionnaire de prononciation, dans lequel figurent pour chaque mot une ou plusieurs séquences de phonèmes représentant les variantes rencontrées dans la prononciation de ce mot. En parallèle, le signal de parole est traité par un système de reconnaissance automatique de la parole destiné à identifier dans le signal une suite de « phones » définis indépendamment du contexte (et qui sont donc assimilables à des phonèmes) ou bien en fonction du contexte (en s'apparentant alors aux différentes variantes allophoniques relatives à chaque phonème). Par le biais de ce système de reconnaissance, un alignement temporel est alors accompli entre le signal de parole et la transcription phonémique de l'énoncé.

Figure 1.2. *Principaux éléments d'un aligneur basé sur un système de reconnaissance automatique de la parole.*



Dans un traitement automatique de ce type-là, le niveau phonémique constitue donc le niveau de référence, et c'est par l'intermédiaire de ce premier découpage que

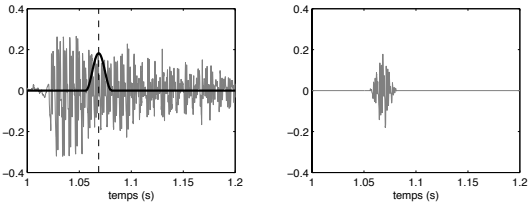
4. Sur cette figure, l'exemple présenté est le mot *piquet*, prononcé par le locuteur LD dans l'enquête PFC faite à Douzens. L'alignement automatique a été réalisé par Martine Adda-Decker.

des analyses seront ensuite réalisées sur des unités de niveau inférieur ou supérieur. Dans la section suivante, nous entreprendrons de brièvement passer en revue les critères utilisables pour délimiter les segments associés à des phonèmes dans le signal de parole. Comme tout espèce de découpage, celui-ci est assujetti à un certain nombre de présupposés théoriques et il renvoie à une approche segmentale et linéaire de la forme sonore du langage qui a été remise en question avec l'avènement de la phonologie post-générative (voir par ex. [DUR 90]). Il est important de ne jamais perdre de vue le fait que les outils actuels de segmentation automatique de la parole se présentent à leur manière comme l'implémentation de ces présupposés théoriques déjà anciens, tout en essayant de tirer le meilleur parti de ce que ces outils ont à nous offrir.

1.4. Analyse spectrale

Nous commencerons par quelques rappels d'ordre pratique dont l'objectif sera de présenter la manière dont les paramètres d'une analyse spectrale sont généralement établis ainsi que les principaux problèmes rencontrés dans les mesures faites sur un spectre[5]. Comme cela est illustré sur la figure 1.3, une analyse spectrale s'effectue à l'intérieur d'une certaine fenêtre temporelle, sur la portion de signal délimitée par elle. L'emplacement de la fenêtre peut être déterminé de manière automatique ou manuelle, en positionnant un curseur à l'endroit voulu dans l'éditeur de signal utilisé.

Figure 1.3. *À gauche : onde sonore pour le mot* pâte*, locuteur JP, enquête PFC/Douzens ; ligne verticale en pointillés : curseur placé dans la voyelle* /ɑ/ *; courbe noire : fenêtre de Hamming de 25 ms centrée sur le curseur. À droite : portion de signal délimitée par la fenêtre.*



La fenêtre apparaissant ici se présente sous la forme d'une courbe en cloche. Cela entraîne une distorsion du signal liée à une atténuation progressive de l'amplitude vers les bords de la fenêtre, mais cette distorsion est rendue nécessaire par l'impact que la fenêtre possède en soi sur la forme du spectre obtenu. En effet, comme la fenêtre et le signal sont multipliés l'un par l'autre, l'analyse spectrale est affectée par les caractéristiques de la fenêtre tout autant que par celles du signal lui-même. Par opposition

5. Le lecteur est renvoyé à [JOH 02] et à [COL 05] pour des introductions non-techniques à l'analyse spectrale de la parole, et à [BOI 00] pour une présentation de niveau avancé.

à une fenêtre en cloche, une fenêtre carrée (c'est-à-dire dont la valeur est de 1 sur toute la durée de l'intervalle temporel analysé, et de 0 partout ailleurs) ne provoque pas de distorsion dans la structure temporelle du fragment de signal correspondant, mais ses deux bords latéraux s'assimilent à des discontinuités qui vont introduire dans le spectre des composantes de haute fréquence [HAR 99]. Il est ainsi généralement recommandé d'utiliser une fenêtre non-carrée, tout en sachant que sur le plan spectral, celle-ci va minimiser la contribution des éléments d'information situés au début et à la fin de l'intervalle d'analyse. Les outils les plus répandus offrent un choix entre des fenêtres de différentes formes (carrée, Hamming, Hanning, gaussienne, etc.), parmi lesquelles la fenêtre de Hamming est peut-être la plus populaire.

La durée de la fenêtre doit s'accorder bien sûr avec celle des phénomènes analysés, et elle sera courte ($\approx$ 20 ms) voire très courte ($\approx$ 5 ms) pour des événements acoustiques brefs, tels que le bruit d'explosion d'une occlusive, ou bien plus longue (entre 50 et 100 ms) pour des événements plus stables dans le temps, tels que la partie centrale d'une voyelle. Une relation importante existe entre la durée de la fenêtre d'analyse et la précision du spectre en fréquence : plus la fenêtre est courte, moins cette précision est bonne. En d'autres mots, la résolution fréquentielle est inversement proportionnelle à la résolution temporelle. Un compromis est donc à établir entre la précision désirée dans le domaine fréquentiel, d'une part, et dans le domaine temporel, d'autre part. Un exemple simple nous permettra d'introduire les quelques règles de calcul à connaître pour prendre les décisions requises. Une fenêtre d'analyse de 32 ms appliquée à un signal échantillonné à 16000 Hz comprendra 512 échantillons (16000×0.032)[6]. Pour des raisons sortant du cadre de ces quelques rappels, le spectre obtenu à partir de cette portion de signal par l'intermédiaire d'une transformée de Fourier comportera 257 composantes ($(512/2) + 1$), espacées à intervalles égaux entre 0 Hz et la fréquence de Nyquist, 8000 Hz dans le cas présent. L'écart en fréquence entre deux composantes spectrales adjacentes sera alors d'environ 31 Hz ($8000/256$). Cet écart, qui représente la résolution du spectre en fréquence, peut se calculer plus directement en divisant la fréquence d'échantillonnage par le nombre d'échantillons contenus dans la fenêtre d'analyse. À fréquence d'échantillonnage égale, il ressort ainsi que la résolution en fréquence devient de plus en plus fine lorsque la fenêtre d'analyse s'allonge.

1.5. Segmenter la parole : aspects méthodologiques d'une expertise manuelle

1.5.1. *Définition et objectifs*

Il convient, en premier lieu, de faire une mise au point terminologique. Les termes *segmentation*, *alignement* et *annotation* sont souvent employés pour désigner un même

6. À strictement parler, il y aura $(16000 \times 0.032) + 1 = 513$ échantillons, mais nous ferons abstraction de ce détail ici.

résultat : le marquage temporel des segments phonétiques positionnés sur le signal de parole. Si le résultat est celui-ci, les processus mis en œuvre sont différents pour des raisons à la fois historiques et disciplinaires. La *segmentation manuelle* désigne le travail d'expertise humain consistant à marquer les frontières de segments phonétiques par des étiquettes de début et de fin. Le terme *alignement* est issu du traitement automatique et consiste, par le biais de modèles acoustiques, à faire correspondre une suite de symboles phonétiques (issue d'une transcription) avec le signal de parole. Enfin, l'*annotation phonétique* fait référence aux résultats soit de la segmentation, soit de l'alignement. Ce terme est utilisé en référence à d'autres niveaux d'annotation comme les niveaux morpho-syntaxique, prosodique, discursif, voire gestuel pour le signal vidéo. Il s'agit donc d'un terme générique dont le niveau phonétique ne forme qu'une partie du domaine d'application. Nous utiliserons donc le terme *segmentation* pour désigner le travail d'un expert humain. *Alignement* correspondra au résultat d'un traitement automatique et *correction manuelle de l'alignement* au travail, désormais fréquent des experts, de déplacement des étiquettes posées automatiquement. Enfin, nous utiliserons éventuellement *annotation* pour indiquer, de façon générale, le fait que le signal de parole est étiqueté phonétiquement.

Segmenter le signal revient à apposer sur le signal de parole des étiquettes temporelles représentant approximativement le début ou la fin d'une unité linguistique. C'est évidemment dans le terme « approximativement » que se trouve tout le débat sur la segmentation de la parole, tout particulièrement en ce qui concerne les unités phonétiques. Il est désormais clair que toutes les composantes de la réalisation d'un segment phonétique ne peuvent être contenues entre deux frontières quelle que soient leur position. L'ensemble de nos connaissances sur la coarticulation (voir par ex. [FAR 97]) et sur la propagation des traits phonétiques sur de longues séquences [HAW 03] ont montré qu'il est illusoire de vouloir représenter les segments phonétiques comme une suite d'éléments contenus entre deux frontières. S'il est indéniable que la production des phonèmes est ordonnée dans le temps, les multiples articulateurs en jeu et leurs mécanismes complexes de synchronisation et d'anticipation suggèrent que la sortie acoustique observable comporte de multiples indices répartis au-delà de la manifestation saillante du phonème sur le signal de parole. Le signal acoustique n'est alors qu'un reflet, incomplet, des mécanismes articulatoires sous-jacents. Cette relation complexe entre niveaux linguistique, articulatoire et acoustique nous rappelle que les frontières de segments ne doivent pas être confondues avec des frontières de phonèmes [FAN 73]. En effet, la notion abstraite de « phonème » ne peut être mise en relation directe avec des discontinuités du signal acoustique [ROS 90]. D'une certaine façon, l'alignement automatique force cette relation complexe entre unités abstraites (les phonèmes) et variations continues du signal de parole dans la mesure où les annotations sont composées d'une suite de début et de fin de segments phonétiques. Dans les années 1980, un rapprochement entre phonéticiens et spécialistes du traitement de la parole a permis d'aborder le problème différemment. La complexité de la relation unité abstraite / signal continu a conduit certains chercheurs à évacuer les problèmes

posés par la segmentation en proposant un étiquetage au centre des segments accompagné d'une description des divers événements acoustiques qui interviennent entre deux étiquettes centrales [ABR 85b, ABR 85a]. Cette solution élégante permet à la fois de localiser le phonème, mais aussi de répertorier l'ensemble des événements acoustiques qui sont, eux, évidemment plus identifiables sur le signal de parole et qui sont porteurs d'information pour les analyses. Une autre solution [AUT 85] consiste à effectuer un étiquetage hiérarchisé dans lequel des étiquettes sont déclinées selon trois niveaux : les macro-classes (voyelle orale, consonne occlusive, consonne approximante, etc.), les Phases (établissement, tenue et relâchement), les Attributs (vocalicité, consonanticité, nasalité, souffle, closion, etc.). Les segments phonétiques sont ainsi étiquetés entre des frontières de macro-classe et sont composés de phases et d'attributs (pour une revue de ces procédures, voir Meunier, [MEU 94]). Notons que ces débats ont animé les années 1980 et ont permis de mettre en évidence l'intérêt de marquer et de coder les événements acoustiques présents sur le signal indépendamment ou en parallèle avec les étiquettes ou frontières de phonèmes[7].

Quoi qu'il en soit, chacun s'accorde désormais pour admettre que les étiquettes de début et de fin des unités phonétiques, qu'elles soient apposées automatiquement ou manuellement, ne représentent pas des frontières réelles de phonèmes. Il s'agit juste de repères permettant de procéder à des analyses sur le signal de parole. Car la segmentation n'est pas un but en soi, mais plutôt un outil qui doit donc répondre à des questions préalables.

Pourquoi segmente-t-on le signal de parole ? Une première réponse, très générique sans doute, est que l'on segmente pour maîtriser l'aspect continu de la parole. Dès lors que l'on étiquette ou que l'on segmente le signal de parole, on se donne les moyens d'y avoir accès et d'en faire, par conséquent, un objet d'analyse exploitable. La question n'est donc pas de « bien » localiser les phonèmes, donc de trouver les « bonnes » frontières, mais de savoir si les critères que l'on applique vont nous permettre de répondre aux questions scientifiques posées.

La plupart du temps, la segmentation sert à délimiter des segments au travers desquels des analyses acoustiques seront menées. Les étiquettes de début et de fin vont alors permettre d'extraire des durées, des valeurs formantiques, la fréquence fondamentale, l'énergie, etc., ces mesures étant prises sur la totalité ou sur un point précis du segment.

Enfin, la tâche de segmentation pourra dépendre de l'unité que l'on souhaite étiqueter. Notamment, l'exigence de précision ne sera pas la même s'il s'agit d'étiqueter

7. L'alignement automatique est revenu, de fait, à un marquage des unités phonétiques en « début-fin » sans identification des événements acoustiques. Une solution d'avenir pourrait ainsi proposer un alignement automatique des unités phonétiques, complété par une identification des événements acoustiques.

des phrases ou bien s'il est question d'étiqueter les différentes phases de réalisation d'une occlusive. Toutefois, la taille de l'unité à segmenter ne résout pas forcément la difficulté de la tâche. Par exemple, une séquence /je/ sera aussi difficile à segmenter, qu'elle se trouve en interne (/pje/ « *pied* ») ou en frontière de mot (/pagajɛnɔʁm/ « *pagaille énorme* »). Une segmentation en phrase pourra ainsi être très délicate si les phrases ne sont pas séparées par des pauses. Il n'en reste pas moins que la précision peut s'avérer moins importante pour de larges unités dans la mesure où cette imprécision aura de moindres répercussions sur les analyses d'unités plus longues.

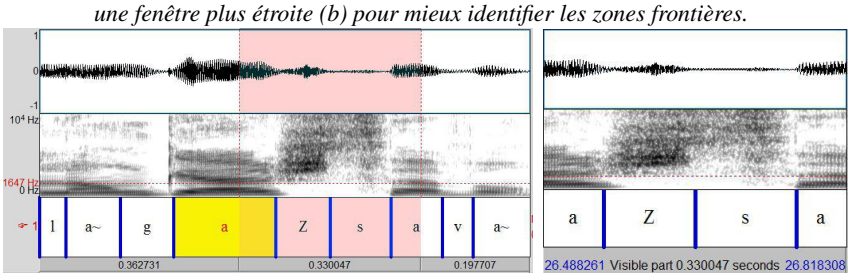
1.5.2. *Environnement de travail pour la segmentation manuelle*

Le plus couramment, deux types de représentation sont disponibles pour la segmentation manuelle. Une représentation temporelle (intensité/temps), l'oscillogramme, et une représentation fréquentielle (fréquence/intensité/temps), le spectrogramme (figure 1.4). Ces deux représentations sont nécessaires et offrent des informations complémentaires. La représentation temporelle sera de préférence choisie pour la pose précise d'une étiquette en raison de la grande précision du déroulement des événements acoustiques (en particulier si l'on choisit de positionner une étiquette à un passage par zéro). Elle permet en outre un meilleur repérage des discontinuités de l'amplitude du signal qui apparaissent plus diffuses sur le spectrogramme. La représentation fréquentielle offre en revanche une vision globale plus claire des discontinuités fréquentielles et donc des changements articulatoires sous-jacents. Cette représentation facilite l'interprétation du signal de parole.

L'identification des discontinuités se fera en deux phases. La première consiste à observer le signal de parole sur une fenêtre assez large (figure 1.4a) de façon à identifier clairement les unités produites (macro-classes) et les transitions entre unités (phases et attributs). La deuxième phase consiste à se focaliser (en zoomant) sur les zones de transition spécifiques (figure 1.4b) de façon à identifier les événements pertinents qui permettront de poser une étiquette de frontières.

L'environnement idéal pour un expert en segmentation n'est pas aisé à trouver et n'est pas disponible sous Praat, comme sous la plupart des outils qui permettent de marquer des événements acoustiques. La segmentation manuelle nécessite une vision large du signal de parole *parallèlement* à une vision étroite. Une vision large car il est indispensable de voir ce qui a été produit globalement. Les discontinuités spectrales et les changements d'intensité sont plus aisément repérables sur une vision élargie du signal et peuvent devenir invisibles lorsque la portion du signal est zoomée. La vision étroite est nécessaire pour une pose précise d'une étiquette de frontière, mais également pour identifier des discontinuités très ténues. Ces deux visions sont donc indispensables et le travail de localisation des événements pertinents nécessite des allers et venues permanents entre ces représentations large et étroite. La configuration

Figure 1.4. *Exemple d'environnement de travail pour l'annotation du signal de parole sur Praat [BOE 01]. De haut en bas : le signal temporel de parole, le spectrogramme et la Tier d'annotation. Après un repérage global des segments phonétiques à partir d'une fenêtre large (a), l'expert peut zoomer sur une fenêtre plus étroite (b) pour mieux identifier les zones frontières.*



a) séquence « langage savant » (extrait de phrases lues) annotée manuellement avec sélection sur la zone /aʒsa/.

b) zoom sur la sélection /aʒsa/, dévoisement du segment /ʒ/.

idéale serait donc que ces deux représentations (fenêtres) soient *synchronisées* (c'est-à-dire que les curseurs soient synchronisés) de façon à pouvoir localiser rapidement des événements fins sur la fenêtre large et inversement.

L'apport de la perception auditive dans la tâche de segmentation du signal est assez paradoxal. D'une certaine façon, la perception est indispensable car elle aide au repérage en macro-classes et permet de savoir ce qui a réellement été produit. En effet, certaines réalisations articulatoires particulières (*cricky voice*, réalisations approximantes, dévoisements, etc.) ont besoin d'être repérées auditivement de façon à être ensuite interprétées correctement lors du repérage des macro-classes. De même, lors de phases de transition très graduelles entre deux segments (dans /wa/ par exemple), l'écoute progressive de portions de signal (comme dans une expérience de *gating*, par exemple) permet de s'assurer de la présence ou de l'absence d'un phonème, ce qui évite de grosses erreurs. En revanche, l'oreille n'est pas d'une grande aide pour identifier des séquences ambiguës ou pour caractériser une transition d'une très courte durée. Nous avons en effet besoin d'un minimum de durée pour interpréter correctement le signal. De ce fait, il vaut mieux ne pas sélectionner des portions très courtes de signal (inférieures à 50 ms) pour une vérification auditive. Le plus raisonnable, lorsqu'une séquence courte pose problème, est de se positionner sur cette zone et d'écouter de longues portions à droite, puis à gauche de façon à interpréter correctement la séquence. Enfin, notre système de perception est conçu pour *interpréter* le signal, c'est-à-dire sélectionner les informations pertinentes pour construire un message linguistique. Cette sélection nous rend sourd à des discontinuités non pertinentes mais toutefois présentes sur le signal. Finalement, l'oreille est l'outil fondamental de la transcription et, dans une certaine mesure, du repérage des macro-classes. Dans une tâche plus fine où il est question de manipuler des portions de parole très courtes ou

ambiguës (identification de phases ou d'événement acoustiques très courts), l'oreille peut, en revanche, nous induire en erreur.

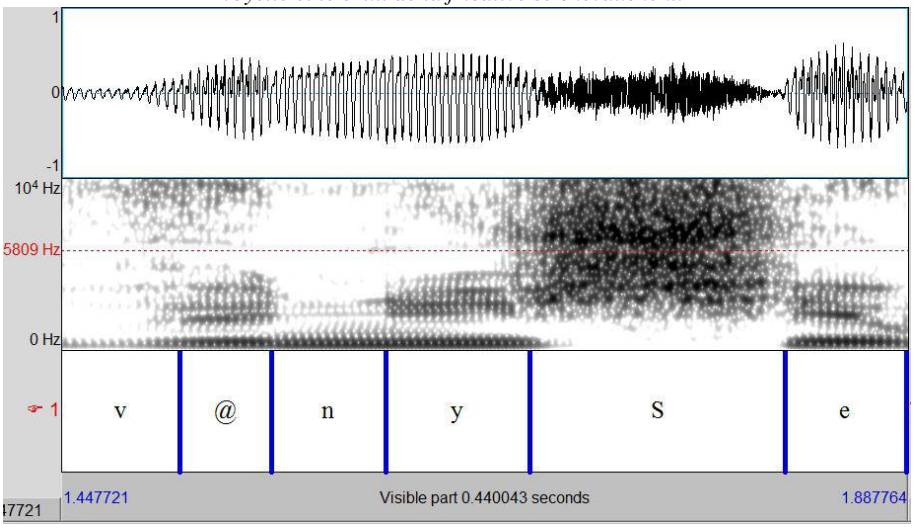
1.5.3. *Discontinuités du signal et indices pertinents*

Le signal de parole n'est pas continu et les discontinuités y sont présentes partout. Toutefois, ces discontinuités ne sont pas systématiquement interprétables comme des indices pertinents pour marquer la frontière d'un segment. Le travail d'un expert consiste alors à choisir parmi les discontinuités présentes, celles qui sont pertinentes pour l'identification d'une frontière. Inversement, le passage d'un segment au suivant n'est pas nécessairement marqué par une discontinuité identifiable. Les indices acoustiques marquant la distinction des phonèmes sont-ils, à ce titre, d'égale pertinence ? Fant [FAN 73] remarque que les changements de mode de production (articulation et phonation) semblent plus appropriés pour marquer les frontières de segments, tandis que les modifications de lieu d'articulation entraînent des changements plus graduels pouvant se manifester par des modifications sur plusieurs segments. Cette remarque ne peut être interprétée sans nuance dans la pratique de la segmentation. En effet, les phénomènes d'assimilation de mode d'articulation ou, le plus souvent, de phonation, peuvent rendre ces indices inexploitables et ce sont alors les indices de lieu d'articulation qui s'avèrent fort utiles pour identifier les frontières (figure 1.4b). Il faut toutefois hiérarchiser les indices pertinents pour le processus de segmentation, cette hiérarchie pouvant être adaptée en fonction de situations. On aurait ainsi par ordre descendant de pertinence pour la segmentation : le mode d'articulation, le mode de phonation et le mode d'articulation. D'une façon générale, s'il est indispensable de bien connaître les traits acoustiques distinctifs de chaque phonème, il faut considérer que cette description est théorique et que les réalisations effectives dévient de cette description en raison des phénomènes d'assimilation et de coarticulation. Il est donc nécessaire de bien observer ce qui est effectivement réalisé et de bien connaître les phénomènes de variation les plus courants.

Il faut également noter que ces trois types d'indices ne présentent pas les mêmes caractéristiques de zone frontière. Par exemple, le début et la fin du voisement sont assez simples à déterminer. En effet, même si l'atténuation de la périodicité se fait graduellement, la première ou la dernière période sont assez simples à identifier. De même, l'apparition ou la disparition de bruit, de silence ou de structure formantique (caractéristiques, respectivement, des fricatives, des occlusives et des segments vocaliques) ne posent généralement pas de problèmes majeurs. En revanche, les discontinuités peuvent s'étendre dans le temps pour des changements de lieu d'articulation. C'est le cas surtout pour des suites de fricatives (/sf/ dans « *asphalte* » par exemple) ou des suites de segments vocaliques (/rwa/ dans « *rois* » ou /lj/ dans « *lien* »), autrement dit, des séquences pour lesquelles le mode de production n'est pas distinctif. Dans ces cas, le passage d'un segment au suivant est graduel et aucun autre indice (mode d'articulation ou de phonation) ne peut aider à la segmentation.

Pour résumer, on notera deux types de discontinuité : 1) des discontinuités brèves *souvent* caractéristiques du mode de production (articulation et phonation), 2) des transitions graduelles *souvent* attribuées au lieu d'articulation[8]. Dans ces deux cas, la segmentation peut être difficile. Dans le premier cas, même si les discontinuités sont brèves, les indices pertinents peuvent se chevaucher. C'est le cas, par exemple, dans une séquence fricative + voyelle (figure 1.5) dans laquelle la structure formantique de la voyelle commence alors que le bruit de la fricative n'est pas terminé. L'expert pourra alors choisir de privilégier un indice (le bruit ou les formants) pour déterminer la frontière[9] ou bien de décider que la frontière se situe au centre de la partie de chevauchement. Dans le cas d'une transition graduelle, l'aspect continu du passage d'un segment à l'autre rend souvent impossible une localisation temporelle précise de la frontière. La solution la moins risquée semble alors être de positionner la frontière au centre de la zone transitoire. Machac & Skarnitzl [MAC 09] proposent ainsi une hiérarchisation des indices pour en déduire des critères de segmentation.

Figure 1.5. *« venue chez moi » (extrait de texte lu) ; la partie finale de la voyelle et le bruit de la fricative se chevauchent.*



8. Nous précisons *plus souvent* dans la mesure où des cas contraires peuvent évidemment être trouvés (discontinuité brèves pour le lieu d'articulation et transition graduelle pour le mode de production).

9. Cette solution présente l'inconvénient de ne pas être adaptable à toutes les situations. Le bruit d'une fricative peut s'étendre sur toute la durée d'une voyelle et une fricative peut être réalisée comme une approximante et donc présenter une structure formantique (voir § 1.6.2).

1.5.4. *Alignement automatique et/ou segmentation manuelle ?*

Evidemment, la question de l'alignement automatique (AA) et de la segmentation manuelle (SM) ne se pose pas réellement en terme de choix. Les analyses phonétiques portant sur de très grands corpus se basent désormais le plus souvent sur les résultats de l'alignement automatique, dans la mesure où il n'est pas envisageable d'annoter manuellement plusieurs centaines de milliers de phonèmes. De ce fait, l'expertise manuelle se raréfie sur ce type de corpus et la question est désormais de savoir si l'AA nous fournit un balisage du signal de parole suffisamment fiable pour y effectuer des analyses phonétiques de qualité [MAC 09]. Une question subsidiaire serait ensuite de savoir quels types d'analyse phonétique sont envisageables avec l'AA et lesquels ne le sont éventuellement pas. Pour répondre à ces questions, les résultats de l'AA ont été corrigés par des experts humains (correction manuelle de l'alignement, CMA) de façon à faire apparaître les avantages et inconvénients respectifs des deux approches.

L'avantage en termes de coût et de rapidité revient évidemment à l'AA qui permet l'annotation phonétique de très grand corpus oraux, ce qui est quasiment impossible à envisager pour la segmentation manuelle. Concernant la qualité et la régularité de l'annotation produite, les avis sont partagés. On considère souvent que l'expertise manuelle est plus précise et qu'elle permet une certaine adaptabilité en fonction des productions réelles des locuteurs. Toutefois, certains travaux [WES 96, PIT 05] ont pu mettre en évidence une forte variabilité inter-experts. Ces résultats ne surprendront pas si l'on considère que les critères de segmentation utilisés par les experts sont le résultat de choix concernant les indices acoustiques identifiés dans le signal de parole. En revanche, lorsque les experts s'accordent sur les choix des critères de segmentation, les écarts sont assez faibles (entre 1 et 3 ms en moyenne ; Volin et al., 2008, cité par [MAC 09]). Ainsi, l'écart inter-expert ne serait pas forcément dû à une localisation approximative des indices acoustiques du signal de parole (même si cela reste possible), mais plutôt à des choix différents concernant l'affectation des indices acoustiques pertinents pour l'identification de zones frontières.

Une comparaison entre AA et CMA a pu être effectuée au cours de plusieurs travaux en français portant sur de larges corpus de parole peu prototypique ou altérée (parole conversationnelle[10], d'une part, et parole dysarthrique[11], d'autre part). Notons toutefois que la segmentation manuelle, dans ce cas, correspond à une correction de l'alignement (déplacement, insertion ou retrait des étiquettes de l'AA) et non à un processus de segmentation, ce qui est assez différent (voir dernière partie). Dans les deux cas, des décalages ont pu être observés entre les étiquettes de l'AA et celles issues de la correction de l'alignement. Notamment, les travaux sur la parole conversationnelle [BER 08] font apparaître une sous-estimation systématique des segments

10. Dans le cadre du projet ANR OTIM (resp. : Philippe Blache).

11. Dans le cadre du projet ANR DesPhoAPaDy (resp. : Cécile Fougeron).

vocaliques pour l'AA se traduisant par un marquage plus tardif du début des voyelles ainsi qu'un marquage plus précoce de la fin. En utilisant un autre aligneur, les travaux portant sur la parole dysarthrique ont observé le même type de décalage entre AA et SM [AUD 10]. Dans les deux cas, les études ont pu montrer que cela n'avait pas d'incidence sur les analyses phonétiques des voyelles [BER 08, FOU 10]. De même, il apparaît que, globalement, les différences de position des étiquettes entre AA et SM sont assez régulières. En conséquence, les analyses portant sur les variations de durée des segments ne semblent pas affectées par le type d'annotation.

À travers ces résultats, il est possible d'entrevoir la complémentarité entre AA et SM en vue d'une exploitation pour des analyses phonétiques. Lorsque les analyses phonétiques portent sur l'ensemble des segments et notamment sur leur durée globale ou sur leur tenue segment (extraction de formants au centre d'une voyelle ou centre de gravité du bruit d'une fricative, par exemple), l'annotation fournie par l'AA semble satisfaisante et ne nécessite pas de correction manuelle. En revanche, lorsque les analyses portent sur des zones transitoires dont les localisations sont précises (événements acoustiques tels que l'explosion d'une plosive, des zones d'assimilation de voisement ou des phénomènes de coarticulation), l'annotation fournie par l'AA apparaît trop approximative.

Il semble donc qu'en fonction des besoins des analyses, l'AA sera utilisable tel quel ou devra être corrigé par des experts. Il est donc envisageable que l'AA rende possible de larges études quantitatives portant sur de très grands corpus de parole dès lors que les analyses portent sur des segments dont les zones frontières peuvent être approximatives. En revanche, des études portant sur des segments très courts et précis sont pour l'instant limitées à une faible quantité de données dès lors qu'elles nécessitent une correction manuelle.

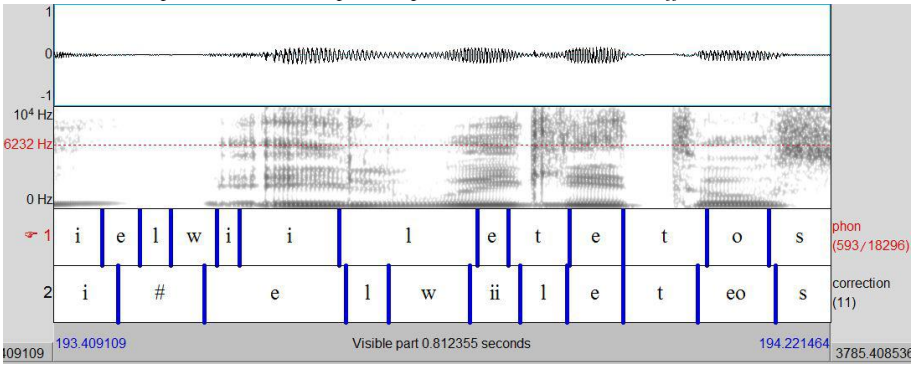
Enfin, la correction manuelle est évidemment indispensable dans les cas d'erreur importante, c'est-à-dire lorsque l'aligneur propose une annotation déconnectée du signal de parole (figure 1.6). Cela peut être le cas lorsque la transcription est trop éloignée de la réalisation réelle du locuteur, ou encore lorsque la production du locuteur est très altérée (ce qui est le cas chez des patients dysarthriques, mais aussi en parole naturelle lors de production très fortement hypo-articulées, cf. § 1.7).

1.6. Segmentation des macro-classes du français

Il s'agit dans cette partie de fournir des indices utilisés très couramment pour caractériser des frontières de segments[12] que chacun pourra utiliser selon leur pertinence vis-à-vis des questions posées. Pour chaque macro-classe, des critères génériques sont

12. Pour des critères plus précis, se reporter à Meunier [MEU 94].

Figure 1.6. *Séquence « […] et Louis il était au […] » (extrait d'un corpus de parole spontanée). Exemple d'erreurs massives de l'aligneur (« phon ») comparées avec la correction d'un expert (« correction »). Les étiquettes de phonèmes ne sont plus en phase avec la réalisation effective.*



d'abord présentés, puis quelques cas spécifiques sont abordés (parfois en contradiction avec les critères génériques). Rappelons que la segmentation se fait en deux temps : 1) le repérage des macro-classes (forme globale des segments), 2) le repérage des zones frontière (zones de transition). Le positionnement d'une étiquette de frontière se fait ensuite selon le type de transition rencontrée :

1) Présence d'une discontinuité sur la zone frontière : l'étiquette est posée sur cette discontinuité ;

2) Chevauchement des indices : a) primauté d'un indice ou b) étiquetage au centre du chevauchement ;

3) Transition graduelle : étiquetage au centre de la transition.

Les macro-classes choisies ici sont au nombre de quatre [MEU 94]. Les phonèmes sont indiqués en Alphabet Phonétique International, puis en code SAMPA[13] (utilisé pour certaines figures dans ce chapitre) :

13. Inventaire du code SAMPA pour le français : `http://www.phon.ucl.ac.uk/home/sampa/french.htm`.

	API	SAMPA
1 Occlusives	/p t k b d ɡ/	/p t k b d g/
2 Fricatives	/f s ʃ v z ʒ ʁ/	/f s S v z Z R/
3 Consonnes vocaliques	/l ʁ m n j ɥ w/	/l R m n j H w/
4 Voyelles	/i y e ø ɛ œ a ɔ o u ɛ̃ œ̃ ɑ̃ ɔ̃/	/i y u e 2 o E 9 @ a A a~ o~ 9~ e~/

Ces regroupements sont justifiés par le type de zone transitoire caractéristique de chaque segment phonétique. Ainsi, la classe des consonnes vocaliques comporte des consonnes très variées d'un point de vue articulatoire mais qui présentent des transitions semblables dans la mesure où toutes ces consonnes sont caractérisées par une structure de formant. Notons également que le phonème /r/ apparaît, à la fois dans la classe des fricatives (/ʁ/) et dans celle des consonnes vocaliques (/ʀ/) en raison de réalisations contextuelles très différentes.

Afin d'éviter une redondance inutile, nous aborderons, pour chaque type de segment, les transitions qui lui sont spécifiques. Il ne s'agit pas de faire un inventaire exhaustif de toutes les combinaisons possibles. Nous avons fait de choix d'aborder chaque macro-classe en nous focalisant sur les caractéristiques de ses phases transitoires [MAC 09].

1.6.1. *Occlusives*

Les plosives sont caractérisées par deux événements ordonnés dans le temps : la tenue (période d'occlusion) et l'explosion (relâchement des articulateurs au lieu d'occlusion). Ces événements sont, la plupart du temps, facilement identifiables sur le signal de parole et le repérage de cette macro-classe est assez aisée car sa représentation acoustique se distingue correctement des autres macro-classes.

Le **début** des plosives est caractérisé par du silence pour des plosives non-voisées et par un signal périodique simple pour des plosives voisées. Notons toutefois qu'en contexte vocalique voisé, le début des plosives non-voisées est souvent voisé, le voisement n'étant pas pertinent pour délimiter les segments (figure 1.7). On observe même des réalisations dans lesquelles le voisement perdure jusqu'au bruit d'explosion.

La **fin** de la plosive est déterminée par la fin du bruit d'explosion. En contexte vocalique, ce critère est assez facilement applicable. Toutefois, dans des séquences du type OCC+FRI, la limite entre le bruit de l'occlusive et celui du segment bruité peut s'avérer extrêmement difficile à distinguer (figure 1.8). Il s'agira alors d'identifier les deux types de bruit. En cas d'impossibilité, on appliquera le critère qui conduit à poser une étiquette au centre d'une zone transitoire.

Figure 1.7. *« me parlait » (extrait de texte lu). La partie initiale du* /p/ *est voisée.*

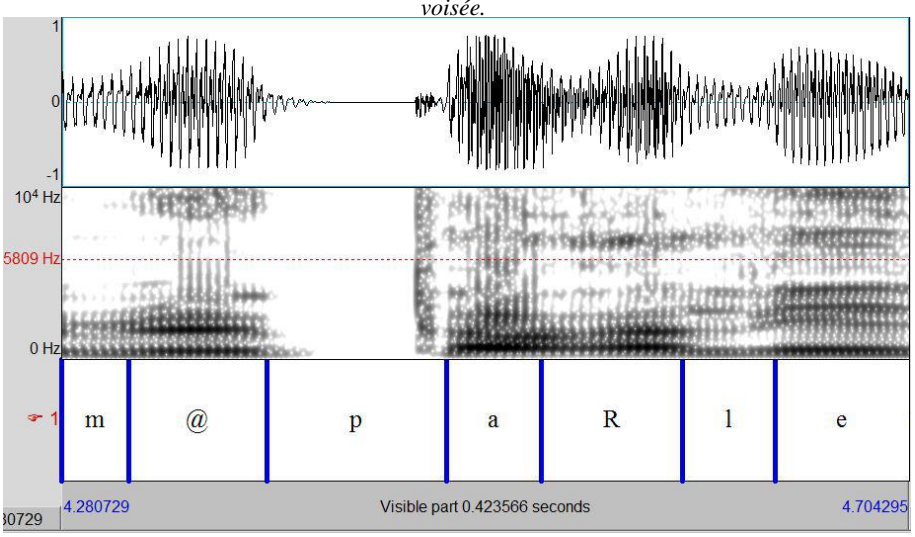
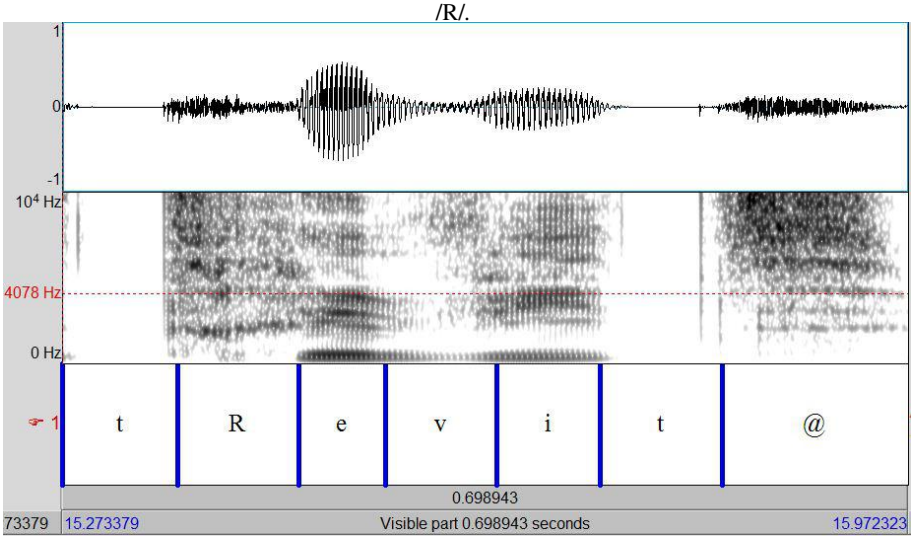


Figure 1.8. *« très vite » (extrait de texte lu). Segmentation difficile entre* /t/ *et* /R/.



Notons enfin que les bruits d'explosion ont des représentations acoustiques variées selon le lieu d'articulation : pour les alvéolaires (/t/ et /d/) le bruit sera court et intense, pour les labiales (/p/ et /b/), le bruit est court et de faible intensité, pour les vélaires (/k/

et /g/) le bruit est intense et souvent multiple. Enfin, le bruit d'explosion est d'intensité plus faible pour les occlusives voisées. Le bruit des bilabiales voisées est ainsi parfois impossible à identifier.

1.6.2. *Fricatives*

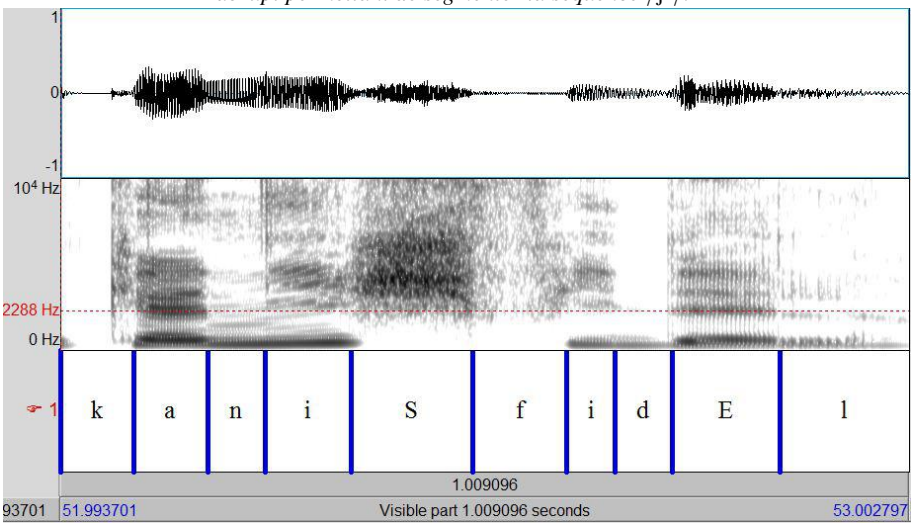
Les **fricatives** sont marquées par la présence de bruit. C'est donc cet indice prédominant qui va déterminer leur identification. Les fricatives sont une macro-classe facile à identifier mais dont les limites sont parfois peu précises. En effet, il est assez fréquent que le début ou la fin de la zone bruitée soit concomitant avec l'établissement ou le relâchement vocalique (particulièrement au contact de voyelles fermées, /i/, /y/ et /u/, figure 1.5). Il est en général admis que la présence de bruit est le critère déterminant pour les frontières de fricatives. Toutefois, il est des cas où le bruit de friction peut accompagner la totalité de la voyelle. Il faudra alors tenter de distinguer le bruit spécifique de la fricative de celui qui accompagne la voyelle.

Une autre difficulté concerne la réalisation des fricatives voisées qui sont parfois réalisées, au moins en partie, comme des approximantes (/ava/ réalisé [awa] sans zone bruitée). Les critères à appliquer se rapprochent alors de ceux utilisés pour segmenter les consonnes vocaliques.

Même si leur fréquence d'apparition est plutôt rare, les séquences de fricatives présentent des caractéristiques spécifiques. Les fricatives ont des fréquences de bruit spécifiques et la frontière sera localisée au moment où le bruit change de fréquence (figure 1.9). Ce changement peut être abrupt ou graduel. En cas de changement graduel, la frontière pourra être posée au milieu de la zone transitoire.

Le /r/ est probablement le phonème du français qui présente le plus de variantes allophoniques. Il peut être réalisé fricatif et non voisé en contexte non voisé (/ʁ/), mais aussi voisé et vocalique en contexte voisé (/ʀ/). De ce fait, les transitions entre cette consonne et les autres segments phonétiques sont elles-mêmes assez variées. En contexte non voisé (obstruante telle que /t/, /p/, /k/, /f/), il prend l'aspect d'une fricative non voisée. Il devient alors parfois délicat de distinguer le bruit de l'obstruante de celui du /ʁ/ (figure 1.8), surtout lorsque le bruit d'explosion de l'occlusive est très faible. Notons toutefois que les fréquences respectives des fricatives et du /r/ sont assez distinctes (les fréquences du bruit du /r/ sont généralement bien plus graves). Enfin, il faut remarquer que la position du /ʁ/ dans ce contexte (non voisé) détermine l'étendue du dévoisement. Il sera en général totalement dévoisé lorsqu'il est consécutif à l'obstruante (/tʁo/, fʁap/), mais plutôt partiellement dévoisé lorsqu'il la précède (/aʁp/, /maʁʃ/).

Figure 1.9. *« caniche fidèle » (extrait de phrases lues). Changement spectral abrupt permettant de segmenter la séquence /ʃf/.*



1.6.3. *Consonnes vocaliques*

Dans cette catégorie sont regroupés des phonèmes très différents dans leur structure articulatoire et acoustique. La raison de ce regroupement tient dans leur réalisation vocalique, c'est-à-dire que tous ces phonèmes sont réalisés avec une structure de formants. De ce fait, ils présentent tous une configuration comparable dans leur zone frontière.

Contrairement à ce qui est souvent remarqué, les frontières des consonnes vocaliques ne sont pas toujours impossibles à déterminer. Certes, dans certains cas (glides + voyelles) il sera difficile, voire impossible, d'identifier une discontinuité brève permettant de poser une étiquette de frontière. Mais les consonnes vocaliques présentent des structures formantiques très différentes les unes des autres par différents aspects : la densité harmonique, la répartition spectrale des formants et la complexité de la structure formantique.

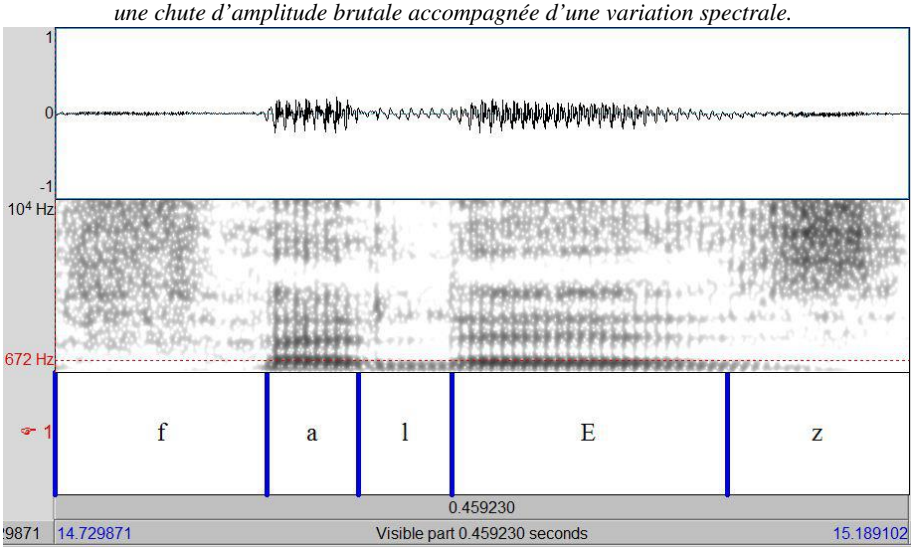
Les consonnes **nasales** sont probablement les plus simples à segmenter parmi les consonnes vocaliques. Leur structure formantique est assez simple et la distinction avec les segments frontières (le plus souvent vocaliques) se fait grâce à un changement d'amplitude souvent assez facile à identifier.

A contrario, les **glissantes** sont probablement les segments qui posent le plus de problème pour la segmentation. À l'inverse des nasales, leurs zones frontières présentent des changements très graduels et il est quasiment impossible d'identifier un

point précis caractéristique du changement. La solution la plus adaptée consiste sans doute à localiser la zone frontière dans son ensemble, puis à poser arbitrairement la frontière au milieu de cette zone. On peut estimer que c'est la façon la plus régulière de segmenter ce type de segment très problématique.

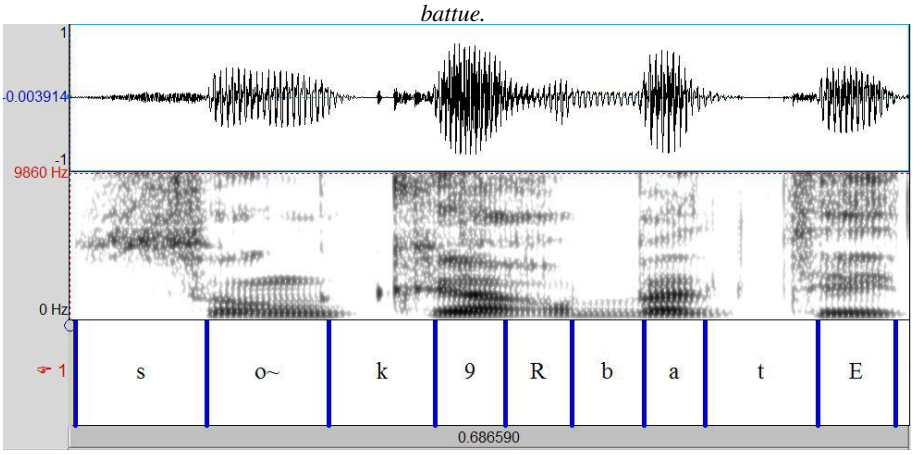
La **liquide** /l/ ne pose pas de problème spécifique et se distingue en général facilement du début des voyelles grâce à un changement brutal de l'amplitude du signal (figure 1.10). En contexte non voisé (occlusive ou fricative), /l/ est souvent partiellement dévoisé et se pose alors le problème de la délimitation du bruit (explosion ou friction). Le plus souvent, la fréquence des zones bruitées est assez distincte (voir § 1.6.2). Notons également que le /l/ peut parfois présenter une sorte d'explosion au cours de sa tenue. Il est probable que ce bruit soit la conséquence du décollement de la langue sur la partie alvéolaire de cette liquide latérale.

Figure 1.10. *« falaise » (extrait de texte lu). Segmentation aisée du /l/ grâce à une chute d'amplitude brutale accompagnée d'une variation spectrale.*



Le /ʀ/ en position intervocalique prend une forme vocalique présentant une structure de formant. Il est simple à identifier, mais ses zones frontières peuvent être assez floues. La frontière sera, dans ce cas, posée au milieu de la zone frontière (cf. les glissantes). Par ailleurs, en contexte voisé (occlusive et fricative), cette consonne prend une forme battue présentant plusieurs phases (deux parties vocaliques séparées par une partie battue, figure 1.11). Les parties externes étant vocaliques, les critères de segmentation des autres consonnes vocaliques (/l/ ou glissantes) seront appliqués ici.

Figure 1.11. *« son cœur battait » (extrait de texte lu). Au contact de la plosive voisée /b/ le /R/ est constitué d'une partie vocalique précédée d'une partie battue.*



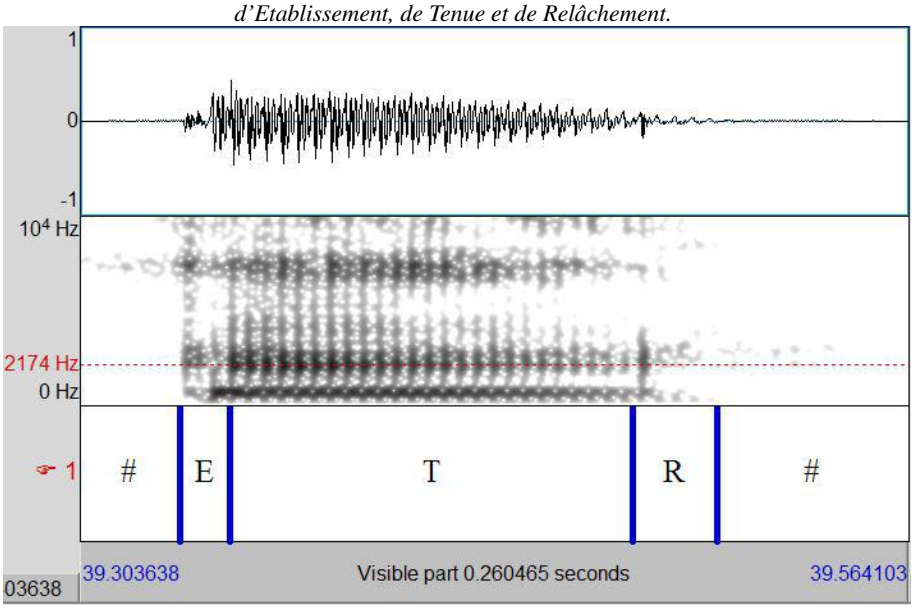
1.6.4. *Voyelles*

Étrangement, la segmentation des voyelles (c'est-à-dire leur zone frontière) est souvent mieux explicitée et détaillée dans la littérature que celle des consonnes, comme si la parole était une succession de voyelles et de consonnes[14], ces dernières étant supposées être homogènes. Il est classique de distinguer plusieurs phases dans la réalisation des voyelles [AUT 85] : une phase d'établissement (E), une phase de tenue (T) et une phase de relâchement (R) (figure 1.12). Le début et la fin de la phase de tenue sont marqués par l'apparition et la disparition du deuxième formant. Il est convenu que la phase de tenue est obligatoire tandis que les phases d'établissement et de relâchement sont fonction du contexte phonétique. La segmentation sera donc déterminée ainsi. Les phases E et R se caractérisent pas une périodicité assez simple (pas de F_2) de faible amplitude (croissante ou décroissante selon E et R). En théorie, elles sont absentes en contexte vocalique (consonnes vocalique et autres voyelles) mais sont présentes en contexte « vide » (début ou fin d'énoncé) ou obstruante (occlusives et fricatives). Dans les faits, il faut être un peu plus nuancé. Les phases E et R sont présentes en contexte occlusif sauf lorsque la voyelle précède l'occlusive (figure 1.7). Dans ce cas, il sera difficile de distinguer le relâchement de la voyelle du début de l'occlusive (périodicité simple), même lorsque l'occlusive est sourde (cf. occlusives). En contexte fricatif voisé, les phases E et R sont rarement présentes. Donc, pour résumer, les phases E

14. Ce qui n'est pas totalement faux car les statistiques portant sur des corpus de parole continue montrent, pour le français, une très forte dominance de la structure CV [MEU 12].

et R sont généralement présentes lorsqu'une voyelle est en contexte isolé, en contact avec des fricatives non voisées, ou lorsqu'elle suit une occlusive.

Figure 1.12. *La voyelle /ɛ/ prononcée isolément. Identification des phases d'Etablissement, de Tenue et de Relâchement.*



Par ailleurs, en contexte vocalique (consonnes vocaliques et voyelles), les voyelles posent les mêmes problèmes de segmentation que ceux qui ont été évoqués pour les consonnes vocaliques : gradualité et longueur des zones transitoires, manque d'indices précis au niveau temporel.

1.7. Au-delà de la parole de laboratoire

Les critères évoqués dans la partie précédente sont fondés sur un style de parole très contrôlé, correspondant à la lecture de mots ou de phrases courtes produits dans des conditions d'enregistrement optimales. La question des styles de parole doit être posée dans une gradualité qu'il est parfois complexe de cloisonner. Toutefois, nous ferons ici une distinction (même si celle-ci se montre en partie arbitraire) entre, d'une part, une parole préparée (PP) dont le périmètre s'étend de ce qui est lu à ce qui est appris et produit de mémoire et, d'autre part, une parole non préparée (PNP) qui recouvre des styles de parole correspondant à des interviews, des conversations, des entretiens libres, etc. Évidemment, la limite entre ces deux catégories peut être assez floue et la PP comme la PNP contiennent des types de parole très variés. Notre attention, dans

cette partie, portera sur des productions non préparées dans la mesure où, depuis plusieurs années, la création de corpus s'oriente vers ce type de parole[15]. Les aspects non préparés de la parole entraînent des réalisations parfois (mais pas systématiquement) différentes de ce que l'on trouve en parole contrôlée et ceci peut rendre caduques les transitions observées entre segments en parole contrôlée.

Dans une autre perspective, de nombreux travaux actuels s'orientent vers la description acoustico-phonétique de parole perturbée par une pathologie (dysarthries, dysphonie, etc.). La comparaison avec une parole non-pathologique nécessite une annotation phonétique qui devra être adaptée en fonction du degré de perturbation dû à la pathologie.

On peut ainsi se demander si l'utilisation exhaustive des outils d'alignement automatique ne répond pas, en partie, à notre désarroi face à un signal de parole qui échappe aux représentations établies dans des critères de segmentation propres à de la parole contrôlée. Ce réflexe n'est pas totalement absurde. Pourquoi ne pas laisser à des techniques efficaces dans leur régularité le soin de segmenter un signal face auquel l'expert est dérouté ? Le problème est que, comme l'expert, les techniques d'alignement automatique s'appuient sur des modèles acoustiques conçus pour de la parole contrôlée et s'avèrent donc parfois inopérantes pour des types de parole peu prototypiques. Ainsi, la compréhension des mécanismes de production de la parole dans des conditions non contrôlées ne peut réellement faire l'économie d'une expertise humaine.

1.7.1. *Identification des phénomènes spécifiques et problèmes de segmentation*

La parole non préparée fait apparaître des séquences phonétiques très spécifiques encore peu ou pas abordées. Le phénomène phare de l'étude des corpus de parole non préparée est la *réduction*[16]. Sous ce terme, on entend aussi bien les élisions que les altérations des segments phonétiques. Johnson [JOH 04] rapporte que dans le corpus *Buckeye* de conversation en anglais américain [PIT 05], l'élision d'un phonème est présente dans 20% des mots et celle de deux phonèmes dans 5% des mots. Plus de 60% des mots sont altérés sur au moins un phonème et 28% sur au moins deux phonèmes. Toujours dans cette même étude, les catégories lexicales sont différemment affectées puisque 4.5% des mots fonction, contre 6% pour les mots de contenu, perdent au moins une syllabe. Les phénomènes de réduction sont ainsi souvent exprimés sous forme d'altération paradigmatique affectant un phonème (ou une syllabe)

15. On peut citer le *Buckeye Corpus* [PIT 05], le *Corpus of Interactional Data* [BER 08] ou encore le *Nijmegen Corpus of Casual French* [TOR 10].

16. On notera, sur cette thématique, la tenue d'un workshop international en 2008 à Nimègue (Pays-Bas), *The First Nijmegen Speech Reduction Workshop*.

et se produisant à l'intérieur d'un mot. Ces phénomènes sont bien entendu présents, mais nous souhaiterions attirer l'attention du lecteur sur des réalisations moins balisées dont les caractéristiques semblent plus difficiles à expliciter d'un point de vue paradigmatique. Ces réalisations posent évidement des problèmes spécifiques pour la segmentation. On utilise le terme terme *métaplasme* pour rendre compte de certaines de ces forme de réduction (« *je crois que c'est quelque chose* » réalisé [ʃʁwaksekʃoz]) en soulignant que l'AA manque actuellement d'une description détaillée de ces phénomènes. Toutefois, la transcription comme la segmentation donnent une image très binaire des formes de réduction (présence ou absence du segment), parfois loin de ce qui est réellement produit en parole non préparée.

La production des sons en parole conversationnelle[17] laisse apparaître des phénomènes de réduction dont les implications phonétiques, articulatoires et perceptives sont encore à éclairer [MEU 12]. Certains de ces phénomènes de fusion concerne des formes dites « figées » et sont assez stéréotypés. Ils sont alors aisément identifiables car reproductibles. C'est le cas de la forme canonique « *je ne sais pas* » /ʒənəsɛpa/, très fréquemment réalisée en conversation [ʃɛpa]. De même, certaines formes peuvent être stéréotypées chez un locuteur. La séquence « *tu vois* » /tyvwa/ est très fréquemment réalisée [tɥa] chez l'un des locuteurs du CID. Mais ces formes sont, d'une certaine façon, attendues, car elles affectent des séquences de mots apparaissant souvent dans la conversation selon le même patron (formes *figées*). Toutefois, la réduction affecte également des séquences de mots qui ne sont pas toujours reproductibles dans la conversation. Nous appelons ainsi *conglomérat* une séquence non identifiable sous forme de suite d'unités phonétiques sur le signal de parole alors qu'elles sont perçues dans un contexte élargi par le transcripteur (figure 1.13). La caractéristique de ces séquences, à la différence des phénomènes de réduction typique comme les élisions claires ou les altérations (du type *assimilations*), est l'impossibilité d'identifier les phonèmes réalisés et ceux qui sont omis. La séquence apparaît comme une sorte de fusion syntagmatique des éléments qui la composent. Cette fusion peut se produire en interne de mots comme aux frontières[18]. Il s'agit probablement d'unités de programmation articulatoires non segmentales répondant à des contraintes de la parole conversationnelle qu'il convient d'analyser plus finement. Les aligneurs, comme les experts phonéticiens, ne proposent pas de solutions satisfaisantes dans la mesure où les deux approches reposent sur une représentation phonétique discrète et paradigmatique du signal de parole. Or, les conglomérats ne peuvent être décrits en ces termes. Ces séquences soulèvent, pour les phonéticiens, des questions théoriques très intéressantes. La figure 1.13 montre que la sortie de l'alignement donne une succession de

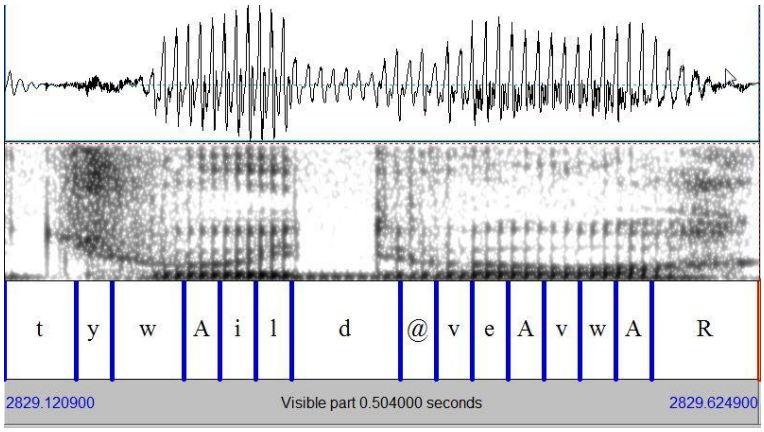
17. Suite à une observation détaillée du *Corpus of Interactional Data* (Bertrand et al., [BER 08]).

18. Notamment, une grande partie des productions lexicales dans ce corpus étant des mots fonction monosyllabiques, ces fusions apparaissent souvent dans des suites de monosyllabes.

phonèmes d'une durée minimale (24 ms) répartis dans le signal sans que cette répartition corresponde à de quelconques événements phonétiques. L'un des objectifs des phonéticiens pourrait être de fournir une interprétation articulatoire mais aussi linguistique de ces phénomènes de réduction tout en proposant des solutions pour une annotation phonétique adéquate.

La tendance est forte de voir dans ces types de réduction des contraintes purement physiologiques dues à l'augmentation du débit de la parole. Cette tendance est sans doute accentuée par l'impossibilité d'en rendre compte dans un format traditionnel de notation phonologique (transcription, élision, assimilation). Toutefois, ces phénomènes, caractéristiques d'une parole très peu contrôlée, sont probablement à relier avec les propriétés de la conversation, du discours, de la prosodie, etc. Il est probable que cette réduction n'apparaît pas n'importe où dans la conversation et que des détails phonétiques ténus (*Fine Phonetic Detail*, [HAW 03]) permettent la préservation et l'intelligibilité de la chaîne parlée.

Figure 1.13. *Exemple de conglomérat « tu vois il devait avoir », transcrit* [tywAild@veAvwaR] *(en SAMPA). La réalisation effective est bien en deçà de la transcription. Notamment la séquence* [@veAvwa] *est insegmentable et ne peut pas être représentée sous la forme d'une suite d'unités discrètes dans laquelle certaines unités seraient réalisées et d'autres non.*



1.7.2. *Nouveaux défis pour l'annotation phonétique*

Avant l'arrivée des systèmes d'alignement automatique, les corpus de parole étaient segmentés à la main, par des experts. Ce travail, coûteux, était toutefois possible dans la mesure où les corpus enregistrés étaient de taille restreinte. Cette expertise ne nécessitait pas de transcription préalable, d'autant que la plupart des corpus correspondaient

à de la lecture de logatomes, de mots, de phrases ou de textes. Les analyses phonétiques s'intéressent désormais à des types de parole très variés et moins prototypiques. La forte variabilité présente dans ces productions non contrôlées, et souvent contextualisées, nécessite un très grand nombre de données de façon à pouvoir extraire des analyses pertinentes d'un point de vue statistique. Cette masse de données rend donc impossible une annotation simplement manuelle du signal de parole.

Il est probable que le travail initial de *segmentation manuelle* évolue désormais vers un travail de *correction manuelle de l'alignement*. En effet, on voit se généraliser l'utilisation des aligneurs[19] même sur des corpus contrôlés très courts car le gain de temps est toujours important et le déplacement d'une étiquette est souvent considéré comme moins coûteux qu'une segmentation sur un signal brut. Dans un avenir proche, il est donc probable que très peu de corpus seront encore segmentés manuellement dans leur ensemble.

Cette évolution n'est pas anodine dans la mesure où la correction manuelle de l'alignement se fait toujours en référence à l'annotation de l'aligneur et, même si l'expert utilise ses propres critères, ses choix sont influencés par la localisation des étiquettes déjà présentes sur le signal. Notamment, cette approche favorise une représentation en segments discrets (phonèmes balisés par une étiquette de début et de fin) qui peut être éloignée ou, du moins, peu informative par rapport aux réalisations effectives des locuteurs. De plus, dans la pratique de la correction de l'alignement, l'expert va éviter de déplacer une étiquette lorsqu'il doute (frontière ambiguë ou chevauchement) ou lorsque la correction ne porte que sur quelques millisecondes. Cette situation favorise le résultat de l'alignement par rapport à la segmentation manuelle sur un signal brut. Toutefois, l'influence de l'alignement sur l'annotation des experts humains peut entraîner une régularité dans les pratiques qui finalement peut être souhaitable. En retour, le regard des experts humains sur l'alignement automatique et ses imperfections peut permettre une amélioration de ces outils au travers d'une meilleure définition des modèles acoustiques utilisés. Notamment, l'identification de phénomènes spécifiques à certains types de parole doit être exploitée dans le développement des aligneurs. On peut ainsi envisager une complémentarité entre les approches manuelles et automatiques à la fois dans leur exploitation actuelle, mais aussi en vue d'une amélioration des techniques et des connaissances sur la langue parlée.

19. Cette tendance est renforcée par le fait que les jeunes générations de chercheurs maîtrisent de mieux en mieux les outils de traitement automatique de la parole et, particulier, les outils d'alignement automatique.

1.8. Bibliographie

[ABR 85a] ABRY C., AUTESSERRE D., BARRERA C., BENOÎT C., BOË L.-J., CAELEN J., CAELEN-HAUMONT G., ROSSI M., SOCK R., VIGOUROUX N., « Propositions pour la segmentation et l'étiquetage d'une base de données des sons du français », *Actes des XIVèmes Journées d'Études sur la Parole*, Paris, p. 156–163, 10-13 juin 1985.

[ABR 85b] ABRY C., BENOÎT C., BOË L.-J., SOCK R., « Un choix d'événements pour l'organisation temporelle du signal de parole », *Actes des XIVèmes Journées d'Études sur la Parole*, Paris, p. 133-137, 10-13 juin 1985.

[AUD 10] AUDIBERT N., FOUGERON C., FREDOUILLE C., MEUNIER C., « Évaluation d'un alignement automatique sur la parole dysarthrique », *Actes des XXVIIIèmes Journées d'Études sur la Parole*, Mons, Belgique, p. 353–356, 2010.

[AUT 85] AUTESSERRE D., ROSSI M., « Propositions pour la segmentation et l'étiquetage de la base de données acoustiques du G.R.E.C.O. Parole », *Actes des XIVèmes Journées d'Études sur la Parole*, Paris, p. 147–151, 10-13 juin 1985.

[BER 08] BERTRAND R., BLACHE P., ESPESSER R., FERRÉ G., MEUNIER C., PRIEGO-VALVERDE B., RAUZY S., « Le CID – Corpus of Interactional Data – Annotation et exploitation multimodale de parole conversationnelle », *Traitement Automatique des Langues*, vol. 49, p. 105–134, 2008.

[BOE 01] BOERSMA P., « Praat, a system for doing phonetics by computer », *Glot International*, vol. 5, p. 341–345, 2001.

[BOI 00] BOITE R., BOURLARD H., DUTOIT T., HANCQ J., LEICH H., *Traitement de la parole*, Presses Polytechniques et Universitaires Romandes, Lausanne, Suisse, 2000.

[BÜR 08] BÜRKI A., GENDROT C., GRAVIER G., LINARÈS G., FOUGERON C., « Alignement automatique et analyse phonétique : comparaison de différents systèmes pour l'analyse du schwa », *Traitement Automatique des Langues*, vol. 49, 2008.

[COL 05] COLEMAN J., *Introducing Speech and Language Processing*, Cambridge University Press, Cambridge, UK, 2005.

[DUR 90] DURAND J., *Generative and Non-Linear Phonology*, Longman, London, UK, 1990.

[FAN 73] FANT G., *Speech, Sounds and Features*, MIT Press, Cambridge, MA, 1973.

[FAR 97] FARNETANI E., « Coarticulation and connected speech processes », HARDCASTLE W., LAVER J., Eds., *The Handbook of Phonetic Sciences*, p. 371–404, Blackwell, 1997.

[FOU 10] FOUGERON C., AUDIBERT N., FREDOUILLE C., MEUNIER C., GENDROT C., PANSERI O., « Comparaison d'analyses phonétiques de parole dysarthrique basées sur un alignement manuel et un alignement automatique », *Actes des XXVIIIèmes Journées d'Études sur la Parole*, Mons, Belgique, p. 365–368, 2010.

[HAR 99] HARRINGTON J., CASSIDY S., *Techniques in Speech Acoustics*, Kluwer Academic Publishers, Foris, Dordrecht, 1999.

[HAW 03] HAWKINS S., « Roles and representations of systematic fine phonetic detail in speech understanding », *Journal of Phonetics*, vol. 31, p. 373–405, 2003.

[JOH 02] JOHNSON K., *Acoustic and Auditory Phonetics*, Blackwell, Oxford, UK, 2002.

[JOH 04] JOHNSON K., « Massive reduction in conversational American English », YONEYAMA K., MAEKAWA K., Eds., *Proceedings of the 10th International Symposium Spontaneous Speech : Data and Analysis*, Tokyo, p. 29–54, 2004.

[LAD 97] LADEFOGED P., « Instrumental techniques for linguistic phonetic fieldwork », HARDCASTLE W., LAVER J., Eds., *The Handbook of Phonetic Sciences*, p. 137–166, Blackwell, Oxford, UK, 1997.

[MAC 09] MACHAC P., SKARNITZL R., *Principles of Phonetic Segmentation*, Epocha Publishing House, Prague, 2009.

[MAL 03] MALFRÈRE F., DEROO O., DUTOIT T., RIS C., « Phonetic alignment : speech-synthesis-based versus Viterbi-based », *Speech Communication*, vol. 40, p. 503–517, 2003.

[MEU 94] MEUNIER C., Les groupes de consonnes. Problématique de la segmentation et variabilité acoustique, PhD thesis, Université de Provence, 1994.

[MEU 12] MEUNIER C., « Contexte et nature des réalisations phonétiques en parole conversationnelle », *Actes des 29èmes Journées d'Études sur la Parole*, p. 1–8, 2012.

[PIT 05] PITT M., JOHNSON K., HUME E., KIESLING S., RAYMOND W., « The Buckeye corpus of conversational speech : Labeling conventions and a test of transcriber reliability », *Speech Communication*, vol. 45, p. 89–95, 2005.

[ROS 90] ROSSI M., « Segmentation automatique de la parole : pourquoi ? Quels segments ? », *Traitement du Signal*, vol. 7, p. 315–326, 1990.

[TAR 03] TARRIER J.-M., « L'enregistrement et la prise de son », DELAIS-ROUSSARIE E., DURAND J., Eds., *Corpus et variation en phonologie du français : Méthodes et analyses*, p. 187–212, Presses Universitaires du Mirail, Toulouse, 2003.

[TES 01] TESTON B., « L'enregistrement numérique de la voix et la parole : problèmes et méthodes », *Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence*, vol. 20, p. 219–232, 2001.

[TOR 10] TORREIRA F., ADDA-DECKER M., ERNESTUS M., « The Nijmegen corpus of casual French », *Speech Communication*, vol. 52, p. 201–212, 2010.

[WES 96] WESENICK M. B., KIPP A., « Estimating the quality of phonetic transcriptions and segmentations of speech signals », *Proceedings of ICSLP 1996*, p. 129–132, 1996.