Revue de l'Université de Moncton

Un système à base de connaissances pour une communication parlée personne-système multilingue

Sid-Ahmed Selouani

La gestion de l'information
Volume 36, numéro 2, 2005

URI : id.erudit.org/iderudit/014499ar
DOI : 10.7202/014499ar

Aller au sommaire du numéro

Éditeur(s)

Revue de l'Université de Moncton

ISSN 0316-6368 (imprimé)
1712-2139 (numérique)

Découvrir la revue

Citer cet article

Selouani, S. (2005). Un système à base de connaissances pour une communication parlée personne-système multilingue. *Revue de l'Université de Moncton*, *36*(2), 53–84. doi:10.7202/014499ar

Résumé de l'article

La tâche de reconnaissance automatique de la parole (RAP), qui est au coeur de la communication parlée Personne-Système, peut être vue comme une gestion de l'information issue de la microstructure acoustique du signal vocal pour la transformer en une information représentée par la macrostructure phonétique implicite. La correspondance avec le moins d'erreurs possible de ces deux structures nécessite une intégration de connaissances a priori sur la macrostructure phonétique dans des systèmes dédiés à la gestion de l'information acoustico-phonétique. Dans cet article, nous abordons des aspects liés tant à la gestion de l'information phonétique véhiculée par le signal vocal qu'à la topologie de systèmes experts capables de conduire des processus de reconnaissance phonémique multilingue. La démarche que nous proposons consiste à enrichir la base de connaissances de ces experts par des indices représentatifs de la majorité des langues humaines afin de rehausser les performances d'identification des macro-classes et des traits phonétiques divers. Les résultats obtenus sur des corpus de logatomes et de phrases en langues française et arabe montrent qu'il est possible d'orienter la conception des systèmes vers une unification du processus de reconnaissance pour l'adapter à une identification phonémique multilingue.

Revue de l'Université de Moncton, vol. 36, n° 2, 2005, p.53-84.

UN SYSTÈME À BASE DE CONNAISSANCES POUR UNE COMMUNICATION PARLÉE PERSONNE-SYSTÈME MULTILINGUE

Sid-Ahmed Selouani
Université de Moncton

Résumé

La tâche de reconnaissance automatique de la parole (RAP), qui est au cœur de la communication parlée Personne-Système, peut être vue comme une gestion de l'information issue de la microstructure acoustique du signal vocal pour la transformer en une information représentée par la macrostructure phonétique implicite. La correspondance avec le moins d'erreurs possible de ces deux structures nécessite une intégration de connaissances *a priori* sur la macrostructure phonétique dans des systèmes dédiés à la gestion de l'information acoustico-phonétique. Dans cet article, nous abordons des aspects liés tant à la gestion de l'information phonétique véhiculée par le signal vocal qu'à la topologie de systèmes experts capables de conduire des processus de reconnaissance phonémique multilingue. La démarche que nous proposons consiste à enrichir la base de connaissances de ces experts par des indices représentatifs de la majorité des langues humaines afin de rehausser les performances d'identification des macro-classes et des traits phonétiques divers. Les résultats obtenus sur des corpus de logatomes et de phrases en langues française et arabe montrent qu'il est possible d'orienter la conception des systèmes vers une unification du processus de reconnaissance pour l'adapter à une identification phonémique multilingue.

Abstract

Automatic Speech Recognition (ASR) is at the heart of Man-Machine speech communication. It can be seen as a management of the information emanating from the speech

acoustical microstructure. This process aims to transform this information in such a way that it can be represented by the phonetic implicit macrostructure. The effective matching between the two structures requires the integration into expert systems, of an *a priori* knowledge about the phonetic macrostructures. These expert systems are dedicated to the management of acoustic-phonetic information. This paper investigates aspects linked either to the management of phonetic information contained in the speech signal, or to the topology of expert systems that are capable of conducting a multilingual phonemic recognition process. The proposed method consists of feeding the knowledge base of these expert systems with indicative parameters representing the major human languages in order to enhance the identification performance of phonetic macro-classes and features. The results of experiments carried out on corpora composed of both French and Arabic utterances show that it is possible to conceive systems based on the concept of unified recognition processes dedicated to multilingual phonetic identification.

Introduction

La communication parlée Personne-Système constitue un défi technologique majeur qui vise par le développement d'interfaces interactives basées sur des agents conversationnels intelligents, une accessibilité naturelle à l'aide de la voix à divers systèmes d'information (centre d'appel, portails Internet vocalisés, dispositifs de communication mobile, etc.). Des efforts très importants sont consentis dans des domaines tels que la reconnaissance automatique de la parole, la synthèse vocale ainsi que l'identification du locuteur et du langage. Des progrès notables ont été réalisés grâce au développement fulgurant des moyens de calculs, qui ont offert la possibilité de traiter de grandes quantités de données et de combiner des techniques variées et complexes (Oviatt, 2002). Une gestion efficace de l'information complexe véhiculée par le signal vocal humain et du flux de données généré par le traitement de ce signal, permet d'envisager aujourd'hui la réalisation de systèmes à coût raisonnable, fonctionnant en un temps proche du temps réel et dont les performances

sont jugées satisfaisantes pour bon nombre d'applications (Deng et Huang, 2004).

Dans la perspective de l'intégration des différentes langues dans un environnement multimédia et multimodal, qui s'étend grâce au réseau Internet à l'échelle de la planète, il devient primordial de ramener les systèmes de communication Personne-Système au même niveau de performance pour toutes les langues, quelles que soient leurs particularités. Paradoxalement, au moment où l'on observe un regain d'intérêt pour les systèmes multilingues tels que ceux préconisés dans le projet C-STAR III (C-STAR III, 2003), et une explosion de ce qui est appelé « *les industries langagières* », nous observons une quasi-absence de produits basés sur une approche multilingue. Ceci est certainement dû à la grande diversité phonétique et phonologique à laquelle doit faire face tout système dédié à un fonctionnement multilingue (Saijayaram, Ramasubramanian et Sreenivas, 2003). En effet, il serait aisé de mesurer cette difficulté en considérant, par exemple, le cas d'un système d'analyse-reconnaissance dont l'apprentissage est effectué pour une langue donnée (exemple : latine), mais dont on voudrait étendre l'utilisation à une autre langue cible ayant par exemple, des paramètres prosodiques sémantiquement pertinents (mandarin, serbo-croate, tchèque, arabe, hébreu, etc.). Dans la perspective d'une reconnaissance phonémique multilingue par un système unique, nous pouvons citer les travaux réalisés dans le cadre du projet Globalphone (Globalphone, 2000) visant la portabilité vers une langue cible en utilisant une modélisation acoustico-phonétique globale sur des corpus multilingues. Cette modélisation recèle de nombreux avantages dont l'optimisation de l'utilisation de ressources linguistiques, en exploitant le fait qu'il existe de nombreuses caractéristiques phonétiques communes dans les langues du monde. Une utilisation efficace d'un tel système requiert que le processus de reconnaissance soit robuste aux changements intempestifs, délibérés ou involontaires du rythme d'élocution, tout en tenant compte des spécificités phonétiques et prosodiques de chaque langue. En effet, il faudra considérer le fait que la normalisation temporelle, nécessaire dans le cas de beaucoup de langues (le français ou l'anglais par exemple), devienne préjudiciable pour d'autres (arabe, hébreu, mandarin, etc.). Les approches visant la portabilité se basent sur le fait que le processus de reconnaissance ne peut être conduit simplement par une gestion statistique

des données. Il doit tenir compte également de l'information qui permet une caractérisation et donc une modélisation plus précise des phonèmes. Dans ce contexte, l'expérience a montré qu'un système markovien très performant dans la normalisation temporelle a des capacités de discrimination limitées (Jelinek, 1997). De plus, comme il est mentionné par Hasegawa-Johnson *et al.* (2005), les systèmes basés sur les modèles de Markov cachés sont très vulnérables aux interférences acoustiques, aux variations de style et requièrent des ressources linguistiques très importantes pour leur apprentissage. Plusieurs alternatives aux modèles de Markov ont été proposées comme celles qui considèrent des systèmes de classification automatique basés sur les réseaux de neurones formels, dont l'aspect discriminant constitue la caractéristique dominante. Cependant ceux-ci restent incapables d'effectuer une normalisation temporelle (Takuya et Shuji 1994; Haton, 1995).

Dans cet article, nous proposons d'utiliser, dans le contexte d'une application de reconnaissance phonémique multilingue, un système basé sur les connaissances formalisant le savoir et le savoir-faire de l'expert l'humain sous forme de règles exhaustives et explicites. Le raisonnement de l'expert humain est disséqué et toutes ses tâches d'inférence, d'induction ou de déduction sont intégrées dans des procédures, protocoles, etc. (Allen, 1994). Nous traitons plus explicitement de la question de l'identification des traits et macro-classes phonétiques par les systèmes basés sur les connaissances conçus selon le modèle des réseaux d'états finis et intégrant diverses sources de connaissances acoustiques, articulatoires et phonétiques sous forme de règles. Une fusion entre le savoir et le savoir-faire représentant respectivement les faits acoustiques et les règles de production, y est effectuée. Sur le plan de la réalisation technique, nous avons opté pour une modularité des réseaux d'états finis afin de pouvoir leur assigner des sous-tâches de classification spécifiques qui tiennent compte des particularités phonétiques. Bien que la division de la tâche globale d'identification en sous-tâches conduites par des experts spécifiques apporte le risque d'aboutir à des solutions sous-optimales (Jacobs, 1995; Waterhouse et Cook, 1996), nous pensons que dans le cas qui nous intéresse, à savoir l'identification des traits et des macro-classes phonétiques, nous pouvons déboucher, avec l'utilisation d'une stratégie de fusion adéquate, sur des configurations avantageuses pouvant couvrir la majorité des particularités des langues ciblées. Cette approche « à base de

règles » gagnerait à être intégrée à celle du « tout automatique » qui prévaut actuellement dans toutes les solutions présentes sur le marché, où l'intervention de l'humain se réduit à enrichir et à diversifier le plus possible des corpus d'apprentissage afin d'atteindre des performances de généralisation acceptables. Notons que l'idée d'une hybridation entre méthodes statistiques et systèmes à base de règles devient de plus en plus populaire (Hasegawa-Johnson *et al.*, 2005; Yu et Waibel, 2004), dans la mesure où malgré leur prédominance sur le marché depuis bientôt trente années, les systèmes basés sur les modèles markoviens ne sont pas encore parvenus à réaliser le rêve du dialogue naturel avec la machine.

Dans cet article, nous décrirons tout d'abord les sources d'information utilisées par la base de connaissance des experts. Nous détaillerons par la suite la configuration des systèmes experts sous forme de réseaux d'états finis et la base de règles qu'ils utilisent dans leur processus d'identification. Nous présenterons ensuite, la solution choisie pour l'implémentation des réseaux et nous terminerons par une discussion des résultats expérimentaux obtenus.

Constitution de la base de connaissances

Les systèmes experts que nous proposons permettent de prendre en compte des sources de connaissances diverses et évolutives. Ils permettent également de modéliser plusieurs niveaux d'abstraction, hiérarchisés ou non. Le principe de base de ces systèmes consiste à retarder la décision en faveur d'une solution à un problème donné tant qu'il n'y a pas assez de connaissances pour l'inférer. Des systèmes basés sur ce principe ont été proposés dans la littérature. Parmi eux, mentionnons : les systèmes « multi-experts » (DeMori, Lam et Gilloux, 1987), les systèmes « sociétés de spécialistes » (Gong, 1988) et le système « multi-opérateurs » (Spalanzani et Selouani 1999). Pour tous ces systèmes, les spécialistes sont des entités de calcul quelconques permettant de résoudre une partie d'un problème d'optimisation.

Le système que nous présentons est composé d'experts structurés en réseaux d'états finis à base de règles et de connaissances organisé autour d'un module de segmentation en phones homogènes. Ces derniers sont des unités infra-phonémiques qui présentent des caractéristiques stables

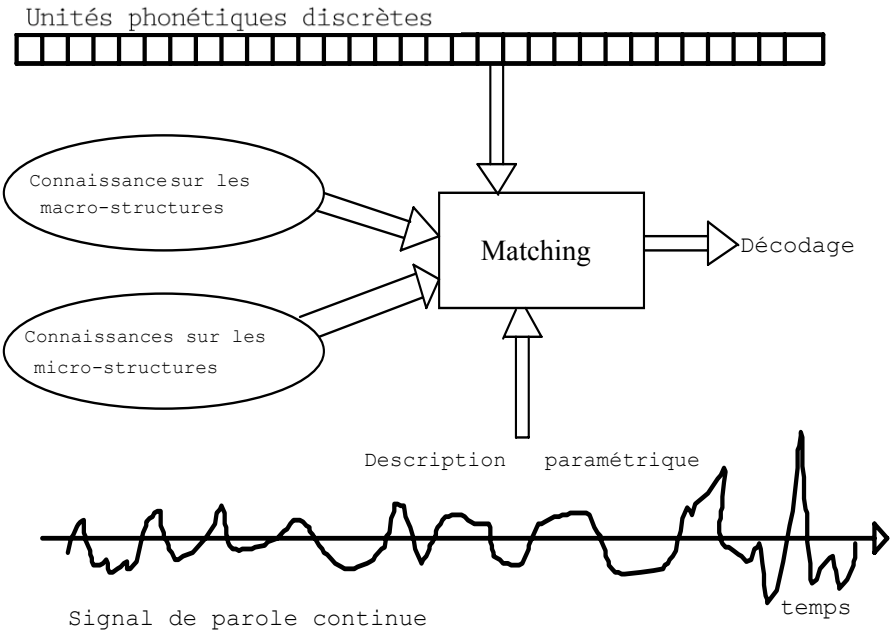
sur le plan acoustique. Le choix du phone comme unité de base est dicté par les considérations suivantes :

- une segmentation infra-phonémique est plus avantageuse dans la mesure où l'on s'attend à un nombre très faible d'omissions. En effet, partant du fait qu'un phonème est la concaténation de plusieurs phones qui correspondent à des phases acoustiques distinctes, une segmentation infra-phonémique est assurée de détecter au moins un phone dans un phonème. De plus, au stade de l'étiquetage, une redondance de l'information est souvent souhaitable car elle renforce la robustesse de l'identification;
- théoriquement, les phones jouissent d'une stabilité prononcée des différents paramètres et indices acoustiques. Ceci facilitera la pose des marqueurs qui les délimitent.

Il faut cependant relever le fait que le phénomène de co-articulation est mal représenté par les unités infra-phonémiques. Pour pallier cet inconvénient, le phone est considéré dans la majorité des cas avec son contexte (gauche et droit) et en incorporant dans la base de connaissances des paramètres différentiels.

Cette approche est une généralisation de la technique proposée par Caelen et Tattegrain (1988). Celle-ci consiste telle que schématisée en figure 1, à mettre en correspondance (*matching*) la microstructure acoustique du signal et la macrostructure phonétique au moyen de leurs modèles sous-jacents respectifs. En utilisant les experts et en manipulant une base de connaissances explicites, nous sommes amenés à définir des modèles phonétiques sous forme de réseaux phonétiques. Les transitions entre les états sont formulées à l'aide de règles adéquates. Le *matching* revient à parcourir convenablement les réseaux. Après segmentation, chaque segment est étiqueté par un ou plusieurs attributs phonétiques (traits, macro-classes, phonème). Le treillis phonétique retenu correspond aux meilleurs candidats issus du décodage. Le cheminement dans tous les réseaux phonétiques est guidé par des contraintes acoustiques, phonétiques et linguistiques. Dans cette stratégie entièrement ascendante, les hypothèses émises sont suffisamment larges pour ne pas remettre en question le niveau de décision précédent. Ceci évitera le difficile problème de *backtracking* (retour en arrière) qui est le propre des systèmes qui raisonnent sur des connaissances incomplètes.

Figure 1 : Processus de décodage interne conduit par un expert



1. La paramétrisation

La parole est un signal réel, continu, d'énergie finie et non stationnaire. Sa structure est complexe et variable dans le temps. Il peut être pseudopériodique, comme pour les sons voisés, ou aléatoires comme pour les sons fricatifs, ou encore pulsionnel comme c'est le cas pour les plosives. Considérant la redondance et la complexité qui caractérisent ce signal, les techniques d'analyse du signal vocal ont toujours visé un double objectif : extraire des paramètres pertinents et stables mais dans un espace de représentation réduit. Les paramètres acoustiques utilisés sont calculés sur des trames de 10 ms et comprennent l'énergie ainsi que le taux de passage par zéro. La fréquence fondamentale est estimée et corrigée par la technique de l'ambiguïté modifiée. Cette technique se base sur la détermination de régularités (périodicités) sur le signal vocal non-

linéairement codé dans le domaine temporel et fréquentiel (Selouani et O'Shaughnessy, 2002).

L'extraction des formants est effectuée sur le spectre lissé obtenu après une modélisation spectrale autorégressive du signal (O'Shaughnessy, 2001). Un modèle auditif est utilisé dans notre système pour extraire les indices acoustiques.

2. Les indices acoustiques auditifs statiques

Le modèle (auditif) de cochlée utilisé dans notre système a été établi par Caelen (1979) pour la partie périphérique de l'oreille (oreille externe, moyenne et interne). Il a été intégré avec succès dans une configuration récente (Tolba, Selouani et O'Shaughnessy, 2005). Il s'agit d'un modèle passif comprenant deux étages : une préaccentuation et un banc de filtres couplés. Ce banc modélise la vibration de la membrane basilaire. Il est obtenu à partir des équations différentielles décrivant son mouvement. Le tableau 1 indique pour une fréquence d'échantillonnage de 16 KHz, les fréquences centrales des 24 canaux constituant le banc de filtres.

Tableau 1 : Fréquences centrales des canaux modélisant la membrane basilaire

Canal	Fréquence/Hz	Canal	Fréquence/Hz	canal	Fréquence/Hz
1	180	9	760	17	2350
2	215	10	880	18	2700
3	260	11	1000	19	3100
4	320	12	1130	20	3550
5	380	13	1340	21	4000
6	450	14	1550	22	4500
7	540	15	1790	23	5000
8	650	16	2060	24	5600

L'attrait qu'a toujours constitué la caractérisation des phonèmes par des traits acoustiques a conduit les chercheurs à élaborer des indices acoustiques comparables aux indices formels utilisés par les phonéticiens. Rappelons que Jakobson, Fant et Halle dans leurs *Preliminaries* (1963), après une étude exhaustive des langues les plus parlées dans le monde, ont conclu que douze indices acoustiques sont suffisants pour caractériser acoustiquement toutes ces langues. Cependant, ils précisent qu'il n'est pas nécessaire, pour la majorité des langues, d'utiliser les douze indices. Le modèle auditif calculatoire de Caelen a permis de quantifier ces indices acoustiques, ce qui a ouvert la voie à leur incorporation dans de nombreuses configurations pratiques (Caelen, 1985; Selouani, Tolba et O'Shaughnessy, 2003).

Dans notre application, chaque indice acoustique est exprimé en dB et est calculé par une combinaison linéaire particulière des énergies de sortie des 24 canaux. Nous avons retenu les indices suivants : Aigu/Grave (AG), Fermé/Ouvert (FO), Tendu/Lâche (TL), Écarté/Compact (EC), Bémolisé/Diésé (BD), Doux/Strident (DS), Continu/Discontinu (CD), Nasal (NZ) et le Noyau vocalique. L'énergie de sortie d'un canal *i* sera notée *Wci.* Les indices acoustiques sont calculés sur chaque trame et leurs caractéristiques sont les suivantes :

- Indice Aigu/Grave (AG) : du point de vue articulatoire, la gravité d'un phonème est générée par un volume plus important de la cavité buccale. Dans notre cas, l'indice aigu/grave sera mesuré par la différence d'énergie entre les basses fréquences (50-400 Hz) et les hautes fréquences (3800-6000 Hz), l'indice sera donc négatif si le signal est aigu et positif si le signal est grave :

$$AG=W_{c1}+W_{c2}+W_{c3}+W_{c4}+W_{c5}-W_{c20}-W_{c21}-W_{c22}-W_{c23}-W_{c24}$$

- Indice Fermé/Ouvert (FO) : sur le plan articulatoire, cet indice caractérise un rétrécissement ou une ouverture en un point d'articulation donné; ce qui réduit ou bien accentue la contribution des cavités situées derrière le lieu d'articulation. Sur le plan acoustique, le phonème sera considéré comme fermé si l'énergie des très basses fréquences (230-350 Hz) est importante par rapport à l'énergie des moyennes fréquences (600-800 Hz), l'indice sera donc négatif si le signal est fermé et positif si le signal est ouvert :

$FO=W_{c8}+W_{c9}-W_{c3}-W_{c4}$

- Indice Tendu/Lâche (TL) : acoustiquement, cet indice estime la différence d'énergie entre les hautes fréquences (2700-5000 Hz) et les fréquences qui s'étendent entre (1000-2060 Hz). L'indice sera donc négatif si le signal est lâche et positif si le signal est tendu :

$$TL=W_{c11}+W_{c12}+W_{c13}+W_{c14}+W_{c15}+W_{c16}-W_{c18}-W_{c19}-W_{c20}-W_{c21}-W_{c22}-W_{c23}$$

Indice Bémolisé/Diésé (BD) : le signal sera considéré comme diésé si les hautes fréquences (2200-3300 Hz) sont plus importantes que les moyennes fréquences (1900-2900 Hz) :

$BD=W_{c17}+W_{c18}+W_{c19}-W_{c11}-W_{c12}-W_{c13}$

- Indice Écarté/Compact (EC) : les phonèmes compacts se caractérisent par la prédominance de la région formantique centrale. Un signal sera considéré comme compact si les moyennes fréquences (800-1050 Hz) sont plus importantes que les fréquences les entourant (300-700 Hz et 1450-2550 Hz). L'indice sera donc négatif si le signal est écarté et positif si le signal est compact :

$$EC=W_{c10}+W_{c11}-(W_{c4}+W_{c5}+W_{c6}+W_{c7}+W_{c8}+W_{c13}+W_{c14}+W_{c15}+W_{c16}+W_{c17})/5$$

- Indice Doux/Strident (DS) : les phonèmes stridents sont principalement caractérisés par un bruit qui est dû à une turbulence située au niveau du point d'articulation. Le signal sera considéré comme strident si les très hautes fréquences (3800-5300 Hz) sont plus importantes que les hautes fréquences (1900-2900 Hz). L'indice sera donc négatif si le signal est doux et positif si le signal est strident :

$$DS=W_{c21}+W_{c22}+W_{c23}-W_{c16}-W_{c17}-W_{c18}$$

- Indice Continu/Discontinu (CD) : cet indice mesure la variation du spectre par rapport au spectre de la trame précédente. Il sera faible s'ils sont semblables et important s'ils sont très différents. Le signal sera qualifié respectivement de continu et de discontinu. L'algorithme qui suit décrit les étapes de calcul de cet indice :

```
Wm = 0          // variable de travail
```

Pour i=1 à 24 Faire

Wm = Wm + $W_{ci}(t)$ $W_{ci}(t-1)$ // t: trame courante, t-1: trame précédente

FinPour

$W_m = W_m/24$ //Wm : énergie moyenne des 24 canaux du spectre actuel

CD =0

Pour i=1 à 24 Faire

CD = CD + | $W_{ci}(t)$ -$W_{ci}(t-1)$ - W_m |

FinPour

La base de connaissances utilise des indices supplémentaires dits de deuxième niveau. Il s'agit des indices nasal et vocalique. Le calcul de ces indices ne se suffit pas de l'énergie des 24 canaux modélisant la cochlée. Ils sont quantifiés par des calculs logiques et algorithmiques et se réfèrent à des patrons phonétiques simples.

- Indice Nasal : Cet indice est calculé en s'inspirant de la caractérisation théorique de la nasalité établie par Rossi, Nishinuma et Mercier (1983). Celle-ci consiste à déterminer le nombre de maxima d'énergie dans la bande fréquentielle [1300 Hz-2800 Hz] qui correspond dans notre cas aux canaux 13, 14, 15, 16, 17, et 18. Cette mesure est sous-tendue par le fait que la nasalité se caractérise par l'ajout de formants et d'anti-formants. Dans notre cas, nous ajoutons pour caractériser les consonnes nasales, la détection d'un formant de nasalité en basse fréquence (formant nasal grave) en dessous de 400 Hz.
- Indice vocalique/non vocalique : il se traduit la présence (resp. l'absence) d'une structure formantique nette produite par l'excitation (resp. sans excitation) au niveau de la glotte. Cet indice sera quantifié par le comptage du nombre de formants par phone.
- Noyau vocalique : le noyau vocalique (Noy_Voc) est un paramètre important; il sera calculé linéairement en sommant les énergies des canaux centraux du spectre :

$$\text{Noy\_Voc} = \sum_{i=7}^{19} \text{Wc}_i$$

3. Les indices et paramètres dynamiques

L'importance du phénomène de coarticulation rend le codage des variations du signal capital pour la reconnaissance vocale. La prise en compte des variations spectrales permet une amélioration appréciable du taux de reconnaissance. D'autre part, les expériences sur l'identification des syllabes ont montré que les zones de transition correspondent à des « points perceptifs critiques » et sont nécessaires à l'identification des syllabes. Il a également été montré que les zones stationnaires et les zones de pauses sont moins cruciales pour cette identification (Nagarajan et Murthy, 2004).

Notre système introduit une information sur la dynamique temporelle et spectrale du signal en utilisant les coefficients différentiels du premier et deuxième ordre des indices acoustiques et autres paramètres (énergie, TPZ). Ces coefficients appelés coefficients delta (1er ordre) et coefficients detla-delta (2ème ordre) peuvent être calculés par simple différence de vecteurs. Notons enfin que pour favoriser la détection des variations brutales dans la structure spectrale, l'indice Continu/Discontinu est tout a fait indiqué et joue de ce fait un rôle de première importance dans la base de connaissances du système. Dans la majorité des systèmes de reconnaissance, le calcul des coefficients différentiels nécessite la détermination du nombre optimal de trames de lissage. Dans notre cas, le nombre de trames de lissage est optimisé automatiquement car le calcul des coefficients différentiels est effectué sur les phones après la phase de segmentation. Les phones sont par hypothèse constitués de trames en nombre variable mais acoustiquement homogènes. Le lissage y est donc intrinsèquement effectué de manière optimale.

4. Élaboration de la base de règles

Globalement, les sons peuvent être subdivisés grossièrement en macro-classes phonétiques dont les principales sont : les voyelles, les fricatives sourdes et sonores, les consonnes liquides, les consonnes nasales, les

occlusives sourdes et les occlusives sonores. La tâche assignée au module d'identification consiste tout d'abord à opérer une classification grossière en macro-classes phonétiques. Une classification plus fine est opérée ensuite pour tenter de déceler le trait d'emphase et/ou de gémination sur toutes les macro-classes.

Description des experts et des règles de transition

À chaque macro-classe est associé un réseau phonétique ou expert représentant la connaissance sur la macro-classe. Celui-ci est appliqué indépendamment sur la suite de phones homogènes. Un réseau est constitué d'un ensemble d'états et d'un ensemble de transitions. Les états représentent toutes les réalisations possibles des différentes phases acoustiques des macro-classes phonétiques. À chaque transition (ou arc) est associée une liste de contraintes (règles) à vérifier, une liste d'actions à effectuer en cas de succès et enfin un score à chaque passage par la transition. Un phone peut être étiqueté par un ou plusieurs réseaux, comme il peut être rejeté par tous. La figure 2 donne un exemple de cheminement dans le réseau des voyelles. Dans ce cas, si pour le phone courant, un 'établissement' est observé, on tentera d'associer au phone suivant la même phase ou bien les phases 'demie tenue' ou 'tenue orale'. On opère de la même manière pour les phones suivants en tenant compte des passages permis dans le réseau, et ce, jusqu'à en sortir en cas de solution. Plusieurs étiquetages sont possibles grâce au processus de retour en arrière. Seule la solution présentant le score le plus élevé sera validée.

Un réseau phonétique (expert) noté Rj est défini par le 5-tuplet :

$$R_j = \{j, S(j), T, s_{oj}, s_{ej}\}$$

Avec j : identificateur du réseau; S(j) : ensemble des états possibles; T: ensemble des transitions; s_{oj} : état initial; s_{ej} : état final.

S(j) représente toutes les réalisations possibles des phases acoustiques pour une macro-classe donnée. Cet ensemble représente aussi toutes les possibilités d'étiquetage offertes.

Chaque transition notée ti est définie par l'ensemble suivant :

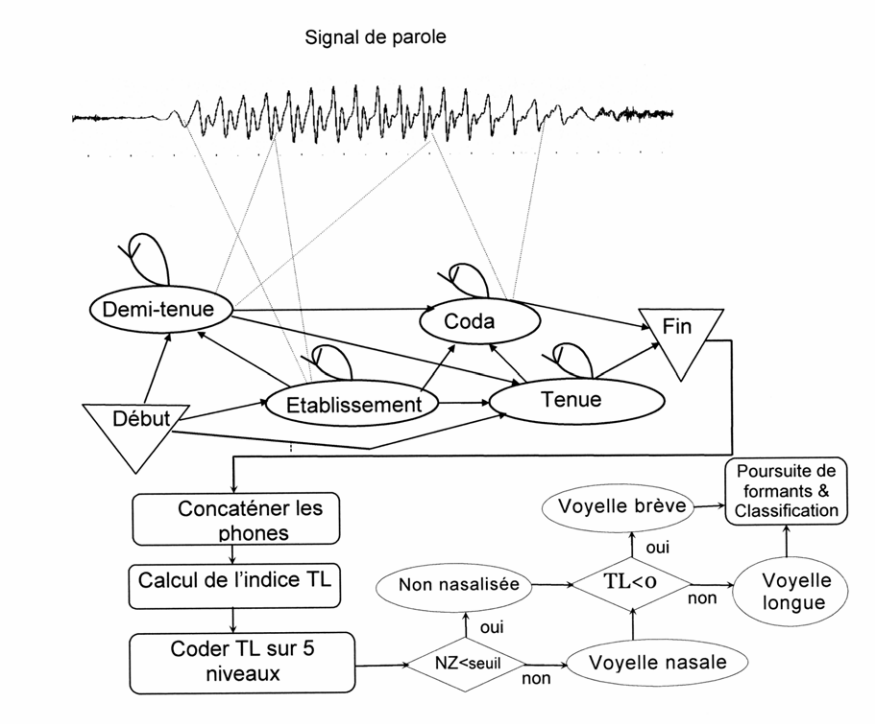
ti = {sk, sl, pi, Ci, Ai}

Avec sk et sl: extrémités de l'arc; pi: score réalisé une fois la transition effectuée; Ci: l'ensemble des contraintes devant être vérifiées avant

d'opérer la transition; Ai : l'ensemble des actions devant être exécutées dans le cas d'une sortie réussie. Celles-ci sont des procédures qui évaluent des prédicats, calculent des paramètres ou déclenchent d'autres règles prioritaires.

Figure 2 : Expert identifiant la macro-classe voyelle et effectuant la discrimination

Voyelle longue/Voyelle brève



Les contraintes, quant à elles, peuvent être subdivisées en trois catégories, à savoir : les conditions de réalisation d'un phone donné; les contraintes relatives au contexte précédent et les conditions générées par le contrôleur du réseau lors de l'exploration de celui-ci. Notons que lors de la progression en profondeur dans le réseau, il est proposé des

hypothèses suffisamment larges afin d'éviter l'annulation de la décision prise au niveau antérieur.

Le parcours d'un réseau phonétique consiste à associer le phone courant à une des phases constituant le réseau en tenant compte des conditions acoustiques de sa réalisation et des décisions prises précédemment. Si aucune phase ne lui est assignée, le parcours d'un autre réseau est enclenché. Quant au passage d'un phone à un autre, il est régi par les règles de production où les contraintes correspondent aux prémisses et les actions aux conclusions. Une fois l'étiquetage du phone voisin effectué, il est alors assemblé avec les phones voisins pour constituer le phonème correspondant à la macro-classe représentée par le réseau considéré.

Le réseau des voyelles

Les voyelles lorsqu'elles se réalisent normalement sont généralement caractérisées par la phase d'établissement où une croissance importante d'énergie est observée. L'amplitude du signal doit dépasser un seuil préalablement fixé. Ce dépassement doit être observée pendant une durée relativement longue afin de différencier cet état de la consonne voisée. Le nombre de pics spectraux doit y être relativement élevé. Afin d'assurer plus de robustesse à l'identification de la phase d'établissement, le parcours dans le réseau est effectué en utilisant des règles mettant en jeu le paramètre de durée, le nombre de formants, l'indice CD ainsi que la vérification de la présence de l'indice vocalique dans le contexte précédent. Ce parcours à base de règles est illustré en figure 3.

La tenue constitue dans la voyelle, l'état le plus caractéristique. Ceci justifie le fait que le score assigné soit relativement élevé (il est de deux) lors de la réalisation de cet état. Ce dernier est caractérisé par : une énergie globale importante, particulièrement au milieu du spectre; une structure formantique nette et stable et une durée très longue pendant laquelle le signal dépasse un seuil empirique. La demie-tenue est analogue à la tenue avec la différence que les caractéristiques qui caractérisent cette phase sont moins marquées que celles de la tenue. La coda décrit la phase de détente de la voyelle. Elle est caractérisée par une décroissance de l'énergie. La durée pendant laquelle l'amplitude du signal dépasse une certaine valeur doit être courte afin d'éviter un bouclage dans cet état.

Figure 3 : Parcours dans un phone produisant un établissement de voyelle

```
SI (Delta_sur_phone(Énergie)>ESO
                        ET état précédent = 'None')
    OU (pente(Énergie)>0
                        ET contexte précédent = 'Vocalique'
                        ET État précédent 'Néant')
        OU (État précédent = 'Établissement Voyelle'
                        ET indice(CD)< '-'
                        ET Delta_on_phone(Energy)>ESO)
    ET Durée (ou le signal>THV)>THl
    ET Energie >=B+3*(S/B)/4
    ET ΣDurées <0.5*Rate*MDV
    ET Nombre (formants)>2
ALORS État= 'Établissement Voyelle';score=score+1
    ET limite de début par défaut =@ phone courant
    ET action procédurale (règle Établissement Voyelle
        OU règle Tenue Voyelle
        OU règle Demi-Tenue Voyelle)
FIN
```

S/B : rapport signal sur bruit ; B étant le niveau de bruit en dB
ESO : valeur minimale de la pente d'énergie
THV, TH1 : seuils sur la durée en fonction du débit d'élocution
MDV : durée moyenne d'une voyelle courte

Nous devons noter la présence, dans le système vocalique français par exemple, du trait de nasalité alors que ce trait est absent dans le système vocalique arabe (Boé et Tubach, 1986; Boudraa, Selouani, Boudraa et Guérin, 1994). La détection automatique du trait de nasalité permet ainsi de rendre le système capable d'identifier les voyelles nasales si elles existent (ce qui est le cas dans beaucoup de langues) sans pour autant réduire son efficacité dans le cas ou ce trait est absent.

La concaténation de tous les phones étiquetés est opérée conformément aux scores obtenus par les différents états phonétiques. Une fois la macro-classe voyelle délimitée, l'indice nasal est déterminé par comparaison à un seuil du nombre de formants et anti-formants. Si ce nombre dépasse le seuil alors la voyelle est déclarée nasale. Par la suite, la

moyenne de l'indice tendu/lâche (TL) est calculée sur les phones constituant la voyelle. Un des cinq niveaux de codage non linéaire ('-', '--', '0', '+', '++') est assigné à la moyenne brute de l'indice calculé. Si le code assigné dépasse '0', cela signifie que la voyelle est tendue et elle sera étiquetée 'longue'. Dans le cas contraire, elle sera déclarée 'brève'.

Le réseau des fricatives

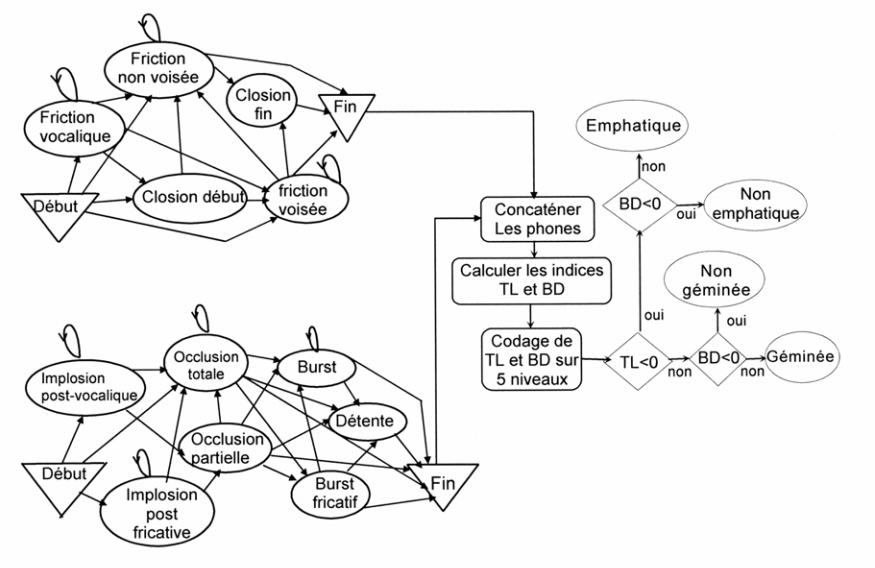
Le réseau des fricatives représenté en Figure 4, effectue l'identification des fricatives voisées et non voisées. La gémination et l'emphase sont également détectées par ce réseau s'ils sont présents. La friction vocalique, 'cloison début', la friction sonore et la friction sourde et enfin 'cloison fin' sont les états possibles parcourus par ce réseau. L'état friction vocalique décrit la phase de début d'une fricative. Dans cet état, l'ordre de grandeur de l'énergie doit varier dans une certaine gamme et le signal doit être aigu. La cloison de début produit une décroissance d'énergie suivie par une courte stabilisation. La friction sonore est caractérisée par la présence de la fréquence fondamentale (F0) et d'une structure formantique contrairement à une friction sourde. La cloison de fin est un état similaire à la cloison de début mais apparaît en fin de fricative. Quand la macro-classe fricative sourde ou sonore est délimitée, la moyenne des indices tendu/lâche et bémolisé/diésé est calculée afin de déceler les traits de gémination et d'emphase respectivement (Bonnot, 1979). Une valeur parmi les cinq niveaux de codage non linéaire est assignée à chacun des indices. Dans le cas où le code tendu est supérieur à '0', la fricative sera étiquetée comme étant une géminée. Si le code de l'indice bémolisé est supérieur à '0' alors la fricative sera considérée comme emphatique.

Le réseau des plosives

Comme indiqué en Figure 4, la règle de début du réseau des plosives permet dans un premier temps de fixer le contexte phonétique puis d'entrer ensuite dans le réseau par l'un des états suivants : implosion post vocalique, occlusion totale ou implosion post fricative. L'état implosion post vocalique décrit l'implosion d'une occlusive sourde après la réalisation d'une voyelle. La pente de l'énergie doit être élevée. La durée de cet état doit être en deçà d'une certaine limite. Cette dernière contrainte

sert à prévenir une inclusion de phases contiguës. La phase d'occlusion totale se présente quand le conduit vocal est complètement obstrué.

Figure 4 : Experts phonétiques pour les fricatives et les occlusives (plosives)



Par conséquent l'énergie est très basse. Il faut s'assurer cependant, que la durée de cet état soit courte afin de ne pas le confondre avec la pause. Le réseau des plosives doit inclure l'explosion (*burst*). Celle-ci correspond à l'explosion observée sur le signal lors de l'ouverture brusque de la cavité buccale. Le *burst* est essentiellement caractérisé soit par un pic d'énergie, soit par une croissance d'énergie accompagnée d'une acuité prononcée du signal. Une discontinuité importante peut être observée. Par ailleurs, l'énergie doit être circonscrite entre deux valeurs de manière à éviter une fausse détection qui générerait une occlusion partielle. Une contrainte sur la durée est établie de façon à prévenir un bouclage excessif dans cet état. Pour le *burst* fricatif, une condition supplémentaire concernant l'indice de friction est prévue. Quant à la friction post vocalique, celle-ci est similaire à l'implosion post vocalique mais avec une pente d'énergie plus faible.

L'occlusion partielle est détectée en vérifiant les mêmes conditions que celles de l'occlusion totale, excepté le fait que l'énergie soit plus importante. Cette phase reflète la réalisation incomplète de l'occlusion qui est due soit à un effet de voisement, soit à une friction héritée de l'état précédent. Notons que le débit d'élocution est un élément important qui est pris en compte lors de l'identification de cet état. Lors de la phase de détente, une importante discontinuité de l'énergie est attendue car cet état décrit aussi le début de la voyelle contiguë. Par conséquent, la durée doit être courte et l'indice vocalique positif. Comme pour les fricatives, la gémination et l'emphase sont respectivement détectées au moyen des indices non linéairement codés tendu/lâche et bémolisé/diésé.

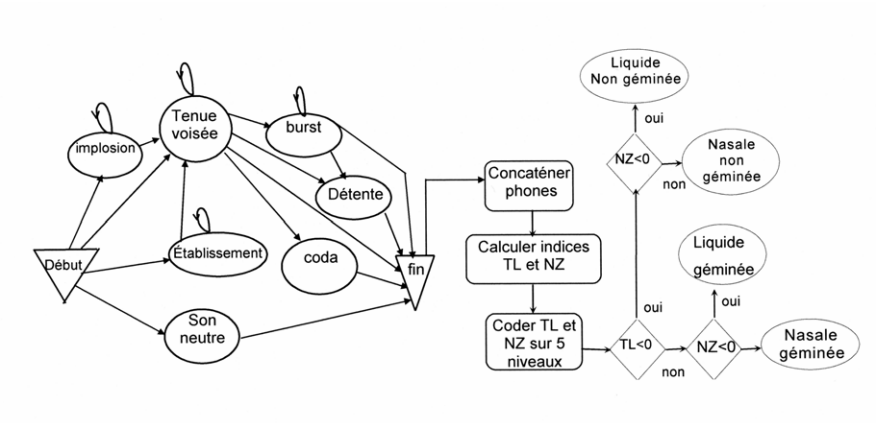
Le réseau des liquides et des nasales

Les nasales et les liquides sont identifiées par le même réseau : le réseau des consonnes (voir Figure 5) mais la discrimination finale est réalisée par l'indice nasal. Le réseau est entièrement parcouru si les conditions et contraintes de ces états sont vérifiées : implosion consonantique, établissement consonantique, tenue voisée et le son neutre. Une implosion consonantique survient après une voyelle (contexte précédent). Par conséquent, l'énergie décroît mais reste à un niveau suffisamment élevé. La structure formantique est préservée. Contrairement au cas précédent, l'établissement consonantique survient après une pause. Ainsi, la pente d'énergie est positive et la structure formantique n'apparaît pas clairement. La tenue voisée consonantique est quant à elle caractérisée par la réalisation complète de la structure formantique et sa stabilisation. Le noyau vocalique et la moyenne de l'énergie varient en dessous d'un seuil préalablement établi. L'indice friction est négatif et une structure harmonique commence à apparaître ($F0 \neq 0$). Comme cet état est très caractéristique le score est incrémenté de 2. Le *burst* consonantique est identifié si une contrainte liée au nombre de phones est vérifiée. Il est aussi caractérisé par un pic dû à une explosion énergétique ou par la croissance des amplitudes des composantes hautes fréquences. L'état détente consonantique survient si le contexte subséquent est une voyelle. Une croissance du paramètre énergie est observée et l'indice vocalique est positif. Dans le cas où le contexte subséquent est une pause, une coda est réalisée. Ce dernier état décrit la phase terminale d'une consonne avant une pause.

Le son neutre est un état particulier qui reflète l'ambiguïté caractérisant les liquides et les nasales dans un contexte vocalique. Dans des conditions particulières, la structure formantique, la fréquence fondamentale et les valeurs de l'énergie et du noyau vocalique peuvent être similaires à celles caractérisant les voyelles. Les liquides et les nasales fondent dans le contexte vocalique et perdent leurs caractéristiques. Le son neutre est prévu afin de limiter cette confusion.

Quand la macro-classe consonne (liquide et nasale) est délimitée et identifiée, la moyenne des indices acoustiques nasal et tendu est calculée sur les phones constituant la consonne. Dans le cas où l'indice nasal est supérieur à '0', la consonne est étiquetée 'nasale' sinon elle sera étiquetée comme étant une 'liquide'. La gémination est détectée grâce à l'indice tendu/lâche.

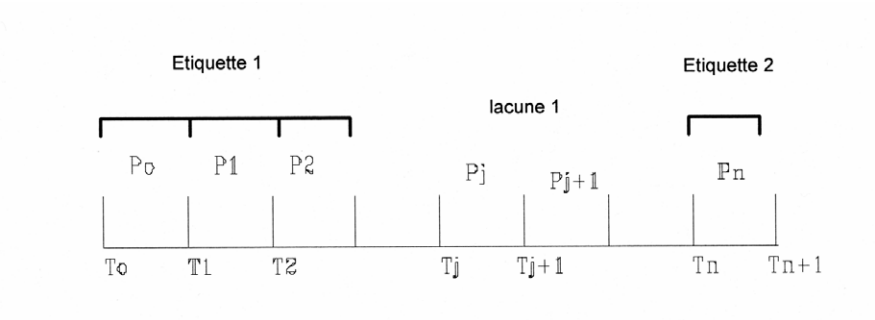
Figure 5 : Expert consonne regroupant les macro-classes nasale et liquide



5. Localisation des solutions d'identification et implémentation

Les réseaux phonétiques (experts) seront appliqués indépendamment sur la suite de phones qui constituent le continuum du signal vocal. Un phone peut être étiqueté par un ou plusieurs réseaux comme il peut être rejeté par tous. La figure 6 illustre le processus global de génération d'une solution d'identification.

Figure 6 : Le processus de génération de l'identification



Soit la suite de phone allant de P*o* jusqu'à P*n*. À l'instant T*o*, on rentre dans les réseaux par le phone P*o*. Deux situations peuvent se présenter pour un réseau donné :

- le processus conduit à la sortie du réseau avec une solution à l'instant T*j* avec $j \leq n$. La suite de phones comprise entre P*o* et P*j* constitue un phonème associé à la macro-classe représentée par ce réseau. Le processus est réinitialisé à l'instant $j+1$ si $j < n+1$;
- le processus se bloque dans le réseau, le phone courant soit P*o* n'est pas reconnu. Le processus est réinitialisé à l'instant T_1. P*o* étant rejeté par ce réseau, il peut être reconnu par un autre réseau.

À la fin du processus, nous aurons les résultats suivants :

- une alternance de portions reconnues et de portions non reconnues pour chacun des réseaux; une portion étant composée d'un ou de plusieurs phones;

- un treillis de solutions pouvant se recouvrir pour l'ensemble des réseaux avec un éventuel rejet collectif d'une portion. Ceci conduit à une configuration lacunaire du processus d'identification du signal.

Après la phase de segmentation, le signal vocal est perçu comme une séquence de phones homogènes, chacun constituant une phase acoustique (Caelen et Tattegrain, 1988). Le nombre de phones n'est pas connu au préalable et l'étiquetage d'un phone de rang *N* ne se fait que si les *N-1* phones qui le précédent ont été étiquetés. L'accès se fait donc séquentiellement. Cette organisation séquentielle du signal vocal, justifie l'utilisation d'une liste linéaire chaînée pour la modélisation. Ainsi, chaque maillon de la liste représente un phone et contient les informations relatives à celui-ci. La liste linéaire chaînée est bidirectionnelle afin de permettre le retour arrière lors de l'exploration du réseau en profondeur. Les règles du réseau sont de la forme :

SI (condition $_i$) ALORS (Action $_i$ et Note_passage_réseau = Note_passage_réseau + Bonus $_i$).

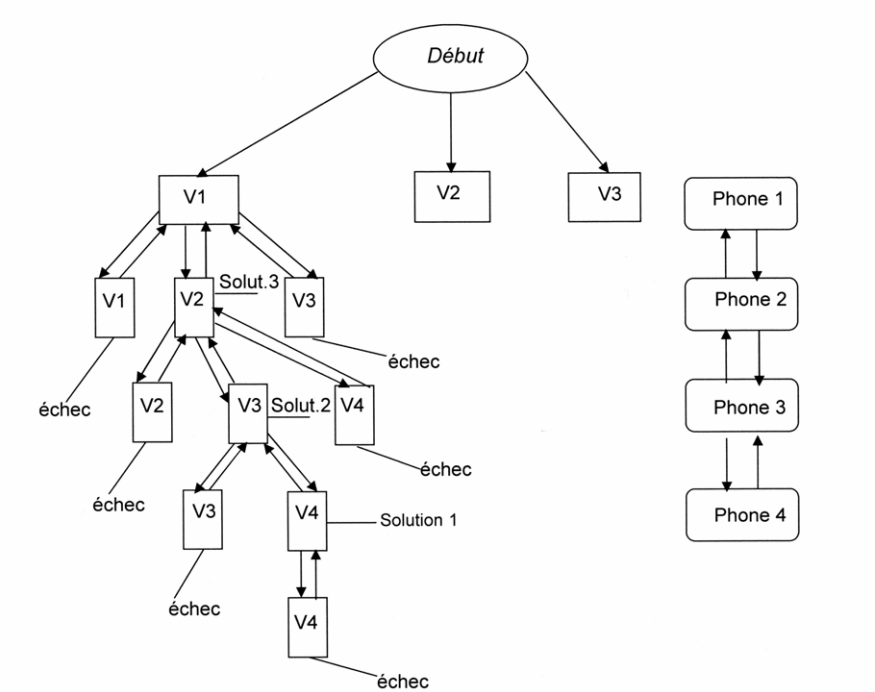
L'évaluation de la *condition* $_i$ renvoie à des calculs de fonctions et de prédicats. Si la condition est vérifiée alors l'action correspondant à cette condition est activée. Pour une séquence de phones (composant une voyelle) constituée des phases acoustiques d'établissement, de demie-tenue, de tenue orale et de coda, le cheminement dans le réseau des voyelles est illustré en Figure 7. Les actions V1, V2 et V3 vont elles-mêmes activer d'autres actions. Nous développons le réseau en profondeur avec la solution V1. L'action V2 ne sera envisagée que lorsque toutes les solutions de l'action V1 seront épuisées. De même, l'action V3 ne sera envisagée que si toutes les solutions de l'action V2 seront épuisées. Lorsque toutes les solutions de l'action V3 sont épuisées, on remontera à la règle Début**,** qui activera la règle V2 (demie-tenue).

Les solutions trouvées par le réseau des voyelles pour la séquence précédente de phones sont :

- Solution 1 avec un score de 2;
- Solution 2 avec un score de 4;
- Solution 3 avec un score de 5.

La solution retenue est la solution 3, car elle représente le score le plus élevé, avec comme phases acoustiques : établissement; demie-tenue; tenue-orale; coda-voyelle. On recommencera le processus de recherche dans le réseau à partir du cinquième phone, et ce, jusqu'à épuisement de tous les phones.

Figure 7 : Implémentation de l'expert Voyelle



6. Évaluation

Les performances du système sont évaluées en cherchant pour chaque phonème étiqueté le nombre de solutions des macro-classes correspondantes correctement localisées. Par ailleurs, les solutions proposées sont considérées plus robustes si leurs durées de couverture du phonème étiqueté sont plus longues.

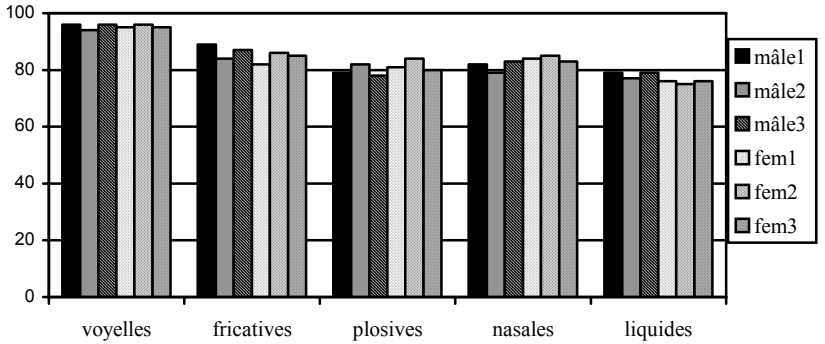
Le corpus de test a été prononcé par 6 locuteurs (3 hommes et 3 femmes). Ces mêmes locuteurs ont participé à l'apprentissage. Les stimuli sont constitués de 40 occurrences (logatomes) VCV et de 20 phrases : 10 phrases en langue arabe et 10 phrases en langue française où les fréquences d'apparition des phonèmes sont respectées. Au total, le corpus de test est composé de 852 voyelles (332 longues et 520 brèves), 384 fricatives, 248 plosives, 164 nasales et 168 liquides. Les semi-voyelles sont assimilées aux voyelles qui leur correspondent. Une séquence additionnelle de 108 occurrences VCV dont la consonne est une géminée a été ajoutée au test. Le choix de tenter la détection de la gémination indépendamment des autres consonnes est délibéré, dans la mesure où phonétiquement il n'existe pas de consonnes géminées (les utiliser dans le corpus général déséquilibrera phonétiquement celui-ci). Le nombre des consonnes emphatiques (fricatives et plosives) testées est de 104.

Il est intéressant de connaître pour chaque phonème étiqueté si la macro-classe correspondante est détectée à la bonne place, par rapport à l'étiquetage manuel. Ceci permettra d'évaluer l'efficacité d'identification. Pour ce faire, nous avons défini le taux de recouvrement qui est calculé de la façon suivante :

Taux=longueur $(Z \cap Z^*)$/minimum(longueur (Z), longueur (Z*)), Z* étant la localisation de solution trouvée par le système et Z la localisation du phonème étiqueté manuellement. Un résultat sera jugé robuste si le taux de recouvrement est supérieur à 50 %.

Dans une première étape, le système global opère une détection des macro-classes. Les résultats obtenus par locuteur sont donnés en Figure 8. Le taux moyen d'identification correcte est de 96 % pour les voyelles si l'on exige uniquement l'étiquetage en macro-classe. Dans une seconde étape, l'identification complète des voyelles est opérée au moyen d'un algorithme de poursuite de formants ainsi que la discrimination longue/brève. Les fricatives et les plosives (sans discrimination d'emphase ni de gémination) sont reconnues avec des taux moyens respectifs de 87 % et 81 %.

Figure 8 : Scores moyens par locuteur



Les résultats montrent que les problèmes viennent surtout des consonnes liquides dont le taux moyen obtenu est de 75 %. Ceci était prévisible car la variabilité de ce type de consonne dépend fortement du contexte. De la même façon, les nasales semblent mal répondre à une description en phases acoustiques bien que celle-ci favorise la détection d'événements brefs. Plus particulièrement, c'est la consonne nasale /m/ dont le niveau énergétique est relativement élevé qui entraîne le plus souvent une confusion avec la classe des voyelles. Les nasales dont détectées avec un taux moyen de 78 %. Certaines fricatives sonores sont prises pour des sourdes. En effet, les fricatives voisées sont faibles énergétiquement ce qui induit une grande difficulté dans l'estimation de la fréquence du fondamentale. Une analyse des résultats par genre (sexe) du locuteur montre que grossièrement les résultats sont équivalents. Nous noterons cependant, une légère supériorité dans les performances obtenues par les locuteurs féminins, notamment dans le cas des voyelles, fricatives et les liquides.

Il est important de signaler que les réseaux qui ont opéré sur les deux corpus sont identiques et que seuls les seuils changent en fonction du locuteur (après apprentissage). Le taux moyen de détection correcte des macro-classes est de 81 %. Les fricatives emphatiques sont décelées avec

un taux de succès moyen de 81 %. Le modèle d'audition pour le calcul des indices acoustiques a permis de déceler la gémination, trait éminemment complexe, avec un taux de succès de 77 %. La distinction voyelle longue-brève est réalisée avec succès dans 79 % des cas. Grâce à un algorithme de poursuite des formants (sur spectre LPC), le système opère une classification totale des voyelles avec un taux moyen de 89 %. Ce dernier taux est calculé en supposant que le système ne fait pas d'erreur dans le cas où la voyelle simple est détectée ou sa correspondante longue. Les résultats sont présentés sous forme de matrice de confusion par genre de locuteur.

La matrice de confusion est une matrice dont la première colonne, représente la voyelle à reconnaître et la première ligne, la voyelle avec laquelle elle est confondue. Les chiffres apparaissant dans cette matrice représentent le nombre de phonèmes pour lesquels cette voyelle est confondue avec une autre. La dernière colonne (ND), représente le nombre de cas qui n'ont pu être identifiés, à cause de l'effet de coarticulation qui influe directement sur les valeurs des formants.

Tableau 2 : Matrice de confusion des voyelles

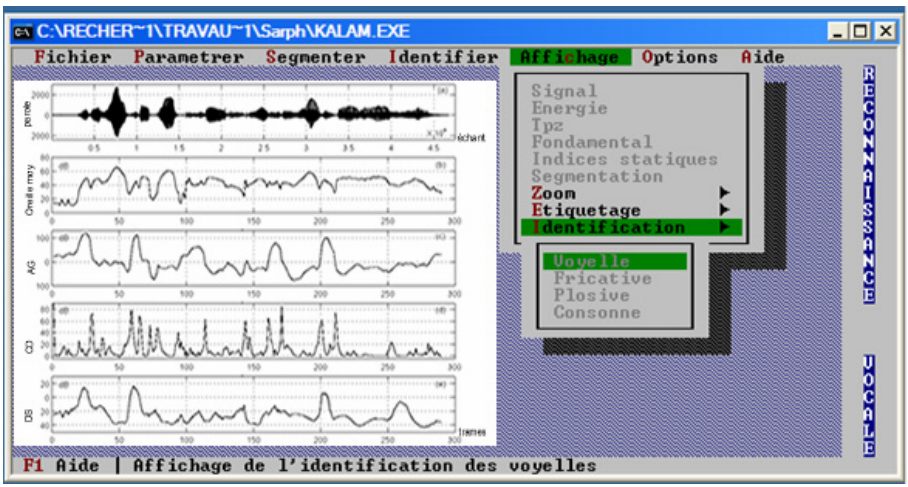
	a	aa	i	ii	u	uu	ND
a	181	30	3	0	7	0	13
aa	24	91	0	1	0	4	0
i	4	1	110	24	8	2	7
ii	0	7	25	71	1	5	0
u	3	0	9	0	97	21	4
uu	1	5	0	2	25	69	1

La confusion des voyelles brèves avec les voyelles longues (par exemple un /a/ est pris pour un /aa/ ou le contraire) est causée par l'élocution propre à chaque locuteur. Il est clair qu'un locuteur ne peut pas normaliser sa voix. La confusion d'une voyelle avec d'autres voyelles est la conséquence du problème de coarticulation causé principalement par les consonnes emphatiques ou les géminées, qui influent sur les voyelles voisines (abaissement du premier formant et augmentation du second, ou

rapprochement des deux premiers formants) et rendant ainsi la déduction d'un modèle de référence très difficile. Notons que les expériences présentées ici ne rendent pas compte des résultats obtenus pour les voyelles nasales françaises car le corpus utilisé n'en contient pas en nombre suffisant pour pouvoir déboucher sur des résultats significatifs. Par ailleurs, signalons que le taux de recouvrement moyen est de 86 % pour les deux sexes. Ceci nous conduit à conclure que ces taux sont largement satisfaisants pour attester de la correspondance phonème/classe et que les seuils utilisés dans les réseaux sont robustes.

Nous prévoyons dans une nouvelle série d'expériences sur les langues anglaise et française analyser plus finement le comportement de l'expert voyelle dans le cas des diphtongues en général, et des voyelles nasales en particulier.

Figure 9 : Capture d'images du système à base de connaissances



Conclusion

Dans cet article, nous avons abordé la problématique posée par l'identification dans un même système à base de règles, de phonèmes et de traits phonétiques complexes de différentes langues. Le développement de tels systèmes visant l'unification de la modélisation acoustico-

phonétique dans une perspective d'application multilingue recèle de nombreux avantages tels que : (i) la réduction de la complexité des systèmes grâce au partage des modèles et de la base de connaissances; (ii) l'intégration de fait de la fonctionnalité d'identification du langage; (iii) la possibilité d'adapter les modèles à une langue cible non apprise par le système.

Cependant, la trop grande diversité des réalisations phonétiques et phonologiques des différentes langues pose des défis quant à l'élaboration du système à base de connaissances dédié à l'identification des traits et macro-classes phonétiques. L'approche que nous avons proposée se caractérise par le fait que son efficacité à traiter un langage donné ne dépend que des règles et des informations acoustico-phonétiques et phonologiques intégrées dans la base de connaissances des experts. Par ailleurs, l'expérimentation du système à base de connaissances nous a permis d'établir trois faits concernant les systèmes vocalique et consonantique qui peuvent s'appliquer à différentes langues, à savoir :

- l'indice tendu/lâche permet d'arriver à une discrimination entre voyelle longue et brève. Cette caractéristique très importante pour les langues sémitiques sur le plan sémantique;
- l'emphase est décelée efficacement au moyen de l'indice bémolisé/diésé;
- la pertinence de l'indice tendu/lâche pour distinguer les consonnes géminées de leurs correspondantes simples;
- la possibilité de déceler le trait de nasalité sur les voyelles. Cette caractéristique est très importante pour la langue française mais non pertinente pour la langue arabe ou la langue anglaise.

Ceci confirme le fait que la machine est capable de *percevoir* les traits phonétiques complexes de plusieurs langues, à condition de la doter d'*oreilles* assez fines (le modèle d'audition) et de lui transmettre le savoir-faire nécessaire sous forme de règles de production.

Il faut relever le fait que la base de connaissances utilise des seuils empiriques qui dépendent des conditions d'expérience. Les tâches de normalisation au locuteur, au rapport signal sur bruit, au débit, etc., sont très ardues et nécessitent une gestion draconienne d'un nombre important

de paramètres et de seuils qui compliquent l'utilisation en temps réel du système. S'il est vrai que l'apport du système soit indéniable du fait qu'il constitue un moyen de valider des hypothèses phonétiques permettant de caractériser plusieurs langues, il reste encore sur le plan pratique (notamment en reconnaissance automatique) à relever le défi qui consiste à concevoir une base de connaissances couvrant toutes les particularités phonétiques des différentes langues. En traitant le phénomène de la parole, nous touchons à un fait qui caractérise la raison de l'être humain et qui constitue l'expression la plus éclatante de son intelligence. De ce fait, hormis les considérations techniques qui restent essentielles, les aspects informationnels, phonologiques, cognitifs, culturels, et même philosophiques doivent être pris en compte si l'on veut éviter l'échec qui a été celui de la vision purement « techniciste » qui a prévalu jusque-là. Ceci justifie le caractère pluridisciplinaire et diversifié des questions qui nous préoccupent ainsi que des objectifs que nous désirons atteindre dans des travaux futurs.

Bibliographie

Allen, J.B. (1994). How do humans process and recognize speech ? *IEEE transactions on speech and audio processing*. 2:4.567-577.

Boé, L.J., et Tubach, J.P. (1986). Des matrices phonétiques aux matrices phonologiques et vice versa. *Bulletin de l'institut phonétique de Grenoble*.15.135-155.

Bonnot, J.F. (1979). Étude expérimentale de certains aspects de la gémination et de l'emphase en arabe. *Travaux de l'institut phonétique de Strasbourg*. 11.109-118.

Boudraa, B., Selouani, S.A., Boudraa, M., et Guérin, B. (1994). Matrices phonétiques et matrices phonologiques arabes. *Actes des XXèmes JEP*. 345-350. Trégastel. France.

Caelen J. (1979). *Un modèle d'oreille, analyse de la parole continue, reconnaissance phonémique*. Thèse de doctorat d'état de l'Université de Toulouse.

Caelen, J. (1985). Space/time data-information in the A.R.I.A.L Project Ear model. *Speech communications*. 4.163-179.

Caelen, J., et Tattegrain, H. (1988). Le décodeur acoustico-phonétique dans le projet DIRA. *Actes des XIIèmes JEP*. 115-121. Nancy. France.

C-STAR III (2003). *Consortium for Speech Translation Advanced Research*. [http://www.c-star.org/]. Consulté le 10 avril 2006.

DeMori, R., Lam, L., et Gilloux, M. (1987). Learning and plan refinement in a knowledge based system for automatic speech recognition. *Transactions IEEE-PAMI*.9:2.289-305.

Deng, L., et Huang X. (2004). Challenges in adopting speech recognition. *Communications of the ACM*. 47:1.69-75.

Globalphone (2000). *The Globalphone Project for multilingual speech recognition and understanding systems*. Consulté le 12 avril 2006. [http://www.cs.cmu.edu/~tanja/GlobalPhone/].

Gong, Y. (1988). *Interprétation des signaux incertains*. Thèse de 3e cycle de l'Université de Nancy.

Hasegawa-Johnson, M., Baker, J., *et al.* (2005). Landmark-based speech recognition : report of the 2004 Johns Hopkins summer workshop. *IEEE International Conference on Audio Speech and Signal Processing*.1:213-216.

Haton, J.-P. (1995). Modèles neuronaux et hybrides en reconnaissance de la parole : état des recherches. *In* Méloni H. (dir.). *Fondements et perspectives en traitement automatique de la parole*. Paris : AUPELF-UREF.139-154.

Jacobs, R.A. (1995). Methods for combining experts probability assessments. *Neural computation*.7:5.867-888.

Jakobson, R., Fant, G.M., et Halle, M. (1963). *Preliminaries to speech analysis : the distinctive features and their correlates*. Cambridge : MIT press.

Jelinek, F. (1997). *Statistical methods for speech recognition*. Cambridge : MIT press.

Nagarajan, T., et Murthy H. A. (2004). Language identification using parallel Syllable-like unit recognition. *IEEE International Conference on Audio Speech and Signal Processing*. 1:401-404. Montréal, Canada.

O'Shaughnessy, D. (2001). *Speech Communication : Human and Machine.* IEEE Press.

Oviatt, S. (2002). Breaking the robustness barrier : recent progress on the design of the robust multimodal systems. *In* Zelkowitz, M. (dir.). *Advances in computers*. San Diego, CA : Academic press. 305-341.

Rossi, M., Nishinuma, Y., et Mercier G. (1983). Indices acoustiques et indépendants du contexte pour la reconnaissance automatique de la parole. *Speech Communication*. 215-217.

Saijayaram, A.K.V., Ramasubramanian, V., et Sreenivas, T.V. (2003). Language identification using parallel sub-word recognition. *IEEE International Conference on Audio Speech and Signal Processing*. 1:32-35.

Selouani, S.A., et O'Shaughnessy, D. (2002). A hybrid HMM/Autoregressive time-delay neural network automatic speech recognition system. *European Signal Processing Conference IEEE-EUSIPCO*. 587-590. Toulouse, France.

Selouani, S.A., Tolba, H., et O'Shaughnessy, D. (2003). Auditory-based acoustic distinctive features and spectral cues for robust automatic speech recognition in Low-SNR car environment. *Proceedings of Human Language Technology Conference of the North American Association for Computational Linguistics*. 91-94. Edmonton, Canada.

Spalanzani, A., et Selouani, S.A, (1999). Improving robustness of connectionist speech recognition systems by genetic algorithms. *In proceedings of IEEE Conference on information and intelligence systems*. Washington, DC.

Takuya, T., et Shuji, S. (1994). Simplified Sub-Neural-Networks for accurate phoneme recognition. *Proceedings of International Conference on Signal and Language Processing*. 1571-1574. Yokohama, Japon.

Tolba, H., Selouani, S.A., et O'Shaughnessy, D. (2005). Towards the improvement of automatic speech recognition by integrating dynamic and static Auditory-Based Acoustic Distinctive Features and spectral Cue. *International Conference on Modelling and simulation*. Cancun, Mexico.

Waterhouse, S.R., et Cook G.D. (1996). Ensembles for phoneme classification. *Advances in Neural Information Processing Systems*. Cambridge, MA : MIT press.

Yu, H., et Waibel, A. (2004). Integrating thumbnail features for speech recognition using conditional exponential models. *IEEE International Conference on Audio Speech and Signal Processing*. 1:893-896. Montréal, Canada