Table des matières

Table des figures

"Le silence de l'homme est plus proche de la vérité que ne le sont ses paroles."
Khalil Gibran

1 Préambule sur le signal de la parole

La parole est un signal émis par suite complexe d'actions que l'on fait sans en avoir réellement conscience. Il y a tout d'abord la *génération d'une énergie ventilatoire* qui va servir à mettre en mouvement oscillatoire les cordes vocales afin de générer un bruit. La *vibration des cordes vocales* donne naissance à tous les sons voisés, soit 80% du temps de phonation. La *réalisation d'une disposition articulatoire* dans ce qu'il est commode de désigner sous le nom de cavités supra-glottiques, termine le processus de phonation. La transcription de la parole se fait à l'aide des phonèmes, qui sont les unités théoriques de la langue, et qui permettent de distinguer deux sons différents. On utilise parfois les diphones ou les triphones qui sont des éléments sonores caractéristiques de la transition entre deux ou trois phonèmes s'étendant de la partie stable d'un phonème à la partie stable du phonème suivant ou précédent. En effet la production d'un phonème diffère selon le phonème qui le suit ou qui le précède. Cette définition n'est cependant pas parfaitement rigoureuse puisque la parole n'est jamais un processus totalement stationnaire.

Les voyelles sont classées selon différentes méthodes. Tout d'abord on peut distinguer 3 groupes selon le lieu d'articulation, déterminé par la position horizontale de la masse linguale dans la cavité buccale. On distinguera :

- les voyelles antérieures [i,y,e,ø,œ]
- les voyelles centrales [a]
- les voyelles postérieures [u,o].

Il y a également la classification selon la position de la langue, haute, basse, ou moyenne. La classification des consonnes est beaucoup plus délicate. Outre la division selon la présence ou non de vibration des cordes vocales, c'est-à-dire en voisées, ou non-voisées, on peut distinguer différents modes d'articulation :

- les occlusives [p,t,k,b,d,g]
- les fricatives [f,s,ʃ,v,z]
- les nasales [m,n]
- les liquides [l]
- les semi-voyelles [j,y,w]

On peut aussi les différencier selon leur point d'articulation, ou la vitesse du mouvement des articulateurs.

En plus de ces classifications, un même phonème sera prononcé différemment selon l'état de la personne qui parle. En effet l'émission de la parole varie selon la fatigue, le stress... Et bien sûr la parole changera aussi selon le locuteur, c'est-à-dire la personne qui parle.

Pour d'avantage de précision, sur la production de la parole, on se réfère à [Calliope89] et à [Bartkova99].

2 Introduction au problème de détection de la parole

Dans de nombreux domaines de traitement de la parole, il importe de bien savoir détecter les segments de parole dans le signal. Les besoins de détection de la parole dans le signal diffèrent selon le problème considéré. Pour la reconnaissance automatique de la parole, on cherche à différencier les segments de parole et de bruit, on parle aussi de détection de début et de fin de mots. On cherche parfois une détection plus fine des segments de parole, en segments voisés et segments non-voisés. Pour la transmission, ou en codage, le problème se nomme détection d'activité vocale, on cherche ici une segmentation plus grossière que pour la reconnaissance. En effet il est plus facile à un être humain de reconnaître des mots même tronqués, que pour un système de reconnaissance. On ne développera cependant ici, que peu ce dernier type de détection.

Selon la détection recherchée, il y a différentes façons d'évaluer la segmentation. Pour la détection d'activité vocale, l'évaluation peut se faire en visualisant le signal et les segments détectés comme de la parole, on peut aussi établir les statistiques des segments tronqués par rapport à une segmentation manuelle. Pour la reconnaissance, ces types d'évaluation restent valable. À partir de la segmentation, on peut également représenter les erreurs rejetables par un algorithme de reconnaissance en fonction des erreurs définitives. Les erreurs rejetables sont composées des insertions et des détections de non parole, tandis que les erreurs définitives sont composées des omissions, des regroupements et des fragmentations de parole. On peut de plus représenter des résultats sur les systèmes de reconnaissance, pour un type de détection. Cette méthode permet d'évaluer un système de détection de la parole, conditionnellement à un système de reconnaissance. On représente alors les faux rejets (i.e. mot du vocabulaire rejeté) en fonction des erreurs de substitution (i.e. mot du vocabulaire pris pour un autre mot du vocabulaire), ou des fausses acceptations (i.e. mot hors vocabulaire ou bruit pris pour un mot du vocabulaire).

Remarquons qu'il y a d'autres segmentations possibles que celles présentées ci-dessous, comme par exemple la segmentation silence - non-silence détaillée dans [De souza83]. Le fait d'être dans un état de non-silence n'indique pas qu'il y ait de la parole.

3 Différents contextes de détection

Toutes les approches font appel à différentes caractéristiques de la parole qui sont d'ordre soit structurelles, soit statistiques, ou bien d'un mélange des deux. L'approche structurelle est définie par un apprentissage ou le plus souvent est déduite de connaissances a priori. Quelques exemples de règles sur la structure de la parole :

- Les maxima d'énergie de la parole ne durent pas plus de 2s.

[Lynch et al.87]

- Un silence intra-mot (tenue de plosive) ne dure pas plus de 150 ms.

[Lynch et al.87]

- Un mot est constitué d'un ou plusieurs maxima d'énergie.

[Lamel et al.81]

- Plus la distance entre deux maxima est grande, moins il est probable qu'ils appartiennent au même mot.[Lamel et al.81]

- La durée minimale d'une voyelle est de 20ms.[Dermatas et al.91]

Notons aussi que d'après [Un et al.80] la parole ne représente environ que 40% des communications téléphoniques.

L'approche statistique consiste à discriminer le signal de parole, avec des outils statistiques appliqués à des caractéristiques du signal. Nous étudierons les différentes caractéristiques utilisées dans une autre section.

L'approche préconisée par [Rabiner et al.75], dans le cadre de la reconnaissance, est une approche fondée sur la reconnaissance des formes. Il écrit : "The problem of locating endpoints of an utterance in these backgrounds of silence essentially is one of pattern recognition. The way one would attack the problem by eye would be to acclimate the eye (and brain) to the 'typical' silence waveform and then try to spot some radical change in the pattern."

Les algorithmes des systèmes de détection doivent vérifier quelques critères importants, qui diffèrent selon les auteurs. Ils doivent être simples, fiables, et applicables à diverses conditions d'utilisation, selon [Rabiner et al.75], indépendants du locuteur et du vocabulaire, selon [De souza83], robustes, en temps réel, et n'utilisant pas les connaissances a priori sur le bruit selon [Savoji89].

Nous présentons ci-dessous différentes approches de détection : la détection d'activité vocale, qui sera développée plus en détail dans un document ultérieur, la détection Bruit/Parole, qui est appliquée en reconnaissance, et la détection des segments voisés/non voisés/silence.

3.1 Détection d'activité vocale (DAV)

De nombreux algorithmes de détection d'activité vocale ont été élaborés. On peut citer [Freeman et al.89] et [Braun et al.90] qui présentent une méthode de détection, destinée à la transmission GSM, utilisant deux systèmes de détection en parallèle, une DAV permettant d'estimer spécifiquement les statistiques du bruit, l'autre les intégrant pour la segmentation. Ces algorithmes sont utilisés pour la transmission en milieu très bruité (avec un faible rapport signal à bruit (SNR)). C'est le cas de plus en plus couramment avec les téléphones portables utilisés en tout environnement. Un autre algorithme de DAV [Watson et al.97] utilise deux systèmes de détection en parallèle, pour le codage GSM.

3.2 Détection Bruit/Parole (DBP)

La détection Bruit/Parole ou détection des début et fin de mots, est utilisée par exemple, pour délimiter un mot à reconnaître. Le système de reconnaissance de mots isolés présuppose que l'on ait isolé le mot dans une communication qui peut être bruitée. La qualité du système de reconnaissance dépend donc de la DBP. Dans le cas de la reconnaissance continue, en milieu bruité, une bonne DBP est également nécessaire. Depuis une trentaine d'années que le problème s'est posé, les techniques évoluent pour être de plus en plus robustes face aux milieux bruités.

[Lamel et al.81] font référence à trois sortes de systèmes de DBP en vue d'appliquer la segmentation obtenue à un système de reconnaissance. Il y a tout d'abord les systèmes *explicites*, qui sont des systèmes de segmentation en amont des systèmes de reconnaissance (par exemple dans [Rabiner et al.75]). C'est le cas de la plupart des systèmes de DBP. Les systèmes *implicites* sont au contraire inclus dans le système de reconnaissance (par exemple dans [Takeda et al.95]). Il y a ainsi interaction entre les deux systèmes. On trouve ensuite les systèmes *hybrides* qui réalisent une partie de la segmentation avant le processus de reconnaissance, puis en inter-agissant avec le système de reconnaissance. Il y a donc une première estimation des débuts et fins de mots par un système explicite, puis correction par un système implicite (par exemple dans [Lamel et al.81]). C'est à dire qu'un système hybride est la combinaison entre un système explicite et un système implicite.

[Lamel et al.81], [Savoji89], [Junqua et al.94] et [Van gerven et al.97] utilisent avant tout l'énergie du signal, dans leur algorithme de DBP. Le pitch dans [Hamada et al.90] est aussi utilisé comme critère de détection. Des méthodes de réseaux neuronaux [Heon et al.98], du maximum de vraisemblance [Arslan et al.98], ou des méthodes utilisant l'entropie [Abdallah et al.97a] et

[Shen et al.98] sont également employés. Dans ce dernier exemple, l'entropie de la densité spectrale normalisée considérée comme une densité de probabilité est utilisée comme principal discriminant.

La DBP utilisée au CNET est fondée sur un automate à cinq états, utilisant l'énergie, ou d'autres critères, pour le passage d'un état à l'autre (cf [Mauuary94] et [Karray98a]). On détaillera dans un autre paragraphe les travaux du CNET.

3.3 Détection de segments voisés/non voisés/silence

Cette détection plus fine qui segmente la parole en voisée ou non voisée, est utilisée en codage de la parole et en synthèse. Pour la reconnaissance la séparation des trames de parole en voisée et non-voisée n'est pas indispensable.

Les sons voisés sont produits par la vibration des cordes vocales. Les voyelles sont intrinsèquement voisées, tandis que les consonnes peuvent l'être ou non (cf [Calliope89]). On peut donc considérer qu'un mot est constitué d'une suite de segments voisés, de segments non-voisés et de silences brefs. Cependant toute suite de ces trois segments de base ne correspond pas à un mot, du bruit peut être constitué par des sons voisés. Un des paramètres de voisement est le pitch, qui est le terme anglais qui couvre la fréquence des cordes vocales, la fréquence laryngienne si l'on veut faire référence au processus de génération articulatoire et la fréquence fondamentale si l'on se place dans le domaine acoustique. On peut retrouver une étude comparative des méthodes d'extraction du pitch dans [Hess83], et plus récemment dans [Bagshaw94]. Cependant les méthodes d'extraction du pitch ne sont pas toujours performantes dans les environnements bruités. Ici aussi, pour la segmentation, il est utilisé plusieurs critères, jusqu'à cinq dans [Atal et al.76] et dans [Rabiner et al.77b]. [Di francesco90] utilise la divergence de Kullback pour l'obtention de la période du pitch.

L'énergie est également employée, pour la segmentation en voisé/non-voisé/silence, dans [Rabiner et al.75] et dans [Rabiner et al.77a], où il est rajouté une information à partir d'une distance LPC (Linear Predictor Coefficient). Des données statistiques, comme les moments d'ordre 3 et 4 du signal sont utilisées dans [Jacovitti et al.91].

4 Quelques techniques de détection

Nous présentons dans ce paragraphe, différentes analyses du signal, une liste de caractéristiques acoustiques pour la détection de la parole, puis

quelques méthodes de modélisation utilisant ces caractéristiques acoustiques en vue de la détection de la parole.

4.1 Analyse du signal

- **Le spectre**

  Le spectre du signal est le vecteur formé par le module au carré de chaque coordonnée de la transformée de Fourier, pondérée sur une fenêtre de Hamming, par exemple, du signal filtré. Il peut éventuellement être calculé en moyennant le module au carré de la transformée de Fourier, sur plusieurs fenêtres. Le fait de prendre le module, simplifie les calculs, puisqu'ainsi le logarithme est réel. Cette simplification supprime l'information sur la phase, mais l'oreille humaine ne la perçoit que très mal (cf [Bartkova99]). De plus les systèmes de reconnaissance actuels n'utilisent pas la phase du signal de parole. Pour plus de détails on peut se référer à [Calliope89] ou [Rabiner et al.93]. Le spectre du signal permet de fournir de nombreuses caractéristiques du signal.

- **Les coefficients cepstraux**

  Les coefficients cepstraux sont les composants du cepstre, anagramme de spectre. Le cepstre est obtenu en prenant la transformée de Fourier inverse du logarithme du spectre. Le lissage cepstral est une analyse qui vise à séparer la contribution respective de la source et du conduit par déconvolution. Pour cela on fait l'hypothèse que le signal vocal s_n est produit par un signal excitateur g_n traversant un système linéaire passif de réponse impulsionnelle b_n. On a donc $s_n = g_n * b_n$. L'homomorphisme décrit précédemment, au niveau de la définition du spectre, nous permet de déconvoluer s_n, en écrivant le produit de convolution dans un nouvel espace comme une somme. Ceci est décrit plus en détails dans [Calliope89], ou [Rabiner et al.93]. En notant $S(w)$, la densité spectrale, on peut écrire :

$$\log S(w) = \sum_{n=-\infty}^{+\infty} c_n \, e^{-inw},$$

  où $c_n = c_{-n}$ sont les coefficients cepstraux du signal. On a ainsi :

$$c_n = \int_{-\pi}^{+\pi} \log S(w) e^{inw} \frac{dw}{2\pi},$$

  c_0 ayant les caractéristiques que le logarithme de l'énergie du signal, il lui est parfois associé. On obtient ainsi un ensemble de paramètres

plus ou moins importants, selon la précision voulue. En effet, plus on considérera de paramètres c_n, mieux on approchera $\log S(w)$. Il est cependant impossible, dans la pratique, d'interpréter chaque coefficient c_n comme représentant telle ou telle caractéristique du signal (type de phonème, pitch...). [Heon et al.98] utilisent les informations cepstrales avec l'aide d'un réseau de neurones, pour élaborer un algorithme de segmentation.
[Hamada et al.90] définissent une distance pondérée des coefficients cepstraux, à laquelle ils associent l'information du pitch. Cette distance comparée à un seuil adaptatif, leur sert de critère pour établir une détection de la parole.

- **L'analyse LPC (Linear Predictive Coding)**

La méthode LPC est fondée sur les connaissances de la production de la parole et suppose que le modèle de production soit linéaire. Ce modèle se décompose en deux parties : la source, active, et le conduit, passif. L'onde vocale est modélisée comme la sortie d'un filtre passe-bas. Le conduit vocal est représenté par un filtre tout pôle autorégressif d'ordre 2M. Le modèle du conduit nasal est un filtre pôle zéro ARMA et le rayonnement aux lèvres se modélise par un filtre tout zéro MA. L'ensemble du conduit se comporte donc comme un système linéaire ARMA. Pour plus de détails on peut se reporter à [Calliope89]. Dans un premier temps, on calcule les valeurs d'autocorrélation :

$$r(m) = \sum_{n=0}^{N-1-m} \tilde{x}(n)\,\tilde{x}(n+m) \quad m = 0, 1, ..., p,$$

où p est l'ordre de l'analyse LPC, N le nombre d'échantillons pris en compte (selon la taille de la fenêtre), et $\tilde{x}$ le signal sur une fenêtre donnée. La méthode de Durbin nous donne :

$$E^{(0)} = r(0),$$

$$k_i = \frac{1}{E^{(i-1)}} \left\{ r(i) - \sum_{j=i}^{L-1} \alpha_j^{i-1} r(|i-j|) \right\} ,$$

avec $1 \leq i \leq p$,

$$\alpha_i^{(i)} = k_i,$$

$$\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \alpha_{j-1}^{(i-1)}, \text{ pour} j < i$$

$$E^{(i)} = (1 - k_i^2) E^{(i-1)},$$

où $E^{(i)}$ est la variance de l'erreur de prédiction d'ordre i et L est le nombre de trames du signal. On obtient ainsi les coefficients LPC : $a_m = \alpha_m^{(p)}$, les coefficients PARCOR (PARtial CORrelation) : k_m et les coefficients du logarithme du rapport des aires :

$$g_m = \log\left(\frac{1-k_m}{1+k_m}\right).$$

Le premier coefficient de prédiction linéaire ainsi que l'autocorrélation avec décalage d'un échantillon sont utilisés par [Atal et al.76], parmi cinq paramètres. Les coefficients PARCOR et les coefficients du logarithme du rapport des aires associés à une distance sur les coefficients LPC, sont employés par [Rabiner et al.77a] pour une segmentation voisée/non-voisée/silence. La distance a la forme :

$$D(j) = \frac{(a-m_j)(\phi)(a-m_j)^t}{(a\phi a^t)},$$

où a est le vecteur des coefficients LPC, m_j la moyenne du vecteur pour la $j^{\text{ème}}$ classe du signal ($j=1$ représente le silence, $j=2$ la parole non-voisée, et $j=3$ la parole voisée) et ϕ est la matrice d'autocorrélation pour la trame courante. $D(j)$ est essentiellement une covariance pondérée de coefficients LPC. Cette distance est combinée avec une distance sur l'énergie par un produit ou une somme. Le but de cette distance est d'utiliser l'information spectrale du signal. Notons que la méthode LPC est un lissage plus régulier du signal, que le lissage obtenu à partir des coefficients cepstraux.

4.2 Caractéristiques acoustiques

De nombreuses caractéristiques de la parole sont employées pour segmenter le signal de parole. Bien souvent, elles sont couplées à des règles heuristiques fondées sur les connaissances structurelles de la parole. Dans ce qui suit, nous présentons les principales caractéristiques de la parole rencontrées et les plus prometteuses.

- **L'énergie**

  L'énergie est le critère le plus utilisé, mais dans les environnements très bruités, le rapport signal à bruit peut devenir très faible, voire négatif. De plus une énergie forte ne représente bien que les voyelles et quelques consonnes. C'est pourquoi, même si elle reste discriminante, elle est de plus en plus employée avec d'autres critères.

- **Le taux de passage par zéro**

  Souvent utilisé avec l'énergie, le taux de passage par zéro du signal est un paramètre calculé facilement et qui donne une information importante au niveau de la représentation spectrale du signal.

  Le taux de passage par zéro permet de détecter les fricatives faibles aux frontières des mots. Citons [Rabiner et al.75], [Junqua et al.94] et [Savoji89] où sont développés des algorithmes faisant appel à l'énergie et au taux de passage par zéro. Dans [Savoji89], ces deux paramètres sont combinés par des mesures utilisant des connaissances statistiques du signal. L'algorithme de [Atal et al.76], fait intervenir le taux de passage par zéro parmi cinq paramètres, optimisé par [Rabiner et al.77a].

- **Le pitch**

  Le pitch souvent confondu par abus de langage à la fréquence fondamentale, représente la périodicité du signal, et sa structure harmonique. C'est en ce sens que l'on peut penser qu'il est un paramètre discriminant de la parole et du bruit. Le problème principal est l'extraction de la valeur du pitch dans le signal, déterminant l'existence de segments périodiques. [Hess83] et [Bagshaw94] en donnent une étude comparative. [Di francesco90] présente une méthode d'extraction du pitch à l'aide de la divergence de Kullback, pour une segmentation voisé/non-voisé/silence. [Hamada et al.90] utilisent une méthode d'autocorrélation pour extraire le pitch. La valeur du pitch est ensuite comparée à un seuil, les faibles valeurs du pitch déterminent les états de non-parole, la parole étant supposée périodique. Le problème réside dans le fait que les périodes de bruit peuvent contenir du voisement, notamment pour les bruits de fonds.

- **Autres caractéristiques**

  [De souza83] utilise en plus du logarithme de l'énergie, du taux de passage par zéro et de l'autocorrélation, le nombre de passages par zéro de la dérivée du signal et une mesure sur le signal s_i. Cette mesure est définie par :

$$D = \log_{10}\left(\frac{\sum_{i=2}^{m} |s_i - s_{i-1}|}{\sqrt{\sum_{i=1}^{m} s_i^2}}\right),$$

  qui représente le caractère "ombré" du signal, que l'on perçoit à l'oeil. Le nombre de passages par zéro de la dérivée du signal donne une information sur le "cisaillement" du signal. Notons que les cinq critères précédents sont utilisés par un test statistique fondé sur la moyenne et la

covariance, pour un algorithme silence/non-silence. Ce même test peut cependant être employé avec d'autres paramètres que ceux précités.

[Andre obrecht et al.93] proposent un algorithme fondé sur l'analyse fréquentielle, ou temporelle. Il est en fait proposé trois méthodes qui prennent en compte des paramètres extraits du signal couplés avec la segmentation statistique du signal.

[Agaiby et al.97] présente la méthode de Tucker élaborée à partir d'une mesure sur la périodicité du signal.

Remarquons que toutes ces caractéristiques sont employées le plus souvent par comparaison à des seuils adaptatifs, qui sont adaptés par apprentissage. Ces seuils peuvent être initialisés par des connaissances heuristiques du signal.

Notons aussi que l'on trouve dans [Rabiner et al.77b] l'algorithme de [Atal et al.76] testé avec 70 paramètres différents, pour l'optimisation des cinq paramètres à utiliser. Il a été utilisé 12 coefficients LPC, 12 coefficients de corrélation, 12 coefficients PARCOR, 12 termes d'erreur partielle LPC, etc.

4.3 Quelques critères

- **L'entropie**

On peut définir de différentes manières l'entropie. [Abdallah et al.97a] implémentent un algorithme fondé sur l'entropie de Shannon. Soit une séquence $\{ x_{[0,N-1]}(n)\}$ pour $0 \leq n \leq N-1$ de N échantillons du signal $x(k)$, par rapport à la base $\{ X_{[0,N-1]}(k)\}_{0 \leq k \leq N-1}$ des coefficients de la transformée de Fourier discrète, c'est-à-dire, pour $0 \leq k \leq N-1$:

$$X_{[0,N-1]}(k) = \frac{1}{\sqrt{N}} \sum_{n=0}^{n=N-1} x_{[0,N-1]}(n) e^{-\frac{2\Pi i}{N} nk}.$$

L'entropie de Shannon est définie par :

$$E_{[0,N-1]} = -\sum_{k=0}^{N-1} |X_{[0,N-1]}(k)|^2 \log |X_{[0,N-1]}(k)|^2.$$

Ils définissent ensuite le critère entropique local, qui est une fonction sensible aux variations du spectre du signal, et qui est plus précisément une mesure sur les variations de la concentration d'énergie du spectre à court terme du signal :

$$CEL(n) = \frac{E_{[0,N-1]} - (E_{[0,\frac{N}{2}]} + E_{[\frac{N}{2},N-1]})}{|E_{[0,\frac{N}{2}]} + E_{[\frac{N}{2},N-1]}|};$$

où n est le milieu de la fenêtre $[0, N-1]$. La dépendance temporelle est obtenue en faisant glisser la fenêtre d'analyse point par point sur le signal. La séquence $\{ x_{[0,N-1]}(k)\}$ pour $0 \leq k \leq N-1$ sera considérée comme *entropiquement instable* et son milieu n constituera un point d'instabilité si :

$$CEL(n) > 0 \quad (\Leftrightarrow E_{[0,N-1]} > E_{[0,\frac{N}{2}]} + E_{[\frac{N}{2},N-1]}),$$

dans le cas contraire la séquence sera considérée comme *entropiquement stable* . Plusieurs changements dans le spectre à court terme du signal se traduisent par la génération de plusieurs points d'instabilité contigus constituant une *zone d'instabilité*. Il est fait l'hypothèse qu'une zone d'instabilité traduit un changement ou une rupture dans l'évolution du signal. Cette mesure permet de localiser les parties entropiquement stables du signal dans le bruit.

[Shen et al.98] définissent une entropie spectrale calculée sur le spectre normalisé considéré comme une densité de probabilité. La fonction densité de probabilité est définie par :

$$p_i = \frac{s(f_i)}{\sum_{k=1}^{N} s(f_k)}, \quad i = 1, ..., N,$$

où $s(f_i)$ est l'énergie spectrale pour la composante fréquentielle f_i et N est le nombre total de composantes fréquentielles dans la transformée de Fourier. Les auteurs restreignent les composantes fréquentielles entre 250 et 6000 Hz. Ainsi :

$$s(f_i) = 0, \text{ si } f_i < 250 \text{ Hz ou } f_i > 6000 \text{ Hz}.$$

De plus, ils restreignent la densité de probabilité :

$$p_i = 0, \text{ si } p_i < \delta_2 \text{ ou } p_i > \delta_1,$$

où les bornes δ_1 et δ_2 sont calculées empiriquement. δ_1 est utilisé pour éliminer le bruit contenu dans certaines bandes de fréquences spécifiques, tandis que δ_2 est utilisé pour oublier les bruits de valeur de densité spectrale presque constante, comme le bruit blanc. L'entropie spectrale est alors définie par :

$$H = -\sum_{k=1}^{N} p_k \log p_k.$$

Un ensemble de facteurs pondérants w_k est ensuite appliqué pour ajuster la composante fréquentielle à l'entropie spectrale. On a alors :

$$H = -\sum_{k=1}^{N} w_k p_k \log p_k.$$

L'ensemble des facteurs pondérants w_k est estimé statistiquement sur un grand nombre de signaux de parole. Cette dernière méthode est d'après les auteurs très performante dans les environnements bruités, pour la détection des débuts et fins de mots.

- **autres critères**

Le critère statistique du rapport de vraisemblance est utilisé par [Karray98a], pour discriminer les courbes représentatives des distributions de l'énergie du bruit et de la parole. Ceci entraîne une supposition forte, les deux distributions doivent être gaussiennes, ou laplaciennes. En notant, respectivement, H_0 et H_1, les hypothèses d'état dans le bruit et dans la parole, et en les supposant équiprobables, le seuil de passage d'un état à l'autre est obtenu par la résolution de l'équation :

$$P(x/H_0) = P(x/H_1),$$

où $P(x/H_i)$ est la probabilité conditionnelle de l'observation x sous l'hypothèse H_i. Cette équation est du second degré, dans le cas gaussien.

Les cumulants d'ordre supérieur 3 et 4 donnent des informations sur la symétrie et l'aplatissement du signal. Le cumulant d'odre 3, normalisé, (ou skewness) est nul pour une densité symétrique, c'est le cas des gaussiennes, pour une gaussienne le cumulant d'ordre 4, normalisé, (ou kurtosis) vaut 3. Les cumulants pouvant s'exprimer en fonction des moments, [Jacovitti et al.91] intègrent les moments normalisés dans un algorithme en vue d'une segmentation voisée/non-voisée/silence.

[Heon et al.98] utilise les réseaux neuronaux dans le domaine cepstral pour réduire le bruit.

Les performances des différentes approches de détection, seront améliorées d'autant plus que les caractéristiques de la parole utilisées représenteront le signal de départ. C'est pourquoi il est important de pouvoir améliorer ces caractéristiques.

5 Prétraitement pour l'amélioration des caractéristiques du signal

Pour augmenter la robustesse de la DBP, ou de la DAV, il peut être intéressant d'employer des techniques de prétraitement, qui améliorent les caractéristiques du signal. Ainsi la soustraction spectrale ou la soustraction cepstrale (cf [Karray98c]) ont donné de bonnes améliorations. L'utilisation de différents filtres (cf [Doukas et al.97]) permet également un débruitage du signal, par séparation de sources. Bien sûr l'emploi de plusieurs microphones (cf [Agaiby et al.97] et [Ney81]) permet d'exploiter un critère supplémentaire, la cohérence spatiale. On peut ainsi discriminer les différentes sources émettrices de sons. Cette technique n'est cependant pas applicable à tous les contextes, notamment, dans le cas de la téléphonie.

La soustraction spectrale

Certain auteurs ont montré que la soustraction spectrale est un bon prétraitement pour le débruitage. On considère le signal observé $x(t)$ comme un signal $s(t)$ dégradé dans le domaine temporel par un bruit additif $b(t)$, on a :

$$x(t) = s(t) + b(t).$$

On suppose que les signaux $s(t)$ et $b(t)$ sont stationnaires à l'ordre 2 et non-corrélés. On obtient la relation suivante sur les densités spectrales de puissance :

$$\gamma_x(f) = \gamma_s(f) + \gamma_b(f).$$

La trame de densité spectrale de la parole propre associée peut donc être estimée par :

$$\hat{\gamma}_s(f) = \hat{\gamma}_x(f) - \hat{\gamma}_b(f),$$

et donc

$$|\hat{s}(f)| = \sqrt{\hat{\gamma}_x(f) - \hat{\gamma}_b(f)}.$$

En pratique, $\hat{\gamma}_b(f)$ est calculée à partir des périodes de non parole dans le signal observé, et $\hat{\gamma}_x(f)$ est calculée de telle sorte que $\hat{\gamma}_x(f) - \hat{\gamma}_b(f)$ reste positif. Remarquons que l'hypothèse faite sur la stationnarité du bruit n'est plus toujours vérifiée, notamment pour les bruits dus à l'utilisation du réseau cellulaire. Pour d'avantage de précisions sur la soustraction spectrale, on se réfère à [Rabiner et al.93].

6 Méthodes de tests et études comparatives réalisées

La comparaison des méthodes de tests proposées par les auteurs des méthodes exposées est parfois délicate. En effet, la segmentation dépend beaucoup des conditions de la prise de son. Elle sera d'autant plus aisée que l'environnement sera peu bruité. Certains corpus utilisés ont été enregistrés dans un environnement calme, puis bruités par des corpus de bruit, voire par des bruits artificiels. Tandis que d'autres sont composés d'enregistrements dans différents environnements. De plus, il est préférable que le corpus servant pour les tests comprenne suffisamment de mots (prononcés par un grand nombre de personnes, hommes et femmes). Lorsqu'il s'agit de segmentation appliquée pour la reconnaissance, les comparaisons sont parfois faits directement sur les performances de la reconnaissance. Or ceci prend en considération d'autres erreurs dues à la reconnaissance, ou au contraire une correction des erreurs de segmentation par le système de reconnaissance. [Junqua et al.91] comparent trois algorithmes de segmentation à l'aide de deux systèmes de reconnaissance. Prenant deux algorithmes, l'un peut être plus robuste que l'autre appliqué à un système de reconnaissance, alors que ce sera le contraire s'ils sont utilisés par l'autre système de reconnaissance.

On cite ci-dessous quelques articles comparant différentes méthodes.

[Lamel et al.81], après implémentation de trois types d'algorithmes, un implicite, un explicite et un hybride, concluent que l'algorithme le plus performant est l'hybride. Notons tout de même que cet algorithme est celui des auteurs et qu'il reste très proche d'un algorithme de type explicite.

[Junqua et al.91] comparent ensuite les travaux de [Lamel et al.81], corrigés par [Reaves91] avec un algorithme de [Hamada et al.90], fondé sur l'information du pitch et deux nouveaux algorithmes. Cette étude étant toujours faite dans le cadre de la reconnaissance, deux systèmes de reconnaissance différents, son employés. Un système est fondé sur les chaînes de Markov cachées, l'autre est fondé sur les principes d'alignement temporel de formes acoustiques (DTW dynamic time warping) (cf [Jouvet88]). La comparaison est faite sur les performances de la reconnaissance, avec un corpus à SNR variable. L'étude est complétée par [Junqua et al.94], avec un cinquième algorithme fondé sur le taux de passage par zéro, et sur un paramètre temps-fréquence qui représente l'énergie dans la bande de fréquence 250-3500 Hz, et le logarithme de la racine carrée de la moyenne des carrés des observations de l'énergie, sur une fenêtre donnée. Ce dernier algorithme est le plus performant dans des environnements bruités.

[Van gerven et al.97] présentent trois méthodes différentes fondées sur le

logarithme de l'énergie. Les différences sont au niveau des seuils et du calcul de l'énergie à court terme et à long terme. Ces algorithmes sont plus ou bien adaptés selon les environnements.

Ces types d'algorithme présentent une approche similaire à celle employée dans [Karray98a] et a été comparé dans [Karray98b], aux algorithme utilisés au CNET.

[Andre obrecht et al.93] proposent trois approches :

Un algorithme fondé sur une analyse temporelle, évaluant les variations de l'abscisse curviligne calculée sur le signal de parole.

Un algorithme fondé sur une analyse fréquentielle paramétrique du signal, consistant à calculer les cumuls des variations d'énergie entre différentes bandes de fréquence.

Et un algorithme fréquentiel non paramétrique partant des résultats de la segmentation automatique pour calculer sur chaque segment, deux modèles AR gaussiens. La décision est prise par comparaison de la vraisemblance des deux modèles AR obtenus.

7 Le système de DBP du CNET

Le sytème de DBP du CNET a été créé en vue de l'appliquer à un système de reconnaissance. Pour plus de détails sur le système de reconnaissance du CNET on peut se reporter à [Jouvet88] et à [Sorin et al.95].

Le système de détection Bruit/Parole du CNET utilise un automate à cinq états, qui sont :

silence, présomption de parole, parole, plosive ou silence, et reprise possible de parole.

Description du fonctionnement de l'automate :

Dans une première version les passages d'un état à l'autre sont conditionnés par un seuil sur l'énergie du signal et par des contraintes structurelles de durée (durée minimum d'une voyelle et durée maximum d'une plosive). D'autres méthodes de passage d'un état à l'autre ont été étudiées, et sont présentées ci-dessous. Les passages à l'état *parole* déterminent les frontières de la parole dans le signal. Le système de reconnaissance prend en compte ces données avec une marge de sécurité sur les frontières.

L'état *silence* est l'état initial de l'algorithme. On fait ainsi l'hypothèse que la communication débute par du silence. Le détecteur reste dans cet état tant qu'il n'y a pas de trame énergétique (i.e. une trame dont l'énergie est supérieure au seuil). À la première trame énergétique, le

détecteur passe dans l'état *présomption de parole*. Dans cet état, une trame non énergétique le fait retourner à l'état *silence*. Après être resté un nombre de trames minimum dans l'état *présomption de parole*, le détecteur passe à l'état *parole*. Il y reste tant que les trames sont énergétiques. Il passera à l'état *plosive ou silence*, dès que la trame sera non énergétique. Il faut au moins un certain nombre de trames non énergétiques pour confirmer le silence et retourner dans l'état *silence*, sinon le détecteur passe dans l'état *reprise possible de parole*. Dans cet état, une trame non énergétique le fait retourner dans l'état *plosive ou silence* ou dans l'état *silence* si la durée cumulée du temps passé dans l'état *plosive ou silence* et dans l'état *reprise possible de parole* représente au moins un certain nombre de trames. Après être resté un nombre de trames énergétiques minimum dans l'état *reprise possible de parole*, le détecteur retourne dans l'état *parole*.

Cet algorithme est décrit en Fig. 1.

Pour plus de détails sur le fonctionnement de l'automate, on se réfère à [Mauuary94].

7.1 Critère de détection fondé sur le rapport signal à bruit

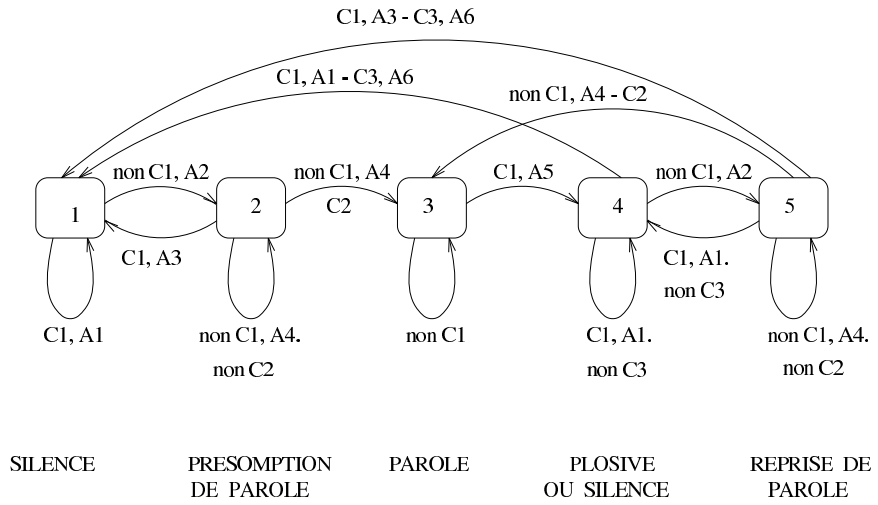
Le seuil sur l'énergie qui permet le passage d'un état à l'autre, est adaptatif. L'énergie intervient en terme de rapport signal à bruit du signal de parole observé. On cherche à comparer des estimations à court-terme et à long-terme de l'énergie du signal. L'énergie à court-terme est la moyenne, sur une fenêtre donnée, par exemple de Hamming, des valeurs du coefficient d'énergie. L'énergie à long-terme (ELT) est calculée, dans les périodes de silence, de façon récursive :

$$ELT(n+1) = ELT(n) + (1-\lambda)(energie - ELT(n)),$$

où λ est un facteur d'oubli (en général fixé à 0.99 pour cet automate), et n est le numéro de l'échantillon dans les périodes de silence. La différence entre l'énergie à long-terme et celle à court-terme est comparée à un seuil donné afin de décider de la présence ou non d'un segment à détecter.

7.2 Critère de détection fondé sur les statistiques du bruit

On fait l'hypothèse que le bruit suit une loi normale de paramètres (μ,σ^2). Les statistiques du bruit sont estimées à chaque passage dans l'état *silence*



CONDITIONS :

C1 : Energie < Seuil

C2 : Durée Parole (DP) >= Parole Minimum

C3 : Durée Silence (DS) >= Silence Fin

ACTIONS :

A1 : DS = DS +1

A2 : DP = 1

A3 : DS = DS + DP

A4 : DP = DP + 1

A5 : DS = 1

A6 : DS = DP = 0

VALEURS INITIALES :

ETAT = SILENCE

DS = DP = 0

Silence Fin est choisi entre 240 ms (mots isolés) et 640 ms (mots connectés),

Silence Fin représente le maximum entre la durée maximum d'une tenue de plosive (240 ms)

et la durée maximum d'une pause entre mots (640 ms).

FIG. 1 - *Automate de détection Bruit/Parole*

de l'automate. La moyenne est estimée par :

$$\mu(n+1) = \mu(n) + (1-\lambda)(energie - \mu(n)),$$

et l'écart-type par :

$$\sigma(n+1) = \sigma(n) + (1-\lambda)(|energie - \mu(n)| - \sigma(n)),$$

où n est le numéro de l'échantillon dans l'état *silence* de l'automate. Le facteur d'oubli λ a été optimisé empiriquement dans [Karray98c]. L'estimation de l'écart-type suppose que le bruit est modélisé par une loi laplacienne, au lieu de la loi gaussienne précédente. Cette hypothèse est faite pour simplifier l'estimation de l'écart-type. L'énergie de chaque trame est considérée, et on cherche à vérifier l'hypothèse que l'on est dans l'état *silence*, qui correspond au bruit seul. La décision sera prise en fonction de l'écart de l'énergie de cette trame par rapport à la moyenne estimée du bruit, c'est-à-dire selon la valeur du rapport critique $r(x) = \frac{x-\mu}{\sigma}$, comparé à un seuil. En prenant un intervalle de confiance de 95 %, ce rapport est comparé à 1.7. (cf [Karray98c] et [Karray98a])

7.3 Critère de détection fondé sur les statistiques du bruit et de la parole

Comme précédemment mentionné, cette approche découle d'une approche Bayésienne. On teste deux hypothèses :

-H_0 : on est dans un état de bruit

-H_1 : on est dans un état de parole (bruitée).

Pour chaque observation x, on cherche à comparer le maximum de vraisemblance $P(H_i/x)$ de chaque hypothèse. C'est-à-dire que le rapport de vraisemblance $r(x) = \frac{P(H_0/x)}{P(H_1/x)}$ est comparé à 1. Si $r(x) \leq 1$ la trame considérée sera alors déterminée comme étant de la parole bruitée, sinon elle sera du bruit. En supposant H_0 et H_1 équiprobables, on a $r(x) = \frac{P(x/H_0)}{P(x/H_1)}$. Les statistiques de la parole sont estimées dans l'état *parole*, de la même façon que pour les statistiques du bruit le sont dans l'état *silence*. Le facteur d'oubli pour la parole a été choisi à 0.95 et à 0.99 pour le bruit. Dans un premier temps on suppose que les deux distributions sont gaussiennes. On a, pour $i = 0, 1$:

$$P(x/H_i) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}}.$$

l'équation $r(x) = 1$ devient donc :

$$(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2})x^2 - 2\left(\frac{\mu_0}{\sigma_0^2} - \frac{\mu_1}{\sigma_1^2}\right)x + \frac{\mu_0^2}{\sigma_0^2} - \frac{\mu_1^2}{\sigma_1^2} + 2\log\frac{\sigma_0}{\sigma_1} = 0$$

Cette équation possède deux solutions, on choisit celle située la plus proche du milieu de $[\mu_0, \mu_1]$, s'il n'existe pas de solution dans cet intervalle, on prend le milieu de $[\mu_0, \mu_1]$. On remarque cependant qu'expérimentalement l'intersection des deux gaussiennes peut se trouver en dehors de cet intervalle. Ces points approchent assez bien, graphiquement, les points réels d'intersection des courbes représentatives de l'énergie. La deuxième solution (la plus grande), la moyenne du bruit étant plus faible que celle de la parole bruitée, permet de donner un seuil de sûreté pour la parole.

Pour simplifier cette équation, on fait l'hypothèse que les distributions sont laplaciennes. On a alors, pour $i = 0, 1$:

$$P(x/H_i) = \frac{1}{\sqrt{2}\sigma_i} e^{-\frac{\sqrt{2}|x-\mu_i|}{\sigma_i}}.$$

En faisant l'hypothèse que $x \in [\mu_0, \mu_1]$, l'équation $r(x) = 1$ devient :

$$\left(\frac{1}{\sigma_0} + \frac{1}{\sigma_1}\right) x = \frac{\mu_0}{\sigma_0} + \frac{\mu_1}{\sigma_1} - \frac{1}{\sqrt{2}} \log \frac{\sigma_0}{\sigma_1},$$

On vérifie expérimentalement, que l'on a bien en général, $x \in [\mu_0, \mu_1]$. Le seuil de décision est donc :

$$s = \frac{\frac{\mu_0}{\sigma_0} + \frac{\mu_1}{\sigma_1} - \frac{1}{\sqrt{2}} \log \frac{\sigma_0}{\sigma_1}}{\frac{1}{\sigma_0} + \frac{1}{\sigma_1}}.$$

Vérifions que $s \in [\mu_0, \mu_1]$, et que s est bien solution de l'équation $r(s) = 1$, on a trois cas possibles.

Premièrement, si $s < \mu_0$, on a aussi $s < \mu_1$, et en remplaçant s par sa valeur, on a :

$$\begin{cases} \sqrt{2}(\mu_0 - \mu_1) - \sigma_0 \log \frac{\sigma_0}{\sigma_1} < 0 \\ \sqrt{2}(\mu_1 - \mu_0) - \sigma_1 \log \frac{\sigma_0}{\sigma_1} < 0 \end{cases};$$

en résolvant l'équation $r(s) = 1$, on obtient :

$$\sigma_1 \log \frac{\sigma_0}{\sigma_1} = \sqrt{2}(\mu_1 - \mu_0),$$

ceci étant expérimentalement incorrect, car $\mu_0 < \mu_1$ et $\sigma_0 < \sigma_1$, d'après les valeurs expérimentales.

Deuxièmement, si $s \in [\mu_0, \mu_1]$, en remplaçant s par sa valeur, on a :

$$\begin{cases} \sqrt{2}(\mu_0 - \mu_1) - \sigma_0 \log \frac{\sigma_0}{\sigma_1} < 0 \\ \sqrt{2}(\mu_1 - \mu_0) - \sigma_1 \log \frac{\sigma_0}{\sigma_1} > 0 \end{cases};$$

dans ce cas s vérifie bien l'équation $r(s) = 1$, et cette condition équivaut à :

$$\mu_0 - \mu_1 < \frac{\sigma_0}{\sqrt{2}} \log \frac{\sigma_0}{\sigma_1},$$

qui est, en général, correcte expérimentalement.

Troisièmement, si $s > \mu_1$, on a aussi $s > \mu_0$, et en remplaçant s par sa valeur, on a :

$$\begin{cases} \sqrt{2}(\mu_0 - \mu_1) - \sigma_0 \log \frac{\sigma_0}{\sigma_1} > 0 \\ \sqrt{2}(\mu_1 - \mu_0) - \sigma_1 \log \frac{\sigma_0}{\sigma_1} > 0 \end{cases},$$

ce qui, en résolvant l'équation $r(s) = 1$, conduit à :

$$\sigma_0 \log \frac{\sigma_0}{\sigma_1} = \sqrt{2}(\mu_0 - \mu_1).$$

Cette condition est fausse en général, d'après les résultats expérimentaux. On a ainsi vérifié que les hypothèses faites sont correctes expérimentalement, et que la solution trouvée vérifie bien l'équation de départ. Cependant l'approximation de la distribution de la parole à une laplacienne est sans doute trop restrictive. Il est noté dans [Karray98a] que les résultats obtenus sont aussi bons en prenant simplement :

$$s = \mu_1 + \alpha(\mu_0 - \mu_1),$$

où α est un facteur d'interpolation ($0 < \alpha < 1$) qui est optimisé empiriquement. Cette approche est cependant moins robuste au changement de corpus et d'environements.

8 Conclusion et perspectives

On a ici présenté quelques méthodes pour la détection de parole dans le bruit. La liste est loin d'être exhaustive. Il n'y a pas à notre connaissance d'étude complète qui compare les différentes méthodes, à l'aide de mêmes corpus et, dans le cas de la reconnaissance, avec un même système de reconnaissance. De nombreuses caractéristiques du signal sont utilisées; il serait intéressant de considérer l'information des moments d'ordre supérieurs sur l'énergie ou l'information du pitch, par exemple, dans l'algorithme existant utilisé au CNET. Il sera aussi nécessaire de considérer les techniques de détection d'activité vocale du type de celles utilisées en transmission, par exemple dans le système GSM. Ceci fera l'objet d'un document ultérieur.

Références

[Abdallah et al.97a] Abdallah (I.), Montrésor (S.) et Baudry (M.). – Speech Signal Detection in Noisy Environment Using a Locoal Entropic Criterion. *European Conference on Speech Communication and Technology*, pp. 2595–2598. – Rhodes, Grèce, septembre 1997.

[Abdallah et al.97b] Abdallah (I.), Montrésor (S.) et Baudry (M.). – Un algorithme récursif pour la segmentation des signaux de parole basé sur un critère entropique local. $4^{ième}$ *Congrès de la Société Francaise d'Acoustique*, pp. 85–88. – avril 1997.

[Agaiby et al.97] Agaiby (H.) et Moir (T.J.). – Knowing the Wheat from the Weeds in Noisy Speech. *European Conference on Speech Communication and Technology*, pp. 1119–1122. – Rhodes, Grèce, septembre 1997.

[Andre obrecht et al.93] André-Obrecht (R) et Puel (J.B.). – Détection des débuts et fin de parole en environnement difficile. $14^{ième}$ *Colloque GRETSI*, pp. 157–160. – Juan-Les-Pins, septembre 1993.

[Arslan et al.98] Arslan (L.M.) et Hansen (J.H.L.). – Likelihood Decision Boundary Estimation Between HMM Pairs in Speech Recognition. *IEEE Transactions on Speech and Audio Processing*, vol. 6, n° 4, juillet 1998, pp. 410–414.

[Atal et al.76] Atal (B.S.) et Rabiner (L.R.). – A Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification with Applications to Speech Recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, n° 3, juin 1976, pp. 201–212.

[Bagshaw94] Bagshaw (P.). – *Automatic prosodic analysis for computer aided pronunciation teaching.* – Thèse de Doctorat, University of Edinburgh, 1994.

[Bartkova99] Bartkova (K.). – *Production, Description, Perception du signal vocal.* – Rapport technique, cours DEA, 1999.

[Braun et al.90] Braun (H.J.), Cosier (G.), Freeman (D), Gilloire (A.), Sereno (D.), Southcott (C.B.) et Van der Krogt (A.). – *Voice control of the Pan-European digital mobile radio system.* – Rapport technique n° 3, CSELT, juin 1990.

[Calliope89] Calliope. – *La parole et son traitement automatique.* – Masson, 1989.

[De souza83] De Souza (P.). – A Statistical Approach to the Design of an Adaptative Self-Normalizing Silence Detector. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 31, n° 3, juin 1983, pp. 678–684.

[Dermatas et al.91] Dermatas (E.S.), Fakotakis (N.D.) et Kokkinakis (G.K.). – Fast Endpoint Detection Algorithm For Isolated Word Recognition Detector. *International Conference on Acoustics, Speech, and Signal Processing*, mai 1991, pp. 733–736.

[Di francesco90] Di Francesco (R.J.). – Real-Time Speech Segmentation Using Pitch and Convexity Jump Models: Application to Variable Rate Speech Coding. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, n° 5, mai 1990, pp. 741–748.

[Doukas et al.97] Doukas (N.), Naylor (P.) et Stathaki (T). – Voice Activity Detection Using Source Separation Techniques. *European Conference on Speech Communication and Technology*, pp. 1099–1102. – Rhodes, Grèce, septembre 1997.

[Freeman et al.89] Freeman (D.K.), Cosier (G) et Southcott, C.B. Boyd (I.). – The Voice Activity Detector for the Pan-European Digital Cellular Mobile Telephone service. *International Conference on Acoustics, Speech, and Signal Processing*, mai 1989, pp. 369–373.

[Hamada et al.90] Hamada (M.), Takizawa (Y.) et Norimatsu (T.). – A Noise Robust Speech Recognition System. *International Conference on Spoken Language Processing*, pp. 893–896. – 1990.

[Heon et al.98] Héon (M.), Tolba (H.) et O'Shaughnessy (D.). – Robust Automatic Speech Recognition by the Application of a Temporal-correlation-based Recurrent Multilayer Neural Network to the Mel-based Cepstral Coefficients. *International Conference on Spoken Language Processing*, pp. 1459–1462. – Sydney, Australie, décembre 1998.

[Hess83] Hess (W.). – *Pitch Determination of Speech Signal.* – Springer-Verlag, 1983.

[Jacovitti et al.91] Jacovitti (G.), Pierucci (P.) et Falaschi (A.). – Speech Segmentation and Classification Using Higer Order Moments. *European Conference on Speech Communication and Technology*, pp. 1371–1374. – Gènes, Italie, septembre 1991.

[Jouvet88] Jouvet (D.). – *Reconnaissance de mots connectés indépendamment du locuteur par des méthodes statistiques.* – Thèse de Doctorat, Ecole Nationale Supérieure des Télécommunications, 1988.

[Junqua et al.91] Junqua (J-C), Reaves (B.) et Mak (B.). – A Study of Endpoint Detection Algorithms in Adverse Conditions: Incidence on a DTW and HMM Recognizer. *European Conference on Speech Communication and Technology*, pp. 1371–1374. – Gènes, Italie, septembre 1991.

[Junqua et al.94] Junqua (J-C), Mak (B.) et Reaves (B.). – A Robust Algorithm for Word Boundary Detection in the Presence of Noise. *IEEE Transactions on Speech and Audio Processing*, vol. 2, n° 3, juillet 1994, pp. 406–412.

[Karray98a] Karray (L.). – *Estimation des Statistiques du Bruit et de la Parole pour une Détection Bruit/Parole plus Robuste.* – Rapport technique n° 8, DT/DIH/DIPS/285, avril 1998.

[Karray98b] Karray (L.). – *Historique et Etat de l'Art des Techniques de Détection de la Parole et du Bruit.* – Rapport technique n° 16, DT/DIH/DIPS/549, juillet 1998.

[Karray98c] Karray (L.). – *Nouveau Critère pour l'Automate de Détection Bruit/Parole*. – Rapport technique n° 3, DT/DIH/DIPS/48, janvier 1998.

[Lamel et al.81] Lamel (L.F.), Rabiner (L.R.), Rosenberg (A.E.) et Wilpon (J.G.). – An Improved Endpoint Detector for Isolated Word Recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, n° 4, août 1981, pp. 777–785.

[Lynch et al.87] Lynch (J.F.), Josenhans (J.G.) et Crochiere (R.E.). – Speech/Silence Segmentation For Real-Time Coding Via Rule Based Adaptive Endpoint Detection. *International Conference on Acoustics, Speech, and Signal Processing*, avril 1987, pp. 1348–1351.

[Mauuary94] Mauuary (L.). – *Amélioration des performances des serveurs vocaux interactifs*. – Thèse de Doctorat, Université de Rennes 1, 1994.

[Ney81] Ney (H.). – An optimization algorithm for determining the endpoints of isolated utterances. *International Conference on Acoustics, Speech, and Signal Processing*, mars 1981, pp. 720–723.

[Rabiner et al.75] Rabiner (L.R.) et Sambur (M.R.). – An Algorithm for Determining the Endpoints of Isolated Utterances. *THE BELL SYSTEM TECHNICAL JOURNAL*, vol. 54, n° 2, février 1975, pp. 295–315.

[Rabiner et al.77a] Rabiner (L.R.) et Sambur (M.R.). – Application of an LPC Distance Measure to the Voiced-Unvoiced-Silence Detection Problem. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, n° 4, août 1977, pp. 338–343.

[Rabiner et al.77b] Rabiner (L.R.), Schmidt (C.E.) et Atal (B.S.). – Evaluation of a Statistical Approach to Voiced-Unvoiced-Silence Analysis for Telephone-Quality Speech. *THE BELL SYSTEM TECHNICAL JOURNAL*, vol. 56, n° 3, mars 1977, pp. 455–482.

[Rabiner et al.93] Rabiner (L.R.) et Juang (B.-H.). – *Fundamentals of Speech Recognition.* – Prentice Hall, 1993.

[Reaves91] Reaves (B.). – Comments on "An Improved Endpoint Detector for Isolated Word Recognition". *IEEE Transactions on Signal Processing*, vol. 39, n° 2, février 1991, pp. 526–527.

[Savoji89] Savoji (M.H.). – A Robust Algorithm for Accurate Endpointing of Speech Signals. *Speech Communications*, vol. 8, 1989, pp. 45–60.

[Shen et al.98] Shen (J-L.), Hung (J-W) et Lee (L-S). – Robust Entropy-based Endpoint Detection for Speech Recognition in Noisy Environments. *International Conference on Spoken Language Processing*, pp. 1015–1018. – Sydney, Australie, décembre 1998.

[Sorin et al.95] Sorin (C.), Jouvet (D.), Gagnoulet (C.), Dubois (D.), Sadek (D.) et Toularhoat (M.). – Operational and Experimental French Telecommunications Services Using CNET Speech Recognition and Text-To-Speech Synthesis. *Speech Communications*, vol. 17, n° 3, 1995, pp. 273–286.

[Takeda et al.95] Takeda (T.), Kuroiwa (S.), Nairo (M.) et Yamamoto (S.). – Top-Down Speech Detection and N-Best Meaning Search in a Voice Activated Telephone Extension System. *European Conference on Speech Communication and Technology*, pp. 1075–1078. – Madrid, Espagne, septembre 1995.

[Un et al.80] Un (C.K.) et Lee (H.H). – Voiced-Unvoiced-Silence Discrimination of Speech by Delta Modulation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, n° 4, août 1980, pp. 398–407.

[Van gerven et al.97] Van Gerven (S.) et Xie (F.). – A Comparative Study of Speech Detection Methods. *European Conference on Speech Communication and Technology*, pp. 1571–1574. – Rhodes, Grèce, septembre 1997.

[Watson et al.97] Watson (S.D.), Cheetham (B.M.G.), Barrett (P.A.), Wong (W.T.K.) et Lewis (A.V.). – A Voice Activity Detector for the ITU-T 8kbit/s Speech Coding Standard G.729. *European Conference on Speech Communication and Technology*, pp. 1571–1574. – Rhodes, Grèce, septembre 1997.