TELECOM PARIS

1994

WAA

Thèse

présentée pour obtenir le titre de docteur

de l'Ecole Nationale Supérieure des Télécommunications

Spécialité : Signal et Images

Claire Waast

Contribution à l'élaboration d'un système de reconnaissance de parole continue à grand vocabulaire

ENST 94 E 026

Soutenue le 31 Janvier 1994 devant le jury composé de

Bernard Robinet	Président
Denis Jouvet	Rapporteurs
Henri Méloni	
Lalit Bahl	Examinateurs
Eric Laporte	
Joseph Mariani	
Bernard Merialdo	
Jean-Pierre Tubach	

Ecole Nationale Supérieure des Télécommunications
Enseignement supérieur des Télécommunications

TELECOM
PARIS

Eric,
Ton message est efficace,
Je m'excuse illico !
J'espère cependant que la
rancune n'aura pas la vie
trop tenace...
Je ne sais comment me faire
pardonner. Claire

1994
WAA

Thèse

présentée pour obtenir le titre de docteur

de l'Ecole Nationale Supérieure des Télécommunications

Spécialité : Signal et Images

Claire Waast

Contribution à l'élaboration d'un système de reconnaissance de parole continue à grand vocabulaire

ENST 94 E 026

Soutenue le 31 Janvier 1994 devant le jury composé de

Bernard Robinet	Président
Denis Jouvet	Rapporteurs
Henri Méloni	
Lalit Bahl	Examinateurs
Eric Laporte	
Joseph Mariani	
Bernard Merialdo	
Jean-Pierre Tubach	

Ecole Nationale Supérieure des Télécommunications
Enseignement supérieur des Télécommunications

REMERCIEMENTS

Je tiens tout d'abord à remercier Jean-Pierre Tubach, qui a bien voulu accepter de diriger cette thèse. Ses conseils et sa disponibilité m'ont été très précieux.

Je remercie Bernard Mérialdo pour l'intérêt constant qu'il a apporté à mon travail, pour sa disponibilité et ses remarques pertinentes.

Je remercie Denis Jouvet et Henri Méloni pour le temps et l'attention qu'ils ont bien voulu consacrer à mon travail, dont ils sont rapporteurs.

Je remercie Bernard Robinet pour m'avoir naguère accueilli au Centre Scientifique d'IBM France et pour me faire honneur aujourd'hui de présider le jury.

Je remercie Joseph Mariani d'avoir accepté d'être examinateur, je lui sais gré des remarques et des suggestions qu'il a su me faire.

Je remercie Lalit Bahl pour ses remarques et critiques fructueuses sur mon travail et pour avoir accepté de se déplacer jusqu'à Paris pour assister à la soutenance.

Je remercie Eric Laporte qui m'a initié à la phonétique. Je suis heureuse qu'il ait accepté de participer au jury.

Merci mille fois à tous mes collègues présents et passés Marc El-Bèze, Jean-Christophe Marcadet, Hélène Cerf, Laurent Diringer, Pierre De la Noue pour m'avoir encouragé et soutenu.

Un dernier remerciement à Gaël et Hugo dont l'attention et la patience ne cessent de m'étonner.

INTRODUCTION

Les techniques de communication parlée cherchent à rendre les ordinateurs accessibles par la voix. Même si la parole ne remplace pas toujours une souris ou un clavier, le langage permet une communication simple et rapide avec la machine. Il peut être un moyen idéal de communication, en particulier, quand l'utilisateur a les mains déjà occupées, quand il est handicapé ou peu failier avec la machine.
Le traitement de la parole fait appel à des connaissances sur les techniques de communication numérique, sur la production et la perception de la parole, sur la linguistique et la phonétique de la langue étudiée. Ce champ d'études est divisé en domaines, la perception et la production, le codage, la synthèse et la reconnaissance de parole. C'est dans le cadre de ce dernier domaine que mes travaux s'inscrivent. Si l'homme a la faculté de comprendre un message vocal provenant d'un locuteur quelconque, dans un environnement bruité, quels que soient son mode d'élocution, la syntaxe et le vocabulaire utilisés, la machine est loin d'en faire autant. Le problème global de la reconnaissance vocale reste ouvert. On ne peut lui apporter que des solutions partielles, solutions de sous problèmes attachés à un ensemble de contraintes. La reconnaissance de mots isolés, le plus souvent mono-locuteur, pour des dictionnaires de quelques dizaines, voire quelques centaines de mots est bien maîtrisée. Des progrès récents ont été réalisés en reconnaissance multilocuteur de petits vocabulaires, dans des conditions sonores difficiles. En reconnaissance mono-locuteur de grands vocabulaires selon un mode d'élocution isolé, quelques systèmes ont été commercialisés pour des tâches de dictée de texte dans un domaine spécifié. Ces systèmes sont le plus souvent fondés sur une modélisation stochastique de la parole, méthode à ce jour la plus performante. En reconnaissance de parole continue, multilocuteur, quelques systèmes ont été commercialisés pour des tâches très restreintes. Mais les difficultés sont encore nombreuses pour passer à des systèmes dont le vocabulaire serait plus volumineux.

Mes recherches s'inscrivent dans le cadre d'un système de reconnaissance de la parole fondé sur une modélisation stochastique de la parole. La reconnaissance est mono-locuteur sur un grand vocabulaire, en environnement sonore calme. Ma contribution porte sur le mode d'élocution que l'on désire continu. Dans le cadre des systèmes fondés sur une modélisation stochastique

de la parole, le passage d'un mode d'élocution isolé à un mode d'élocution continu, ne remet pas en cause fondamentalement la structure des systèmes. Le problème est essentiellement un problème acoustique, la complexité linguistique reste identique. Les algorithmes de décodage acoustique classiques peuvent être utilisés pour décoder de la parole continue. Le problème réside essentiellement dans la prise en compte de la variabilité contextuelle de la parole.
Cette variabilité peut être inter-locuteur. L'accent et le timbre de la voix varient d'un individu à l'autre. La reconnaissance est facilitée si on ne permet à un individu l'utilisation de la machine qu'après avoir enregistré les caractéristiques de sa voix dans une phase d'apprentissage.
La variabilité peut aussi être intra-locuteur. Un mot peut être prononcé de différentes manières selon son contexte dans la phrase, l'humeur du locuteur etc... On peut distinguer ici deux types de variations. Les variations de nature phonologique: liaison, élision, dénasalisation qui modifient la représentation phonétique du mot et les variations de nature acoustico-phonétique: anticipation, retard dans le positionnement des organes articulatoires... Ces dernières ne sont pas toujours perçues mais impliquent d'importantes modifications spectrales.
En l'absence de pause entre les mots, leurs frontières ne sont plus marquées, les sons des uns influencent ceux des autres, et réciproquement. Les variations du second type sont alors très fréquentes.

Le premier chapitre est un état de l'art en reconnaissance de la parole. Une attention particulière sera portée aux systèmes fondés sur une modélisation stochastique.
Le second chapitre est consacré à l'étude des variations de nature phonologique et à la présentation d'un phonétiseur du français avec variantes. Un processus de sélection de la meilleure variante permet d'affiner le modèle phonétique du locuteur en l'adaptant à ses habitudes phonologiques.
Le troisième chapitre présente une méthode qui permet de prendre en compte des altérations plus fines (variantes acoustico-phonétiques) dont la subtilité ne peut être rendue par le phonétiseur. L'approche retenue est celle d'une modélisation des effets contextuels au sein même de l'automate markovien. Pour chaque phonème, on construit un arbre de décision dont les feuilles représentent les allophones du phonème considéré. On construit un automate markovien pour chaque allophone.
Le quatrième chapitre présente une méthode alternative à la méthode classique fondée sur l'algorithme de Baum permettant d'établir une liste de mots candidats. Le continuum entre le chapitre III et le chapitre IV réside dans l'arbre de décision. La structure arborescente permet une recherche très rapide de la liste des mots candidats. Celle-ci représente un avantage appréciable quand on sait que dans le système utilisé, la recherche d'une liste de mots candidats prend 5 fois plus de temps que le décodage acoustique affiné, et 22 fois plus de temps que le décodage linguistique.

CHAPITRE I: ETAT DE L'ART

ETAT DE L'ART

Les deux domaines du traitement de la parole dans lesquels l'étude de la variabilité contextuelle a suscité un certain intérêt sont la synthèse et la reconnaissance. En ce qui concerne la synthèse, il existe sur le marché un certain nombre de produits, comme le système DECTALK développé au MIT [48] ou des systèmes fondés sur les recherches du CNET de Lannion [34]. La parole produite est intelligible mais il lui manque l'aisance d'un parlé humain. Si cet inconvénient fut sans importance pour les premiers systèmes dont le but était la faisabilité, il n'est plus acceptable pour un dialogue homme-machine plus poussé. Un effort important est consacré à l'étude de la prosodie. L'étude des phénomènes phonologiques parait moins urgente dans la mesure où il est possible de calculer les paramètres de synthèse sur des mots isolés et les concaténer pour produire un signal de parole continu. Cette approche n'est cependant qu'approximative dans la mesure où l'image acoustique du mot en parole continue n'est pas identique à celle du mot isolé.

En reconnaissance, il s'agit de comprendre un message dont la forme linguistique est en général unique mais qui acoustiquement présente des formes diverses dépendantes du locuteur, de son humeur, de son origine sociale, de l'influence mutuelle des phonèmes qui le composent, etc. Pour comprendre ce message, il faut être capable de prévoir ces variations. On distingue, en reconnaissance, différents systèmes types en fonction des tâches qu'ils sont capables d'accomplir. Si le dictionnaire est de petite taille et que les mots qui le composent sont phonétiquement éloignés, les variantes phonologiques apportent peu d'intérêt. En effet, même si le système est indépendant du locuteur, les mots sont suffisamment différentiables par leurs différences phonétiques. Si le dictionnaire est plus grand, l'ambiguïté entre les mots est plus importante. Un système capable de gérer plusieurs formes de prononciation d'un mot sera plus performant qu'un système à correspondances graphème-phonème standards, à condition, cependant, que l'on veille à ce que les variantes rares n'introduisent pas trop d'erreurs.

La reconnaissance d'un message vocal s'effectue à travers différents niveaux. A un niveau acoustique, il faut déceler les mots à l'intérieur du flux continu du signal vocal; à un niveau lexical, il faut les identifier comme des mots connus; à des niveaux syntaxiques, sémantiques et pragmatiques: il faut analyser la phrase, éventuellement la replacer dans un contexte pragmatique, pour en extraire la syntaxe et le sens. L'homme a la faculté d'effectuer l'ensemble de ces tâches sans grandes difficultés quel que soit le locuteur et dans des environnements sonores parfois difficiles. Aucun système de reconnaissance n'est encore capable d'en faire

autant. Pour atteindre cet idéal, les chercheurs se sont fixé des contraintes à l'intérieur desquelles le problème de la reconnaissance est moins ardu. Ces contraintes sont appelées à être levées au fil des années de recherche.

LES CONTRAINTES

Le mode d'élocution. Pour contourner le problème de détection des mots dans un flux continu de parole, on contraint l'utilisateur du système à ménager de brèves pauses entre les mots qu'il prononce. Cette contrainte permet de connaître la limite entre deux mots et facilite grandement le travail du décodeur. Ce mode d'élocution est dit "élocution en mots isolés". Si aucune pause n'est ménagée entre les mots, on parle d'"élocution continue".

La taille du vocabulaire. Les systèmes de reconnaissance vocale ne sont capables de reconnaître que les mots présents dans leur dictionnaire. Un dictionnaire de grande taille aura l'avantage d'avoir une bonne couverture[1]. Par contre, la reconnaissance sera plus difficile car les ambiguïtés sur le plan acoustique (les homophones "*sot*" ou "*saut*") ou sur le plan grammatical ("*ils courent*" ou "*il court*") sont plus nombreuses. Le choix du dictionnaire n'est que compromis. Le lecteur trouvera dans [67] une comparaison des performances obtenues selon qu'on utilise un dictionnaire de 10.000 ou de 200.000 mots. L'une des conclusions de cette étude est que ce que l'on perd à cause de l'ambiguïté accrue avec le dictionnaire de 200.000 mots est compensé par ce que l'on gagne grâce une plus large couverture.

La dépendance à l'égard du locuteur. Pour que le système puisse décoder le message d'un locuteur, il faut lui fournir des informations sur les caractéristiques de sa voix à partir desquelles il calcule ce qu'on appelle un modèle acoustique (ou modèle de voix). Demander à un système de reconnaître toute voix potentielle sans adaptation préalable, suppose le calcul d'un modèle commun valable pour tout le monde. Si le système "ne répond qu'à la voix de son maître", on dira qu'il est dépendant du locuteur (ou monolocuteur). Si par contre, il fonctionne pour tout locuteur potentiel sans adaptation préalable, on dira qu'il est indépendant du locuteur (ou multilocuteur). Il est à noter que les techniques de reconnaissance sont très semblables en mono ou multilocuteur, la différence tient dans l'apprentissage et le taux de reconnaissance.

La qualité de l'environnement. Si le système est performant dans un environnement bruité (cockpit d'avion, cabine téléphonique), on dira qu'il est résistant au bruit, sinon on dira qu'il fonctionne dans un environnement sonore protégé.

[1] La couverture d'un texte par un dictionnaire est définie comme étant le pourcentage des mots du texte présents dans le dictionnaire. C'est la limite supérieure du taux de reconnaissance, puisque les systèmes actuels ne peuvent reconnaître que les mots qui existent dans leur dictionnaire.

La nature de la syntaxe. Elle peut être contrainte: les phrases prononcées doivent suivre des schémas grammaticaux prédéfinis; ou libre: sans contraintes grammaticales autres que celles de la langue naturelle.

On distingue essentiellement trois types d'applications correspondant chacun à un ensemble particulier de contraintes:

- Les systèmes de commande vocale.
- Les systèmes de compréhension.
- Les machines de dictée.

LES SYSTEMES DE COMMANDE VOCALE

Le portrait robot d'un système de commande vocale pourrait être le suivant. Un système à petit vocabulaire, dont le mode d'élocution est isolé ou continu. Il est résistant au bruit et dépendant ou indépendant du locuteur. On trouve dans ce domaine un grand nombre de produits qui permettent de contrôler l'environnement. Il serait difficile d'en dresser l'inventaire mais on peut citer à titre d'exemple:

Le serveur vocal "Les Baladins" développé par le CNET [44]. Ce serveur diffuse les programmes de cinéma de la région de Lannion. La reconnaissance vocale se fait sur des appels téléphoniques (donc sur un signal vocal dégradé) pour tout locuteur potentiel. La taille du dictionnaire est de 26 mots. Notons un effort ergonomique important pour la réalisation de ce produit, en particulier sur la gestion des menus et des hésitations du locuteur.

Le système multilocuteur SPHINX [53]. Il effectue une reconnaissance sur un dictionnaire d'un millier de mots en mode d'élocution continu. Ce module est désormais intégré en standard sur les ordinateurs Macintoch Centris 660AV et Quadra 840AV.

La serrure vocale Sésame du LIMSI. Elle vérifie plus qu'elle ne reconnaît le message qui lui est transmis. Chaque membre du laboratoire possède une carte sur laquelle est enregistrée l'empreinte vocale d'un mot clé. Si le mot prononcé par une personne désirant entrer correspond à l'empreinte vocale de sa carte, Sésame déclenche l'ouverture de la porte.

Ce domaine est en ce moment en pleine expansion commerciale. Le nombre des produits commercialisés pour les serveurs vocaux en témoignent.

LES SYSTEMES DE COMPREHENSION

Si les systèmes de reconnaissance au sens strict conduisent à des applications de type dictée, les systèmes de compréhension cherchent à accéder à la signification du message prononcé. Cette différence ne se manifeste que dans le cadre de projets assez ambitieux. Ces systèmes sont connectés à des modules d'interprétation du message reconnu qui répondent à la

stimulation vocale soit par une réponse vocale soit par une action mécanique. Pour faciliter le dialogue avec le système, la syntaxe est souvent contrainte par des schémas grammaticaux simples et le mode d'élocution est continu. La fermeture du domaine sémantique est acceptable dans la mesure où l'application est bien ciblée. Ces projets ont vu le jour dans les années 70. Comme on ne savait pas très bien identifier les phonèmes, on essaya de reconnaître une phrase, malgré les erreurs du décodage phonétique, en faisant appel à des contraintes relatives à la langue. Cette approche fut soutenue aux Etats Unis de 1971 à 1976 par le Département de la Défense (projet ARPA/SUR). Il s'agissait de développer des systèmes capables de "comprendre" des phrases en parole continue sur un vocabulaire de mille mots. Issus de ce projet, trois systèmes ont vu le jour: Harpy de l'Université de Carnegie Mellon (Pittsburg) [60], Hearsay II du CMU [57] et HWIN de BBN [90]. A la même époque la recherche française était très active dans le domaine. Citons les travaux de J.P. Haton et J.M. Pierrel avec Myrtille I et II au CRIN (Nancy) [74][39], ceux de G. Mercier du CNET avec Keal [69], ceux de J. Mariani et J.S. Lienard du LIMSI avec Esope [63]. Citons aussi, la réalisation d'un système de reconnaissance et de traduction automatique présenté par J.P. Tubach [85] au 18e congrès des JEP; et le projet à très long terme d'ATR-Japon [49] d'un système de reconnaissance de la parole continue par téléphone, couplé à un système de traduction automatique et à un système de synthèse qui permettrait à deux interlocuteurs de langues différentes de communiquer en temps réel sans contraintes.

LES SYSTEMES DE DICTEE AUTOMATIQUE

Les systèmes de dictée automatique ont pour objet de transcrire un texte dicté par un locuteur. La compréhension n'est pas requise, ce qui signifie que ces systèmes n'ont aucune idée du sens des mots qu'ils transcrivent. S'ils l'avaient les performances seraient certainement meilleures mais la construction d'un module d'analyse sémantique sur des phrases dont le vocabulaire est très étendu est à ce jour un problème ouvert. D'autre part, il s'agit bien de dictée, le système n'est pas capable de transcrire une conversation riche en hésitations, reprises ou retours en arrière. La dictée suppose l'utilisation d'un dictionnaire très étendu (quelques dizaines voire des centaines de milliers de mots) et la gestion d'une grammaire la plus libre possible. Dans l'état actuel des recherches, les systèmes qui gèrent des vocabulaires de grande taille ne sont pas indépendants du locuteur. Ils nécessitent un apprentissage préalable à l'utilisation et ne peuvent pas encore supporter un mode d'élocution continu[2]. Le portrait robot d'un tel système serait celui d'une reconnaissance en mode isolé sur un dictionnaire de quelques dizaines de milliers de mots, dépendante du locuteur et dans un environnement sonore calme. A la fin des années 70, la tendance est à un retour à une recherche sur le décodage acoustico-phonétique. Certains travaux se situent au niveau du traitement de signal [70], d'autres autour de systèmes fondés

[2] Avec un taux d'erreur de l'ordre de 5%

sur des connaissances phonétiques [91][92], d'autres encore reposent sur une modélisation markovienne. Cette dernière approche de type auto-organisatrice, introduite par J. Baker [13] et F. Jelinek [41] et dont l'influence resta longtemps limitée, va prendre au cours des années 80 et 90 une place de plus en plus importante en tant qu'alternative à la programmation dynamique en particulier aux Bell Labs. [76]. En 1983, le centre scientifique d'IBM France s'intéresse avec Parsyfal [67][69] à un système de dictée en syllabes isolées sur un dictionnaire de très grande taille (200.000 mots). Citons aussi dans le cadre des très grands vocabulaires, les travaux de l'INRS (86.000 mots[3]) dont on trouvera des résultats récents dans [56]; et pour les langues asiatiques, le projet Mandarin (60.000 mots) [59]. Un des premiers systèmes commercialisés fut le système Dragon-Dictate [14] qui fonctionne en mots isolés avec un dictionnaire de 16.000 mots extensible à 30.000 mots. Sa particularité réside en l'absence d'apprentissage préalable, l'apprentissage se fait en cours d'utilisation par validation au clavier des mots reconnus. Le taux de reconnaissance sur les premières phrases prononcées est assez faible, le locuteur corrige les erreurs, et au fur et à mesure des corrections, le taux de reconnaissance s'améliore. Cette adaptation doit être faite correctement, sinon le système diverge. Plus récemment, IBM a commercialisé un système de reconnaissance multilingue (anglais, français, italien, espagnol, allemand) en mots isolés nommé ISSS plus connu sous le nom de prototype Tangora. En France, outre la version française de Tangora, le LIMSI a démontré dans le cadre d'un projet ESPRIT [87], la faisabilité d'un système de reconnaissance monolocuteur multilingue sur un grand vocabulaire (10.000 et 50.000 mots), selon un mode d'élocution isolé et celle d'un système de reconnaissance de parole continue (1000 mots).

Les systèmes de dictée peuvent se découper en trois modules: Le traitement du signal, le décodage acoustico-phonétique et l'analyse syntaxique. Ce découpage est cependant arbitraire, puisque dans les derniers travaux de J.L. Gauvain, la distinction entre les deux derniers niveaux n'existe pas. Le traitement simultané du décodage phonétique et de l'analyse syntaxique s'avère adapté au problème de la reconnaissance de parole continue, en particulier au traitement des liaisons. Cette stratégie peut cependant poser problème si l'on fait croître la taille du dictionnaire.

Les méthodes de traitement du signal sont pour la plupart issues des techniques de codage, dont les plus courantes sont l'analyse spectrale, l'analyse par prédiction linéaire [2] et l'évaluation des coefficients cepstraux [20].

Le décodage phonétique consiste à mettre en correspondance le signal et une suite d'unités phonétiques prédéfinies. Il nécessite la localisation et l'identification des phénomènes qui décrivent le phonème. La localisation repose sur une analyse de l'évolution de certains

[3] sur l'anglais.

paramètres du signal [58][65][66]. L'identification repose sur des techniques de classification (quantification vectorielle, analyse de données, techniques issues de la reconnaissance des formes...). Notons que la programmation dynamique [64][81], la modélisation markovienne, les réseaux neuronaux à délai temporel [90] permettent d'effectuer conjointement ces deux opérations. L'identification d'une information permet de localiser sa position dans le signal.

La prise en compte de la structure syntaxique du langage permet de réduire, à chaque instant, le nombre de mots envisageables à un sous ensemble de mots syntaxiquement acceptables, et donc, le temps de réponse et le nombre des erreurs. Par contre, une grammaire trop contrainte aura comme effet d'ajouter des erreurs en cas de non-respect de la syntaxe. On trouvera dans [18] et [40] différents exemples de réseaux syntaxiques.

Actuellement, les principaux axes de recherche dans le domaine de la dictée portent sur le mode d'élocution que l'on voudrait plus naturel, sur l'adaptation rapide au locuteur voire l'indépendance à son égard, et sur la taille des vocabulaires. La technologie dominante est une technologie fondée sur des automates markoviens. Il existe d'autres approches concurrentes ou complémentaires, soit connexionistes soit de type intelligence artificielle, mais elles restent encore marginales.

C'est dans le cadre de la dictée automatique et du progrès vers un mode d'élocution non contraint que mes recherches s'inscrivent. Pour pouvoir décoder la parole continue il faut non seulement être capable de déceler les mots dans le flux continu de parole, mais encore gérer toutes les déformations que subissent les mots. En effet, en supposant le système doté d'un algorithme capable de trouver la limite entre les mots, si ces mots ont été modélisés par une forme canonique correspondant à une prononciation isolée, le système ne sera pas capable de décoder le message reçu car il s'évertuera à chercher des formes canoniques qui ne s'y trouvent pas. En effet, un mot "pris en contexte", c'est-à-dire dans un flux de parole, entouré d'autres mots, n'a pas la même allure (spectralement ou même phonétiquement) que le même mot "pris hors contexte". Pour résoudre ce problème il faudrait donc chercher les formes canoniques des mots pris en contexte. Mais des contextes possibles, il y en a beaucoup trop! Une autre approche consiste à quantifier l'information en intégrant dans le système certaines connaissances que les experts phonéticiens ont construites au fur et à mesure qu'ils étudiaient la variabilité de la parole. Avant d'aborder les problèmes liés à la variabilité je vais décrire le système de reconnaissance qui est à la base de mon travail.

CADRE DE TRAVAIL

LE CANAL ACOUSTIQUE

Le support physique de la parole est une onde acoustique produite par le conduit vocal. Cette onde riche en informations doit être traitée de façon à réduire la redondance, augmenter le contraste entre des informations voisines pour finalement ne garder qu'un certain nombre de paramètres acoustiques pertinents. Dans les systèmes Tangora ou Parsyfal, l'extraction des paramètres acoustiques est réalisée par une carte prototype dotée d'un processeur spécialisé en traitement du signal. Le signal issu du microphone est échantillonné à une fréquence de 20 kHz. Il est découpé en fenêtres de 512 échantillons. Le décalage entre deux fenêtres est d'une centiseconde. Chaque fenêtre est multipliée par une fenêtre de Hamming. Et sur chaque fenêtre, on effectue une transformée de Fourier. Le spectre entre 200 et 8.000 Hz est découpé en vingt bandes de fréquences établies d'après des critères perceptifs. Des algorithmes de modélisation de l'oreille et de traitement du bruit par l'intermédiaire de modèles adaptatifs garantissent une certaine résistance au bruit [71][72]. A partir des vecteurs à vingt composantes issus du spectre, on réalise une quantification vectorielle qui consiste à classifier les vecteurs en 200 zones acoustiques. Cette opération est faite à l'aide de l'algorithme des K-moyennes (ou K-means) dont voici une rapide description.

- Faire enregistrer au locuteur la totalité du corpus d'apprentissage et calculer l'ensemble des vecteurs à vingt composantes selon les étapes précédemment décrites.
- Choisir 200 vecteurs au hasard parmi tous les vecteurs du corpus d'apprentissage. Ce sont les "prototypes initiaux". A chaque prototype on associe une classe pour le moment composée d'un seul élément, son prototype initial.
- Pour chaque vecteur du corpus, calculer la distance euclidienne qui le sépare de chacun des prototypes, puis l'affecter à la classe du prototype dont il est le plus proche.
- Recalculer les prototypes, comme étant les centroïdes de leur classe.
- Itérer les deux étapes précédentes jusqu'à convergence, c'est-à-dire jusqu'à ce que le nombre de migrants soit inférieur à un seuil fixé.

On appelle "Dictionnaire de spectres" l'ensemble des prototypes obtenus à la fin de la quantification vectorielle. La taille du dictionnaire de spectres utilisé ici est de 200, ses éléments sont numérotés de 1 à 200. On les désignera désormais par leur numéro (appelé

numéro, label ou observation acoustique). Par la suite, toute phrase prononcée par le locuteur sera représentée par une suite de numéros compris entre 1 et 200.

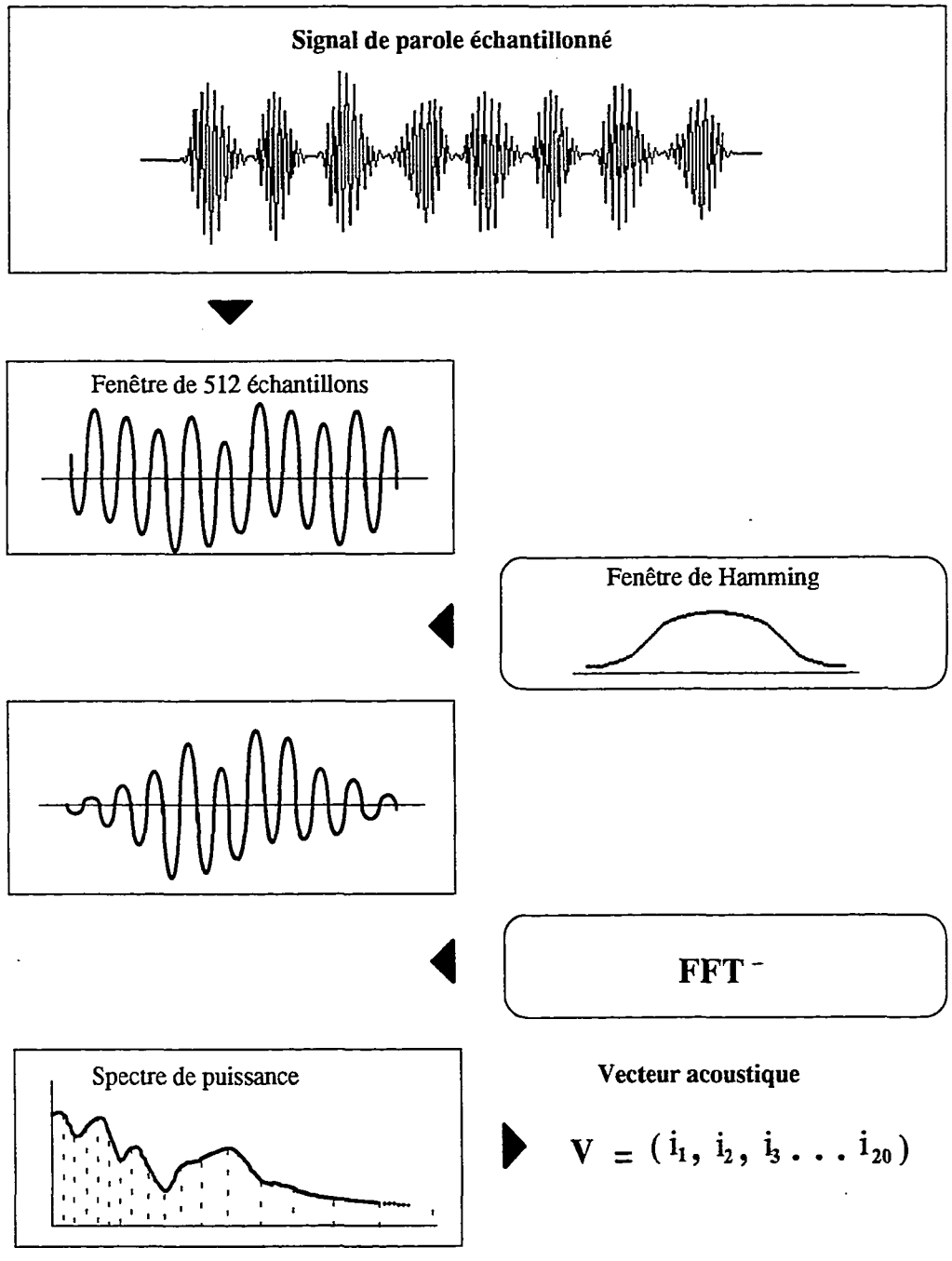

FIGURE 1. Traitement du signal

Le but de la reconnaissance est d'associer à une suite d'observations acoustiques issues de la quantification vectorielle, la suite de mots qui correspond le mieux à ce que le locuteur a prononcé et qui respecte les contraintes de la langue. Ceci se fait couramment en deux étapes. La première cherche à mettre en correspondance la suite d'observations acoustiques et des symboles qui peuvent être des mots, des phonèmes ou même des unités sans valeur linguistique ou phonétique. Ces symboles, si ce ne sont pas des mots, sont reliés aux mots par un lexique.

La seconde étape cherche parmi l'ensemble des mots proposés par l'étape précédente, la combinaison de mots qui paraît "grammaticalement" la plus correcte. Lorsqu'on formule le problème de la reconnaissance en termes de théorie de l'information, sous certaines hypothèses, les deux étapes apparaissent clairement comme étant indépendantes.

THEORIE DE L'INFORMATION

L'appréhension de la reconnaissance de parole à travers la théorie de l'information fut une idée de F. Jelinek et de J. Baker [41][13] dans les années 70. Le nombre d'adeptes qui y ont recours aujourd'hui, témoigne de son efficacité.

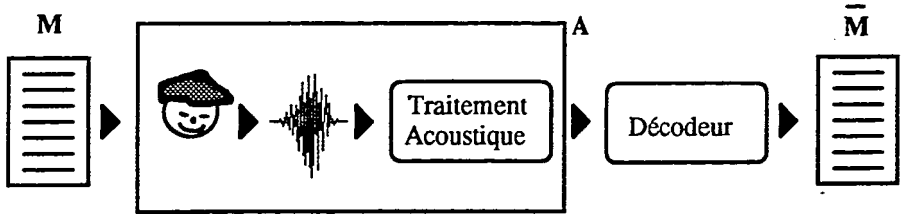

FIGURE 2. Théorie de l'information

Le principe en est le suivant. De la suite des mots prononcée par un locuteur, on extrait une suite $A = (a_1, a_2...a_n)$ d'observations acoustiques. Il s'agit de trouver, connaissant A, la suite de mots $\overline{M}$ la plus probable parmi toutes les suites de mots possibles. Ceci s'exprime de la façon suivante.
On cherche la suite $\overline{M}$ telle que:

$$P(\overline{M}|\,A) = \max_{M \in \mathcal{D}} P(M|\,A)$$

Où $\mathcal{D}$ est l'ensemble des suites de mots du dictionnaire. En utilisant la règle de Bayes, on a:

$$P(\overline{M}|\,A) = \max_{M \in \mathcal{D}} \frac{P(A|M).P(M)}{P(A)}$$

- P(A) est la probabilité d'occurrence de la suite acoustique A. En supposant P(A) indépendante de M, le problème revient alors à chercher la suite $\overline{M}$ telle que:

$$\overline{M} = \operatorname*{argmax}_{M \in \mathcal{D}} P(A|\,M).P(M)$$

- P(A| M) est la probabilité d'observer la suite A connaissant la suite de mots M prononcée.
- P(M) est la probabilité que la suite de mots M apparaisse dans le langage considéré. La suite de mots "*lait pool du couvents couve*" aura une probabilité bien moindre que la suite

de mots phonétiquement identiques "*les poules du couvent couvent*". Pour une grammaire stochastique la combinaison de mots qui paraît "grammaticalement" la plus correcte est celle qui produit la plus grande probabilité.

Cette formulation fait apparaître les trois composantes d'un système de reconnaissance probabiliste: le module qui extrait du signal vocal les observations acoustiques A, le modèle acoustique qui permet de calculer la probabilité P(A| M), le modèle de langage qui permet de calculer la probabilité P(M). Elle fait apparaître que la reconnaissance ne peut se faire sans apprentissage préalable. Il faut connaître P(A| M) et P(M) pour calculer $\overline{M}$. Pour ce faire, on peut considérer, d'une part, les observations acoustiques comme des symboles émis par une source (acoustique) différente pour chaque valeur possible de M et, d'autre part, les mots comme des symboles émis par une autre source (linguistique). En théorie de l'information, les sources sont traditionnellement représentées par des processus de Markov. On parlera alors de source de Markov. Notons que l'utilisation d'une quantification vectorielle conduit à considérer des modèles markoviens discrets. Cette approche a été retenue pour les systèmes Tangora et Parsyfal, mais il pourrait en être autrement. L. Rabiner, par exemple, travaille directement sur les vecteurs spectraux, ce qui le conduit à considérer des modèles markoviens continus. Le lecteur trouvera dans [37] une comparaison intéressante de ces deux types d'approche.

LE MODELE ACOUSTIQUE

Il n'est pas réaliste d'avoir, dans le cas d'un dictionnaire de grande taille, une source de Markov pour chaque mot. En effet, pour calculer les paramètres de la source de Markov, il faudrait demander à l'utilisateur de prononcer plusieurs fois l'ensemble des mots du dictionnaire (c'est-à-dire n fois 20.000 mots). L'idée est de définir des unités (ou symboles) plus petites, qui concaténées, formeront les mots. Ces unités peuvent être des phonèmes, des diphones, des syllabes ou des entités sans connotations phonétiques comme l'automate fénonique dont on parlera plus loin. Le choix de l'unité est un compromis entre la qualité de la reconnaissance et la taille du corpus d'apprentissage. Nous reviendrons sur ce point au chapitre III.

SOURCE DE MARKOV

Une source de Markov est un automate probabiliste d'états finis qui se définit par un ensemble fini d'états $\mathcal{E}$, comprenant un état initial e_0 et un état final e_F, et dans le cas d'une approche discrète, un alphabet fini de symboles $\mathcal{A}$ comprenant le vide: $\varnothing$. Les éléments $t = (e_i, a_k, e_j)$ où e_i et e_j sont des états et a_k un symbole, sont appelés des transitions. L'élément t désigne ici le passage de l'état e_i à l'état e_j avec émission du symbole a_k.

- Si $e_i = e_j$, on dira que t est une boucle
- Si $a_k = \varnothing$, on dira que t est une transition nulle, sinon on parlera de transition non-nulle

• Un chemin à travers un ou plusieurs automate(s) de Markov est la donnée d'une suite finie (t_i) i=1 à n de transitions telles que:

l'état de départ de t_1 est e_0

l'état d'arrivée de t_i est égal à l'état de départ de t_{i+1}

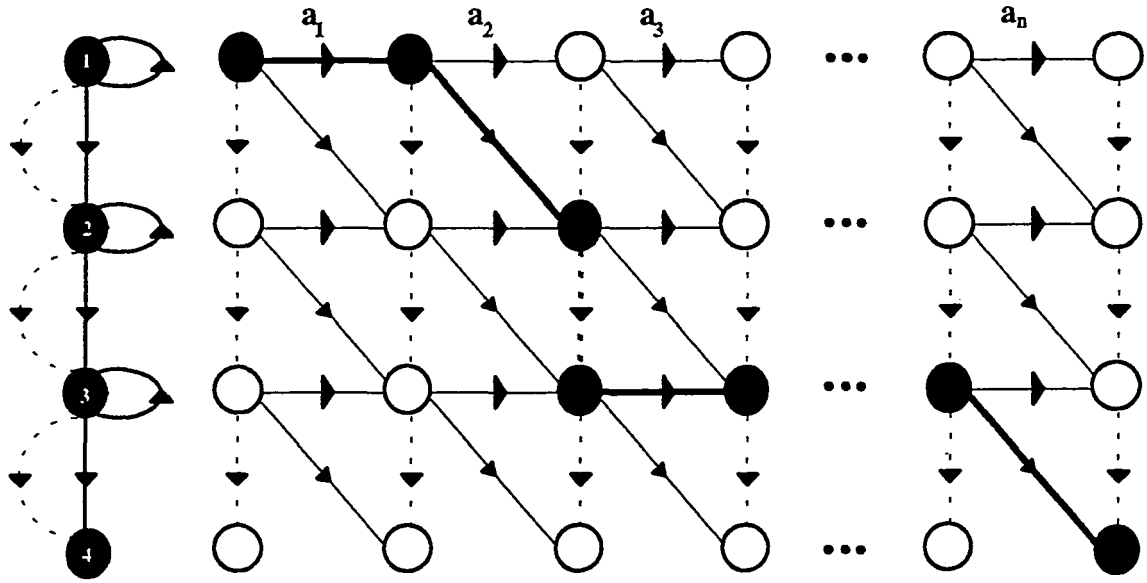

FIGURE 3. Treillis sur trois automates fénoniques concaténés

Le chemin noté en gras sur le treillis des trois automates concaténés, correspond à la suite des transitions (t_1, t_2, t_3 ... t_m), où,

- t_1=(1,a_1,1) est une boucle sur l'état 1, avec émission du symbole a_1.
- t_2=(1,a_2,2) est transition non-nulle de l'état 1 à l'état 2 avec émission du symbole a_2.
- t_3=(2,∅,3)est une transition nulle de l'état 2 à l'état 3.
- t_4 est une boucle sur l'état 3, avec émission du symbole a_3.
- t_m est transition non-nulle de l'état 3 à l'état 4 avec émission du symbole a_n.

A chaque transition t est associée la probabilité q(t) d'émettre le symbole a_k lors du passage de l'état e_i à l'état e_j. Si la transition est non-nulle, q(t) peut se décomposer en la probabilité a_{ij} de transiter de l'état e_i à l'état e_j et la probabilité b_{ijk} d'émettre le symbole a_k sachant que l'on a pris la transition particulière t.

$$q(t) = a_{ij} \cdot b_{ijk}$$

Pour une transition nulle t, q(t) sera noté v_{ij}. Une chaîne de Markov d'ordre p est une structure où le passage par un état dépend de façon probabiliste des p états le précédant. Pour un modèle d'ordre 1 la probabilité du chemin C = (t_i) i=1 à n est égale au produit des probabilités des transitions qui le composent.

$$P(C) = \prod_{ti \in C} q(ti)$$

Remarquons qu'à un chemin donné ne correspond qu'une suite de symboles. Par contre, une suite de symboles peut être produite par beaucoup de chemins.

ESTIMATION DES PARAMETRES

L'estimation des paramètres a_{ij}, b_{ijk} et v_{ij} d'une source de Markov se fait lors d'une étape appelée "apprentissage". Elle consiste à chercher les paramètres qui maximisent, selon un critère de maximum de vraisemblance, la probabilité que cette source ait produit un ensemble d'observations relatives à l'unité modélisée par la source. Par exemple, si le modèle est phonétique et que la source considérée est celle du phonème [a], on cherche les paramètres qui maximisent la probabilité que l'automate phonétique associé à [a], produise l'ensemble des suites d'observations acoustiques relatives au [a]. Le problème de cette méthode est que le calcul d'une telle probabilité nécessite de connaître les paramètres de l'automate. On résout ce problème avec une procédure itérative de l'algorithme de Baum [15] qui converge vers un optimum local. Les principes de base de l'algorithme de Baum sont les suivants.

On calcule la probabilité $\alpha(e_j, x)$ que les x premières observations aient été produites par tous les chemins qui partent de l'état initial et aboutissent à e_j. Symétriquement on calcule la probabilité $\beta(e_i,x')$ que les T-x' dernières observations aient été produites par tous les chemins allant de l'état e_i à l'état final e_F, T étant la longueur totale de la suite d'observations. On a alors:

La passe-avant (forward pass)

$$\alpha(e_0,0) = 1$$
$$\alpha(e_j,0) = 0 \ \forall \ e_j \neq e_0$$
$$\alpha(e_j,x) = \sum_{ei} \left[\alpha(e_i,x-1) \cdot a_{ij} \cdot b_{ija(x)} + \alpha(e_i,x) \cdot v_{ij}\right]$$

Où a(x) est le $x^{ième}$ symbole de la suite des T observations.

La passe-arrière (backward pass)

$$\beta(e_F,T) = 1$$
$$\beta(e_i,T) = 0 \ \forall \ e_i \neq e_F$$

$$\beta\,(e_i,x') = \sum_{e_j} [\beta\,(e_j,x'+1)\,.\,a_{ij}\,.\,b_{ija(x'+1)} + \beta\,(e_j,x')\,.\,v_{ij}\,]$$

La probabilité qu'une suite donnée d'observations acoustiques ait été produite par l'ensemble des chemins allant de l'état initial à l'état final vaut:

$$\alpha\,(e_F,T) = \beta\,(e_0,0) = \sum_{e_i} \alpha\,(e_i,x)\,.\,\beta\,(e_i,x)$$

Après une itération de l'algorithme de Baum, les paramètres des automates markoviens sont réestimés et ce nouveau modèle devient le modèle initial de l'itération suivante. Les formules de réestimation des paramètres sont les suivantes.

$$a'_{ij} = \frac{\sum_x \alpha(e_i, x-1).\ a_{ij}.\ b_{ija(x)}.\ \beta(e_j, x)}{\sum_p \sum_x [\alpha(e_i, x-1).\ a_{ip}.\ b_{ipa(x)}.\ \beta(e_p, x) + \alpha(e_i, x).\ v_{ip}.\ \beta(e_p, x)\ ]}$$

$$b'_{ijk} = \frac{\sum_{x:\ a(x)=a_k} \alpha(e_i, x-1).\ a_{ij}.\ b_{ija(x)}.\ \beta(e_j, x)}{\sum_x \alpha(e_i, x-1).\ a_{ij}.\ b_{ija(x)}.\ \beta(e_j, x)}$$

$$v'_{ij} = \frac{\sum_x \alpha(e_i, x).\ v_{ij}.\ \beta(e_j, x)}{\sum_p \sum_x [\alpha(e_i, x-1).\ a_{ip}.\ b_{ipa(x)}.\ \beta(e_p, x) + \alpha(e_i, x).\ v_{ip}.\ \beta(e_p, x)\ ]}$$

L'algorithme n'a pas besoin de connaître les limites entre phonèmes ou entre mots. La seule chose dont il ait besoin en dehors du modèle initial, c'est de la suite des états. Si le modèle est phonétique, cette suite d'états est obtenue en phonétisant le corpus d'apprentissage et en remplaçant chacun des phonèmes par son automate de Markov. L'algorithme de Baum gère l'ensemble de la phrase. Il permet l'apprentissage à partir de parole continue. D'autre part, outre le fait que l'algorithme de Baum permet d'estimer les paramètres des automates de Markov, c'est sur lui aussi que peut reposer le décodage au cours duquel seuls les $\alpha\,(e_j,x)$ seront calculés. Un autre type de décodage peut être réalisé avec l'algorithme de Viterbi [86] qui calcule la probabilité d'une suite d'observations acoustiques donnée, non plus sur l'ensemble des chemins, comme c'est le cas pour l'algorithme de Baum, mais le long du chemin optimal. Ce chemin se calcule grâce à une formule récurrente qui diffère de la passe-avant de l'algorithme de Baum par le fait que la somme est remplacée par la fonction maximum.

$$\gamma(e_0,0) = 1$$
$$\gamma(e_j,0) = 0 \ \forall \ e_j \neq e_0$$
$$\gamma(e_j,x) = \max_{e_i} [\ \gamma(e_i,x-1) \cdot a_{ij} \cdot b_{ija(x)} , \gamma(e_i,x) \cdot v_{ij} \ ]$$

AUTOMATES PHONETIQUES ET FENONIQUES

Une alternative à l'approche fondée sur la modélisation du mot, qui pose un problème d'apprentissage pour les systèmes à grand dictionnaire, est d'utiliser des unités de parole plus petites que le mot. Tangora fonctionne avec deux approches de modélisation d'unités de taille inférieure au mot. La première est fondée sur le concept de phonème; la seconde, sur les segments acoustiques. Cette seconde approche, introduite par Bahl et al., pour la modélisation markovienne discrète [3] et par Lee, Soong et Juang, pour la modélisation markovienne continue [52], consiste à modéliser l'acoustique plutôt que la phonétique. Les automates acoustiques sont appelés "automates fénoniques" ("fenonic"[4], en anglais).

L'unité phonétique, dans Tangora, est représentée par un automate de Markov à sept états dont la structure est illustrée par la figure suivante.

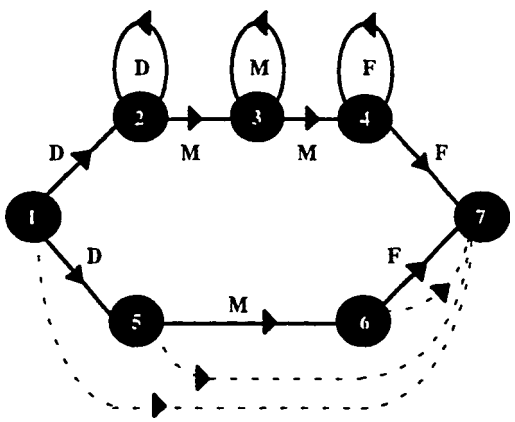

FIGURE 4. Automate phonétique

L'alphabet est constitué des numéros des prototypes du dictionnaire de spectres de taille 200. Une flèche pleine correspond à 200 transitions non-nulles, chacune entraînant l'émission de l'un des 200 numéros du dictionnaire de spectre. Une flèche en pointillé correspond à une transition nulle. L'axe vertical illustre deux modes de prononciation du phonème. Les six transitions de la partie inférieure de la machine correspondent à une prononciation rapide, les sept transitions du haut à une prononciation plus lente. L'axe horizontal correspond à l'axe temporel et peut être découpé en trois segments relatifs à la prononciation du phonème. Une phase transitoire: l'attaque (**D**). Une phase stable: le coeur du phonème (**M**). Une phase transitoire: la chute (**F**).

[4]Cette dénomination provient du terme "Front End" qui désigne l'observation acoustique en anglais.

Les unités fénoniques sont représentées par des automates à deux états dont la structure est la suivante.

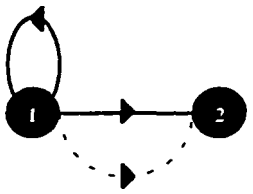

FIGURE 5. Automate fénonique

L'avantage de l'approche phonétique est, qu'en raison du petit nombre de phonèmes, l'estimation des paramètres ne pose pas de problèmes. Avec un texte d'apprentissage équilibré, les automates apparaissent suffisamment de fois pour être bien appris. Cependant, les performances d'un système dont le modèle acoustique est uniquement phonétique sont assez mauvaises. Ceci s'explique par le fait que les phonèmes sont des entités abstraites auxquelles correspondent différentes réalisations acoustiques, dont la diversité n'est pas prise en compte par les automates phonétiques. L'approche fénonique ne pose pas ce problème. Par contre, elle nécessite un double apprentissage. Le premier, réalisé une fois pour toutes, permet de déterminer pour chaque mot du dictionnaire sa forme[5] fénonique. Le second est un apprentissage classique qui a pour but d'estimer les paramètres des automates.

Pour réduire le nombre des paramètres à estimer, les distributions d'émission des observations acoustiques peuvent être liées. Dans le cas des automates phonétiques, les distributions sont liées de la façon suivante.

D $b_{12k} = b_{22k} = b_{15k}$.

M $b_{23k} = b_{33k} = b_{34k} = b_{56k}$.

F $b_{44k} = b_{47k} = b_{67k}$.

Les trois distributions **D**, **M** et **F** correspondent au découpage temporel de l'automate phonétique décrit plus haut. Les probabilités de transitions restent indépendantes les unes des autres. Le nombre de paramètres à estimer pour un automate à sept états s'élève[6] alors à $7 + (N_{ds}-1 = 199)\times 3 = 604$, au lieu de $7 + 199\times 10 = 1997$ si les distributions n'étaient pas liées.

---

[5]La forme fénonique d'un mot est constituée d'une suite des automates fénoniques. De même, la forme phonétique d'un mot est constituée d'une suite des automates phonétiques. La forme phonétique d'un mot est obtenue par phonétisation de ce mot.

[6]Nds est la taille du dictionnaire de spectres qui résulte de la quantification vectorielle.

Le deux centième paramètre est déterminé car on impose les deux contraintes suivantes: $\sum_k b_{ijk} = 1 \ \forall i,j$ et $\sum_j (a_{ij} + v_{ij}) = 1 \ \forall i$.

Dans le cas des automates fénoniques, les deux distributions d'émission sont liées par $b_{11k}=b_{12k}$ ce qui réduit le nombre de paramètres à estimer de $2xN_{ds}$ à $N_{ds}+1$.

LE MODELE DE LANGAGE

Le décodeur linguistique a pour but de définir, par l'attribution d'un score probabiliste à chaque combinaison de mots possibles, un ordre sur les suites de mots proposées par le décodeur acoustique. On verra comment décodage acoustique et décodage linguistique s'associent pour résoudre le problème de la reconnaissance dans la section suivante. Le calcul de ce score se fait par l'intermédiaire d'un modèle dit "de langage" (ou modèle linguistique). La probabilité d'une suite de mots $M = (m_1, m_2...m_n)$ s'écrit comme le produit des probabilités conditionnées de chaque mot connaissant le début de la phrase.

$$P(M) = P(m_1).\prod_{i=2}^{n} P(m_i \mid m_1 \ldots m_{i-1})$$

Lorsque n est grand, l'estimation des probabilités pose problème car le corpus d'apprentissage nécessaire à leur calcul doit être gigantesque. Un moyen de réduire le nombre des paramètres à calculer est de partitionner le passé $(m_1 \ldots m_{i-1})$ en classes d'équivalences. On peut, par exemple, considérer comme équivalentes toutes les phrases dont les p mots qui précèdent m_i sont identiques. On a alors:

$$P(m_i \mid m_1 \ldots m_{i-1}) \cong P(m_i \mid m_{i-p+1}\ldots m_{i-1})$$

Un tel modèle est appelé "p-grammes" et quand p vaut trois, comme c'est le cas dans Tangora, le modèle est dit tri-gramme. D'autres partitions consistent à définir les classes d'équivalences comme des classes grammaticales ou encore comme des classes de lemmes [29][33]. Selon le formalisme adopté en théorie de l'information, on considère que les mots sont des symboles émis par une source de Markov pour laquelle les états sont les classes, les lemmes ou les mots du dictionnaire et les probabilités de transition sont les fréquences relatives $f(m,C_n)$ d'occurrence du mot m dans la classe d'équivalence C_n. Les processus de Markov utilisés dans Tangora sont des processus d'ordre 0 (uni-gramme), 1 (bi-gramme) ou 2 (tri-gramme) combinés pour obtenir le score associé aux suites de mots décodées. Le lecteur trouvera dans [33] une description détaillée de ces différents modèles.

LA STRATEGIE DE DECODAGE

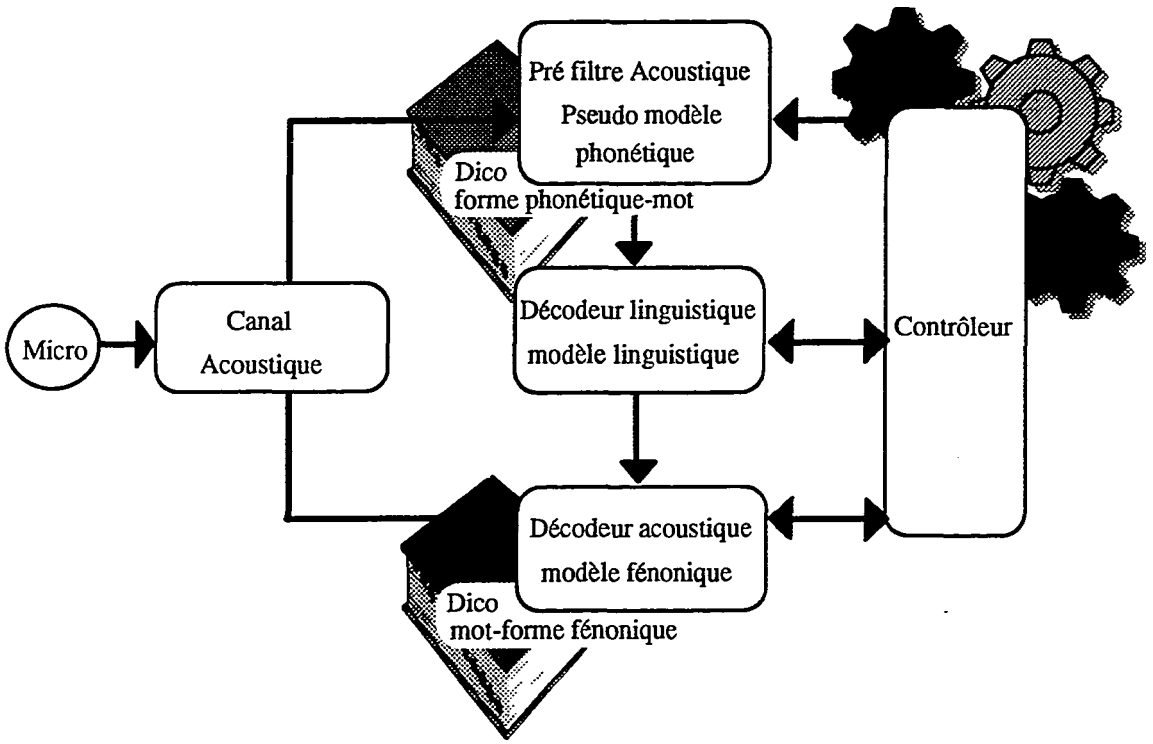

FIGURE 6. Stratégie de décodage

- Le canal acoustique transforme le signal vocal en une suite dobservations acoustiques.
- Le préfiltre acoustique[7] [5] effectue un décodage acoustico-phonétique et détermine une première liste des mots candidats. Pour ce faire, il calcule pour chaque mot du dictionnaire la probabilité que sa forme phonétique produise la suite d'observations acoustiques fournie par le canal acoustique. Les mots du dictionnaire sont classés par ordre de probabilité décroissante et on ne garde de cette liste que les n premiers mots.
- Le décodeur linguistique calcule, à l'aide d'un modèle de langage tri-gramme, pour chaque mot de la liste fournie par le préfiltre acoustique, la probabilité du mot connaissant les deux mots qui le précèdent. Ce score linguistique permet de réorganiser la liste fournie par le préfiltre acoustique.
- Le décodeur acoustique gère par l'intermédiaire d'un lexique la transcription en formes fénoniques des mots de la liste fournie par le décodeur linguistique. Puis pour chaque mot de la liste, il calcule à l'aide de l'algorithme de Baum et d'un modèle acoustique fénonique, la probabilité que sa forme fénonique produise la suite d'observations acoustiques fournie par le canal acoustique. Les n mots de la liste sont alors classés par ordre de probabilité décroissante et on ne garde de cette liste que les $m < n$ premiers mots.
- Le contrôleur gère l'ensemble de ces tâches, organise les données à l'échelle de la phrase et décide de la phrase finale.

[7] plus connu sous sa terminologie anglaise: "Fast match"

L'algorithme de décodage de Tangora est l'algorithme dit de décodage à pile (ou stack decoding [43]). On construit un treillis des mots candidats. Pour chaque nouvelle liste de mots candidats, on étend les chemins les plus probables. La probabilité des chemins partiels est recalculée et les chemins triés selon un ordre de plus grande probabilité. La phrase décodée est celle qui correspond au chemin de plus grande probabilité.

LA VARIABILITE DE LA PAROLE

La variabilité phonologique sous-entend plusieurs types de variabilité. Elle peut être d'ordre physiologique et exprime alors l'instabilité articulatoire de nos systèmes phonatoires. En effet, contrairement à l'idée défendue par la phonétique classique selon laquelle une certaine position des organes de la parole caractérise un son, ces organes sont en constant mouvement et leurs mouvements diffèrent lors de prononciations successives d'une même séquence de parole. De plus, la constitution des organes phonatoires varie d'un individu à l'autre. Par exemple, le conduit vocal féminin est en moyenne de 15% plus court que le conduit vocal masculin ce qui entraîne un décalage formantique vers les hautes fréquences [18]. La variabilité peut être d'ordre sociologique. On entend par là essentiellement deux types de variantes. D'une part les variantes régionales qui sont liées aux habitudes d'un parler local. D'autre part, les variantes idiolectales qui reflètent les différences d'âge, de sexe, de milieu social. Enfin la variabilité peut être de nature contextuelle et s'explique par le fait que le langage est composé de sons juxtaposés qui s'influencent les uns les autres et se modifient. Parmi les variantes contextuelles, on peut distinguer, comme le fait Giachin [35], deux types d'altérations: celles de nature phonétique et celles de nature allophonique. Les altérations allophoniques représentent les différentes réalisations acoustiques d'un même phonème. Elles sont souvent sans importance pour l'oreille humaine mais peuvent être significatives pour un système de reconnaissance. Les variantes phonétiques sont, par contre, plus perceptibles. Il peut s'agir de la substitution d'un phonème par un autre ou même de sa disparition. Elles sont moins fréquentes que les altérations allophoniques et par là même, à cause du manque de données d'apprentissage, plus difficiles à modéliser dans un système auto-organisateur. On peut néanmoins aider le système à les prévoir en lui insufflant une connaissance phonétique extérieure.

Notons que si d'un point de vue méthodologique, il est nécessaire de distinguer différents types de variabilité, ils sont cependant très liés. La séquence "*la fenêtre décorée*" prononcée par un locuteur pressé sera perçue comme [la fnɛd dekore], séquence dans laquelle le [r] final de "*fenêtre*" a chuté et le [t] s'est assimilé au [d] de "décorée". Alors que dans la même séquence prononcée par un locuteur méridional, tous les phonèmes seront pleinement prononcés avec sans doute l'ajout d'un [ø] final à "*fenêtre*" et l'ouverture du [o] de "*décorée*". Une étude complète des variantes phonétiques engloberait l'étude des habitudes langagières ou des habitudes liées au milieu social comme la prononciation ou non du [t] final dans la séquence de

mots "*en fait*" ou dans "*exact*"; l'étude des phénomènes d'élision de phonèmes lors d'une élocution rapide; l'étude de la place des silences... Il apparaît clairement qu'on ne peut pas s'attaquer globalement au problème de la variabilité. Son traitement au niveau phonétique nécessite l'aide d'experts phonéticiens, son traitement automatique au niveau allophonique nécessite une base de données particulièrement représentative de tous les types de variabilité traités. Dans un cas comme dans l'autre, mon champ de prospection pourra sembler restreint au regard de l'étendue des possibilités. Mon choix, quant aux variantes phonétiques, porte sur celles qui rendent compte des principaux phénomènes de co-articulation. Quant aux altérations allophoniques, elles seront calculées par une méthode automatique dont la validité repose sur la quantité et la diversité des données d'apprentissage.

CONCLUSION

Dans un système de reconnaissance probabiliste tel que Tangora, l'information phonétique est présente à deux niveaux. D'une part en phase d'apprentissage: le calcul des paramètres d'un automate phonétique se fait en confrontant la suite des observations acoustiques à une suite d'états. Cette suite d'états est obtenue en *phonétisant* le corpus d'apprentissage et en remplaçant chacun de ces phonèmes par son automate markovien. D'autre part en phase de reconnaissance: le préfiltre acoustique calcule pour chaque mot du dictionnaire la probabilité que sa *forme phonétique* produise la suite de numéros fournie par le canal acoustique. Si le corpus d'apprentissage est phonétisé selon une prononciation standard, le locuteur n'ayant pas nécessairement respecté les contraintes de cette prononciation standard, les paramètres de son modèle acoustique seront approximatifs et son modèle peu performant. De même si le dictionnaire ne propose que des formes phonétiques standard, le système ne sera pas capable de décoder un message ne suivant pas les standards de cette phonétisation.

Contraindre l'utilisateur d'un système de reconnaissance vocale à une prononciation standard est une vaine requête, le système se doit de prendre en compte les différentes prononciations possibles d'un mot. Une première approche pour résoudre ce problème peut se faire à un niveau phonétique, par la construction d'un phonétiseur de texte par règles, capable de générer, pour un mot, plusieurs formes phonétiques. Je présenterai dans le second chapitre, la mise au point d'un tel phonétiseur. Les variantes phonétiques d'un mot sont ensuite soumises à un processus de décision qui permet de sélectionner la variante la plus probable étant donnée la suite des observations acoustiques. Ce processus de décision s'insère dans la logique markovienne du système de reconnaissance.
Si un phonétiseur par règles s'avère efficace pour traduire les transformations des phonèmes pris en contexte (insertions, disparition ...), il devient insuffisant pour rendre compte de la finesse des altérations allophoniques. Des règles construites à partir de spectrogrammes seraient sans doute plus efficaces, mais leur mise au point est longue et difficile. L'approche retenue est une modélisation des effets contextuels au sein même de l'automate markovien. Elle sera exposée dans le troisième chapitre.

NOTATIONS

Les conventions de notations phonétiques adoptées dans ce manuscrit sont les suivantes.

- Les phonèmes seront notés entre crochets.
  exemple: [e] [ɛ]
- Les sons sans distinction d'aperture seront notés entre deux barres verticales.
  exemple: |e|
- Les réalisations acoustiques d'un phonème seront notées entre accolades.
  exemple: {e} {ɛ}
- L'automate markovien représentant le phonème sera noté entre deux barres obliques.
  exemple: /e/ /ɛ/

CHAPITRE II: VARIABILITE PHONETIQUE

INTRODUCTION

Historiquement, les premiers phonétiseurs automatiques de texte en français furent conçus pour être incorporés à des systèmes de synthèse de parole à partir de texte. Au LIMSI le phonétiseur de D. Teil (1969) comportait environ deux cents règles et une liste d'une vingtaine de mots pour lesquels les règles générales étaient insuffisantes. Depuis d'autres phonétiseurs ont été construits sur le même principe. Ils diffèrent les uns des autres par le nombre de mots qu'ils phonétisent correctement. Au C.E.A., L. Poirot et X. Rodet (1976) ont mis au point un phonétiseur doté d'un outil qui permettait d'obtenir, par la recherche du nombre d'occurrences d'une chaîne de caractères dans un dictionnaire de 65.000 mots, la liste des mots susceptibles d'être concernés par une règle donnée. Au CNET, M. Divay et M. Guyomard et au LIMSI, B. Prouts ont élaboré des phonétiseurs dont la facilité de lecture des règles a permis une mise au point et une mise à jour souple et efficace. Ce souci de souplesse et de lisibilité s'explique quand on considère la hiérarchie complexe que forment les règles et leurs exceptions. Par exemple, une règle générale R qui exprime que la lettre 's' se prononce [s] possède une exception E qui précise que la lettre 's' pris entre deux voyelles se prononce [z]. L'exception E a elle même des exceptions comme celle qui stipule qu'après certain préfixe, comme 'para', le 's' entre voyelles se prononce [s], comme dans "parasol". Une règle et ses exceptions ne peuvent pas être modifiées indépendamment les unes des autres. Lorsqu'on corrige, ajoute ou supprime une règle, on modifie le traitement de plusieurs mots: ceux qui relèvent de la règle modifiée et éventuellement ceux qui relèvent des règles en relation hiérarchique avec cette règle. C'est pourquoi il est souvent difficile d'évaluer les conséquences de la modification d'une règle. La phonétisation par dictionnaire contourne ce problème, l'élément modifiable est l'entrée lexicale. La correction, l'ajout ou la suppression d'une entrée ne remet pas en question les autres entrées lexicales. De plus, le dictionnaire peut comporter des informations grammaticales et syntaxiques faciles à consulter et utiles pour le traitement des liaisons, pour désambiguïser des homographes non homophones ou pour déterminer les valeurs des paramètres prosodiques en synthèse de parole. Cependant, l'élaboration d'un tel dictionnaire se fait à l'aide d'un phonétiseur par règles, les modifications ayant lieu ensuite sur les entrées lexicales. Les méthodes algorithmiques et lexicales de phonétisation de texte furent l'objet de nombreux travaux, citons ceux de E. Laporte du CERIL et ceux de B. Prouts, F. Néel, M. Eskenazi, J. Mariani, A. Lacheret-Dujour au LIMSI qui chacun ont apporté une attention particulière dans

leurs travaux, à la prise en compte de la variabilité de la parole. Le phonétiseur par règles n'est pas toujours la meilleure approche du problème de phonétisation, son efficacité dépend de la langue utilisée. Dans la version anglaise du système de reconnaissance de la parole Tangora, la phonétisation des mots repose sur le calcul de la probabilité qu'une lettre (ou un ensemble de lettres) produise un phonème (ou un ensemble de phonèmes) [9]. Ces probabilités sont conditionnées au contexte graphique dans lequel apparaît la lettre à phonétiser. Cette technique est utilisée dans la version française de Tangora pour calculer la forme phonétique des mots inconnus que l'utilisateur veut ajouter au dictionnaire.

Au sein du centre scientifique d'IBM France, l'objectif de la construction d'un phonétiseur avec variantes était double. Il s'agissait, d'abord, d'enrichir le phonétiseur existant de règles lui permettant de prévoir un certain nombre de déformations subies par les mots quand ils sont pris dans le flux d'une parole continue. Ce nouveau phonétiseur est divisé en deux modules séquentiels. Le premier associe à la graphie d'entrée une forme phonétique intermédiaire. Le second produit à partir de la forme intermédiaire un ensemble de variantes phonétiques. La division du phonétiseur en deux modules permet de traiter séparément les problèmes liés à l'ambiguïté de la forme graphique et les problèmes de variabilité phonétique.
Le phonétiseur fut, ensuite, testé dans un environnement de reconnaissance vocale. Le calcul du modèle acoustique de chaque locuteur se fait à partir d'un corpus phonétique d'apprentissage adapté aux habitudes phonatoires du locuteur [19]. Après avoir décrit le phonétiseur avec variantes, j'exposerai le calcul du modèle acoustique adapté.

VARPHO: UN PHONETISEUR AVEC VARIANTES

Le point de départ de la construction du phonétiseur avec variantes VARPHO est un phonétiseur normalisé (qui à une graphie donnée fait correspondre une phonétique normalisée) développé au centre scientifique d'IBM France [19].
Le principe de transcription repose sur l'exploitation d'un module de règles contenant les règles de prononciation et leurs exceptions. Schématiquement, une règle est composée d'une chaîne à traiter, d'un résultat, éventuellement des contextes droit et gauche et d'une condition grammaticale.

Contexte droit (Chaîne à traiter) Contexte gauche → Résultat. [Condition Grammaticale]

Du fait de l'exploitation séquentielle du module, les règles doivent être ordonnées. Pour la transcription de la graphie en forme intermédiaire, les règles d'exceptions sont classées avant les règles plus générales, de façon à appliquer en premier les règles les plus restrictives.
Pour la transcription de la forme intermédiaire en variantes phonétiques, l'application des règles se fait de façon à exploiter toutes les possibilités de la phonétique combinatoire. Par exemple, pour obtenir toutes les variantes de la séquence "*je te dis*", la règle provoquant la chute du [ə] de "*je*" devra s'appliquer avant celle du dévoisement des obstruantes sonores au contact d'obstruantes sourdes, permettant ainsi d'avoir la variante [ʃtdi].

La transcription d'une forme en une autre, consiste à examiner la chaîne à transcrire de la gauche vers la droite et d'y chercher une partie qui coïnciderait avec la "chaîne à traiter" d'une des règles du module. L'exemple qui suit donne une idée du processus de transcription graphème-phonème pour mot "*parasol*".

1 Chaîne à transcrire		2 Lecture des règles concernant la transcription du :	3 Phonème
'parasol'	'p'	règle (p) → [p], la règle générale s'applique.	[p]
'arasol'	'a'	règle (a) → [a], la règle générale s'applique.	[a]
'rasol'	'r'	règle (r) → [r], la règle générale s'applique.	[r]

'asol'	'a'	règle (a) → [a], la règle générale s'applique.	[a]
'sol'	's'	Ordre d'application: R: règle générale (s) → [s] 4. (*piste*) E: exception de R V(s)V → [z] 3. (*base*) E': exception de E para(s)V → [s] 2. (*parasol*) Exception de E' para(s)it → [z] 1. (*parasite*) La règle 2 s'applique.	[s]
'ol'	'o'	Ordre d'application: R: règle générale (o) → [o] 4. (*polémique*) Exception de R (o)mmeS → [ɔ] 3.(*pomme*) Exception de R (o)lS → [ɔ] 2. (*sol*) Exception de R (o)lleS → [ɔ] 1. (*colle*) La règle 2 s'applique.	[ɔ]
'l'	'l'	règle. (l) → [l], la règle générale s'applique.	[l]

V désigne une voyelle graphique et S une fin de mot éventuellement au pluriel.

L'indépendance du module de règles et du processus de transcription permet d'ajouter ou de supprimer des règles sans affecter la structure du traducteur.

DE LA GRAPHIE A LA FORME INTERMEDIAIRE

Avant d'être transcrite, la graphie est découpée en syllabes. Cette étape n'est pas nécessaire mais s'avère commode pour différentier, par exemple les prononciations de "*un*" et "*une*".
La méthode de syllabisation suit les règles de découpage syllabique suivantes.

- On admet une seule voyelle (diphtongue ou voyelle non sécable) par syllabe.
- Deux, ou trois consonnes qui se suivent appartiennent à deux syllabes différentes, sauf dans le cas des consonnes non sécables.

Le tri des règles se fait d'abord sur la "chaîne à traiter", puis sur le contexte droit, puis sur le contexte gauche, puis enfin sur la condition grammaticale. De cette manière, si deux règles portent sur la même "chaîne à traiter", celle qui a le plus long contexte droit sera placée avant l'autre. Si deux règles ne diffèrent que par la donnée grammaticale, celle qui en possède une sera placée avant l'autre.

Les règles doivent être suffisamment nombreuses pour décrire le plus possible d'exceptions aux règles générales et suffisamment bien conçues pour ne pas engendrer d'erreurs. Les sigles, les noms propres, les mots d'origine étrangère sont autant exceptions qu'il faut traiter, ou réunir dans un index spécialisé. D'autre part, la transcription phonétique d'une phrase soulève souvent en français des problèmes liés à la syntaxe. Un indice grammatical permet de désambiguïser les

homographes non homophones comme "*couvent*" dans "*les poules du couvent couvent*". La liaison se fera toujours entre un adjectif et un substantif, après un verbe suivi d'un pronom personnel, jamais après un substantif singulier... Une condition grammaticale associée à la règle permet de lever ces ambiguïtés. Par contre, aucune information relative aux dérivations ou aux compositions de mots n'est prise en compte, ce qui peut conduire à des erreurs comme la phonétisation en [z] du 's' de "*désolidariser*". Pour ce type de mots une liste de règles particulières a été écrite. Notons que si un 's' précédé de certains préfixes comme 'anti' ("*antisocial*"), 'aéro' ("*aérosol*") se prononce obligatoirement [s], il n'en est pas de même avec d'autres préfixes. On a vu qu'à la suite du préfixe 'para' un 's' peut se prononcer [s] ("*parasol*") ou [z] ("*parasite*"), il en est de même avec le préfixe 'dé' ("*désolidariser, désambiguïser*") ou avec le préfixe 'pré' ("*présupposer, présider*").

PRE TRAITEMENTS

LE TRAITEMENT DES LIAISONS

Le phénomène de liaison à lieu lorsqu'une consonne muette en position finale est suivie d'un mot commençant par une voyelle. Les deux lettres s'unissent alors pour ne former qu'un son.

Les consonnes graphiques génératrices de liaisons sont les sivantes:

finale	d	g	n	p	s	t	x	z
liaison	[t]	[k]	[n]	[p]	[z]	[t]	[z]	[z]

« c » et « f » donnent des liaisons en [k] et [v] dans les cinq cas suivants: "*porc-épic*", "*croc-en-jambe*", "*neuf heures*", "*neuf hommes*", "*neuf ans*". Notons que la liaison en [n] peut s'accompagner d'une dénasalisation qui a lieu lorsqu'un adjectif terminé par une voyelle nasale précède un mot avec lequel il fait liaison. La dénasalisation consiste en la transformation de la voyelle nasale en voyelle orale.

exemple	**voyelle nasale**	**voyelle dénasalisée**
un b**on** ami	'on' ([ɔ̃])	[ɔ]
div**in** enfant	'in' ([ɛ̃])	[i]
cert**ain** espoir	'ain' ([ɛ̃])	[ɛ]
pl**ein** air	'ein' ([ɛ̃])	[ɛ]

La règle n'est cependant pas générale. Il n'y a pas de dénasalisation pour "***un***", "***en***", "***on***", "*aucun*", "*rien*", "*bien*". La dénasalisation est libre pour les possessifs, "***mon***", "***ton***"... et discutée pour "***non***".

La réalisation d'une liaison étant fortement dépendante de son environnement syntaxique, la génération automatique de liaisons nécessite en générale un analyseur syntaxique. Néanmoins, le phonétiseur Graphon du LIMSI n'en possède pas. Il traite partiellement le problème des liaisons en utilisant une liste de mots grammaticaux appelés *mots-outils*. Ici, l'analyse syntaxique est faite à l'aide du modèle de langage tri-classe qui permet d'identifier la classe des mots du texte [25][26]. On trouvera dans [33] une description détaillée des règles de liaisons possibles, interdites et facultatives qui sont utilisées par VARPHO. D'autres informations sur le sujet pourront être consultées dans [51] et [50]. Les marques[8] de liaisons sont générées automatiquement sur le texte à phonétiser par un module conçu par Marc El-Bèze [33]. Les règles de liaisons facultatives sont pondérées d'un coefficient de réalisation de manière à ce que, grâce au réglage d'un seuil, on puisse inhiber certaines d'entre elles.

Notons que l'étiquetage grammatical du texte ayant été fait pour le traitement des liaisons, l'information grammaticale sera conservée pour être utilisée quand l'application d'une règle l'impose.

SIGNES PARTICULIERS ET MODIFICATIONS

Les modifications apportées au phonétiseur d'origine pour transcrire la forme graphique en forme intermédiaire porte essentiellement sur le module de règles. Ces modifications sont de quatre ordres. Elles concernent l'aperture des sons |e|, |ø| et |o|; la présence d'un e caduc; les mots sujets à la fois à synérèse et diérèse et enfin le traitement des liaisons.

PHONEMES ATONES

Les résultats de la phonétisation d'un texte par le phonétiseur d'origine étaient assez irréguliers en ce qui concerne l'emploi des voyelles ouvertes et fermées [e][ɛ], [o][ɔ], et [ø][œ]. Ainsi "*aéroport*" était phonétisé [aerɔpɔr], transcription phonétique pour laquelle le choix du premier |o| ouvert, comme dans "*corps*", est contestable. Pour rendre plus homogène la transcription des trois sons |e|, |ø| et |o|, trois phonèmes atones ont été ajoutés pour représenter ces sons lorsqu'ils sont variables en aperture et situés en syllabe non finale. Ces trois phonèmes sont notés: [e̊] pour le son |e|, [ø̊] pour le son |ø| et [o̊] pour le son |o|. Les règles d'utilisation de ces phonèmes sont les suivantes:

[8] un tiret souligné entre les mots qui font liaison.

1. Lorsqu'un son |e|, |ø| ou |o| se trouve en *syllabe non finale*

son	**transcription**	**exemple**
|e|	[e̊]	*séparer*
|ø|	[ø̊]	*mesurer*
|o|	[o̊]	*coller*

2. Lorsqu'un son |e|, |ø| ou |o| se trouve en syllabe finale ou devant les groupes de lettres suivantes *C+ement, C+erie, C+futur, C+conditionnel*

C désigne une consonne quelconque

son	**transcription**	**exemple**
|e|	[ɛ]	*rêveras*
|ø|	[ø] ou [œ]	*creusera* ([ø]) ou ***heure*** ([œ])
|o|	[o] ou [ɔ]	*brome* ([o]) ou *collerions* ([ɔ])

3. Lorsqu'un son |e|, |ø| ou |o| se trouve en *finale absolue:*

son	**transcription**	**exemple**
|e|	[ɛ] ou [e]	*muet* ([ɛ]) ou *assez* ([e])
|ø|	[ø]	*creux*
|o|	[o]	*tricot*

LE e CADUC

Le e caduc est une voyelle correspondant au son |ø|, qui peut s'élider dans certains contextes phonétiques. Sa présence est très dépendante de la rapidité d'élocution, du niveau de langue et des habitudes langagières du locuteur. Le repérage des e caducs peut se faire en suivant les indications du Grévisse [36] dont nous rappelons quelques règles.

La règle des trois consonnes. Quand il est encadré par deux consonnes, le |ø| tombe dans la mesure ou sa chute n'entraîne pas une suite consécutive de trois consonnes. Cependant:

- Le |ø| se conserve généralement devant [r], [l], [m] et [n].
- Il se conserve également lorsqu'il appartient à la première syllabe d'une phrase ou d'un groupe de mot.
- Dans une suite de plusieurs mots monosyllabiques, on conserve un |ø| sur deux.

Derrière une consonne et devant une voyelle le |ø| tombe en fin de mot, sauf:

- Quand une pause est marquée.
- Dans le pronom personnel 'le' quand il est placé après un impératif (*prends le*).
- Dans 'ce' quand il est placé devant une proposition relative (*ce à quoi je pense*).

Dans la forme intermédiaire, le e caduc est noté par le signe [ə].

DIERESE ET SYNERESE

Aux trois voyelles phonétiques fermées [i], [u] et [y] correspondent trois consonnes [j], [w] et [ɥ] dites semi-consonnes. Les semi-consonnes sont souvent suivies d'une voyelle [wa] dans "***loi***", [ɥi] dans "***lui***", [je] dans "***pied***" avec laquelle elles forment un unique son. Les voyelles fermées [i], [u] et [y] peuvent dans certains mots prendre la place des semi-consonnes. Les deux voyelles en contact forment alors un hiatus phonétique et appartiennent à deux syllabes distinctes comme c'est le cas dans "***lui aussi***" [io], dans "***clouer***" [ue] ou dans "***truand***" [yã]. Lorsque la prononciation se fait en une seule syllabe le phénomène est appelé synérèse. Lorsqu'elle se fait en deux syllabes, le phénomène est appelé diérèse. En poésie le besoin d'ajuster un vers régit le choix entre synérèse et diérèse. Pour énoncer une sentence de façon insistante, on aura tendance à transformer la synérèse en diérèse. Par contre, le passage de la diérèse à la synérèse est fréquent lorsque le débit de parole s'intensifie. Ainsi "***louer***", si l'élocution est lente sera transcrit [lue] alors que dans le cas d'une élocution rapide il sera transcrit [lwe]. Pour les mots susceptibles de faire à la fois diérèse et synérèse trois signes seront introduits : [ij] pour le couple ([i],[j]), [uw] pour le couple ([u],[w]) et [yɥ] pour le couple ([y],[ɥ]). Les règles de diérèse-synérèse retenues sont les suivantes:

1. Si la graphie est précédée d'un groupe de consonnes *obstruante-liquide*

graphie	**exemple**	**débit lent**	**débit rapide**	**signe**	
'ui'	*fluide*	[yi] ou [ɥi]	[yi] ou [ɥi]	[yɥ]	(1)
'ou', 'u', 'i'	*clouer, affluer, plier*	[u], [y], [ij]	sans changement		

2. Si la graphie est précédée d'une *consonne*

graphie	**exemple**	**débit lent**	**débit rapide**	**signe**	
'ou'	*bouée*	[u] ou [w]	[w]	[uw]	(2)
'ui'	*puis*	[ɥi]	sans changement		
'u'	*suave*	[y] ou [ɥ]	[ɥ]	[yɥ]	(3)
'i'	*lier*	[ij] ou [j]	[j]	[ij]	(4)

(1) En élocution lente "*fluide*", "*altruiste*" et leurs dérivés ainsi que "*truisme*" sont transcrits [yi]. Pour ces mots on utilisera le signe [yɥ]. Dans les autres cas la transcription est [ɥi] quelque soit le débit.

(2) Certains mots font exception: "***boueux***", "*hindouisme*", "***noueux***". Ils ne font que la diérèse en débit lent. En débit rapide les deux sont acceptées.

(3) "***buanderie***", "***duo***", "*nuage*"... font exception. Ils ne font que la diérèse en débit lent. Synérèse et diérèse sont acceptées en débit rapide.

(4) Des exceptions encore: "*fiancer*", "***hier***", "*skier*", "*souriant*", qui ne font que la diérèse en débit lent.

3. En frontière de mot, il peut y avoir variation dans les cas suivants:
   - En frontière du pronom personnel: "tu": "***tu as vu***".
   - En frontière du relatif "qui", quand il est sujet et inaccentué: "*C'est Jean qui **a** tout fini*".
   - Dans la locution "***demi-heure***".
   - En frontière du pronom adverbial "y" en position inaccentuée, c'est-à-dire non post-verbiale: "***il y a***" pour lequel trois prononciations sont possibles: [ia], [ija] et [ja].

Le lecteur trouvera dans [51] et [50] une description plus complète de ces phénomènes.

PREPARATION DES LIAISONS

Les liaisons sont marquées dans le texte à phonétiser par un tiret souligné entre les deux mots qui font liaison. Voici l'extrait d'un texte traité par le programme de génération automatique de liaison:

"Les résultats des_élections législatives en_Israël paraissent répondre aux_espoirs des partisans de la paix au proche orient."

Le premier module de VARPHO se contente simplement de noter les signes de liaisons, ainsi que toutes les terminaisons susceptibles de faire liaison. C'est au second module qu'il incombe de faire la liaison. Ce choix relève de l'organisation des deux modules. Le premier à l'image du phonétiseur d'origine, traite le texte à phonétiser mot par mot, le second a une appréhension plus large de la chaîne à traiter puisqu'il modélise des phénomènes qui peuvent s'appliquer sur plusieurs mots. Etant donnée cette organisation, il était plus naturel de traiter les liaisons dans le second module. Le premier ne se charge que de véhiculer l'information concernant les liaisons, même si toute cette information n'est pas utile. Par exemple, la forme intermédiaire du groupe de mots "*Les résultats des_élections*" sera:

Le symbole [**] traduit le tiret souligné dans la forme intermédiaire.

graphie	forme intermédiaire	marque de liaison
les	[le(z]	[(z] inutile
résultats	[rẽzylta(z]	[(z] inutile
des	[de(z]	[(z] utile
_	[**]	[**] utile
élections	[ẽlɛksjõ(z]	[(z] inutile

DE LA FORME INTERMEDIAIRE AUX TRANSCRIPTIONS PHONETIQUES

Il est assez rare dans la parole de rencontrer des sons isolés. Ce que nous disons, ce que nous entendons sont des chaînes de sons qui, juxtaposés, s'influencent et parfois se masquent. La prononciation se fait selon un principe d'économie décrit par Malmberg [62] en ces termes: "En prononçant les sons du langage, l'homme a tendance à obtenir le maximum d'effet avec un minimum d'effort". Si par exemple, on doit prononcer deux [t] consécutifs comme dans [sɛt tɛt], la prononciation du premier [t] sera complète avec une occlusion suivie d'une explosion. Pour celle du second, on se contentera de prolonger un peu l'occlusion du premier [t]. On évite ainsi d'avoir à empêcher (occlusion) puis permettre (explosion) le passage de l'air entre les deux [t]. Il en est de même quand on prononce deux consonnes consécutives de nature différente, par exemple [t] et [d] dans [tɛt dø linɔt], à la différence près que l'unique occlusion change de nature en son milieu: les cordes vocales se mettent à vibrer pour rendre le voisement du [d]. Le plus souvent il se produit une assimilation consonantique entre le [t] et de [d] qui fait que l'on n'entend que le [d]. L'étude de ces phénomènes est généralement désignée par les termes de phonétique combinatoire ou de co-articulation. L'ensemble des règles qui vont suivre tente d'en exprimer les principaux points. Ces règles, ainsi que celles du premier module, ont été élaborées grâce aux compétences d'Eric Laporte[9] [51].

La plupart des règles du second module proposent pour une situation particulière plusieurs solutions. Une règle s'écrit alors de la façon suivante:

Contexte droit (Chaîne à traiter) Contexte gauche → Résultat1. Résultat2... RésultatN

Les contextes droit et gauche ainsi que la "chaîne à traiter" peuvent porter sur plusieurs mots. Chaque application d'une règle donne lieu à une nouvelle forme phonétique de la phrase

[9]Professeur en Informatique et Phonétique. Laboratoire M. Gross, Université de Paris 7.

initiale. Le premier module appliqué à la séquence de mots "*les petits*" donne la forme intermédiaire: [le pøti]. Sur cette forme, le second module applique les règles suivantes.

Le signe ∅ signifie l'absence de phonème.

forme intermédiaire	**Règle**	**Nature de la règle**	**variantes produites**
[l]			
[e]	([e]) → [e]ou [ɛ]	[e]ou [ɛ] en finale	[e]ou [ɛ]
[p]			
[ə]	([ə]) → [ø] ou ∅	[ə] caduc	[ø] et ∅
[t] et [i]			

Les quatre variantes phonétiques de la séquence "*les petits*" sont les suivantes.

[le pøti]	[le pti]
[lɛ pøti]	[lɛ pti]

La structure du second module est comparable à celle du premier. Elle diffère en deux points: sur la gestion des résultats, puisque à chaque étape plusieurs résultats sont possibles; et sur l'organisation des règles. Les règles sont réparties en quatre fichiers distincts qui s'appliquent à tour de rôle sur les formes dont on cherche les variantes.

LES REGLES DE PHONETISATION AVEC VARIANTES

Dans ce paragraphe, nous présenterons:

- trois catégories de règles qui, par l'exploitation des signes particuliers introduits par le premier module, permettent de générer plusieurs formes phonétiques du texte. Ces règles s'occupent du traitement des [ə], de celui des mots sujets à synérèse et diérèse et de l'exploitation des signes de liaisons.
- des règles particulières, comme celle qui permet que le phonème [œ̃] puisse être substitué au phonème [ɛ̃] ("*parfum*": [parfœ̃] ou [parfɛ̃])
- les règles permettant de corriger les choix faits sur les sons [e], [ø] et [o] lors de l'étape précédente
- enfin, en matière de co-articulation, notre attention s'est portée sur les trois phénomènes suivants: l'assimilation consonantique, la nasalisation des occlusives et la chute des consonnes finales

LE e CADUC: [ə]

Le signe [ə] donne lieu aux deux variantes [ø] et ∅, sauf dans les six exceptions suivantes:

[ə] ne chute pas lorsqu'il est immédiatement *(c'est-à-dire dans le même mot)*

précédé du groupe:	*et suivi par le groupe:*	**exemple**
consonne-liquide	consonne	*autrement*
occlusive-occlusive	consonne	*il optera*
occlusive-fricative	consonne	*il axera*
consonne	fricative - occlusive	*restratifier*
consonne	la semi consonne [j]	*cueillir*

DIERESE ET SYNERESE

Trois règles permettent d'exploiter les signes de diérèse-synérèse introduits par le premier module:

Condition	**Transformation**	**Exemple**
présence du signe particulier	[ij] → [ij] ou [j]	"*plier*"
présence du signe particulier	[uw] → [u] ou [w]	"*louer*"
présence du signe particulier	[yu] → [y] ou [u]	"*tuer*"

[e̊], [ø̊] ET [o̊] DEVANT UNE LIQUIDE

En milieu de mot, lorsque [e̊], [ø̊] ou [o̊] est suivi d'un groupe de *consonnes* dont la première est une *liquide*, il est alors remplacé par son homologue ouvert.

Condition	**Transformation**	**Exemple**
[e̊] -liquide-consonne	[e̊] → [ɛ]	"*perdu*"
[ø̊] -liquide-consonne	[ø̊] →[œ]	"*meurtrier*"
[o̊]-liquide-consonne	[o̊] → [ɔ]	"*colmater*"

EXPLOITATION DES SIGNES DE LIAISON

La liaison sera faite dans la mesure où coexistent, côte à côte, dans la forme intermédiaire une marque de liaison potentielle comme [(z] et un signe [**] qui indique que la liaison doit être faite. Par exemple, la forme intermédiaire de "*Les résultats des_élections*" étant [le(z rẽzylta(z de(z ** ẽlɛksjõ], sa forme phonétique après exploitation des liaisons sera [le rẽzylta de zẽlɛksjõ].

CAS PARTICULIERS

Quelques règles particulières traitent de phénomènes isolés. Le tableau qui suit en fait l'inventaire:

Condition	Transformation	Exemple

1 En *finale* le son |e| donne lieu aux deux variantes : ouvert [ɛ] et fermé [e]

finale en |e|	[e] → [e] ou [ɛ] [ɛ] → [e] ou [ɛ]	"*les*" ou "*fait*"

2 Variante sur le phonème [œ̃]: il donne lieu aux deux variantes [œ̃] et [ɛ̃]

présence du phonème	[œ̃] → [œ̃] ou [ɛ̃]	"*parfum*"

3 Phénomène de dénasalisation

3a	cas répertoriés	[õ] → [ɔn]	"***bon** ami*"
3b	cas répertoriés	[ɛ̃] → [in]	"*divi**n** enfant*"
3c	cas répertoriés	[ɛ̃] → [ɛn]	"*certai**n** espoir*"

4 |ø| en *finale*. Après écoute des textes du corpus d'apprentissage lus par différents locuteurs, il a semblé nécessaire de réintroduire des [ø] en finale absolue.

	finale du mot traité	début mot suivant	transformation	exemple
4a	obstruante	fricative-obstruante	∅ → [ø̊]	"*plante splendide*"
4b	CC	CC	∅ → [ø̊]	"*gourde pleine*"

où C désigne une consonne.

L'ASSIMILATION CONSONANTIQUE

Les modifications que subissent les sons au contact d'autres sons peuvent aller jusqu'à atteindre leurs qualités essentielles. Si l'on prononce rapidement "*bec de gaz*", le [k] aura tendance à être prononcé [g]. Chaque fois qu'un son se rapproche d'un autre son, en ce qui concerne son point d'articulation, on dira qu'il y a assimilation. On peut classer les phénomènes d'assimilation en deux groupes selon que l'assimilation se fait en contact ou à distance. Dans l'exemple "*bec de gaz*" le [k] subit une assimilation régressive au contact de la consonne sonore [g] et devient à son tour sonore. Si l'on prend l'exemple de "jusque", l'influence de la chuintante [ʒ] sur la sifflante [s] peut amener à la transformer en chuintante [ʃ]. Mais ce sont surtout les voyelles qui s'influencent à distance. Le caractère

fermé du premier |e| de "*aimer*" s'explique par l'influence régressive du second. Le |e| de "*aimons*" ne subit pas cette influence, il est et reste ouvert. Pour les consonnes, c'est essentiellement sur la sonorité que l'assimilation se fait. Les assimilations ici considérées concernent l'influence régressive des obstruantes les unes sur les autres.

LES OBSTRUANTES

Les obstruantes peuvent être regroupées en trois catégories. Les nasales [n m ŋ], les occlusives [p t k b d g] et les fricatives [f s ʃ v z ʒ]. Les deux dernières catégories peuvent être, à leur tour, divisées en deux sous catégories selon leur voisement. Une obstruante sourde dont la prononciation se fait sans vibration[10] des cordes vocales peut par anticipation devenir sonore au contact de l'obstruante sonore qui la suit. Réciproquement, la vibration des cordes vocales d'une obstruante sonore peut disparaître quand l'obstruante qui suit est sourde.

Assimilation régressive d'une obstruante sourde suivie d'une consonne sonore

[p] sonore →[p] ou [b]	[t] sonore →[t] ou [d]	[k] sonore →[k] ou [g]
"*une soupe bien chaude*"	"*tzar*"	"*une anecdote*"
[f] sonore →[f] ou [v]	[s] sonore →[s] ou [z]	[ʃ] sonore →[ʃ] ou [ʒ]
"*nous faisons*"	"*bisbille*"	"*enchevêtrement*"

Assimilation régressive d'une obstruante sonore suivie d'une consonne sourde

[b] sourde →[b] ou [p]	[d] sourde →[d] ou [t]	[g] sourde →[g] ou [k]
"*absorber*"	"*de toute façon*"	"*gangster*"
[v] sourde →[v] ou [f]	[z] sourde →[z] ou [s]	[ʒ]-sourde →[ʒ] ou [ʃ]
"*une veuve fatiguée*"	"*les fraises sauvages*"	"*jeter*"

NASALISATION DES OCCLUSIVES

Les occlusives sont caractérisées par une interruption momentanée du passage de l'air. Lorsque cette occlusion est faite par les lèvres (occlusion bilabiale), on obtient le [p] et le [b]. Lorsqu'elle est faite par la langue contre les dents (occlusion apico-dentale), on obtient le [t] et le [d]. Lorsqu'elle est faite par la langue contre le palais dur (occlusion dorso-palatale) ou contre le palais mou (dorso-vélaire)[11], on obtient le [k] et le [g]. Une consonne nasale est une occlusive en ce qui concerne l'articulation buccale, mais contrairement aux occlusives pour lesquelles le voile du palais ferme les fosses nasales de

[10] on parle d'occlusive non voisée, ou d'occlusive sourde. Réciproquement lorsque la prononciation d'une occlusive se fait avec une vibration des cordes vocales, on parle d'occlusive voisée, ou d'occlusive sonore.

[11] [k] (resp. [g]) est dorso-palatale lorsque la voyelle qu'il précède est une voyelle antérieure (pour la prononciation de laquelle la langue s'élève vers la palais dur). Si par contre la voyelle qui suit est postérieure (pour la prononciation de laquelle la langue s'élève vers la palais mou), [k] (resp. [g]) est dorso-vélaire.

façon à ce que la fermeture du passage de l'air soit complète, la prononciation des nasales se fait en abaissant le voile du palais pour laisser passer l'air par le nez. Ainsi, si on prononce un [b] en ouvrant les fosses nasales, on obtient la consonne bilabiale nasale [m]. Le phénomène de nasalisation des occlusives a lieu lorsque l'occlusive est précédée d'une voyelle nasale et suivie d'une consonne:

Vnas [p] C →[m] ou [p]	Vnas [t] C →[n] ou [t]	Vnas [k] C →[ŋ] ou [k]
"***un petit ouistiti***"	"***ton compte postal***"	"***distincte***"
Vnas [b] C →[m] ou [b]	Vnas [d] C →[n] ou [d]	Vnas [g] C →[ŋ] ou [g]
"***trombe d'eau***"	"***seconde guerre***"	"***langue de bois***"

Le signe Vnas désigne une voyelle nasale: [ɑ̃ ɛ̃ œ̃ ɔ̃].

Les règles de nasalisation doivent être précisées en fonction de la nature de la consonne qui suit l'occlusive et de la configuration dans laquelle l'occlusive apparaît (à l'intérieur d'un mot, en limite de mot, etc.).

Nature de la consonne suivante	Si l'occlusive est sonore	Si l'occlusive est sourde
Obstruante sonore	Le phénomène a lieu	Le phénomène a lieu
Obstruante sourde	Le phénomène a lieu	Le phénomène a lieu
Nasale	Le phénomène a lieu*	Il n'a pas lieu
Liquide	Il n'a pas lieu	Il n'a pas lieu

Les occlusives sourdes ne se nasalisent pas devant des consonnes nasales ou liquides, la nasalisation d'une occlusive sonore devant une consonne nasale dépend de la configuration dans laquelle elle apparaît. La nasalisation peut avoir lieu dans les deux configurations suivantes:

* **a.** L'occlusive est isolée: ...Vnas **O** C..., comme dans "*avant **de** manger*".
* **b.** L'occlusive est accolée, en fin de mot, à la voyelle nasale: ...Vnas**O** C..., comme dans "***langue** naturelle*".

Le lecteur trouvera des informations complémentaires dans [51] et [50].

CHUTE DE CONSONNE FINALE.

Une autre façon de faciliter la prononciation est de supprimer des consonnes, par exemple, la suppression du 'l' dans "*quelque chose*". Ce phénomène est particulièrement fréquent pour les liquides finales lorsqu'elles sont précédées d'une occlusive comme dans "*le peuple*

canadien". On retiendra que lorsqu'un groupe occlusive-liquide en finale absolue est suivi d'un mot commençant par une consonne, la liquide peut tomber.

Notons que lorsque l'on prévoit un phénomène d'assimilation entre deux mots, le silence qui les sépare tombe.

ORDRE D'APPLICATION DES REGLES

Les règles du second module sont réparties en quatre sous ensembles qui s'appliquent tour à tour sur la forme intermédiaire. Dans chacun des sous ensembles, les règles sont triées de façon à appliquer en premier les règles les plus restrictives. La nécessaire division des règles en sous ensembles distincts tient au fait que les conditions d'application de certains phénomènes ne sont réalisées que si d'autres phénomènes ont eu lieu auparavant. Par exemple, l'assimilation du [t] en [d] dans la séquence "*être d'attaque*" ne se fait qu'après chute du [ø] et du [r].

1 Le premier sous ensemble contient l'ensemble des règles concernant:
   - le traitement des liaisons
   - les diérèses et synérèses
   - les modifications de [e̊], [ø̊] et [o̊] devant une liquide
   - le traitement des [ə]
   - l'ensemble des cas particuliers
2 La chute des e caducs peut amener à des aberrations comme la variante imprononçable: [jnlrkõnɛpa] de "*je ne le reconnais pas*". Pour éviter ces variantes indésirables, le second sous ensemble contient des règles capables de les repérer et de les éliminer.
3 Le troisième sous ensemble autorise les chutes de liquides en finale.
4 Le quatrième, contient les règles relatives à l'assimilation consonantique et à la nasalisation des occlusives.

Voici un exemple de quelques règles utilisées pour la phonétisation avec variantes de la phrase "le jardin entoure un petit lac".

Graphie → forme intermédiaire		*forme intermédiaire → variantes phonétiques*		
l(e):	(e)S →[ø̊]	[œ̃]:	([œ̃]) →[œ̃] ou [ɛ̃]	cas particulier
ent(ou)re:	(ou) →[u]	[ə]:	([ə]) →[ø̊] ou ∅	e caduc
p(e)tit:	(e) →[ə]	[p]:	([p]) →[p] ou [m]	nasalisation
petit(t)	(t).. →[(t]			
forme intermédiaire		*Variantes phonétiques*		

[lø ʒardɛ̃ ãtur œ̃ pəti(t lak]	[lø ʒardɛ̃ ãtur œ̃ pøti lak ɛ̃ pti mti]

DISCUSSION

Le phonétiseur avec variantes que nous venons de présenter a permis de phonétiser un corpus composé de deux textes. Le premier (200 phrases) est extrait du corpus équilibré de P. Combescure [23]. Le second (253 phrases) a été constitué par A.M. Derouault dans le cadre de ses travaux sur les phonèmes contextuels [28]. Le tableau suivant donne, pour chacun de ces deux textes le pourcentage des mots donnant lieu à une, deux, trois variantes ou plus:

Nombre de variante(s)	Premier texte (200 phrases) [12]	Second texte (253 phrases)
1	52,7	35,4
2	21,4	32,0
3	13,4	20,0
4	11,1	2,6
≥5	1,4	10,0

La différence des résultats, en particulier en ce qui concerne les deux dernières lignes, illustre les différences de nature des deux textes. Le premier est un corpus phonétiquement équilibré. Le second fut spécialement conçu pour l'apprentissage des phonèmes contextuels établis par A.M. Derouault. Parmi ces phonèmes contextuels, on trouve des associations de phonèmes qui se prêtent volontiers aux variations modélisées par le phonétiseur. Par exemple, une des classes de phonèmes contextuels proposée, concerne les occlusives sourdes dont le contexte droit est une occlusive sonore. Pour ses travaux, A.M. Derouault imposait que chaque classe soit au moins représentée vingt fois dans le texte. Pour chacune de ces occurrences, le phonétiseur proposera, en accord avec les règles d'assimilation consonantique, la variante sonore de l'occlusive sourde. Ceci explique le pourcentage important dans le second texte de mots ayant plusieurs formes phonétiques.

Ce phonétiseur permet de phonétiser les mots d'un dictionnaire en leur donnant plusieurs formes phonétiques, mais il permet aussi d'adapter la phonétisation des corpus d'apprentissage aux habitudes langagière des locuteurs et permet ainsi d'augmenter le taux de reconnaissance du système. En effet, lors de l'apprentissage du modèle acoustique d'un locuteur, si la phonétisation du corpus d'apprentissage est normalisée, les suites d'observations acoustiques et les formes phonétiques ne se correspondent pas nécessairement. Le modèle acoustique du

[12] Pourcentage des mots du texte qui ont 1, 2, 3, 4 ou plus de 4 formes phonétiques.

locuteur est alors "mal" appris. A l'aide d'un processus de sélection capable de faire un choix parmi les variantes, on espère, en remplaçant les formes phonétiques standards par une variante sélectionnée, améliorer les performances du système. Cette approche est développée dans la partie suivante.

ELABORATION D'UN MODELE ACOUSTIQUE

Cette partie contient deux sections. La première expose le principe de sélection de la meilleure variante et la construction du modèle acoustique. La sélection de la meilleure variante se fait à l'aide de l'algorithme de Viterbi sur l'ensemble des variantes phonétiques. Cet ensemble est organisé sous forme de graphe, structure qui s'intègre particulièrement bien à l'algorithme de Viterbi. La construction du modèle acoustique se fait de façon classique avec une procédure itérative de l'algorithme de Baum. La seconde section expose et commente les résultats obtenus.

MEILLEURE VARIANTE ET APPRENTISSAGE

Selon le formalisme adopté en théorie de l'information, la meilleure variante est celle qui réalise:

$$P(\overline{V}|A) = \max_{V \in \mathcal{V}} P(V|A)$$

Où $\mathcal{V}$ est l'ensemble des variantes de la phrase prononcée. En utilisant la loi de Bayes et en supposant P(A) indépendant de V, on a:

$$\overline{V} = \operatorname*{argmax}_{V \in \mathcal{V}} P(A|V).P(V)$$

La sélection de la meilleure variante est donc un problème de décodage. La donnée d'un modèle acoustique initial permet de calculer P(A|V) et la donnée d'une distribution sur les variantes permet de déterminer P(V).

La sélection est faite par l'algorithme de Viterbi auquel on fournit, pour chaque phrase du corpus d'apprentissage, la suite des observations acoustiques lui correspondant et le graphe des variantes phonétiques établi par le phonétiseur. Après avoir, pour chaque phrase du corpus d'apprentissage, sélectionné la meilleure variante, on construit le modèle acoustique du locuteur.

L'ALGORITHME DE VITERBI

Les notations sont identiques à celles adoptées au chapitre I, au cours duquel les algorithmes de Baum et de Viterbi ont été rapidement introduits. L'architecture qui permet l'exécution de ces algorithmes consiste à associer au calcul de la probabilité α (e_j, x) un point (e_j, x) du plan défini par l'axe des états et l'axe temporel. Entre les points du plan, trois sortes de transitions sont possibles:

La boucle (flèche horizontale)
on reste dans le même état: e_i, avec émission du symbole a_k.
La transition nulle (flèche verticale en pointillés)
on passe de l'état e_i à l'état e_j sans émission de symbole.
La transition non nulle (flèche diagonale)
on passe de l'état e_i à l'état e_j, avec émission du symbole a_k

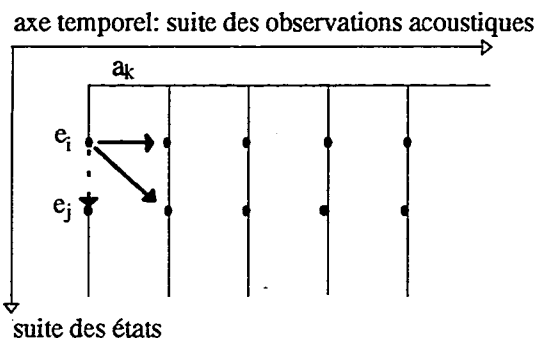

FIGURE 7. Viterbi: opérations élémentaires

Considérons une suite d'observations acoustiques a(0), a(1)...a(x), a(x+1)...a(T) et une suite d'états comprenant un état initial e_0 et un état final e_F. Soit α (e_j,x) la probabilité que les x premières observations aient été produites par tous les chemins qui partent de l'état initial et aboutissent à e_j. Le calcul de α (e_j, x) dépend du calcul des α (e_i, x ou x-1) où e_i un état précédent e_j. On peut avoir:

$e_i = e_j$ La transition de e_i à e_j est une boucle (flèche horizontale: émission de l'observation a(x)) de probabilité $a_{ii}.b_{iia(x)}$

$e_i \neq e_j$ La transition de e_i à e_j peut être nulle (flèche verticale: sans émission) de probabilité v_{ij}

La transition de e_i à e_j peut être non nulle (flèche diagonale: émission de l'observation a(x)) de probabilité $a_{ij}.b_{ija(x)}$

α (e_j,x) vaut alors:

$$\alpha(e_j,x) = \sum_{e_i} [\alpha(e_i,x-1) \cdot a_{ij} \cdot b_{ija(x)} + \alpha(e_i,x) \cdot v_{ij}]$$

Quand le calcul s'effectue sur l'ensemble de tous les chemins, on obtient la passe-avant de l'algorithme de Baum. Pour l'algorithme de Viterbi, seul le chemin qui assure la production

de la suite des observations acoustiques avec une probabilité maximale est retenu. Ainsi la probabilité que les x premières observations aient été produites par le chemin de probabilité maximale allant de l'état initial à l'état e_j vaut:

$$\gamma(e_j,x) = \max_{e_i} [\, \gamma(e_i,x-1) \cdot a_{ij} \cdot b_{ija(x)} \,,\, \gamma(e_i,x) \cdot v_{ij} \,]$$

L'ensemble des états qui à chaque étape réalisent ce maximum, définit le chemin de probabilité maximale.

Remarquons que le parcours en sens inverse du chemin de probabilité maximale permet de segmenter les données. En effet, le chemin passe obligatoirement par l'état initial et l'état final de chaque automate de la transcription phonétique. Ainsi, si l'on conserve une trace des couples (e_j,x) où e_j est un état du chemin de probabilité maximale et x l'instant auquel il a été emprunté, on peut, en notant les instants associés aux états finaux et initiaux de chaque automate, obtenir une segmentation de la suite des observations acoustiques en phonèmes:

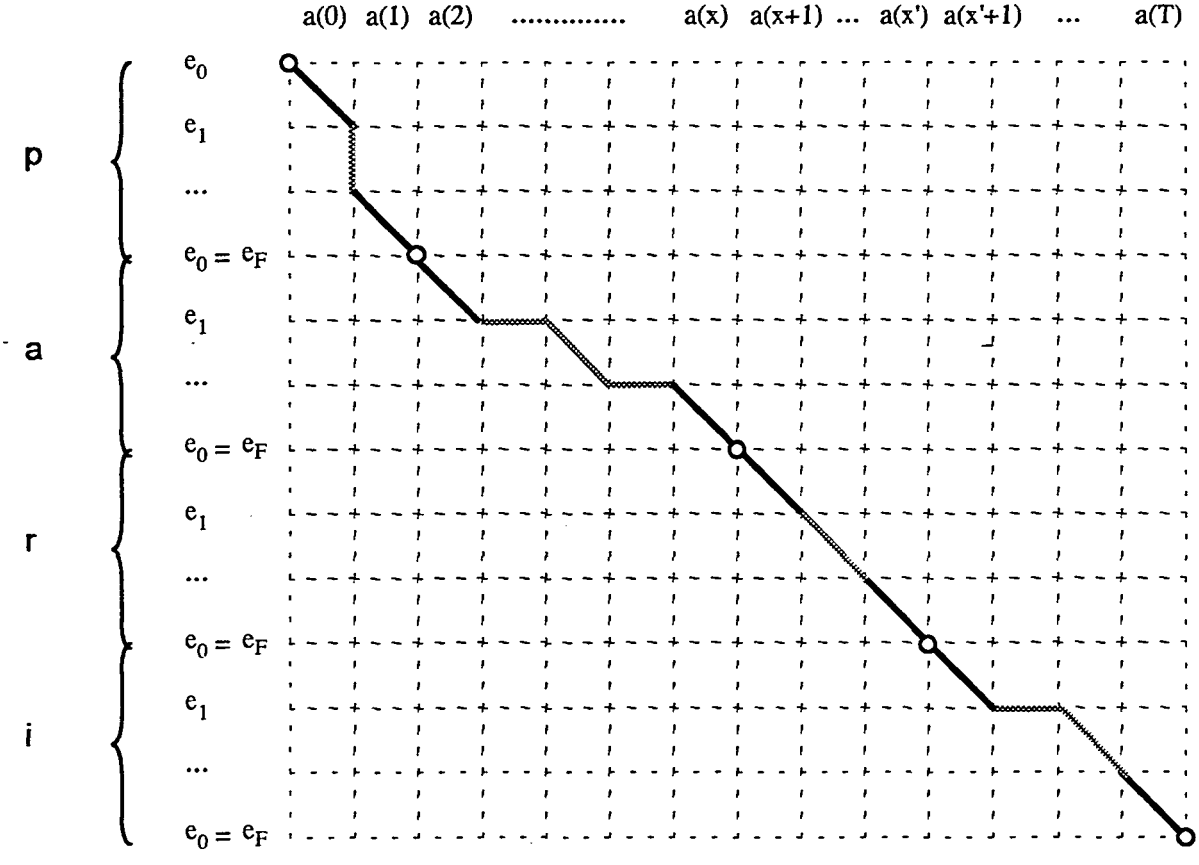

FIGURE 8. Viterbi: treillis

La suite des observations acoustiques de a(0) à a(1) correspond au phonème [p], la suite a(2) à a(x) au phonème [a], a(x+1) à a(x') au [r] et a(x'+1) à a(T) au [i].

VITERBI SUR LE GRAPHE DES VARIANTES

La suite des états et les arcs qui les joignent sont définis par la structure de l'automate markovien utilisé. Par exemple, dans le cas des automates phonétiques, la suite des états de deux phonèmes consécutifs (chacun modélisé par un automate à sept états) est la suivante:

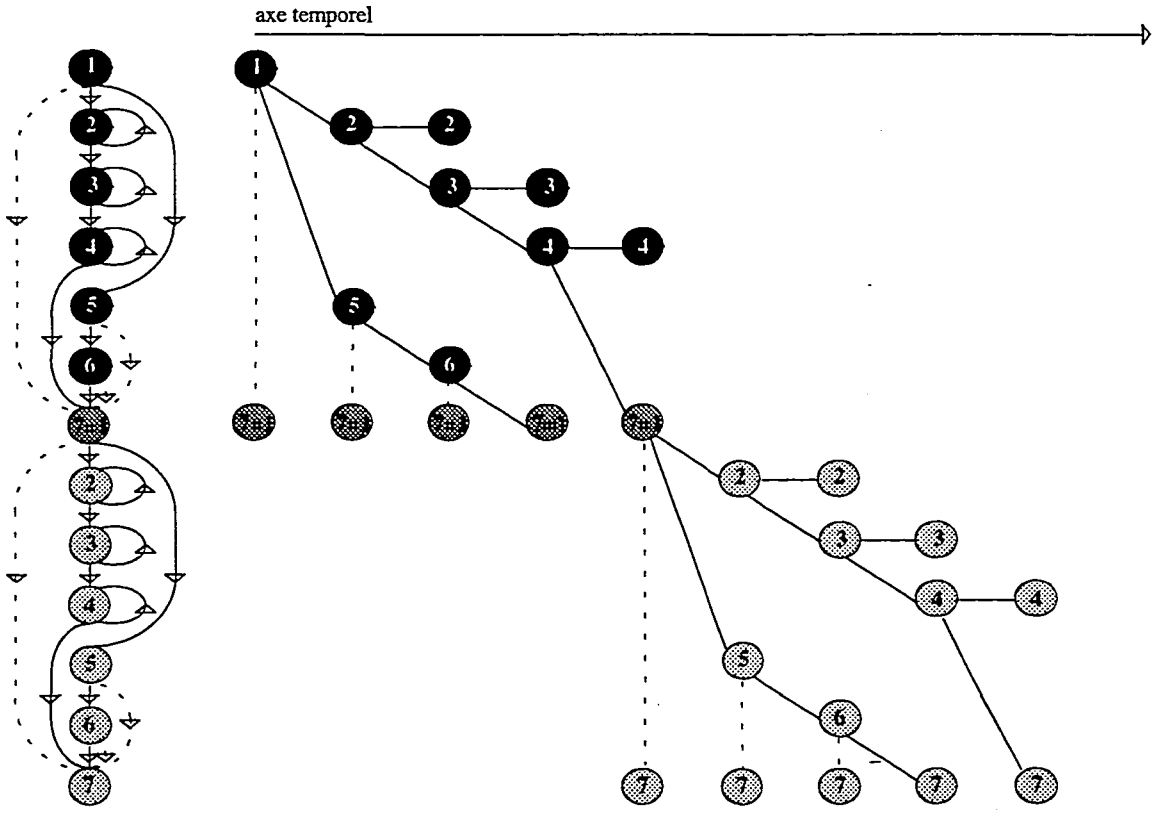

FIGURE 9. Viterbi sur un graphe de variantes 1

L'état final du premier automate est confondu avec l'état initial de l'automate suivant. Pour effectuer l'algorithme de Viterbi sur le graphe des variantes, il suffit, à la fin de chaque automate, de confondre l'état final d'un automate avec tous les états initiaux des automates pouvant le suivre. A cette intersection, le choix d'un automate plutôt qu'un autre se fait selon le chemin de probabilité maximale.

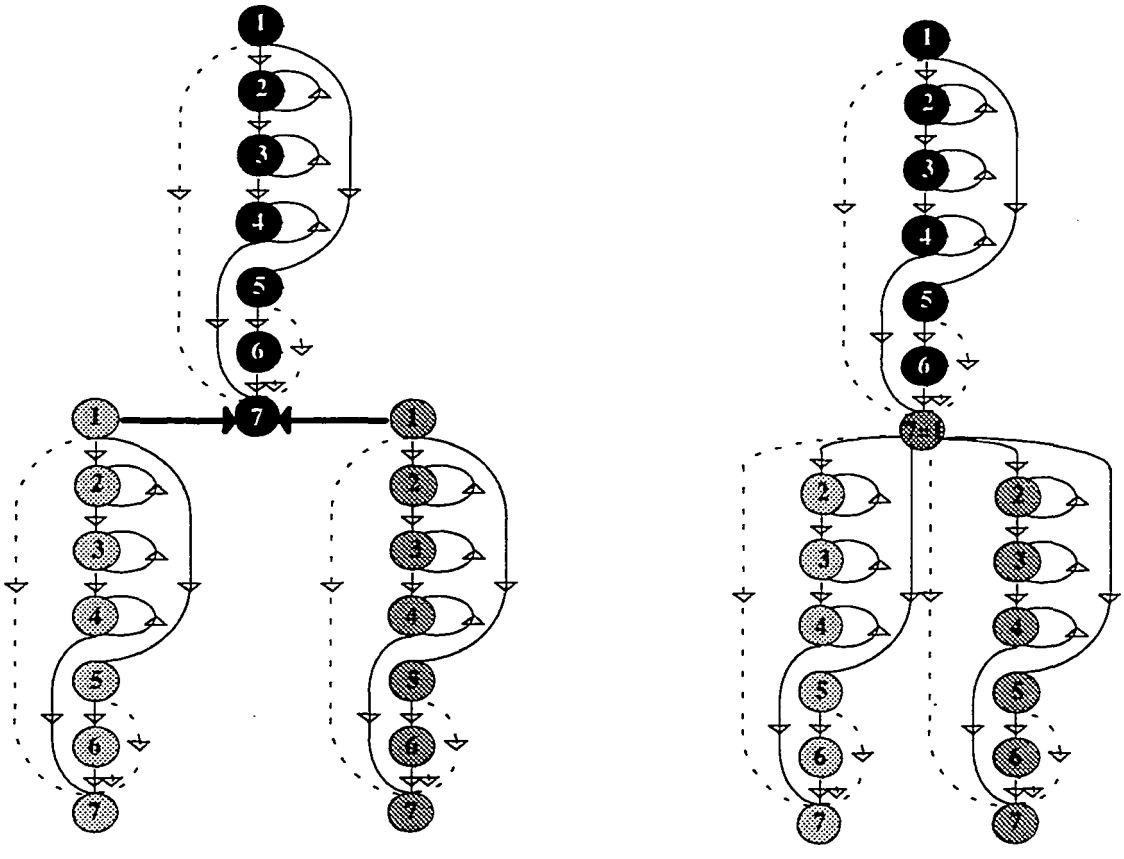

FIGURE 10. Viterbi sur un graphe de variantes 2

Ce procédé peut poser des problèmes si on concatène des automates dont l'état initial est réciproque (c'est à dire sur lequel il peut y avoir une boucle). On pourrait, alors, commencer par emprunter l'état initial d'un automate pour ensuite décider d'aller vers un autre automate. Cette éventualité ne se présente pas pour les automates phonétiques envisagés. Il suffirait, dans le cas contraire, de lier les états initiaux par une transition associée à la probabilité P(V) de la variante considérée.

Nous avons considéré, en première approximation, que les variantes étaient équiprobables. Le problème revient par conséquent à trouver la variante $\overline{V}$ qui réalise:

$$\overline{V} = \underset{V \in \mathcal{V}}{\operatorname{argmax}}\ P(A|\ V)$$

Sous cette hypothèse le choix de l'automate qui suit se fait selon le chemin de probabilité maximale à la manière d'un Viterbi classique.

LE GRAPHE DES VARIANTES

Deux méthodes ont été essayées pour construire le graphe des variantes. La première consiste en un alignement de Viterbi entre une variante élue référence et toutes les autres

variantes. Cette méthode a plusieurs inconvénients. Tout d'abord, pour être rigoureux, il faudrait comparer toutes les variantes deux à deux, chacune étant élue référence tour à tour. Ensuite, lorsque les variantes sont nombreuses, la gestion des arcs ou des noeuds à ajouter au graphe initial devient lourde et délicate. Enfin, et c'est là le point le plus important, le graphe génère des variantes non prévues. Cette approche fut abandonnée au bénéfice de la seconde, désignée ici sous le terme de "graphe naïf et graphe réduit", qui repose sur une simplification et une optimisation de l'algorithme de Moore dans le cas des automates acycliques [1][79].

GRAPHE NAIF ET GRAPHE REDUIT

La construction se fait en deux temps. La première étape consiste en la construction d'un graphe dit "naïf", qui décrit l'ensemble des variantes sans souci de redondance. La première variante engendrée par le phonétiseur constitue la première branche du graphe. Pour la seconde, on parcourt le graphe et dès qu'un phonème de la seconde variante diffère de celui de la première, on crée une nouvelle branche partant du noeud litigieux et allant jusqu'au noeud final du graphe. La seconde étape consiste à regrouper, hauteur par hauteur, les noeuds équivalents. Le graphe obtenu est appelé graphe réduit.

LE GRAPHE NAIF

Voici un exemple du graphe naïf des variantes de la phrase "un petit lac"[13].

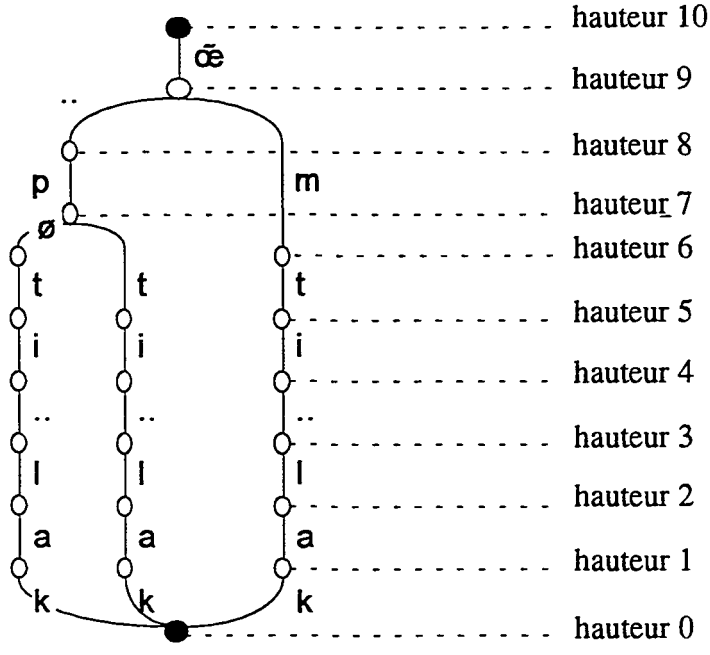

FIGURE 11. Graphe naïf

[13] la variante[ɛ̃] de "un" a été omise pour une meilleure lecture du graphe.

LE GRAPHE REDUIT

La hauteur d'un noeud dans un graphe acyclique est définie comme la longueur du plus long chemin allant de ce noeud à l'état final du graphe. L'ensemble des noeuds de même hauteur i, est noté $\mathcal{H}i$.

En travaillant de bas en haut, c'est à dire par ordre croissant de hauteur, on crée dans chaque ensemble $\mathcal{H}i$ des classes d'équivalences de noeuds. Deux noeuds sont équivalents si:

- ils sont de même nature, c'est-à-dire terminaux ou non terminaux.
- les transitions qui en partent sont porteuses d'une étiquette (ici phonétique) identique, et si elles aboutissent au même noeud.

Une fois ces classes d'équivalences établies, on confond tous les éléments d'une même classe. Après une mise à jour des arcs aboutissant aux noeuds fusionnés et une renumérotation des noeuds traités, on obtient le graphe réduit.

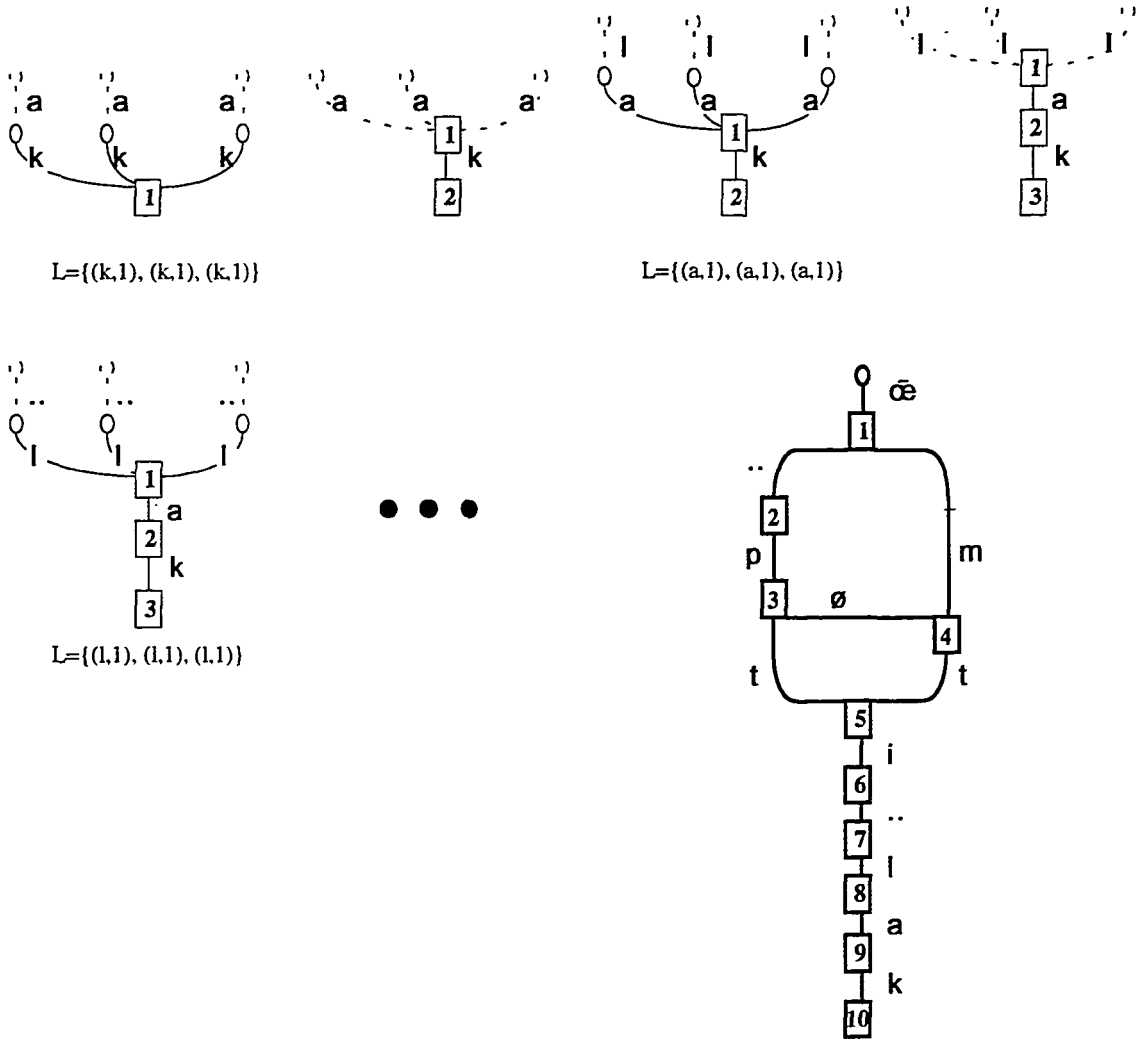

FIGURE 12. Graphe réduit

LES SILENCES

Le silence n'a été considéré jusqu'alors que comme séparateur de mots. Or la place des silences dans le discourt pourrait faire, à lui seul, l'objet d'une thèse. La variante, absence ou présence d'un silence entre deux mots, n'est pas traitée par le phonétiseur, car si on l'incluait dans l'ensemble des règles, le nombre des variantes deviendrait alors très important. Ceci pourrait poser un problème de stockage du graphe naïf en mémoire lors de sa construction. Le traitement des silences s'effectue directement sur le graphe naïf avant réduction, en ajoutant des arcs. Cet ajout nécessite de modifier légèrement le calcul des hauteurs, mais n'altère pas la construction du graphe réduit. Pour les modèles phonétiques, sept automates, dit pseudo-phonétiques ont été envisagés pour décrire le silence.

- Les phonèmes [p], [t] et [k] sont modélisés, chacun, par un couple d'automates: /op/, /ot/ et /ok/ traduisent le phénomène d'occlusion de l'appareil phonatoire qui précède l'explosion. /bp/, /bt/ et /bk/ traduisent l'explosion. Quand un silence précède chacun de ces phonèmes, les automates qui traduisent l'occlusion sont inutiles.
- /.v/ et /.c/ représentent les transitions entre une fin de mot et une pause. /.c/ est utilisé quand le mot finit par une consonne, /.v/ quand il finit par une voyelle.
- /.g/ marque l'impulsion glottale devant toute voyelle précédée d'une pause.
- /../ désigne aussi bien une pause respiratoire qu'un silence de début ou fin de phrase. Contrairement aux autres automates qui possèdent sept états, celui ci n'en possède que deux.

Quand on évite un silence, il faut aussi éviter les automates de transition /.c/, /.v/ et /.g/, et ajouter, si nécessaire, les automates qui traduisent l'occlusion des phonèmes [p], [t] et [k]. Enfin, si l'absence de silence place deux phonèmes identiques côte à côte, on prévoit de pouvoir en omettre un des deux.

ITERATION SUR LA SELECTION

La construction du modèle acoustique d'un locuteur, à partir de la meilleure variante, se fait par une méthode itérative qui utilise l'algorithme de Baum.

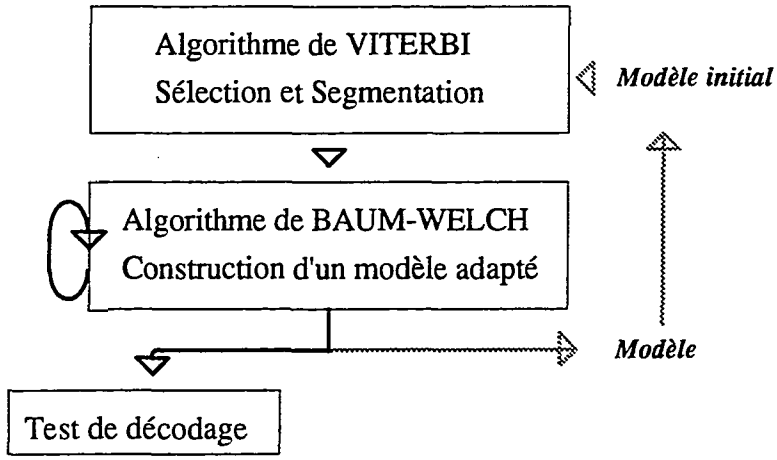

FIGURE 13. Variantes phonétiques: processus de sélection

Après avoir calculé les meilleures variantes sur l'ensemble du corpus d'apprentissage, on construit, en cinq passes de l'algorithme de Baum, un premier modèle qui est utilisé pour recalculer les meilleures variantes. Ce procédé s'arrête quand le décodage ne révèle plus d'améliorations.

Le test de décodage consiste, d'une part, en un décodage de Viterbi sur un texte de test qui détermine la suite d'automates phonétiques qui a la plus grande probabilité de générer les observations acoustiques et d'autre part, en une évaluation qui est faite par un alignement de Viterbi entre la suite de phonèmes décodés et une transcription phonétique référence du texte de test.
Le décodage s'effectue par une confrontation de la suite des observations acoustiques et d'un hyper-automate constitué de tous les automates phonétiques mis en parallèle et d'une transition de bouclage.

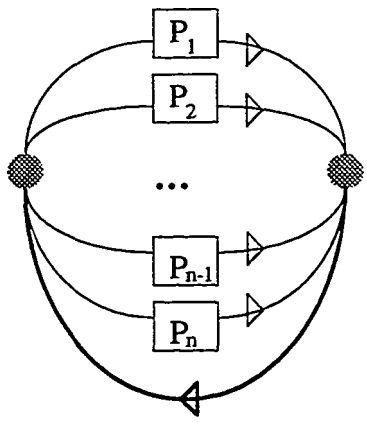

FIGURE 14. Décodage en boucle

Pour tracer un chemin dans cet hyper-automate, à partir de l'état initial, il suffit d'emprunter une transition vers un automate phonétique, puis les transitions de l'automate phonétique, puis la transition de bouclage, puis de nouveau une transition vers un

automate phonétique, etc. Parmi l'ensemble des chemins, celui de plus forte probabilité donne la suite des phonèmes décodés.

La transition de bouclage, allant de l'état final à l'état initial s'emprunte avec une probabilité de 1, les autres transitions sont associées à des probabilités dites "diphones". Dans le cas d'un décodage en mot isolé, ces probabilités peuvent être calculées sur un corpus phonétique en comptant le nombre d'apparitions des couples de phonèmes. Dans le cas de la parole continue, on ne peut pas utiliser la même méthode, car les diphones qui apparaissent à l'intérieur des mots ne décrivent pas l'ensemble des diphones du continu. 35% des diphones rencontrés en continu n'apparaissent pas en isolé [84]. On peut, néanmoins, collecter des diphones fictifs en utilisant un modèle de langage biclasse. L'estimation des probabilités diphones en frontière de mot se fait de la manière suivante.

$$\overline{P}(P_f | P_d) = \sum_{C_d} \sum_{C_f} P(P_d | C_d).\ P(C_d | C_f)\ .P(C_f | P_f)$$

Où $P(P_d | C_d)$ est la probabilité que le phonème P_d apparaisse au début d'un mot de classe C_d et où $P(C_d | C_f)$ est la probabilité que la classe C_d suive la classe C_f. On trouvera dans [33] des résultats comparatifs entre un calcul des probabilités diphones obtenu avec cette méthode et obtenu le compte des diphones sur un corpus préalablement phonétisé prenant en compte les liaisons. De meilleurs résultats sont obtenus avec la seconde méthode. Il serait intéressant de calculer ces probabilités sur une transcription phonétique personnalisée du corpus.

EXPERIMENTATION

Le premier locuteur testé, MOY, a été choisi surtout parce qu'il avait fait l'objet de recherches antérieures sur la parole continue et que j'avais à disposition un modèle initial pour pouvoir calculer les premières meilleures variantes.

Le modèle continu initial a été construit en amorçant la procédure itérative de l'algorithme de Baum avec un modèle en syllabes isolées. Les résultats du décodage acoustique[14] avec ce modèle sont les suivant:

[14] Les résultats sont, dans ce chapitre, toujours ceux d'un décodage acoustique

Exp1: Modèle 0					
nb-d'automates	reconnaissance	substitution	omission	insertion	% erreur[15]
40	71,5	23,7	5	7,4	33,6

Les probabilités diphones utilisées, sont les probabilités diphones calculées sur les syllabes isolées. Si on améliore ces probabilités en utilisant un modèle de langage biclasse pour estimer les probabilités inter mots, on obtient, avec le même modèle, le résultat suivant.

Exp2. Modèle 0, probabilités diphones: modèle biclasse					
nb d'automates	reconnaissance	substitution	omission	insertion	% erreur
40	76	19,7	4,3	8,4	29,8

Le décodage acoustique a l'inconvénient de ne pas donner une idée très juste de la valeur de la reconnaissance. En effet, lors de l'évaluation, on compare la suite de phonèmes décodés à une suite de phonèmes référence, obtenue en phonétisant le corpus de test de façon standard. Cette référence ne correspond pas nécessairement à ce que le locuteur a effectivement prononcé. L'idéal serait d'avoir pour référence la transcription phonétique exacte de ce qui a été prononcé. On peut avoir une première idée du décalage qu'il existe entre une transcription standard et la réalité en comparant les résultats précédents à ceux d'un décodage dont la référence est la transcription phonétique personnalisée du corpus de test, obtenue par le processus de sélection avec le modèle 0.

Exp3. Modèle 0, proba dipho biclasse, référence personnalisée					
nb d'automates	reconnaissance	substitution	omission	insertion	% erreur
40	77,6	18,3	4,2	5,9	26,7
130	80,7	15,7	3,6	5,5	23,5

La construction du modèle personnalisé du locuteur MOY, se fait en cinq passes de l'algorithme de Baum initialisé avec le modèle 0. Les résultats obtenus après cet apprentissage sont les suivants

[15] $\%\,\text{erreur} = \dfrac{\text{Substitutions} + \text{Ommissions} + \text{Insertions}}{\text{Total} + \text{Insertions}}$

Exp4. Modèle 1, proba dipho biclasse, référence personnalisée. Modèle initial= modèle 0					
nb d'automates	reconnaissance	substitution	omission	insertion	% erreur
40	76,9	18,4	4,7	4,5	26,4
130	80,8	15,6	3,6	4,9	22,9

En itérant le processus, avec le modèle 1 comme modèle initial, on obtient:

Exp5. Modèle 2, proba dipho biclasse, référence personnalisée. Modèle initial= modèle 1					
nb d'automates	reconnaissance	substitution	omission	insertion	% erreur
40	77,4	17,8	4,8	4,8	26,1
130	80,9	15,5	3,6	4,5	22,6

Une itération supplémentaire du processus n'apporte aucune amélioration. Bien qu'un progrès soit enregistré, on ne peut pas conclure que la modélisation de la variabilité phonétique, telle qu'elle est faite ici, conduise à une amélioration notable des performances du système. Si on fait une comparaison entre la phonétisation manuelle du corpus de test et la phonétisation issue du processus de sélection, on peut s'apercevoir que les divergences portent massivement sur la position du silence et l'aperture des |e|, |ø| et |o|.

transcription manuelle / transcription personnalisée	40 automates	130 automates
phonèmes égaux	96,3	96,3
insertions	0,5	0,2
substitution	1,7	0,7
élision	- 1,5	1,7

Analyse des erreurs (insertions, substitutions, élisions):

erreurs sur le silence	38,6	31,8
erreurs sur l'aperture des |e|, |ø| ou |o|	37,7	36,2
erreurs sur [ø]	10,5	14,1
erreurs sur les assimilations consonantiques	4,4	5,2
erreurs sur les nasalisations	2,6	2,6
erreurs sur les chutes de consonnes	2,6	5,2
autres	3,6	4,8

L'importance de la place des silences est mise en évidence par l'expérience suivante pour laquelle les seules variantes sont celles relatives aux silences, dont on cherche la place optimale:

Exp6. Modèle 2, proba dipho biclasse, référence standard. Silences optimaux.					
nb d'automates	reconnaissance	substitution	omission	insertion	% erreur
40	77,2	17,7	5,1	4,2	25,9

Trois autres locuteurs, UBU, FAR, et CRI, ont fait l'objet de test, on retrouve pour chacun des résultats similaires à ceux obtenus pour MOY:

UBU Exp3. Modèle initial, proba dipho biclasses, référence personnalisée					
nb d'automates	reconnaissance	substitution	omission	insertion	% erreur
40	65	23,8	11,2	6,5	38,9
UBU Exp5. Modèle 2, après apprentissage et sélection					
40	66,4	23,1	10,5	6,3	37,5
FAR Exp3. Modèle initial, proba dipho biclasses, référence personnalisée					
nb d'automates	reconnaissance	substitution	omission	insertion	% erreur
40	73,4	20,1	6,6	5,7	30,6
FAR Exp5. Modèle 2, après appretissage et sélection					
40	73,1	20	6,9	4,7	30,2
CRI Exp3. Modèle initial, proba dipho biclasses, référence personnalisée					
nb d'automates	reconnaissance	substitution	omission	insertion	% erreur
40	76,7	17,9	5,4	6,5	27,9
CRI Exp5. Modèle 1, après apprentissage et sélection					
40	77,2	17,1	5,1	5,3	26,7

Trois points ressortent de ces résultats: la nécessité d'avoir de bonnes probabilités diphones, l'importance de la place des silences et enfin, les meilleures performances des 130 automates contextuels par rapport aux 40 automates phonétiques. Il est à noter que ces résultats sont ceux d'un décodage acoustique dont j'ai souligné les inconvénients en début de section. Ils donnent cependant une idée de l'évolution du taux de reconnaissance.

CONCLUSION

Nous avons présenté dans ce chapitre un phonétiseur capable de générer plusieurs variantes phonétiques d'une phrase donnée. Ce phonétiseur prend en compte les liaisons, la chute des e caducs, les variations en aperture des |e|, |ø| et |o|, les mots sujets à la fois à synérèse et diérèse, un certain nombre de cas particuliers, les phénomènes d'assimilation consonantique régressive et les phénomènes de nasalisation des occlusives. A partir du graphe des variantes, on calcule à l'aide d'un processus de sélection qui repose sur un alignement de Viterbi, la meilleure variante, étant donnée la suite des observations acoustiques de la phrase prononcée par un locuteur. Sur un corpus d'apprentissage, ainsi phonétisé, on calcule, par une procédure itérative utilisant l'algorithme de Baum, un modèle acoustique adapté au locuteur. Un décodage acoustico-phonétique permet d'évaluer ce modèle. Les résultats obtenus ne conduisent pas à une amélioration sensible du taux de reconnaissance. Ils montrent l'importance des probabilités diphones utilisées au décodage et l'importance de la place du silence dans la phrase.

Avant d'élargir le champ des variations prises en compte, il serait intéressant d'approfondir trois points.

- En premier lieu, le calcul des probabilités diphones pourrait se faire sur un corpus préalablement phonétisé selon la procédure de phonétisation avec variantes et de sélection que l'on vient de présenter. Si on dispose d'un corpus suffisamment important, et de suffisamment de locuteurs, on peut espérer avoir un corpus dans lequel les probabilités diphones du continu sont bien représentées.
- Le second point concerne l'importance de la place du silence. De façon plus générale, il serait intéressant de faire plusieurs expériences ne prenant en compte, chaque fois, qu'une fraction des règles proposées. On pourrait ainsi s'apercevoir des effets de chacune.
- Enfin, la probabilité des variantes, ici uniforme, pourrait être affinée en pondérant les règles de phonétisation du second module selon leurs fréquences d'utilisation. La probabilité d'une variante s'exprimerait, alors, en fonction du produit des poids associés aux règles appliquées pour obtenir la variante considérée. La pondération des règles pourrait être établie de la manière suivante. Après avoir procédé à la phonétisation avec variantes d'un corpus d'apprentissage, puis à la sélection des meilleures variantes pour plusieurs locuteurs, compter, sur ces corpus sélectionnés, le nombre de fois où une règle est appliquée. Ce

nombre, normalisé par le nombre total des règles appliquées, peut donner un estimateur du poids associé à la règle considérée. Ce problème a été abordé différemment par M. Cohen dans [22] et par M. Randolphe dans [78].

Le chapitre suivant propose une autre approche permettant de modéliser les effets de co-articulation. Cette approche est fondée sur un principe de classification des réalisations acoustiques des phonèmes selon un arbre binaire et un critère de séparation des données. Cet arbre est communément appelé "arbre de décision". On construit un arbre par phonème. Les réalisations acoustiques d'un phonème sont récursivement divisées en deux parties. La division se fait en posant des questions sur la nature des contextes dans lesquels le phonème étudié apparaît. La division retenue est celle qui maximise le critère de séparation. A partir des feuilles de l'arbre constituées d'un ensemble de réalisations acoustiques, on construit un modèle dit allophonique.
Le chapitre suivant comporte deux parties. La première expose l'ensemble des étapes nécessaire à la construction de l'arbre. La seconde présente et discute les résultats obtenus par cette méthode.

CHAPITRE III: VARIABILITE ALLOPHONIQUE

INTRODUCTION

Pour modéliser les phénomènes de co-articulation, les recherches se sont orientées vers des modèles acoustiques dont l'unité de base dépend du contexte dans lequel elle apparaît. Ces unités, dans le cadre d'un système de reconnaissance à grand dictionnaire, doivent répondre au double critère suivant:

- Consistance: pour rendre possible la discrimination entre unités, différentes manifestations de la même unité doivent avoir les mêmes caractéristiques.
- Apprentissage: les paramètres des unités doivent pouvoir être estimés à partir d'un corpus de taille raisonnable.

En se référant à l'article de K.F. Lee [53], un tour d'horizon des principales unités considérées dans la littérature, permet d'établir le tableau suivant qui résume l'adéquation d'unités fréquemment utilisées au critère considéré.

Type de modèle	**Consistance**	**Apprentissage**
modèle de mot	oui	mauvais
modèle phonétique	non	bon
modèle multiphone	oui	dépend de la langue
modèle phonétique dépendant du mot	oui	bon sous condition de fusion
modèle phonétique dépendant du contexte	oui	bon sous condition de fusion

- Le modèle de mot est le plus naturel puisque c'est ce qu'on cherche à identifier. Il permet de modéliser la variabilité intra mot. Son inconvénient, on l'a vu au chapitre I, est que chaque mot est appris individuellement. Il est nécessaire, pour construire le modèle acoustique, d'avoir plusieurs prononciations de chaque mot. Si cette approche peut convenir à des systèmes de reconnaissance sur petit dictionnaire [77], elle est par contre, à proscrire pour des grands dictionnaires.
- Le modèle phonétique présente l'avantage de n'avoir qu'un petit nombre d'automates. Il n'y a donc pas de problème d'estimation des paramètres à l'apprentissage. Par contre, la réalisation acoustique d'un phonème est fortement influencée par les phonèmes qui l'entourent. Les unités phonétiques ne sont pas consistantes, elles sont incapables de prendre en compte les phénomènes de co-articulation.

- Le terme de "modèle multiphone" désigne les modèles dont les unités sont constituées d'un groupe de phonèmes, comme l'unité syllabique. Si en français 6.300 syllabes permettent de décrire un dictionnaire de 200.000 formes [33] leur nombre est plus important en anglais [54]. L'intérêt de l'unité syllabique dépend donc de la langue considérée. En particulier, son utilisation en français permet de traiter naturellement les apostrophes et les liaisons et donne accès à des dictionnaires de très grande taille [67]. Ces unités, d'un point de vue acoustique, prennent en compte la variabilité intra unité, le problème de co-articulation entre deux unités consécutives reste entier.
- Le modèle phonétique dépendant du mot est un compromis entre le modèle de mot et le modèle phonétique [21]. Pour ce modèle, les paramètres de l'automate phonétique dépendent du mot dans lequel il apparaît. Comme le modèle de mot, ce modèle prend en compte la variation phonologique intra mot et nécessite aussi un nombre de données d'apprentissage important. Si un mot n'a pas été fréquemment observé, ses paramètres peuvent être interpolés grâce à un modèle phonétique classique.
- Le modèle dépendant du contexte est un modèle similaire au précédent, à ceci près que les paramètres de l'automate phonétique dépendent du contexte phonétique dans lequel il apparaît. Quand le contexte est constitué des phonèmes précédant et suivant le phonème considéré, l'automate contextuel est appelé "triphone". Ainsi un phonème P pris dans deux contextes différents C_1PC_2 et $C'_1PC'_2$ (où $C_1 \neq C'_1$ et/ou $C_2 \neq C'_2$) correspond à deux triphones différents. Les triphones sont très nombreux, différentes méthodes permettent de construire des classes d'équivalences de contextes qui réduisent le nombre des paramètres à estimer et rendent alors possible l'apprentissage. La méthode présentée dans ce chapitre, en fait partie.

Le chapitre III est organisé en deux parties. Dans la première, après avoir présenté quelques autres méthodes de construction des phonèmes contextuels, je développerai une méthode qui utilise les arbres de décision appelés ici arbres phonologiques. La seconde partie expose et discute les résultats obtenus.

LES ARBRES PHONOLOGIQUES

La définition de classes équivalentes de contextes peut se faire de différentes manières. A.M. Derouault dans l'article [27] construit ses classes en accord avec une classification articulatoire des phonèmes. Seuls les contextes droits sont considérés. Dix types de contextes sont retenus. A chaque phonème, est associé un ensemble de phonèmes contextuels. L'ensemble des automates phonétiques contextuels définit un nouvel ensemble d'automates. Les résultats exposés au chapitre précédent prouvent que ce jeu d'automates contextuels, aussi simple soit-il, améliore de façon significative les performances du système. Une approche similaire est proposée par L. Deng dans [24] dans le cadre d'une reconnaissance markovienne à paramètres continus. S'étant fixé un certain nombre de classes de contextes, également établies selon une classification articulatoire des phonèmes, Deng considère trois types de phonèmes contextuels. Ceux pour lesquels on ne considère que le contexte droit, ceux pour lesquels on ne considère que le contexte gauche et enfin les triphones pour lesquels on considère à la fois les contextes droit et gauche.

Une autre approche, sensiblement différente, consiste à définir les classes d'équivalence de phonèmes non plus selon une classification articulatoire, mais selon une mesure de similitude. Cette mesure peut être définie à l'aide de l'entropie croisée [45], du maximum d'information mutuelle [30] ou de la quantité d'information théorique comme le fait K.F. Lee dans [54]. Cette dernière méthode conduit à un nouvel ensemble d'automates appelés triphones généralisés, dont le nombre a été fixé en fonction des performances du système[16] et dont la structure est identique à celle des automates phonétiques indépendants du contexte. Le taux d'erreur entre une reconnaissance faite avec les automates phonétiques indépendants du contexte et une reconnaissance faite en utilisant les triphones généralisés est réduit de moitié. Il y a cependant deux inconvénients à cette approche. En premier lieu, pour éviter une trop grande dépendance au corpus d'apprentissage ou pour palier aux lacunes d'un apprentissage nécessairement réduit, les paramètres des triphones généralisés doivent être lissés. Ce lissage consiste en une combinaison linéaire des paramètres des triphones généralisés et des paramètres des automates phonétiques. Ce lissage est trop fort, en particulier pour les triphones "mal" appris. L'autre

[16] Les performances du système croissent avec le nombre des triphones généralisés jusqu'à ce qu'ils soient trop nombreux pour que l'apprentissage se fasse dans de bonnes conditions.

inconvénient est que, lorsqu'on rencontre dans le corpus de test un triphone ne faisant pas partie du corpus d'apprentissage, on utilise alors l'automate phonétique correspondant. Ceci a pour effet de dégrader sévèrement les performances du système. Une approche alternative est d'utiliser les arbres de décision [55]. Initialement, un automate markovien est appris pour chaque triphone rencontré. Un arbre de décision est utilisé pour opérer une classification de ces automates. Le critère de séparation fait intervenir la même mesure que celle définie pour l'élaboration des triphones généralisés. Comme un noeud entretient une relation de parenté avec le noeud dont il provient, le lissage d'un noeud peut être fait par une méthode d'interpolation [42] faisant intervenir tous ses ancêtres. D'autre part, si un triphone n'a jamais été rencontré dans le corpus d'apprentissage, il sera représenté par l'automate allophonique qui correspond à la feuille obtenue en répondant négativement à toutes les questions posées. L'approche retenue par K.F Lee dans [55] est une approche hybride des deux précédentes. Un arbre de décision peu profond permet d'éviter l'utilisation des automates phonétiques pour le lissage. Le regroupement des feuilles se fait ensuite selon la méthode des triphones généralisés. Une comparaison du taux d'erreur obtenu entre l'utilisation exclusive des triphones généralisés et la méthode hybride montre un léger avantage de la seconde sur la première.

La méthode que nous testons dans ce chapitre a été proposée par L. Bahl dans [8]. Elle s'éloigne de celle proposée par K.F. Lee en ce qu'aucun apprentissage initial n'est effectué. Dans l'approche proposée par K.F. Lee, les arbres de décision sont utilisés pour regrouper et fusionner des automates contextuels, dans celle de L. Bahl, ils sont utilisés pour regrouper des suites d'observations acoustiques à partir des quelles seront calculées des automates contextuels. Les classes d'équivalences sont établies selon un critère de similitude des suites d'observations acoustiques plutôt que selon une similitude d'automates allophoniques. Le calcul des automates allophoniques n'est effectué qu'après avoir construit l'arbre. A partir des suites d'observations réunies dans chaque feuille, on calcule une forme moyenne qui rend compte au mieux de l'ensemble de ces observations. Cette forme est constituée d'une suite d'automates fénoniques concaténés.

Cette technique rencontre dans le domaine de la reconnaissance vocale, un vif intérêt. Elle a été appliquée, tout récemment, par Gupta [38] à un système markovien à paramètres continus. Outre sa simplicité de mise en oeuvre, elle conduit à un bon compromis entre la taille du corpus d'apprentissage et les performances du système. Son point faible réside dans l'importance et la diversité des données qu'elle nécessite pour sa mise en oeuvre.

Cet intérêt est partagé par d'autre domaine que celui de la modélisation acoustique. A titre d'exemple, les méthodes de classification fondée sur les arbres de décision ont été appliquées à la construction de modèles de langage, pour lesquels il est nécessaire de quantifier le contexte

dans lequel un mot apparaît [7], ou au calcul de la forme phonétique d'un mot inconnu, pour lequel on cherche une relation entre la traduction phonétique d'un groupe de lettres en fonction de son contexte graphique [9].

Dans les paragraphes suivants nous décrirons la préparation des données nécessaires à la construction de l'arbre. Nous exposerons ensuite les principes de l'algorithme de construction de l'arbre et enfin le calcul des automates allophoniques.

PREPARATION DES DONNEES

On désire dans cette première étape, obtenir pour chaque phonème, un ensemble important et diversifié de ses réalisations acoustiques. Pour ce faire, il faut d'abord se doter d'un corpus adéquat et le faire prononcer par plusieurs locuteurs. Les formes allophoniques d'un phonème sont calculées une fois pour toutes. Elles doivent être valables pour tout locuteur potentiel. Il est important d'avoir un nombre suffisamment important de locuteurs. On procède, ensuite, à un alignement entre phonèmes et suites d'observations acoustiques. Pour obtenir les meilleurs alignements possibles, les alignements seront calculés avec les modèles phonétiques de chaque locuteur et les observations acoustiques codées dans l'alphabet personnel du locuteur. Pour la construction des arbres, les données sont mélangées. Les observations acoustiques sont alors codées dans un alphabet commun.
Les lignes suivantes résument les principaux points de la préparation des données.

- Du signal on extrait, selon la procédure décrite au chapitre I, des vecteurs spectraux au rythme de cent par seconde. Ces vecteurs sont doublement étiquetés par une quantification vectorielle, une première fois dans l'alphabet personnel, une seconde fois dans l'alphabet commun.
- On calcule les modèles acoustiques propres à chaque locuteur.
- Pour chaque locuteur, on aligne, à l'aide de l'algorithme de Viterbi, ses observations acoustiques à la transcription phonétique du corpus d'apprentissage.
- Pour chaque alignement phonétique, on note:
  1. l'identité du phonème aligné P_0
  2. le contexte phonétique dans lequel il apparaît, c'est-à-dire les N phonèmes précédants et suivants P_0: P_{-N} ...P_{-1} P_1 ...P_N
  3. la suite des observations acoustiques alignée sur le phonème P_0 dans le contexte P_{-N} ...P_{-1} P_1 ...P_N

CONSTRUCTION DE L'ARBRE

On construit un arbre de décision par phonème. Ses feuilles représentent l'ensemble des allophones du phonème considéré. Pour chaque phonème Po, on cherche une partition de l'ensemble O des segments acoustiques associés à Po qui rassemble les segments dont le contexte phonétique est similaire. Cette partition s'obtient en effectuant des divisions binaires successives de O: O est divisé en deux parties disjointes, elles même de nouveau séparées en deux parties disjointes... La division se fait en posant des questions simples sur les contextes de Po, auxquels on répond par oui ou non. Comme l'illustre la figure suivante, à chaque question correspond une partition particulière.

Données pour P_0		q_1		q_2		q_3	
Contextes	Observations acoustiques	le contexte gauche est-il un a ?		le contexte droit est-il une liquide ?		le contexte droit est-il une voyelle nasale ?	
		Oui	Non	Oui	Non	Oui	Non
a P_0 a							
.. P_0 r							
.. P_0 z							
a P_0 l							

Les suites d'observations acoustiques sont symbolisées par des traits de natures différentes

FIGURE 15. Arbres phonologiques principe

Pour construire l'arbre il faut se fixer un critère de séparation. A chaque noeud la question choisie (et la partition retenue) est celle qui maximise ce critère. Il faut aussi se fixer un critère d'arrêt: quand le nombre des données associées à un noeud est inférieur à un seuil fixé θo, ou lorsque la valeur du critère de séparation est inférieure à un seuil[17] θc. Le résultat de cette construction est un arbre binaire dont les feuilles définissent les classes d'équivalences de contexte. A partir des suites d'observations acoustiques contenues dans une feuille, on construit la forme allophonique qui lui est associée. En phase de reconnaissance, étant donné un phonème et son contexte, on utilisera l'arbre de décision qui lui est associé pour déterminer l'automate allophonique à utiliser.

Soit Q l'ensemble des questions considérées. Soit n un noeud de l'arbre, et soit C(n,q) la valeur du critère pour la question $q \in Q$, au noeud n. Pour construire l'arbre on utilise un algorithme sous optimal qui sélectionne à chaque noeud la "meilleure" question $\overline{q}$ de Q:

[17] On peut ajouter une autre clause qui limite le nombre des noeuds de l'arbre.

1. Au départ, toutes les suites d'observations acoustiques (ou échantillons) relatives à Po sont à la racine.
2. Pour chaque noeud n non encore traité:
   2.1. Evaluer C (n,q) pour toutes les questions q$\in$ Q
   2.2. Si le critère d'arrêt est vérifié, transformer n en feuille terminale.
   2.3. Sinon, noter la question $\overline{q}$ qui réalise le maximum[18] de C (n,q) et créer deux noeuds fils, n_g et n_d successeurs du noeud n. A n_g on associe tous les échantillons correspondant à une réponse affirmative à la question $\overline{q}$, à n_d on associe les échantillons restants.

Les deux aspects les plus importants de cet algorithme résident dans le choix de l'ensemble des questions et le choix du critère de séparation. Les deux paragraphes suivants discutent de ces deux points.

CRITERE DE SEPARATION

Le critère de séparation est ici, comme dans [7][8][9], un critère qui maximise la vraisemblance des données associées à chaque noeud après application d'une question q. D'autres critères sont envisageables comme un critère fondé sur la quantité d'information [54] déjà mentionné dans l'introduction.

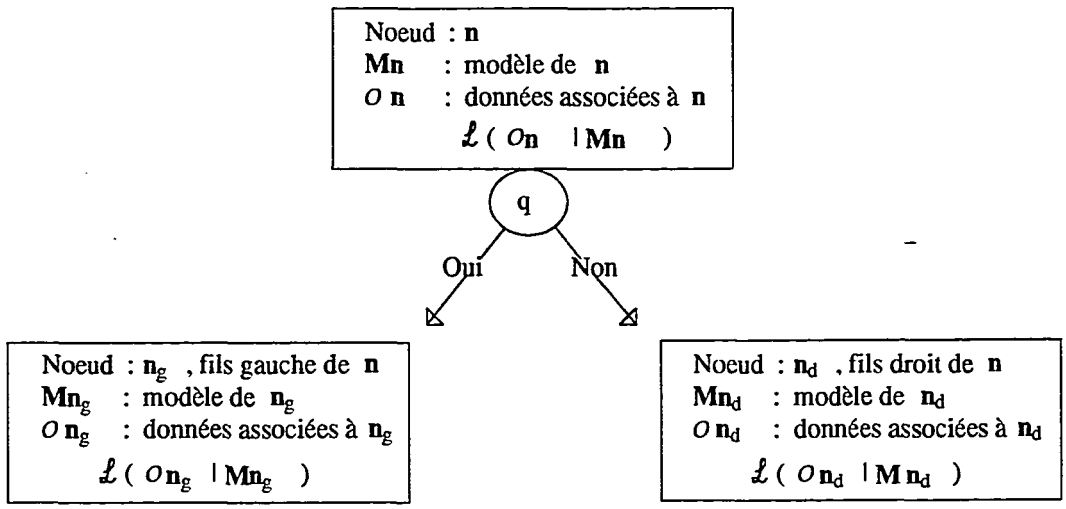

FIGURE 16. Critère de séparation

Soit $\mathcal{P}$ l'ensemble des phonèmes considérés et NP son cardinal. Soit O l'ensemble des réalisations acoustiques d'un phonème Po de $\mathcal{P}$. Soit $\mathcal{M}$ une classe de modèles paramétriques particulière permettant d'associer une probabilité à une suite d'observations acoustiques

[18] ou le minimum, selon le critère considéré.

donnée. Pour tout M∈ $\mathcal{M}$, soit $P_M(\omega)$ la probabilité de la suite d'observations acoustiques ω. Soit O_n l'ensemble des suites d'observations acoustiques associées au noeud n.

$$P_M(O_n) = \prod_{\omega \in O_n} P_M(\omega)$$

mesure l'adéquation du modèle M aux données: O_n.
Soit M_n le modèle de $\mathcal{M}$ qui réalise $P_{Mn}(O_n) \geq P_M(O_n)$ pour tout M∈ $\mathcal{M}$. On dira alors que M_n est le meilleur modèle pour O_n. $P_{Mn}(O_n)$ mesure la similarité des données de O_n. $P_{Mn}(O_n)$ est d'autant plus grand que les suites d'observations acoustiques de O_n sont similaires.

Une question q sépare O_n en deux sous ensembles disjoints $O_{nq,g}$ et $O_{nq,d}$: $O_n = O_{nq,g} \cup O_{nq,d}$.
Soient $M_{nq,g}$ et $M_{nq,d}$ les meilleurs modèles pour $O_{nq,g}$ et $O_{nq,d}$.

$$C(n,q) = \log\left[\frac{P_{Mnq,g}(O_{nq,g}).\ P_{Mnq,d}(O_{nq,d})}{P_{Mn}(O_n)}\right]$$

mesure le gain en similarité qui résulte de la séparation de O_n en $O_{nq,g}$ et $O_{nq,d}$.
Comme le but de la classification est de regrouper dans les feuilles de l'arbre les suites d'observations acoustiques similaires, cette quantité définira le critère de séparation d'un noeud en deux noeuds fils.

Les suites d'observations regroupées dans chaque feuille étant utilisées par la suite pour construire un automate markovien, il semble naturel d'utiliser comme classe de modèles $\mathcal{M}$, la classe des automates de Markov. Ce choix ne sera cependant pas retenu, car il est très coûteux en temps de calcul. La recherche du meilleur modèle M_n nécessiterait pour chaque noeud n, un apprentissage sur O_n et ainsi de suite, quand O_n est divisé en $O_{nq,g}$ et $O_{nq,d}$. Ce sont les modèles de Poisson qui sont retenus, essentiellement parce que la simplicité de leur forme permet d'évaluer rapidement leurs paramètres.

De plus, on fera l'hypothèse simplificatrice que les observations acoustiques de la suite $\omega = a_1, a_2, a_2, a_3 \ldots a_p$ sont indépendantes les unes des autres. Cette hypothèse n'introduit pas une trop grande erreur quand ω représente un seul phonème. Une conséquence de cette hypothèse est que l'ordre dans lequel les observations acoustiques apparaissent au sein du phonème n'a pas d'importance. On peut dès lors, considérer les histogrammes des suites d'observations acoustiques plutôt que les suites elles-mêmes. ω sera désormais représenté par son histogramme: $\nu_{\omega 1}, \nu_{\omega 2}, \nu_{\omega 3} \ldots \nu_{\omega Nds}$ où N_{ds} est la taille du dictionnaire de spectres (ici 200) et $\nu_{\omega i}$ la fréquence d'apparition de l'observation a_i dans ω. Chaque composante $\nu_{\omega i}$ de l'histogramme

est représentée par un modèle de poisson de moyenne μ_i. La probabilité de ω selon ce modèle s'écrit:

$$P_M(\omega) = \prod_{i=1}^{Nds} \frac{\mu_i^{v_{\omega i}} . e^{-\mu_i}}{v_{\omega i}!}$$

La probabilité jointe de l'ensemble des suites d'observations acoustiques de O_n vaut:

$$P_M(O_n) = \prod_{\omega \in O_n} \prod_{i=1}^{Nds} \frac{\mu_i^{v_{\omega i}} . e^{-\mu_i}}{v_{\omega i}!}$$

Le meilleur modèle M est obtenu en prenant comme estimateur de μ_i, l'estimateur du maximum de vraisemblance donné par la fréquence moyenne d'occurrence de l'observation a_i sur l'ensemble O_n :

$$\overline{\mu}_i = \frac{1}{N_{O_n}} \sum_{\omega \in O_n} v_{\omega i}$$

où N_{O_n} est le cardinal de O_n.

Soient $\overline{\mu}_{iq,g}$ et $\overline{\mu}_{iq,d}$ $1 \leq i \leq Nds$ les estimateurs des moyennes des modèles $M_{nq,g}$ et $M_{nq,d}$.
Pour plus de clarté:

- $\overline{\mu}_{iq,g}$, $\overline{\mu}_{iq,d}$ et $\overline{\mu}_i$ seront notés respectivement: μ_{ig}, μ_{id} et μ_i
- $O_{nq,g}$ et $O_{nq,d}$ seront notés O_{ng} et O_{nd}

C (n,q) s'écrit alors:

$$C(n,q) = \left| \begin{array}{l} \sum_{i=1}^{Nds} \sum_{\omega \in O_{ng}} [\, v_{\omega i} \log(\mu_{ig}) - \mu_{ig} - \log(v_{\omega i}!) \,] \\ + \sum_{i=1}^{Nds} \sum_{\omega \in O_{nd}} [\, v_{\omega i} . \log(\mu_{id}) - \mu_{id} - \log(v_{\omega i}!) \,] \\ - \sum_{i=1}^{Nds} \sum_{\omega \in O_n} [\, v_{\omega i} . \log(\mu_i) - \mu_i - \log(v_{\omega i}!) \,] \end{array} \right.$$

En remarquant que:

$$\sum_{\omega \in O_{ng}} \log(v_{\omega i}!) + \sum_{\omega \in O_{nd}} \log(v_{\omega i}!) = \sum_{\omega \in O_n} \log(v_{\omega i}!)$$

et que $\mu_i = \mu_{ig} + \mu_{id}$

on a alors:

$$C(n,q) = \sum_{i=1}^{Nds} [\, NO_{ng}.\mu_{ig}.\log(\mu_{ig}) + NO_{nd}.\mu_{id}.\log(\mu_{id}) - NO_{n}.\mu_{i}.\log(\mu_{i}) \,]$$

où NO_{ng} et NO_{nd} sont les cardinaux de O_{ng}, O_{nd}.

On trouvera dans [10] une étude sur les liens qu'entretient C (n,q) avec un critère de minimisation d'entropie.

QUESTIONS POUR LES ARBRES PHONOLOGIQUES

Une question simple sur la variable $X \in \mathcal{P}$, est une question du type $X \in S$? où S est un sous ensemble de $\mathcal{P}$, à laquelle on répond par oui ou par non. Chaque question est appliquée à chacun des phonèmes (P_{-N}...P_{-1}, P_1...P_N) du contexte phonétique de P_0. Pour chaque question q, on pose donc 2xN questions par échantillon. Si Ns est le nombre de sous ensembles S considérés, sur chaque échantillon on pose $2xN_xNs$ questions. Si on considère tous les sous ensembles non vides de $\mathcal{P}$, alors $Ns = 2^{NP}-1$, soit $2xN_x(2^{NP}-1)$ questions à poser sur chaque échantillon. Ce nombre étant très élevé[19], il faut opérer un choix parmi les sous ensembles retenus.

Une solution est de considérer les sous ensembles faisant référence à une classification articulatoire des phonèmes. On obtient ainsi, un ensemble de questions établies manuellement et fixées d'avance. D'autres choix sont possibles. On peut faire appel à une procédure algorithmique qui détermine automatiquement les sous ensembles à retenir. Lucassen et Mercer [61] propose, dans le cadre de la détermination automatique de formes phonétiques, un algorithme sous optimal en $O(N_P^2)$. La mise au point des questions se fait dans une phase préliminaire à la construction de l'arbre. Elle repose sur la recherche d'une partition S,S' de $\mathcal{P}$ dont l'information mutuelle avec O est maximale. Les questions peuvent aussi être calculées lors de la construction de l'arbre: Nadas et al. proposent dans [73] un algorithme rapide (en $O(N_P)$ sous optimal) qui généralise celui proposé par Brieman [17].

Lors de la construction de l'arbre, la séparation disjointe peut conduire à une importante fragmentation des données ayant pour conséquence la construction de feuilles très similaires. Pour éviter ce problème, on peut se donner la possibilité de fusionner des noeuds de l'arbre. Les questions associées au réseau sont alors dites "complexes". Elles s'expriment sous forme d'une combinaison de conjonctions et disjonctions de questions simples [55][7]. La fusion entre deux noeuds est conditionnée à un seuil: on accepte la fusion à condition que la perte de similarité ne dépasse pas un seuil fixé. Cette perte est mesurée de la même façon que le gain obtenu lors d'une séparation.

[19] Si N=5 et NP=40 le nombre des questions à poser est de l'ordre de 10^{13}.

UTILISATION DES ARBRES EN RECONNAISSANCE

En phase de reconnaissance, la construction de l'automate allophonique d'un mot se fait de la manière suivante. Considérons la forme phonétique du mot. Pour chacun de ses phonèmes P_0, on trace à travers l'arbre phonologique associé un chemin en posant à chaque noeud une question sur le contexte phonétique de P_0. Ceci conduit à une feuille de l'arbre. On modélise le phonème considéré par l'automate allophonique associé à la feuille. Dans le cadre d'un système dont le mode d'élocution est isolé tel que Tangora, le décodeur n'est pas modifié. On remplace uniquement lors du décodage acoustique et de l'apprentissage, les formes fénoniques par les formes allophoniques. Dans le cadre d'un système de reconnaissance de parole continue, le décodeur est légèrement modifié. Pour les derniers phonèmes, dont on ne connaît pas la totalité du contexte droit, on modélise le contexte inconnu par un automate particulier. Quand la reconnaissance progresse et que l'on a des hypothèses sur les mots suivants, on remplace cet automate particulier par l'automate qui convient et on recalcule le score acoustique du mot étant donnée sa forme allophonique [11][8].

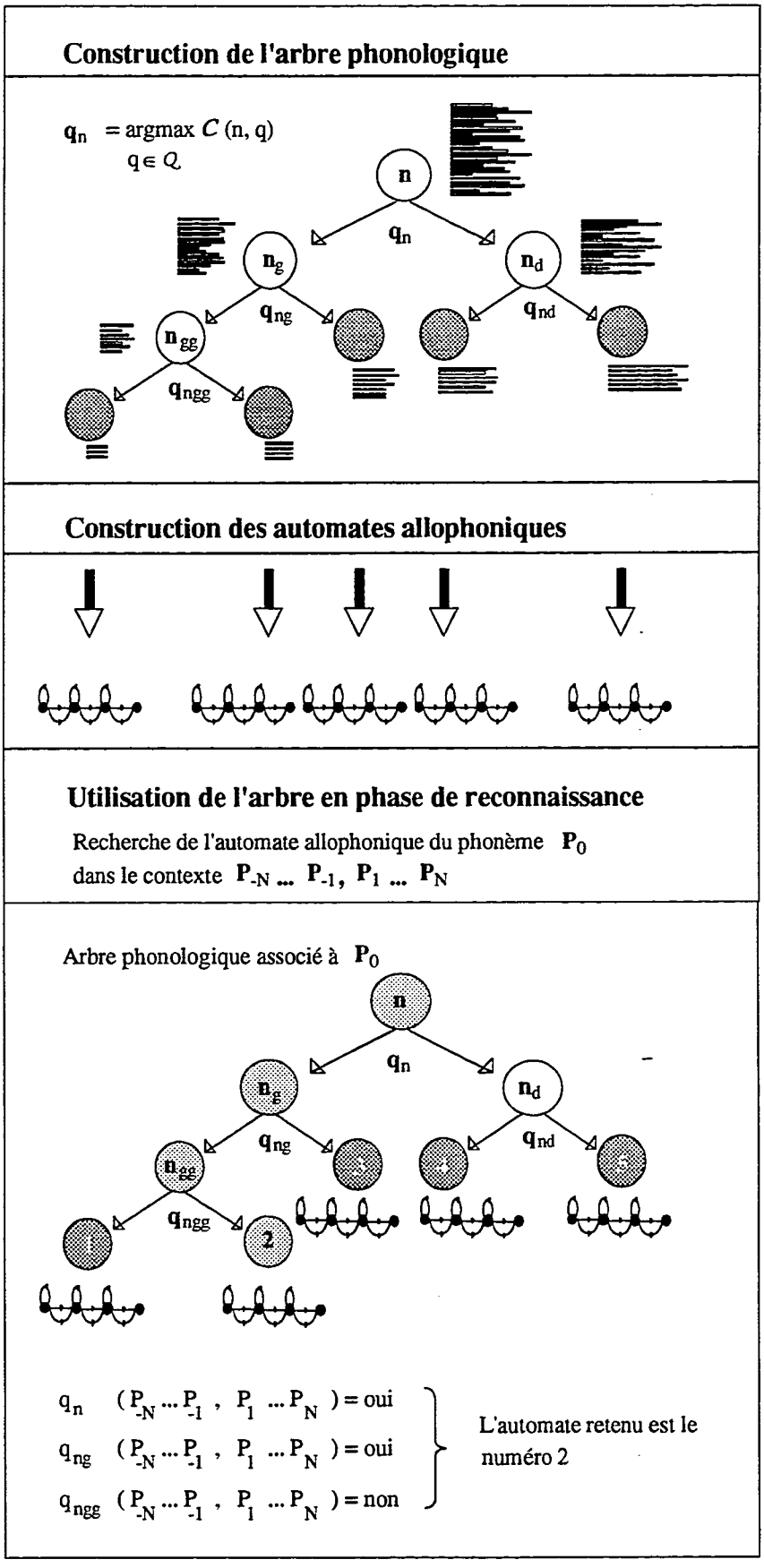

FIGURE 17. Construction et utilisation des arbres

CALCUL DES AUTOMATES ALLOPHONIQUES

La construction des automates allophoniques à partir des suites d'observations acoustiques regroupées dans chacune des feuilles de l'arbre se fait d'une façon analogue à la construction des formes fénoniques des mots du dictionnaire présenté dans [3] et dont je vais rappeler ici les principaux points.

Soit $\omega = a_1, a_2, a_3 \ldots a_p$ une suite d'observations acoustiques produite par le canal acoustique en réponse à la prononciation d'un mot m. On pourrait considérer $a_1, a_2, a_3 \ldots a_p$ comme étant la forme acoustique de m. Cependant, d'autres prononciations du même mot ne conduiront pas exactement à la même suite mais à une suite d'observations similaire. L'idée est de modéliser cette variation en remplaçant chaque observation par un automate markovien, appelé "fenone" ou "automate fénonique". Il y a bijection entre l'alphabet des automates fénoniques et le dictionnaire de spectres définit par le canal acoustique. A chaque observation acoustique correspond un automate fénonique unique. La structure retenue pour l'automate fénonique est un automate à deux états, déjà présenté au chapitre I. La figure suivante donne la structure de la forme fénonique d'un mot.

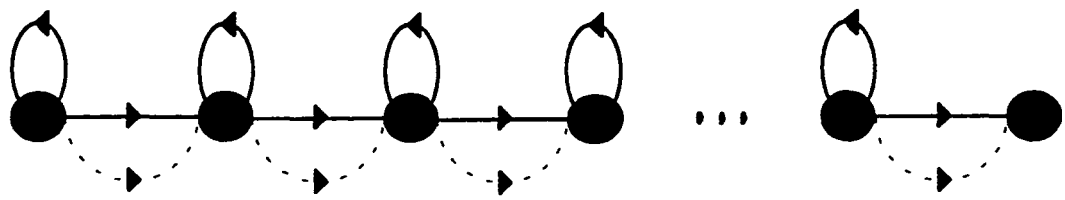

FIGURE 18. Automates fénoniques concaténés

La construction des formes fénoniques des mots du dictionnaire à partir d'une seule prononciation n'étant pas très performante, la forme fénonique d'un mot est calculée à partir de plusieurs prononciations. Le problème est alors de trouver une forme moyenne qui représente au mieux ces différentes prononciations.

Soient $\omega_1, \omega_2, \omega_3 \ldots \omega_n$ les suites d'observations acoustiques correspondant aux n prononciations du mot m. La forme fénonique optimale $\bar{f}$ du mot m est celle qui maximise la probabilité de produire les n suites $\omega_1, \omega_2, \omega_3 \ldots \omega_n$:

$$\bar{f} = \arg\max_{f} P(\omega_1, \omega_2, \ldots \omega_n \mid f)$$

En supposant qu'étant donnée f, ω_i est conditionnellement indépendant de ω_j pour $i \neq j$, on a:

$$\bar{f} = \underset{f}{\operatorname{argmax}} \prod_{i=1}^{n} P(\omega_i | f)$$

Le calcul de la forme optimale pourrait être fait en considérant le problème comme un problème de décodage d'une suite de mots selon un critère de maximum de vraisemblance pour lequel le dictionnaire des mots serait remplacé par l'alphabet des automates fénoniques et pour lequel, le score acoustique total s'exprimerait par le produit des scores acoustiques individuels de chacune des suites ω_i. On chercherait la suite la plus probable des automates fénoniques au lieu de chercher la suite la plus probable des mots.

Une autre façon de calculer $\bar{f}$, sous optimale mais moins coûteuse en calcul, consiste à affiner une forme initiale f_{init}:

0. Sur l'ensemble des suites d'observations acoustiques issues des différentes prononciations d'un mot m, en choisir une et l'élire "forme initiale" f_{init}. Choisir par exemple, celle dont la longueur est la plus proche de la longueur moyenne des suites. Remplacer dans cette suite chaque observation acoustique par son modèle fénonique associé et apprendre cette forme en utilisant les autres prononciations.

1. Aligner à l'aide d'un algorithme de Viterbi, les observations acoustiques des suites $\omega_1, \omega_2, \omega_3 ... \omega_n$ sur les automates de f_{init}

2. Pour chaque automate fénonique F de f' (initialement $f' = f_{init}$):

- Rassembler l'ensemble des suites d'observations acoustiques partielles alignées sur F.

  2.1: *Substitution.* Trouver l'automate fénonique F' qui maximise la probabilité de produire l'ensemble de ces suites partielles. Si $F \neq F'$, remplacer F par F'.

  2.2: *Insertion.* Trouver la paire d'automates fénoniques (F''_1, F''_2) qui maximisent la probabilité de produire ce même ensemble. Si $F''_1F''_2$ est plus probable que F' remplacer F' par $F''_1F''_2$.
- Pour chaque paire G_1G_2 de f', rassembler l'ensemble des suites partielles alignées sur cette paire.

  2.3: *Suppression.* Trouver l'automate fénonique G' qui maximise la probabilité de produire cet ensemble. Si G' est plus probable que G_1G_2 remplacer G_1G_2 par G'.

3. Répéter les étapes **1.** et **2.**(1, 2, 3) avec la nouvelle forme f' obtenue. La procédure s'arrête quand la forme $f'^{(i)}$ (établie lors de la $i^{ème}$ itération) produit l'ensemble des suites $\omega_1, \omega_2, \omega_3 ... \omega_n$ avec une probabilité moindre ou égale à celle de la forme $f'^{(i-1)}$.

La construction des automates allophoniques d'un phonème P_0, se fait de la même façon que celle des formes fénoniques de mots. Les suites d'observations acoustiques $\omega_1, \omega_2, \omega_3 ... \omega_n$ proviennent d'une des feuilles de l'arbre phonologique de P_0.

La phase d'initialisation du calcul de la forme allophonique retenue est légèrement différente de celle que je viens d'exposer. Pour déterminer f_{init}, on cherche la suite d'observations acoustiques qui, vue sous forme de suite d'automates fénoniques, a la plus grande probabilité de produire les autres suites d'observations acoustiques. Ceci se fait en utilisant les modèles fénoniques personnels de chaque locuteur. Par contre pour affiner cette forme, on utilise le modèle commun:

0. Soit $f_i \; 1 \leq i \leq n$ une forme fénonique obtenue en remplaçant chacune des observations acoustiques de ω_i par l'automate fénonique associé. La forme initiale est celle qui, parmi les formes $f_i \; 1 \leq i \leq n$, a la plus grande probabilité de produire toutes les suites $\omega_1, \omega_2, \omega_3 ... \omega_n$.

1.2.3. On affine cette forme initiale selon la procédure présentée plus haut.

EXPERIMENTATION

Le but des premiers tests est de comparer les performances d'une approche allophonique à celles de l'approche fénonique utilisée par le système Tangora. L'idée est de remplacer le dictionnaire des 20.000 formes fénoniques par un dictionnaire de formes allophoniques construit à l'aide d'arbres de décision.

Les données utilisées pour la construction des arbres de décision sont composées de trois textes enregistrés par huit locuteurs. Les deux premiers, sont ceux établis respectivement par A.M. Derouault et P. Combescure dont nous avons parlé au chapitre II, le troisième est un dictionnaire de 6400 syllabes. Ces trois textes ont été prononcés en syllabes isolées. L'ensemble de ces corpus contient 33.180 phonèmes. Le phonème le plus représenté est le [r] (3.222 fois) et le phonème le moins représenté est le [œ̃] (90 fois).
Le signal, une fois traité, est segmenté par l'intermédiaire d'un alignement de Viterbi. Les transcriptions phonétiques utilisées lors de cet alignement sont standards. Pour la parole continue, il serait plus judicieux d'utiliser une transcription personnalisée. La préparation des données conduit à un ensemble de 265.440 segments phonétiques, soit en moyenne 8043,6 segments par phonème, ce qui n'est pas très important si on compare ces quantités à celles utilisées dans [8]. Enfin l'ensemble des questions a été conçu "à la main" selon une classification articulatoire des phonèmes. Cet ensemble est donné en annexe.

Lors de la construction des arbres phonologiques, quatre paramètres peuvent varier. La longueur des contextes, le seuil θ_C sur la valeur du critère de séparation en dessous de laquelle le noeud devient feuille, le nombre maximal θ_N de noeuds de l'arbre et enfin le nombre minimal θ_O d'échantillons par feuille. Etant donnée la nature des données sur lesquelles les arbres vont être construits, les longueurs des contextes droit (lctxd) et gauche (lctxg) ne dépasseront pas 2 phonèmes. Plusieurs valeurs empiriques de θ_C ont été testées. Nous en mentionnerons deux. θ_O et θ_N restent constants, le premier à 20 le second à 300.

	θ_O	θ_N	θ_C	L=lcxtd=lctxg
Exp1	20	300	300	2
Exp2	20	300	300	1
Exp3	20	300	45	2
Exp4	20	300	45	1

Quand θ_C diminue, la croissance de l'arbre est peu contrainte ($\theta_C = 45$), le nombre des feuilles augmente. Quand la croissance de l'arbre est fortement contrainte ($\theta_C = 300$), une augmentation de la longueur du contexte entraîne dans la mesure où les données associées au phonème sont suffisamment importantes, un accroissement régulier du nombre des feuilles. Par contre, quand la croissance de l'arbre est peu contrainte, une augmentation de la longueur du contexte ne produit plus un accroissement régulier du nombre de feuille. En effet, de l'expérience 3 à l'expérience 4 le nombre des feuilles est en moyenne multiplié par 2, mais ce facteur vaut trois pour les obstruantes et moins de 2 pour les voyelles moins sujettes à variations. Ces remarques laissent penser que l'arbre sera d'autant plus apte à décrire la variabilité des phonèmes que le seuil θ_C est faible et que la longueur du contexte est grande. Ce point de vue est renforcé par le fait que lors de la construction du dictionnaire allophonique, pour les arbres très contraints, le nombre de fois où les arbres n'ont pas prévu de solutions particulières à la traduction phonème-allophone (c'est-à-dire le nombre de fois où une feuille "archi-non"[20] a été rencontrée) est plus important que dans le cas d'arbres moins contraints.

Nombre de phonèmes dont le % des feuilles "archi-non" est > 10	Exp1	Exp2	Exp3	Exp4
	8	7	3	4

	Données	feuilles Exp1	feuilles Exp2	feuilles Exp3	feuilles Exp4
/œ̃/	721	3	3	9	9
/ɥ/	1.553	9	8	38	25
/œ/	1.812	9	5	34	30
/ɛ̃/	2.556	10	9	53	42
/ɔp/	3.142	7	5	24	11
/ɔ̃/	3.583	16	12	66	54
/ʃ/	3.690	17	12	66	31
/o/	3.742	21	13	42	64
/ /	4.412	24	14	96	82
/w/	4.428	21	14	98	62
/ʒ/	4.819	15	13	80	29

[20] C'est à dire qui correspond à une réponse négative à toute les questions posées.

/ɑ̃/	5.242	17	11	97	84
/z/	5.397	16	11	83	27
/v/	5.489	21	15	92	32
/ok/	5.532	12	9	49	14
/g/	5.622	27	15	109	47
/f/	5.657	23	14	101	41
/y/	5.737	23	16	103	85
/e/	5.862	13	10	45	24
/ø/	5.906	16	11	42	29
/.g/	6.150	10	8	46	13
/b/	6.452	18	14	122	45
/ɔ/	7.381	24	20	157	116
/j/	8.158	33	26	164	116
/m/	8.194	29	18	114	40
/ot/	8.228	12	8	57	18
/n/	8.441	31	21	141	44
/d/	10.156	28	20	124	43
/bp/	10.167	20	13	105	30
/ɛ/	10.681	26	23	210	150
/bk/	12.085	29	16	112	36
/i/	12.335	38	26	204	144
/bt/	15.760	29	16	119	32
/s/	16.036	43	30	192	72
/l/	18.494	62	41	- 255	129
/a/	19.105	39	27	227	177
/r/	25.775	81	55	301	186
/.v/	37.710	12	12	164	14
/.c/	51.155	32	16	301	18
Moyennes	9.676,3	23,5	16,2	113,9	57,6

Une fois les arbres phonologiques et les automates allophoniques construits, le dictionnaire des formes allophoniques est obtenu à partir du dictionnaire des formes phonétiques en utilisant les arbres phonologiques, comme cela a été décrit dans le paragraphe "Utilisation des arbres en reconnaissance". Les résultats qui suivent sont ceux d'un décodage complet réalisé par le système de reconnaissance Tangora. Pour chacune des quatre expériences, on a calculé un dictionnaire allophonique en utilisant les arbres phonologiques et les automates allophoniques qui leur sont associés. Rappelons que le décodeur de Tangora est composé de trois parties. Un

préfiltre acoustique, dont le rôle est d'extraire du dictionnaire une liste de mots candidats étant donné ce qui a été prononcé. Un modèle de langage qui réorganise cette liste en fonction des probabilités tri-gramme. Un décodeur acoustique dont le rôle est de réduire et de réorganiser la liste en fonction d'un examen acoustique plus précis que celui effectué par le préfiltre. Le décodeur acoustique effectue pour chaque mot de la liste une passe-avant de l'algorithme de Baum entre la suite des observations acoustiques et la forme fénonique du mot. Cette étape nécessite d'avoir, d'une part, un dictionnaire des formes fénoniques des 20.000 mots de Tangora, et d'autre part le modèle fénonique du locuteur à décoder. Le tableau suivant donne les pourcentages d'erreurs obtenus au cours des quatre expériences Exp1, Exp2, Exp3 et Exp4. Le dictionnaire fénonique est remplacé par un dictionnaire allophonique, par contre les modèles utilisés sont les modèles fénoniques. Il aurait été préférable de construire un modèle allophonique pour chacun des locuteurs, ce n'est qu'une question de temps.

	M. fénonique	Exp1	Exp2	Exp3	Exp4
JCM	5,32	5,22	6,02	5,12	5,82
PRF	6,12	7,92	8,22	7,42	7,62
AED	6,92	9,83	9,63	9,43	8,93
Moyenne	6,12	7,65	7,95	7,32	7,45

Le premier locuteur fait partie de ceux qui ont enregistré la base de données des arbres phonologiques.

Ces premiers résultats sont moins bons que ceux obtenus par l'approche fénonique, en particulier sur les locuteurs ne faisant pas partie de la base de données de construction des arbres phonologiques. Il reste à voir ce que donnerait un décodage allophonique avec les modèles allophoniques. Il semble que la fiabilité des arbres repose sur la cohérence et la quantité des données de construction. L'utilisation des syllabes isolées ne permet pas de prendre en compte les phénomènes de co-articulation des mots. La variabilité intra syllabe est assez pauvre. Prenons l'exemple du mot "*cheval*", quand le mot est prononcé en syllabes isolées, le [v] ne donne lieu à aucune variation notable à cause de la découpe en [ʃø val]. Prononcé en mots isolés, on pourra observer la variante [ʃfal] dans laquelle le voisement du [v] disparaît sous l'influence progressive de la fricative sourde [ʃ]. Le mode d'élocution des données de construction doit être cohérent avec celui du décodage. Quant à la quantité des données, les corpus doivent être choisis aussi vastes que possible et être enregistrés par plusieurs locuteurs, féminins et masculins. La cohérence entre le mode d'élocution des données de construction des arbres et celui du décodage est une condition nécessaire à la bonne adéquation des allophones à diversité des contextes possibles. La quantité de ces données est une condition nécessaire au calcul d'une forme allophonique moyenne appropriée au plus grand nombre.

Notons tout de même que les résultats sont meilleurs quand la longueur du contexte vaut deux (et quand le seuil θ_c est faible).

Comparaison des erreurs. entre la référence et Exp3.			
Erreur	JCM	PRF	AED
Acous. communes	12	21	19
Acous. différentes[21]	***9*** - 15	***12*** - 26	***18*** - 34
Lang. communes	17	15	14
Lang. différentes	***3*** - 4	***7*** - 3	***3*** - 6
Total	***41*** - 48	***55*** - 65	***54*** - 73

Parmi les erreurs acoustiques spécifiques à Exp4, on peut noter une augmentation des erreurs sur les mots courts ("par" au lieu de "à", "cars/ cas", "vol/ veulent"...) et des erreurs dues à une mauvaise identification du début de mot ("mini/ émis", "mystère/ espère", "acquitté/ inculpé", "tous/ pour", "milieu/ lieu", "traînée/ freinée", "Echo/ F.O.", "honneur/ promeneur"...). Les erreurs les plus fréquentes sont celles dues à une mauvaise identification des consonnes, en particuliers des occlusives.

[21] Les chiffres en gras et italique sont ceux de la référence.

CONCLUSION

Nous avons présenté dans ce chapitre une méthode de construction d'automates phonétiques contextuels dont le principe repose sur une classification des réalisations acoustiques des phonèmes par l'intermédiaire d'arbres de décision. Son efficacité est conditionnée à la quantité et la qualité des données disponibles pour sa construction. Comme nous venons de le montrer dans ce chapitre le modèle allophonique n'est pas efficace si les données sont incohérentes et en nombre insuffisant. L'incohérence des données induit la construction de l'arbre en erreur. Leur nombre insuffisant pose deux problèmes, celui de la couverture allophonique, et celui de la couverture de la variabilité inter locuteur. Ces premiers résultats ne sont cependant qu'indicatifs: aucun apprentissage allophonique n'a était effectué, les modèles utilisés pour le décodage allophonique sont les modèles fénoniques.

Des expériences complémentaires ont récemment été menées au Centre Scientifique d'IBM France. Pour ces expériences, 16 locuteurs, 8 femmes, 8 hommes, ont enregistré en mots isolés, chacun 1500 à 2500 phrases pour la construction des arbres et 50 phrases de test. Le corpus d'apprentissage est constitué de 1000 phrases communes à tous les locuteurs (avec en moyenne 10 mots par phrase), puis selon les locuteurs, de 500 à 1500 phrases différentes pour chaque locuteur (avec en moyenne 20 mots par phrase). Ces phrases sont phonétiquement équilibrées et ont été conçues de façon à contenir le maximum de phonèmes contextuels. La longueur du contexte est de 5 phonèmes de part et d'autre du phonème étudié. Les seuils θ_C et θ_N sont choisis de sorte à ne pas trop contraindre la construction des arbres. θ_O est inchangé. Les tests portent sur 10 locuteurs ne faisant pas partie de la base de données d'apprentissage. Les résultats de ces expériences montrent une diminution relative du taux d'erreur de 12.7 % par rapport à l'approche fénonique utilisée par le système Tangora. Ces résultats valident la méthode proposée et mettent en évidence la nécessité, pour les techniques basées sur les arbres de décision, d'un volume de données d'apprentissage important et diversifié.

Cette approche peut paraître complexe et coûteuse pour une réduction du taux d'erreur de 12,7%. Elle a cependant le mérite d'être efficace et l'avantage d'être généralisable à la parole continue.

Des améliorations peuvent être amenées sur:

- l'ensemble des questions. Les questions ont été choisies selon une classification articulatoire des phonèmes. D'autres catégories que celles que j'ai choisi peuvent être proposées. Elles ont été conçues "à la main", des méthodes automatiques pourraient être envisagées.
- Le choix du critère. Le critère repose sur le maximum de vraisemblance. D'autres critères sont envisageables reposant par exemple sur le maximum d'information mutuelle ou sur l'entropie croisée. D'autre part, il pourrait être intéressant de définir en plus du critère de séparation, un critère de fusion qui éviterait une trop importante fragmentation des données.
- La représentation des segments acoustiques. On a considéré lors de la construction de l'arbre non pas les suites d'observations acoustiques mais leurs histogrammes. Ce choix simplifie de façon significative le calcul du critère mais n'est pas tout à fait satisfaisant dans la mesure où on cherche à étudier l'influence contextuelle. Pour se débarrasser des histogrammes, la solution idéale est d'utiliser la classe des modèles de Markov. Mais cette solution est très coûteuse en en temps de calcul. Une solution intermédiaire entre la solution idéale est celle proposée serait de découper les segments acoustiques en trois parties (un début, un milieu et une fin de phonème), calculer des histogrammes sur chacune d'entre elles et considérer trois modèles de Poisson, un par partie.

CHAPITRE IV: LISTE DE MOTS CANDIDATS, UNE METHODE ALTERNATIVE

INTRODUCTION

Le coût du décodage en terme de quantité d'opérations à effectuer croît avec la taille du vocabulaire. Dans le cadre d'un système de reconnaissance de la parole utilisant des modèles markoviens, le score acoustique d'un mot peut être évalué avec la passe-avant de l'algorithme de Baum. Cet algorithme nécessite au maximum, pour le décodage d'un mot, $n_t \times n$ opérations élémentaires, où n_t est le nombre de transitions du modèle de mot et n est la longueur du segment acoustique[22]. Pour des systèmes de reconnaissance dont la taille du dictionnaire dépasse la centaine de mots, les modèles de mots sont obtenus par concaténation de modèles de sous unité de mot. Supposons que l'on utilise des modèles phonétiques à cinq états et treize transitions et que le nombre moyen de phonèmes par mot soit de sept. Le calcul du score acoustique d'un mot nécessite alors, pour un segment acoustique d'une seconde, un peu moins de 10^4 opérations[23] élémentaires. Soit 2.10^8 opérations si le dictionnaire contient 20.000 mots et si l'on fait tous les calculs. Si cette approche est viable pour des systèmes dont le dictionnaire est de taille modeste, elle ne l'est plus pour des systèmes à grand dictionnaire. Une solution consiste à construire un préfiltre acoustique[24] dont l'objet est d'extraire du dictionnaire une liste de mots candidats sur laquelle on pourra par la suite effectuer des traitements plus fins et plus gourmands en calculs. Les principales qualités de ce préfiltre doivent être sa rapidité et sa fiabilité. En effet, si le préfiltre ne permet pas d'être plus rapide que l'approche fondée sur la passe avant de l'algorithme de Baum, il n'a plus de raison d'être. Il en est de même s'il ne garantit pas la présence du mot effectivement prononcé avec une marge d'erreur relativement faible ou s'il propose une liste aussi longue que celle du dictionnaire tout entier. Ainsi, les trois paramètres permettant de juger de la qualité d'un préfiltre acoustique sont la rapidité d'exécution, la longueur de la liste moyenne, le pourcentage d'erreurs, c'est-à-dire le nombre de fois en pourcentage, où le mot prononcé ne fait pas partie de la liste des mots candidats. Il y a de multiples façons de construire la liste des mots candidats.

Le préfiltre acoustique utilisé dans le système Tangora est une simplification d'un décodeur fondé sur la passe avant de l'algorithme de Baum. Cette simplification repose sur la constatation que pour réduire les $n_t \times n$ opérations nécessaires au décodage d'un mot, il suffit de réduire n_t. Ceci se fait de la manière suivante. L'unité fondamentale de reconnaissance utilisée

[22] La longueur du segment acoustique fait référence ici au nombre de trames nécéssaires à sa description.
[23] $(n_t = 7 \times 13) \times (n = 100) = 9100$
[24] Rappel: la terminologie anglaise de ce mot est "Fast match"

par le préfiltre est le phonème. Cette unité, représentée par un automate à sept états et treize transitions, est apprise au cours de la phase d'apprentissage. Pour le décodage, l'automate à sept états est réduit à un automate à deux états et deux transitions, dont on trouvera la description précise dans [6] et dont les paramètres sont calculés à partir de ceux de l'automate à sept états. Ainsi, on réduit la valeur de n_t, mais par contre, on s'impose de travailler avec des pseudo modèles[25] et on perd la notion de séquentialité sur la longueur des phonèmes. La perte de la séquentialité est un facteur de rapidité mais elle entraîne aussi une dégradation de la fiabilité du préfiltre. Il faut faire un choix entre une très grande rapidité et une bonne précision. Un arbre phonétique permet de réduire le nombre des calculs. L'utilisation de cet arbre et des pseudo modèles réduit le nombre d'opérations élémentaires de façon significative.

D'autres méthodes sans rapport avec la passe avant de l'algorithme de Baum ont été proposées. Notons celle du vote majoritaire[26] [4] appliquée à l'obtention d'une liste de mots candidats dont l'idée consiste à définir le score d'un mot m par:

$$s_m = \sum_{j=1}^{n} v(a_j, m) + i_m$$

Où $v(a_j, m)$ est la valeur du vote de l'étiquette acoustique a_j pour le mot m et i_m est la valeur initiale de s_m. La liste des mots candidats est obtenue en triant par ordre décroissant les scores s_m des mots du dictionnaire et en ne conservant que les p premiers. Etant données les définitions de $v(a_j, m)$ et i_m proposées dans [4], une telle méthode nécessite toute les centisecondes, une addition de vecteur de dimension: la taille du dictionnaire. Soit $20.000 \times n$ opérations pour un segment acoustique de longueur n et un dictionnaire de 20.000 mots. Cette méthode présente l'avantage d'une parallélisation des calculs que l'on pourra faire simultanément sur plusieurs processeurs. On gagne un facteur 14 en nombre d'opérations par rapport à la méthode du préfiltre acoustique implémentée dans le système Tangora, par contre la perte de la séquentialité se répartit sur l'ensemble du mot.

Je tiens particulièrement à remercier Lalit Bahl, Marc El-Bèze et Bernard Merialdo, dont les idées judicieuses et les conseils précieux m'ont permis de mettre au point la méthode proposée dans les paragraphes suivants. Cette méthode repose sur une pré-classification des mots du dictionnaire. La classification s'effectue à l'aide d'un arbre de décision dont le critère repose sur une analyse discriminante des données d'apprentissages de cet arbre. Chaque feuille de l'arbre représente une liste de mots candidats. L'avantage de la structure arborescente est sa très grande rapidité de consultation. Le fil conducteur entre les travaux présentés dans le chapitre précédent et ceux qui vont l'être dans celui-ci est donc l'arbre de décision. La structure est la même: à chaque noeud est associée une question dont la réponse détermine le chemin à suivre.

[25] En effet, l'automate phonétique réduit n'est pas un "véritable" modèle markovien dans la mesure où la somme des probabilités de transitions partant de l'état initial est supérieur à 1.
[26] dont la terminologie anglaise est "Polling"

Cependant, la question posée doit être compatible avec les données sur lesquelles elle s'exerce: des phonèmes pour les arbres phonologiques, des suites d'étiquettes acoustiques pour l'arbre du préfiltre acoustique. Alors que les questions phonétiques sont d'établissement aisé, la mise au point des questions "acoustiques" est loin d'être immédiate. C'est à l'aide d'une analyse discriminante que nous avons choisi d'établir ces questions. Si dans le cas idéal, poser une question divise la liste des mots candidats par deux, il suffit, pour un dictionnaire de 20.000 mots et des feuilles de 150 mots, de poser 7 questions pour atteindre une feuille, c'est-à-dire d'effectuer $7 \times n$ additions.

PREFILTRE ACOUSTIQUE

En phase de reconnaissance, étant donné un segment acoustique, on désire lui associer rapidement une liste de mots acoustiquement proches. Une méthode rapide consiste à utiliser un arbre de décision. A chaque noeud de l'arbre, la réponse à une question posée sur le segment acoustique permet de tracer un chemin dans l'arbre qui conduit à une feuille constituée d'une liste de mots candidats. Deux problèmes se posent: d'une part, celui de définir des questions de nature acoustiques, d'autre part, celui de définir un critère de séparation des mots du dictionnaire. Pour définir les questions, lors de la construction de l'arbre, on va s'aider de la phonétique des mots. Supposons que l'on ait à disposition un dictionnaire phonétique et pour chacun de ses mots une ou plusieurs suites d'observations acoustiques correspondant à la prononciation du mot, si l'on divise les mots du dictionnaire en deux parties, une première composée des mots possédant un phonème P_0 et une seconde composée de ceux ne le possédant pas, on obtient une partition du dictionnaire et par là même une partition des suites d'observations acoustiques correspondant à leurs prononciations. Sur cette dernière on peut, pour mesurer le pouvoir discriminant de la séparation, appliquer une analyse discriminante. Cette analyse fournit, d'une part, une évaluation de la partition: le pouvoir discriminant, et d'autre part, une frontière qui définit la limite entre les suites d'observations acoustique correspondant aux mots dont la phonétique possède P_0, et celles correspondant aux mots dont la phonétique ne le possède pas.

Voici la méthode que je propose:

Construction de l'arbre:

Soit $\mathcal{T}$ l'ensemble des transcriptions phonétiques des mots du dictionnaire. On cherche une partition $\mathcal{T} = \{ T_{oui} , T_{non} \}$ de $\mathcal{T}$. Parmi l'ensemble des partitions de $\mathcal{T}$, seules sont retenues celles définies par un ensemble de questions Q posées sur les transcriptions phonétiques ϕ de $\mathcal{T}$: $\{ \mathcal{T}_q = \{ T_{oui,q} , T_{non,q} \} \; q \in Q \}$[27]. Si la réponse de q sur ϕ est oui, ϕ sera classée dans $T_{oui,q}$, sinon elle sera classée dans $T_{non,q}$. En itérant ce processus sur les sous ensembles $T_{oui,q}$ et $T_{non,q}$ de $\mathcal{T}$, on obtient un arbre dont chaque feuille contient une liste de mots. Le choix d'une partition parmi l'ensemble des partitions $\mathcal{T}_q$ se fait à l'aide d'un critère qui sera décrit dans le paragraphe suivant.

Décodage:

L'information phonétique étant absente, c'est à partir de l'information contenue dans le segment acoustique que l'on doit déterminer la feuille de l'arbre à utiliser. L'intérêt du critère d'analyse discriminante est qu'il fournit une question acoustique à l'image de la

[27] Ces questions sont, à peu de chose près, les même que celles utilisées pour la construction des arbres phonologiques.

question phonétique posée sur les éléments de $\mathcal{T}$. Ainsi, au décodage, on parcourt l'arbre en posant à chaque noeud la question acoustique qui lui est associée. Sa réponse détermine la direction à prendre. On aboutit finalement à une feuille qui définit la liste des mots candidats.

CRITERE DE SEPARATION

Pour chaque transcription ϕ de $\mathcal{T}$, on collecte un ensemble de segments acoustiques correspondant à sa prononciation. Notons $\mathcal{A}$ l'ensemble des segments acoustiques collectés.
Posons le problème en termes de classification du nuage de points $\mathcal{A}$. Une analyse en composantes principales donne une visualisation du nuage et fournit des indications sur les corrélations entre variables. Notre but étant de délimiter des frontières qui séparent les individus relatifs à un groupe de mots, nous désirons surtout trouver une frontière qui rend compte au mieux de cette séparation, tout en concentrant les représentants de chaque groupe de mots autour de centres polaires. Les groupes, ici au nombre de deux, sont préétablis par les questions q de Q. A chaque partition $\mathcal{T}_q = \{ T_{oui,\,q}\,,\ T_{non,\,q}\}$ de $\mathcal{T}$ correspond une partition de $\mathcal{A}_q = \{ A_{oui,\,q}\,, A_{non,\,q}\}$ de $\mathcal{A}$. L'analyse discriminante semble particulièrement adaptée au problème qui nous préoccupe. L' objectif d'une analyse discriminante est double: donner une représentation graphique résultant des combinaisons linéaires qui discriminent au mieux les classes d'individus (analyse à but descriptif) et décider à quelle classe affecter un nouvel individu (analyse à but décisionnel).

DISCRIMINATION A BUT DESCRIPTIF

Il est clair que la discrimination sera d'autant plus aisée que les classes $A_{oui,q}$ et $A_{non,q}$ sont éloignées l'une de l'autre et que les individus d'une même classe sont proches les uns des autres. Ceci se traduit par la double optimisation: maximiser la variance interclasse et minimiser la variance intraclasse. Le problème revient donc à trouver les axes factoriels discriminant u qui maximisent le rapport de la variance interclasse de u à la variance intraclasse de u. Notons que pour calculer les variances inter ou intraclasse il est nécessaire de travailler avec des données de dimension constante. Les données étant ici des segments acoustiques de longueur variable, nous avons choisi de les représenter par leurs histogrammes dont l'établissement est aisé puisque les segments acoustiques sont issus d'une quantification vectorielle et qu'il suffit, pour en établir l'histogramme, de compter les occurrences de chaque observation acoustique. Ceci entraîne, comme dans le cas du préfiltre acoustique par vote majoritaire, une perte de l'information de séquentialité sur l'ensemble du mot. $\mathcal{A}$ désignera désormais, non plus l'ensemble des segments acoustiques collectés, mais leurs histogrammes et $\mathcal{A}$ une partition de $\mathcal{A}$.

Rappelons les principes de l'analyse discriminante dont on trouvera une description complète dans [80] ou [82]. Le problème est de trouver u tel que:

$$u = \underset{v}{\text{argmax}} \frac{v^t . B . v}{v^t . W . v}$$

où

- B est la matrice de covariance interclasses: $B = \sum_{A \in \underline{\mathcal{A}}} \frac{N_A}{N} (\bar{o} - \bar{z}).(\bar{o} - \bar{z})^t$

  $\bar{z}$ est l'isobarycentre de la classe A de $\mathcal{A}$ et N_A le cardinal de A.
  $\bar{o}$ est l'isobarycentre de $\mathcal{A}$ et N le cardinal de $\mathcal{A}$.

- W est la matrice de covariance intraclasse: $W = \frac{1}{N} \sum_{A \in \underline{\mathcal{A}}} \sum_{z \in A} (z - \bar{z}).(z - \bar{z})^t$

- C est la matrice de covariance totale: $C = \frac{1}{N} \sum_{o \in \mathcal{A}} (o - \bar{o}).(o - \bar{o})^t$

Comme d'après le théorème de Huygens on a: C = B + W

$$u = \underset{v}{\text{argmax}} \frac{v^t . B . v}{v^t . W . v} = \underset{v}{\text{argmax}} \frac{v^t . B . v}{v^t . [C - B] . v} = \underset{v}{\text{argmin}} \left(\frac{v^t . C . v}{v^t . B . v} - 1\right) = \underset{v}{\text{argmax}} \frac{v^t . B . v}{v^t . C . v}$$

Le problème est donc de trouver u tel que:

$$u = \underset{v}{\text{argmax}} \frac{v^t . B . v}{v^t . C . v}$$

Le premier axe factoriel est le vecteur propre de $C^{-1}.B$ correspondant à la plus grande valeur propre.

Dans le cas qui nous intéresse, $\mathcal{A}$ contient deux classes: A_{oui} et A_{non}, de cardinal respectivement N_{oui} et N_{non} et d'isobarycentre respectivement $\bar{z}_{oui}$ et $\bar{z}_{non}$. En notant que:

$$\bar{o} = \frac{N_{oui} . \bar{z}_{oui} + N_{non} . \bar{z}_{non}}{N}$$

et $\quad N = N_{oui} + N_{non}$

On a alors:

$$B = \frac{N_{oui}}{N} (\bar{z}_{oui} - \bar{o}).(\bar{z}_{oui} - \bar{o})^t + \frac{N_{non}}{N} (\bar{z}_{non} - \bar{o}).(\bar{z}_{non} - \bar{o})^t = \frac{N_{oui} . N_{non}}{N^2} (\bar{z}_{non} - \bar{z}_{oui}).(\bar{z}_{non} - \bar{z}_{oui})^t$$

$C^{-1}.B$ admet un vecteur propre unique u associé à la valeur propre λ où:

$$u = C^{-1}.(\bar{z}_{non} - \bar{z}_{oui})$$

$$\text{et} \qquad \lambda = \frac{N_{oui}.N_{non}}{N^2} u^t.(\bar{z}_{non} - \bar{z}_{oui})$$

La forme linéaire u^t est appelée forme de Fisher, λ est le pouvoir discriminant de la forme linéaire u^t. D'un point de vue géométrique, le problème revient à trouver l'axe sur lequel les projections des nuages de points A_{oui} et de A_{non} sont les plus éloignées l'une de l'autre et leurs points les plus rapprochés les uns des autres.

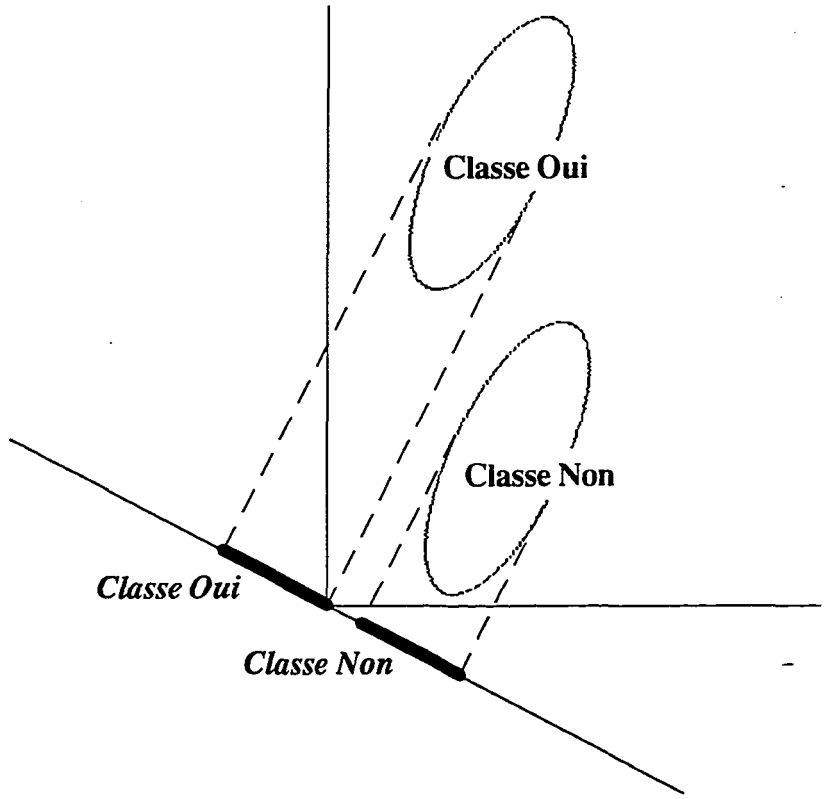

FIGURE 19. Préfiltre acoustique critère

Soit n un noeud de l'arbre et soit C (n,q) la valeur du critère pour la question $q \in Q$. Si $\mathcal{T}n$ est l'ensemble des transitions phonétiques associées au noeud n et si $\mathcal{T}_{qn} = \{T_{oui,qn}, T_{non,qn}\}$ est la partition de $\mathcal{T}n$ induite par la question q, la partition retenue est celle qui maximise le pouvoir discriminant λ_{qn}:

$$\bar{q} = \underset{q}{\text{argmax}}\ \lambda_{qn}$$

Le critère à maximiser est donné par le pouvoir discriminant de Fisher:

$$C(n,q) = \lambda_{qn}$$

D'autres critères pourraient être retenus, comme:

- celui de Mahalanobis: $C(n,q) = \lambda_{qn} = u_{qn}{}^t.(\bar{z}_{oui,qn} - \bar{z}_{non,qn})$
- celui dérivé du critère de Fisher proposé par Marc El-Bèze dans [32].
- ou encore un critère de vraisemblance comme celui utilisé dans le chapitre précédent.

Ce qui précède concerne la discrimination à but descriptif pour laquelle on désire, étant données différentes classes d'individus, mettre en évidence le pouvoir discriminant des variables mesurées sur ces individus. La discrimination à but décisionnel consiste à décider à quelle classe, définie *a priori*, affecter un nouvel individu. La discrimination à but décisionnel est donc une seconde étape, succédant à la première de discrimination descriptive. Cette seconde étape dans le cadre du préfiltre acoustique, constitue la phase de décodage. Il est cependant intéressant pour une bonne compréhension de ce que vont être les questions "acoustiques", d'en décrire le fonctionnement dès à présent.

DISCRIMINATION A BUT DECISIONNEL

Etant donné un nouvel individu I, on cherche la classe A_{oui} ou A_{non} à laquelle il sera affecté. L'axe de direction u étant celui selon lequel la discrimination entre A_{oui} et A_{non} est la plus forte, la décision se fait en fonction de la position de la projection $u^t.I$ de I sur cet axe. Si la distance entre $u^t.I$ et $u^t.\bar{z}_{oui}$ est inférieure à celle entre $u^t.I$ et $u^t.\bar{z}_{non}$, I sera affecté à A_{oui}..
En somme:

$$\text{si} \quad \left| u^t.(I-\bar{z}_{oui}) \right| < \left| u^t.(I-\bar{z}_{non}) \right| \qquad \text{alors } I \in A_{oui}$$

$$\text{sinon } I \in A_{non}$$

ce qui revient à dire que:

$$\text{si} \quad u^t.I < u^t.\left(\frac{\bar{z}_{non} + \bar{z}_{oui}}{2}\right) \qquad \text{alors } I \in A_{oui}$$

$$\text{sinon } I \in A_{non}$$

Posons: $\pi_m = u^t.\frac{\bar{z}_{non} + \bar{z}_{oui}}{2}$. I sera affecté à A_{oui} si sa projection orthogonale sur u est inférieure à π_m et affecté à A_{non} sinon.
Dès lors, à chaque noeud n de l'arbre, la question acoustique retenue $\bar{q}$ est définie par la donnée de la forme de Fisher $u^t_{\bar{q}\,n}$ et du point de séparation $\pi_{m,\bar{q}\,n}$.

CONSTRUCTION ET UTILISATION DE L'ARBRE

La construction s'effectue en deux étapes. La première consiste à collecter les données d'apprentissage de l'arbre, la seconde en son calcul. Les données comprennent d'une part les transcriptions phonétiques des mots du dictionnaire et d'autre part pour chaque transcription,

un échantillon de segments acoustiques correspondant à sa prononciation, organisés sous forme d'histogrammes. L'algorithme de construction est le suivant:

1. Initialement l'arbre est constitué d'un seul noeud, la racine de l'arbre, auquel est associé l'ensemble des transcriptions phonétiques $T_{racine} = \mathcal{T}$ et leurs histogrammes associés $A_{racine} = \mathcal{A}$.
2. Pour chaque noeud n non encore traité:
   2.1. Calculer la matrice de covariance totale C sur An et calculer C^{-1}
   2.2. Si C n'est pas inversible transformer n en feuille terminale.
   Sinon:
   2.2.1. Calculer $C(n,q)$ pour toutes les questions $q \in Q$, c'est-à-dire:
   - Calculer $\bar{z}_{oui,qn}$ $\bar{z}_{non,qn}$ u^t_{qn} $\pi_{m,qn}$ λ_{qn}

   2.2.2. Soit $\bar{q}$ la question qui réalise le maximum de $C(n,q)$
   - Créer deux noeuds n_{oui} et n_{non} successeurs du noeud n.
   - A n_{oui}, on associe tous les données correspondant à une réponse affirmative à la question $\bar{q}$ ($T_{noui} = T_{oui,\bar{q}n}$ et $A_{noui} = A_{oui,\bar{q}n}$).
     A n_{non}, on associe les données restantes ($T_{nnon} = T_{non,\bar{q}n}$ et $A_{nnon} = A_{non,\bar{q}n}$).
   - La question acoustique $\bar{q}_a$, correspondant à $\bar{q}$, est donnée par: $\bar{q}_a = (u^t_{\bar{q}n}, \pi_{m,\bar{q}n})$

En phase de décodage, étant donné un segment acoustique dont on calcule l'histogramme I, la liste de mots candidats est donnée par l'algorithme suivant:

1. n = la racine de l'arbre.
2. Tant que n n'est pas une feuille
   2.1. Calculer $u^t_n.I$
   2.2. Si $u^t_n.I < \pi_{m,n}$ alors $n = n_{oui}$
   Sinon $n = n_{non}$
3. On a atteint une feuille. Elle définit la liste de mots candidats.

Notons que l'histogramme I du segment acoustique s'écrit: $I = \sum_{x=1}^{n} Y_{n(x)}$

Où n est la longueur du segment, n(x) est le numéro du prototype associé à la trame x par la quantification vectorielle et où $Y_{n(t)}$ est un vecteur de dimension la taille du dictionnaire de spectres (N_{ds}) et dont les composantes sont toutes nulles sauf la $n(x)^{ième}$ qui vaut 1.
Si $u^t = [u_1, u_2 \ldots u_{Nds}]$ alors:

$$u^t.I = \sum_{x=1}^{n} u^t.Y_{n(x)} = \sum_{x=1}^{n} u_{n(x)}$$

On effectue ainsi, pour un segment de longueur n, n additions par question posée.

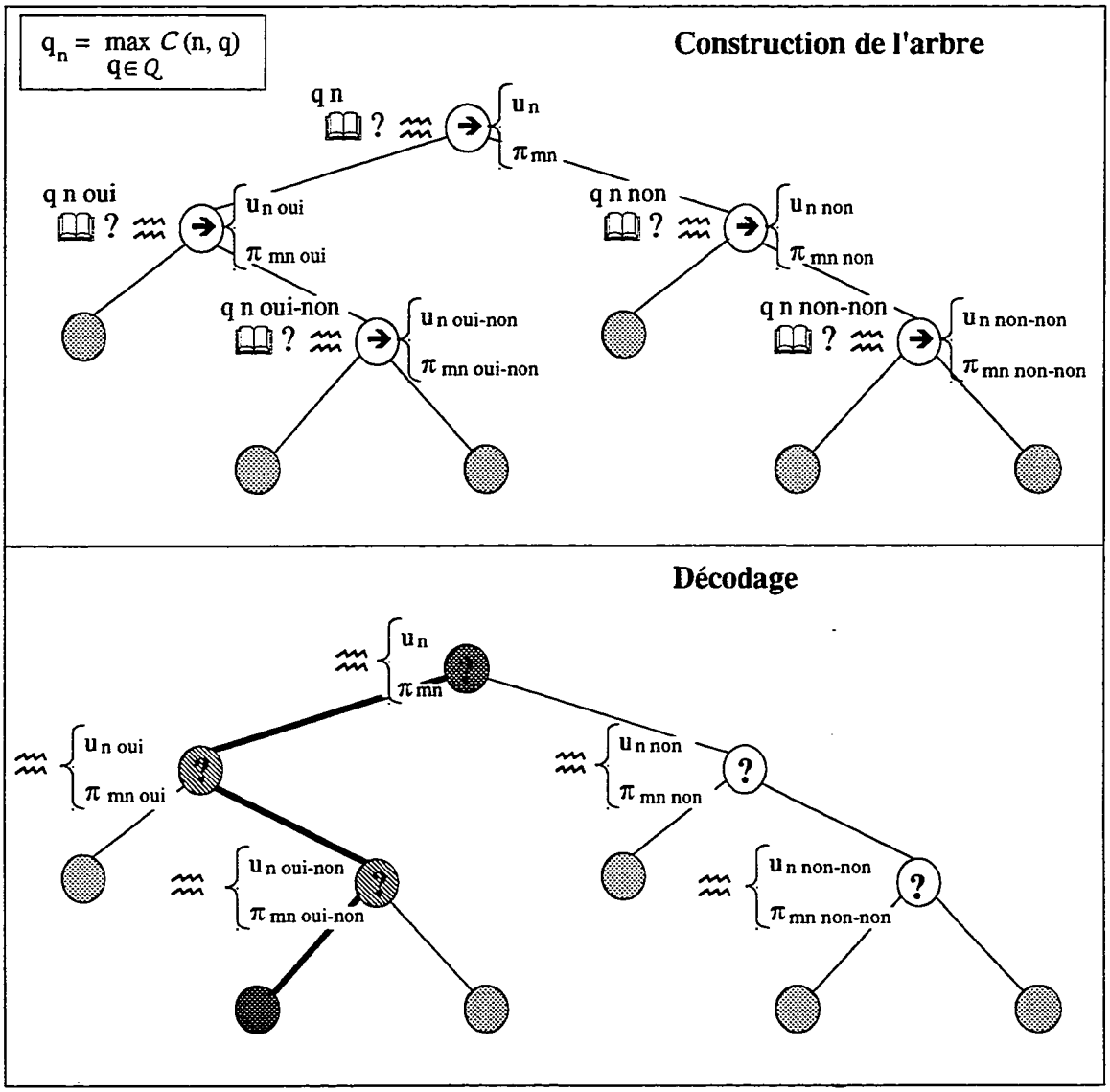

FIGURE 20. Construction et utilisation de l'arbre

Remarquons qu'aucun ordre n'est défini sur les mots composant la feuille. La notion de rang moyen n'existe pas. L'évaluation de l'arbre consiste à se doter d'un corpus de test, à le décoder selon l'algorithme précédent et à l'évaluer en comptant le nombre de fois où les mots du corpus ne font pas partie de la liste des mots candidats proposée par l'arbre. En ramenant ce nombre à 100, on obtient le pourcentage d'erreurs commises par le décodeur.

DISCUSSION

L'analyse discriminante procède à un regroupement des individus sur lesquels elle s'exerce. Elle définit par là même pour les individus d'un même groupe, un portrait robot dont les traits représentent les caractéristiques invariantes de cette population. Cette méthode est appliquée à

des fins très diverses. En parole, citons les travaux de G.S. Sebestyen sur la reconnaissance automatique des chiffres [83]. Plus récemment, Kitazawa utilise dans [46] et [47] une analyse discriminante pratiquée sur une population constituée de plosives pour mettre en évidence les caractéristiques invariantes de chacune d'elles. Cette même analyse utilisée dans un but décisionnel conduit à un taux de reconnaissance des plosives de 87% en français contre 82,5% en japonais. Quatre-vingt-sept fois sur 100 les caractéristiques (indépendantes du locuteur et du contexte phonétique) de la plosive mise à jour par l'analyse discriminante se sont avérées efficaces. Une étude similaire menée par M. El-Bèze [31][32] sur les voyelles et les fricatives pour la mise au point d'une méthode d'assistance automatique à l'articulation des sons stables, conduit à des résultats un peu moins performants. Ceci tient sans doute à la différence de représentation des données sur lesquelles s'exerce l'analyse. Que peut-on espérer d'une analyse discriminante sur les histogrammes d'étiquettes acoustiques? Le fait de considérer les histogrammes plutôt que les segments acoustiques comporte l'inconvénient, déjà mentionné, de la perte de la séquentialité qui a comme conséquence que les anagrammes phonétiques "grappe" et de "Prague" seront classés dans la même feuille. Si dans les expériences menées par Kitazawa, les résultats de l'analyse conduisent à une interprétation claire des données, comme celle d'une opposition entre les consonnes vélaires selon qu'elles sont suivies ou non d'une voyelle labialisée, ici, la séparation des histogrammes n'exprime qu'une différence de formes moyennes mesurée par λ. S'il y avait parfaite adéquation entre sous ensembles d'étiquettes acoustiques et phonèmes, cette séparation garantirait des feuilles d'anagrammes phonétiques et la différence des formes moyennes serait interprétée comme présence ou absence d'un phonème particulier. Mais le fait de travailler sur les étiquettes acoustiques induit un bruit lié aux approximations de la quantification vectorielle dont elles sont issues. L'adéquation des étiquettes acoustiques aux phonèmes ici n'existe pas. La section suivante expose différentes expériences menées afin d'obtenir une séparation toujours plus marquée entre les différentes classes.

EXPERIMENTATION

PROLOGUE

Les expériences suivantes rendent compte du fait que sur la parole, la séparation induite par l'analyse discriminante est rarement parfaite. Nous avons fait varier trois types de paramètres: la frontière de séparation, l'intervalle de recouvrement des données et la structure de l'arbre.

CORPUS ET LOCUTEURS

Pour ces premières expériences, on s'est fixé de tester la méthode sur un dictionnaire dont on possédait un enregistrement pour plusieurs locuteurs. Etant donné le petit volume de données que nous possédons sur les mots isolés, nous avons retenu un dictionnaire de syllabes plutôt qu'un dictionnaire de mots. Le corpus considéré est le même que celui utilisé dans le chapitre précédent, c'est à dire composé d'un dictionnaire de 6.400 syllabes et deux textes mis au point par A.M. Derouault et P. Combescure. L'ensemble de ce corpus contient 12.000 syllabes dont 6400 différentes. Le texte de test contient 1150 syllabes dont 340 différentes. L'arbre dépend du locuteur. En effet, des tests initiaux faits sur des données en provenance de locuteurs différents s'étant soldés par une reconnaissance médiocre, j'ai opté pour la construction d'une analyse par locuteur. Les expériences qui suivent ont été menées sur le locuteur MOY.

AU FIL DES EXPERIENCES

Différents cas de figure ont été étudiés. Nous relatons ici les principales étapes de nos prospections.

La première expérience correspond exactement à la méthode décrite dans les paragraphes précédents. Ses caractéristiques sont les suivantes:

p1:0 point de séparation: $\pi_m = u^t.\frac{\bar{z}_{non} + \bar{z}_{oui}}{2}$

p2:0 A_{oui} contient l'ensemble des histogrammes relatifs aux transcriptions classées dans T_{oui} et A_{non} le reste des histogrammes.

p3:0 On construit un arbre pour l'ensemble des données.

La valeur du *point de séparation* π_m est d'une importance capitale puisque le bon choix de sa valeur détermine le taux de reconnaissance du système. Par exemple, si la question posée est "la transcription phonétique comprend-elle un [t]?", les réponses sépareront les histogrammes en deux groupes et l'analyse discriminante calculera le pouvoir discriminant de cette question. Supposons que cette question obtienne le meilleur score. Il est probable que parmi les échantillons classés dans le groupe "non" certains possèdent une forte proportion d'étiquettes proches de celles relatives au [t]. Les projections de tels éléments sur l'axe factoriel u seront très proches de celles du groupe "oui" et étant donné la définition de π_m, ces éléments seront mal orientés au décodage. On peut améliorer la définition de π_m en tenant compte de la dispersion des échantillons autour de leur centre de gravité. Pour ce faire, on supposera que les éléments de A_{oui} et A_{non} se projettent selon une distribution gaussienne respectivement autour de $u^t.\bar{z}_{oui}$ de variance σ^2_{oui} et autour de $u^t.\bar{z}_{non}$ de variance σ^2_{non}. Le point de séparation est alors défini par:

$$\pi'_m = \frac{(\bar{z}_{oui} + \tau.\sigma_{oui}) + (\bar{z}_{non} - \tau.\sigma_{non})}{2} = \pi_m + \tau.\frac{\sigma_{oui} - \sigma_{non}}{2}$$

où τ est un coefficient statistique déterminé par une table gaussienne. Sous l'hypothèse d'une distribution gaussienne, lorsque $\tau = 2$ A_{oui} et A_{non} sont séparés avec une probabilité de 95%.

Le fait de déplacer la frontière ne résout cependant pas tous les problèmes. La plage de recouvrement entre les éléments projetés de A_{oui} et de A_{non} est rarement vide. Nous avons tenté deux approches. La première consiste à dupliquer les éléments perturbant la séparation dans A_{oui} et A_{non}. La seconde consiste à les réunir, créant ainsi un noeud spécial pour les éléments confus. Enfin, pour estimer ce que la perte de l'information séquentielle pouvait avoir comme répercussion sur le taux de reconnaissance, nous avons construit plusieurs arbres chacun travaillant sur des transcriptions phonétiques de longueur fixée. Lors du décodage les arbres sont mis en parallèle.

EXPERIENCES ET RESULTATS

DESCRIPTION

Exp0:

Base:

p1:0 point de séparation:

$$\pi_m = u^t . \frac{\bar{z}_{non} + \bar{z}_{oui}}{2}$$

p2:0 A_{oui} contient l'ensemble des histogrammes relatifs aux transcriptions classées dans T_{oui} et A_{non} le reste des histogrammes.

p3:0 On construit un arbre pour l'ensemble des données.

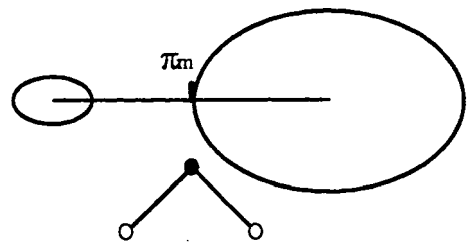

FIGURE 21 Schéma Exp0

Exp1:

Particularité:

p1:1 π_m est remplacé par

$$\pi'_m = \frac{(\bar{z}_{oui} + \tau.\sigma_{oui}) + (\bar{z}_{non} - \tau.\sigma_{non})}{2}$$

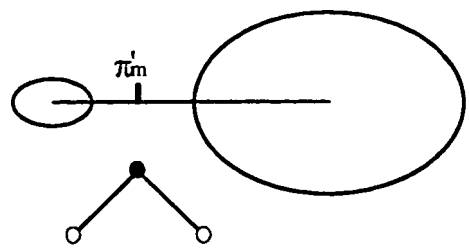

FIGURE 22 Schéma Exp1

Exp2:

Particularité:

p2:1 On définit un intervalle de recouvrement.

Les données se projetant dans cet intervalle sont dupliquées dans A_{oui} et A_{non}.

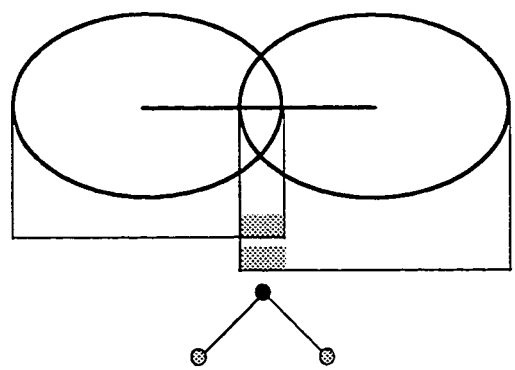

FIGURE 23 Schéma Exp2

Exp3:

Particularité:

p3:1 On construit plusieurs arbres dépendant de la longueur des transcriptions phonétiques.

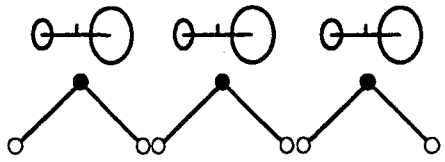

FIGURE 24 Schéma Exp3

Exp4:

Particularité:

p2:2 On définit un intervalle de recouvrement.

Les données se projetant dans cet intervalle sont réunies dans A_{flou}.

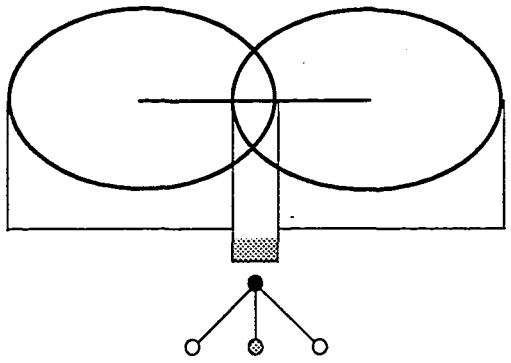

FIGURE 25 Schéma Exp4

Définition de l'intervalle de recouvrement

Initialement l'intervalle de recouvrement est défini quand il existe, par l'espace séparant min_{non} de max_{oui}, où min_{non} est la plus petite valeur des projetés de A_{non} sur u et max_{oui} la grande valeur des projetés de A_{oui} sur u. Les éléments de A_{non} dupliqués dans A_{oui} sont ceux se trouvant entre π_m et max_{oui} et ceux de A_{oui} dupliqués dans A_{non} sont ceux se trouvant entre min_{non} et π_m. Cette manière de dupliquer les données s'étant avérée insuffisante, on s'est imposé de dupliquer, en plus, un pourcentage des histogrammes de $\mathcal{A}$ ne faisant pas partie de ceux de A_n mais[28] dont la projection tombe dans l'intervalle de recouvrement. Dans l'expérience 4, A_{flou} contient l'ensemble des données se projetant dans l'intervalle de recouvrement, et est enrichi d'un pourcentage des données de $\mathcal{A}$.

Définition de la dépendance des arbres aux longueurs des transcriptions

Les transcriptions phonétiques des syllabes du corpus sont codées dans un alphabet de 40 signes. Trente-trois de ces signes sont relatifs aux phonèmes eux-mêmes. Parmi les sept restants, trois désignent l'occlusion précédant l'explosion des occlusives sourdes, trois autres désignent des transitions entre silence et syllabe, et le dernier le silence. Ces signes proviennent d'un codage phonétique des mots dans le jeu des 40 automates phonétiques du système Tangora. Ils ont été conservés car ils véhiculent une information qui peut être utile lors de la

[28] n étant le noeud traité.

construction de l'arbre. La longueur d'une transcription est ici définie comme étant le nombre de signes qu'elle comprend. Ainsi les transcriptions des syllabes "à" et "ta" ont une longueur de trois, "lac" et "hâte" ont une longueur de cinq. Quatre arbres ont été construits. Un premier pour les transcriptions de longueur trois, un second pour les transcriptions de longueur trois ou quatre, un troisième pour les transcriptions de longueur quatre ou cinq et un dernier pour les transcriptions de longueur supérieure ou égale à cinq. Au décodage ces quatre arbres sont mis en parallèle. Les quatre feuilles obtenues sont fusionnées. Leur fusion définit la liste des mots candidats.

RESULTATS

Exp	p1	p2	p3	données	reco.	erreurs	liste	σ	max.	min.
0	0	0	0	1150	94,87	5,13	286	206	677	23
1	1	0	0	1150	84,96	15,04	291	218	677	23
2.0	0	1	0	1150	96,43	3,57	322	175	640	21
2.1	1	1	0	1150	97,04	2,96	468	294	992	23
3.0	0	1	1	1150	98,17	1,83	699	223	1286	172
3.1	1	1	1	1150	98,52	1,48	845	375	1700	162

Point de séparation:	*p1 = 0 original*	*p1 = 1 déplacé*
Intervalle de recouvrement	*p2 = 0 sans*	*p2 = 1 avec*
Dépendance à la longueur	*p3 = 0 sans*	*p3 = 1 avec*

ANALYSE DES ERREURS:

le déplacement de la frontière

	Nombre d'erreurs	EC2E	ED2E	Exemple d'erreur pour l'une mais pas pour l'autre
Exp0	59	44	15	"moi" question 1 [s z]? réponse: non question 2 [a]? réponse: non $u^t.I = 2.20$ $\pi_{m,n} = 1.41$ alors que $\pi_{m,n} = 2.71$ pour Exp1
Exp1	173	44	129	"mie" question 1 [s z]? réponse: non question 2 [a]? réponse: non question 3 [ɛ e]? réponse: non question 4 [i]? réponse: oui question 5 [y o ɔ ɔ̃ w]? réponse: non question 6 [ʃ ʒ]? réponse: oui $u^t.I = 1.78$ $\pi_{m,n} = 6.80$ alors que $\pi_{m,n} = -2.64$ pour Exp0

EC2E: Nombre d'erreurs communes aux deux expériences.

ED2E: Nombre d'erreurs non rencontrées dans l'autre expérience.

En regard des résultats de l'expérience 1 et en comparaison de ceux de l'expérience 0, le déplacement de la frontière de séparation dégrade plutôt qu'il n'améliore la reconnaissance. En fait dans l'expérience 1, deux "mauvaises" questions: [ʃ ʒ]? et [.v]? monopolisent plus de 73% des erreurs, alors que dans l'expérience 0 la seule mauvaise question: [.v]? n'est responsable que de 23% des erreurs. Parmi les autres erreurs beaucoup sont communes aux deux expériences. Pour celles-ci, le déplacement du point de séparation joue parfois en faveur de l'expérience 1, parfois en sa défaveur. La figure suivante donne l'arbre obtenu dans les expériences 0 et 1:

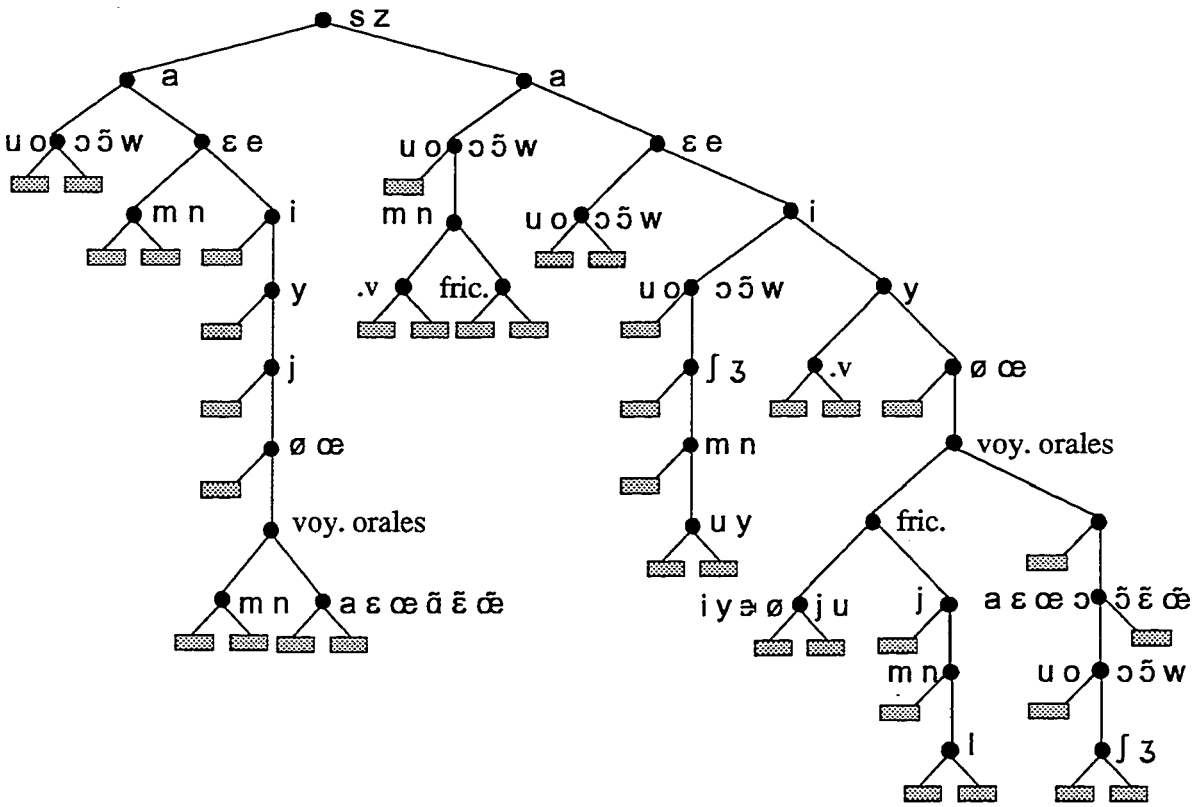

FIGURE 26 Arbre expérience 0

Notons que les questions portent très souvent sur les voyelles et les fricatives. Constatation heureuse, la construction de l'arbre n'est pas dénuée de sens. Pour les expériences suivantes, le déplacement du point de séparation conduit à une amélioration des performances.

	Nombre d'erreurs	EC2E	ED2E
Exp2.0	41	20	21
Exp2.1	34	20	14
Exp3.0	21	13	8
Exp3.1	17	13	4

EC2E: Nombre d'erreurs communes aux deux expériences Exp2.0 et Exp2.1 (resp. Exp3.0 et Exp3.1)

ED2E: Nombre d'erreurs non rencontrées dans l'autre expérience (Exp2.0 versus Exp2.1 et Exp3.0 versus Exp3.1)

Au cours des expériences Exp0 et Exp1 les erreurs les plus fréquentes, hormis celles liées aux mauvaises questions, sont les suivantes:

	exemple	Erreurs Exp0	Erreurs Exp1
raté du [e ɛ]	"se" [e]? →NON	9	6
raté du [a]	"mwa"[a]? →NON	9	4

raté du [ø œ]	"vø" [ø]? →NON	5	3
raté du [y]	"py" [y]? →NON	2	3
confusion [i] ou [ɛ] au lieu de [y]	"gy" [i]? →OUI	2	3
confusion [y] ou [i] au lieu de [j]	"rj " [i]? →OUI	1	5
conf. [a] au lieu de [ɛ][o] ou [ø]	"pro" [a]? →OUI	8	3
confusion [œ] au lieu de [ɔ]	"rjod" [œ]? →OUI	1	2
confusion [ɑ̃] ou [ɛ] au lieu de [ɛ̃]	"bjɛ̃" [ɑ̃]? →OUI	4	3
quelques confusions sur les consonnes	"ʒir" [s z]? →OUI	5	5

Il y a un peu plus de ratés pour l'expérience 0 qu'il n'y en a pour l'expérience 1 (29-17). C'est l'inverse pour les confusions (16-29). Pour les voyelles, une comparaison avec les résultats obtenus dans [32] montre un recoupement sur les confusions, notamment pour le [i] et le [ɛ̃]. Il y a divergence, par contre, sur le [a] jugé dans l'article peu sujet à confusion. Les confusions les plus notables [ɔ̃/ɛ̃][ɔ̃/ɛ̃][ɛ̃/u] mises en évidence dans l'article ne se retrouvent pas ici, mais la question singleton [ɔ̃] n'apparaît qu'une fois et assez profondément. Les questions singletons [ɛ̃] et [u] n'apparaissent jamais.

La duplication des données

Le nombre des erreurs décroît. Il n'y a plus comme précédemment de "mauvaises" questions sur lesquelles se focalisait une forte proportion des erreurs. Les erreurs ont à quelques exceptions près, déjà été rencontrées lors des expériences précédentes. On remarque une prépondérance des ratés sur le [a], les [ɛ] ou [ ] et les [ø] ou [œ] et des confusions, surtout autour des [i], [y] et [j]. Il y a quasiment autant de ratés que de confusions. La duplication rend l'arbre plus robuste. Une erreur d'aiguillage peut être récupérée grâce à la duplication. Par contre, elle augmente de façon non négligeable la taille des feuilles. La figure suivante donne l'arbre obtenu au cours de l'expérience 2:

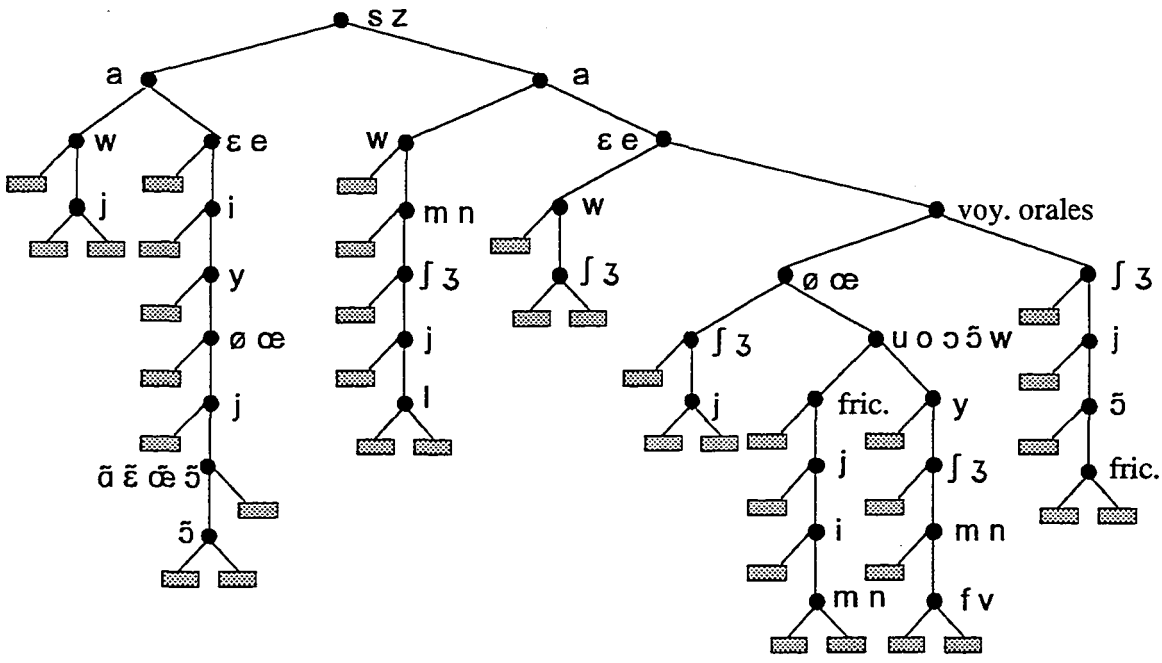

FIGURE 27 Arbre expérience 1

Construction de plusieurs arbres avec duplication

Les confusions et ratés diminuent tout en conservant les proportions de l'étape précédente. Seuls les ratés sur [a] restent importants (4 sur 8 ratés). Cette expérience est de loin celle qui obtient les meilleurs scores de reconnaissance et de loin aussi celle qui présente les feuilles les plus grosses. Bien que l'approche ne soit pas viable tel quel à cause de la taille des feuilles, elle permet de se rendre compte de l'influence de la longueur des syllabes sur la reconnaissance. En effet, si on ne décode que les syllabes de longueur 3 (au nombre de 736 dans le test) on obtient un score de 99,32% pour une liste moyenne de 77 syllabes. Si on décode les syllabes de longueur inférieure ou égale à 4 (au nombre de 1049 dans le test) on obtient un score de 98,85% avec une liste moyenne de 245 syllabes. L'idéal serait de pouvoir choisir le bon arbre en fonction des caractéristiques du segment acoustique. C'est un problème ouvert sur lequel je n'ai pas eu le temps de me pencher. La figure suivante donne les arbres obtenus au cours de l'expérience 3:

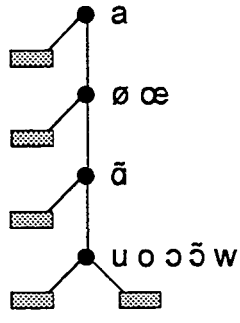

FIGURE 28.a Arbres expérience 3: longueur $l \leq 3$

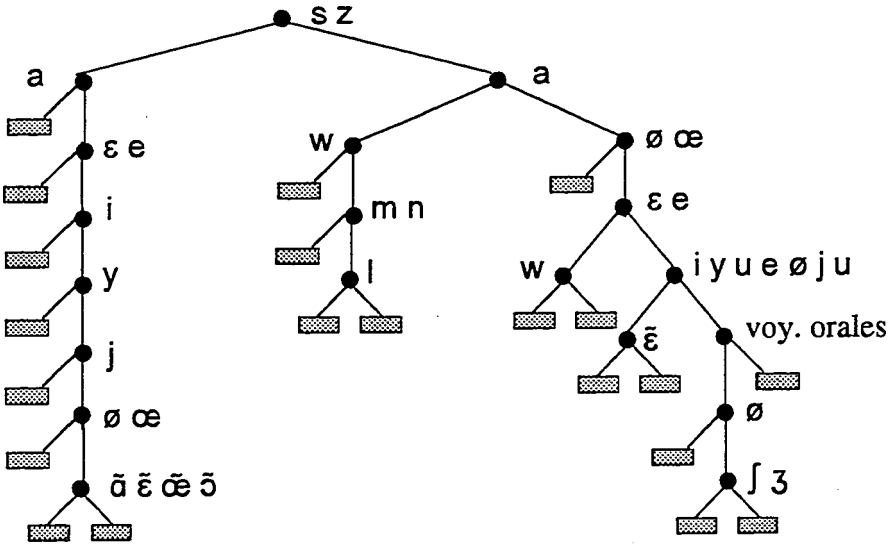

FIGURE 28.b Arbres expérience 3: longueur $3 \leq l \leq 4$

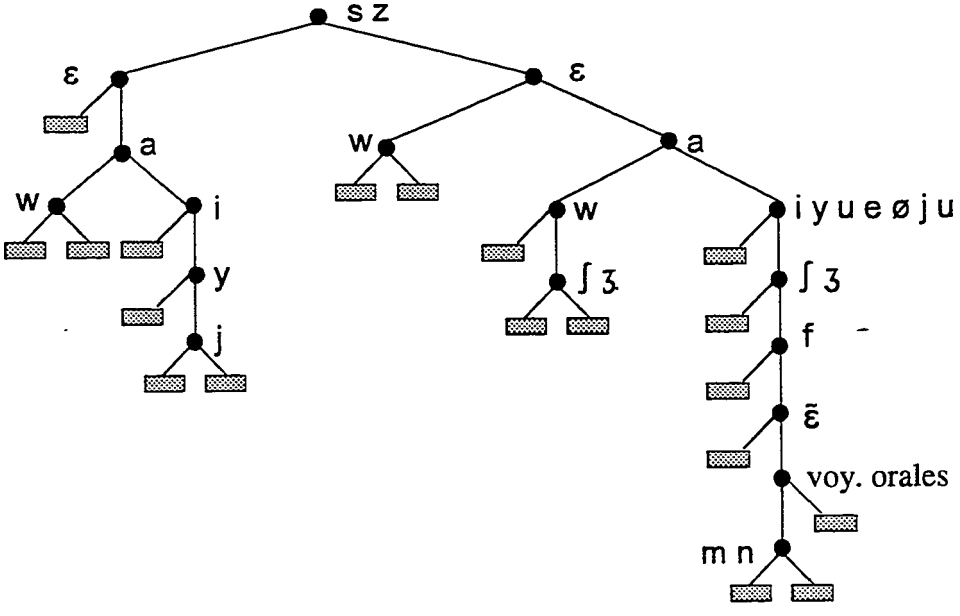

FIGURE 28.c Arbres expérience 3: longueur $4 \leq l \leq 5$

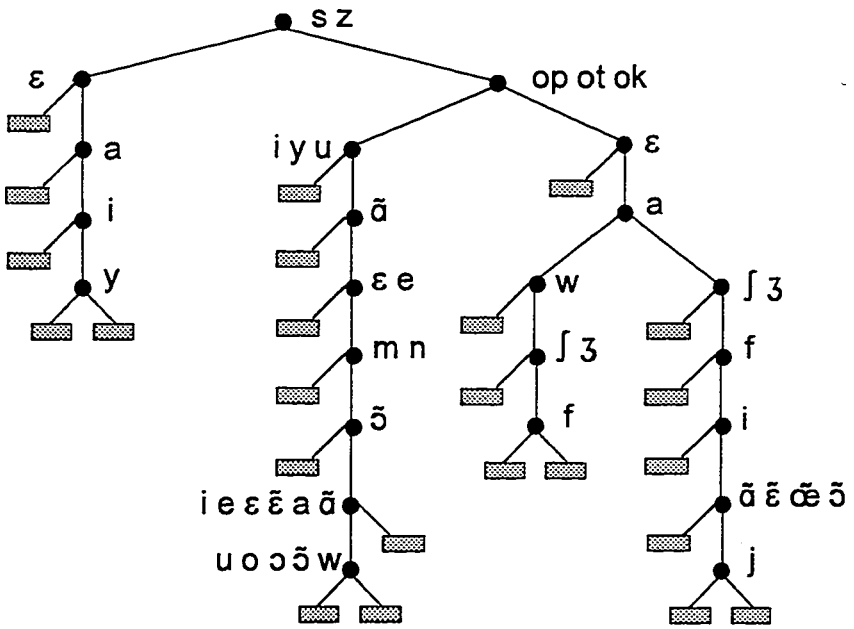

FIGURE 28.d Arbres expérience 3: longueur $l \geq 5$

Expérience 4

L'expérience 4 a été menée sur un sous ensemble de données constitué des transcriptions de longueur inférieure ou égale à quatre. Les résultats sont les suivants:

Exp	p1	p2	p3	données	reco.	erreurs	liste	σ	max	min.
2.1	1	1	0	1049	98	2	247	139	491	23
4.1	1	2	0	1049	86,55	17,45	89	52	177	1

Point de séparation: *p1 = 0 original* *p1 = 1 déplacé*

Intervalle de recouvrement *p2 = 0 sans* *p2 = 1 avec* *p2 = 2 trois branches*

Dépendance à la longueur *p3 = 0 sans* *p3 = 1 avec*

Le taux de reconnaissance se dégrade. Notons cependant, que l'arbre est plus volumineux et que par conséquent les feuilles sont moins chargées. La matrice de covariance s'inverse plus volontiers quand elle est calculée sur les éléments de la branche centrale, c'est-à-dire, sur un ensemble de données très disparates (décorrélées), que quand elle est calculée sur un ensemble de données mixtes (Exp 2), c'est-à-dire, constitué de données propres (corrélées) bruitées par quelques données disparates. Une solution permettant de contourner le problème de la non inversion de la matrice de covariance, qui conduit à des arbres peu profonds (Exp 1, 2, 3), est de considérer la pseudo inverse de Moore-Penrose plutôt que l'inverse de la matrice de covariance. Une autre solution consisterait à pratiquer, avant de calculer la matrice de covariance et son inverse, une analyse en composantes principales qui permettrait de déterminer les descripteurs pertinents de l'ensemble des données associées au noeud traité.

La figure suivante donne les arbres obtenus au cours de l'expérience 4:

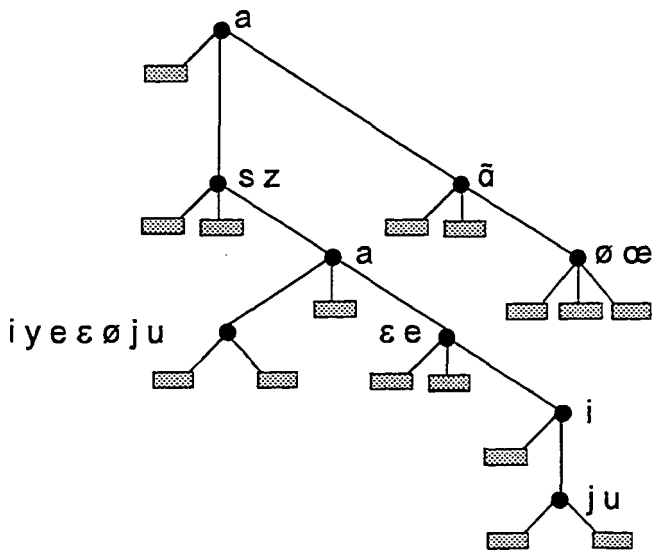

FIGURE 29.a Arbres expérience 4: longueur $l \leq 3$

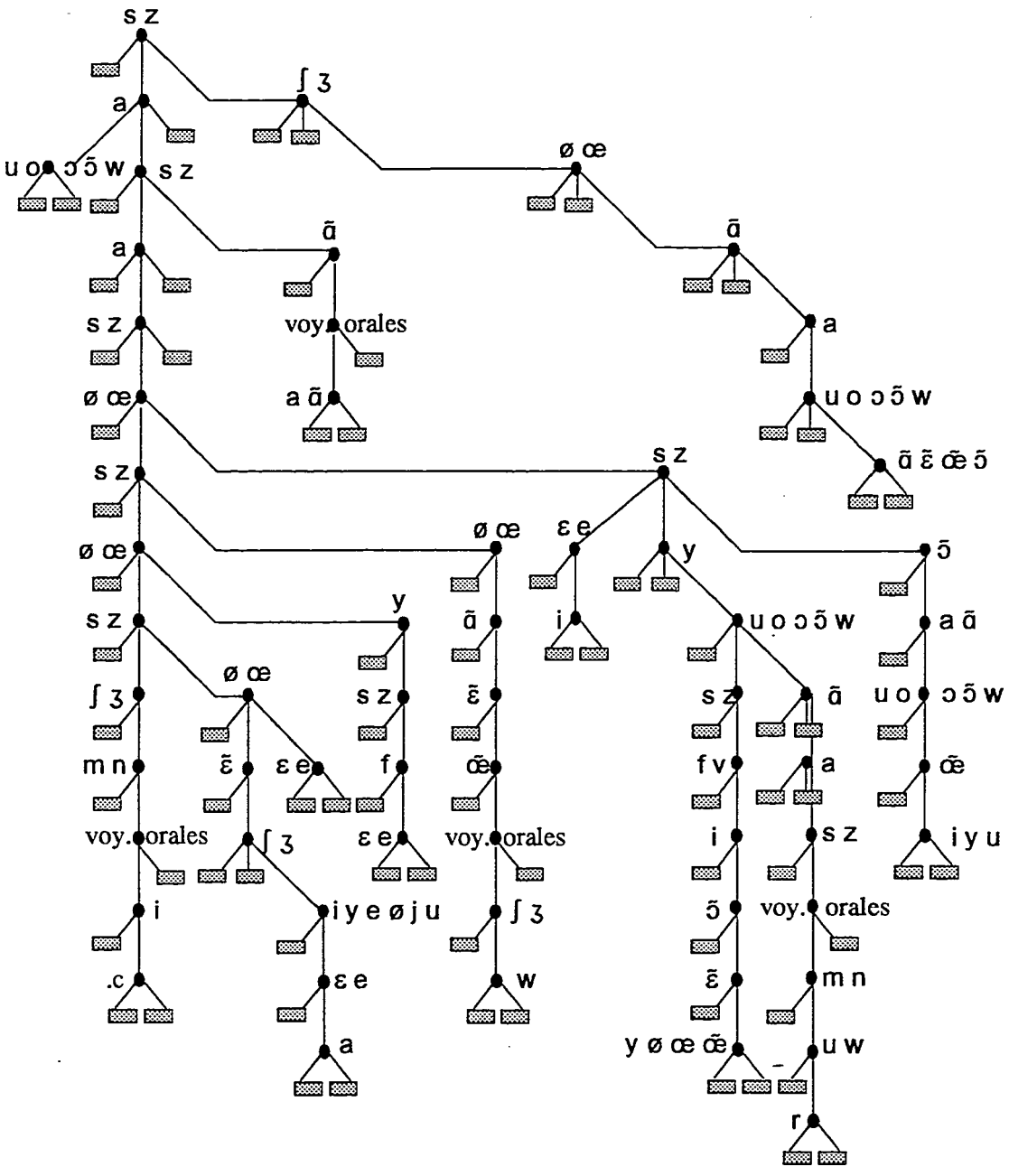

FIGURE 29.b Arbres expérience 4: longueur $3 \leq l \leq 4$

CONCLUSION

Nous avons présenté dans ce chapitre une méthode permettant d'obtenir, étant donné un mot prononcé, une liste de mots candidats. Cette méthode repose sur un arbre de décision. Lors de la construction de l'arbre, des questions de nature phonétique permettent de partitionner le corpus d'apprentissage en classes de mots phonétiquement proches sur lesquelles on applique une analyse discriminante. Cette analyse fournit à la fois une question de nature acoustique image de la question phonétique posée et un critère de séparation. L'avantage de la structure arborescente est sa très grande rapidité de consultation. Dans le cas idéal où une question partage la liste des mots candidats par deux, pour un dictionnaire de 20.000 mots et un segment acoustique de longueur n, il suffit d'effectuer $7 \times n$ additions pour obtenir une liste de 150 mots candidats. La séparation des données d'apprentissage est rarement tranchante. Différentes expériences ont été menées pour tester la séparation. Au cours de ces expériences, trois paramètres varient. La frontière de séparation, le traitement de la plage de recouvrement et la structure de l'arbre. Dans la plupart des expériences le choix d'une frontière pondérée par les écarts types des deux parties séparées s'est avéré fructueux. Le choix de la construction de plusieurs arbres conduit à des taux de reconnaissance satisfaisant mais à des listes moyennes trop importantes. En ce qui concerne la plage de recouvrement, deux approches ont été proposées. La première consiste à dupliquer les données appartenant à l'intervalle de recouvrement, la seconde à créer pour ces éléments, une branche particulière. La première approche est plus robuste, elle permet de récupérer des erreurs d'aiguillage. Par contre elle conduit à un arbre moins profond que la seconde, et par conséquent à des feuilles plus volumineuses. Ce sujet reste ouvert a toute proposition d'amélioration. Nous proposons dans les lignes suivantes quelques axes de recherches.

Perspectives expérimentales.

- Une des critiques que l'on peut faire à cette méthode est que la construction de l'arbre du préfiltre dépend éminemment du dictionnaire. Non seulement, comme dans tout autre système, parce que si le mot prononcé ne fait pas partie du dictionnaire, il ne sera représenté par aucune feuille et par conséquent ne pourra pas être reconnu. Mais surtout, parce que si on désire changer de dictionnaire, il faut reconstruire l'arbre du préfiltre, et pour ce faire, avoir à disposition différentes réalisations acoustiques de chaque mot du

dictionnaire. Cette condition étant irréaliste, une solution consiste à remplacer les histogrammes des segments acoustiques réels par des histogrammes synthétiques. Pour ces histogrammes synthétiques, les fréquences d'occurrences des observations acoustiques pourront être estimées à l'aide des automates phonétiques et/ou fénoniques comme c'est le cas dans [4].

- De plus, pour introduire une dimension linguistique dans ce traitement, il pourrait être intéressant de pondérer les histogrammes par probabilité uni-gramme des mots.
- D'autre part, l'utilisation de transcriptions phonétiques contextuelles plutôt que les transcriptions phonétiques devrait, sous la condition d'une adaptation de l'ensemble des questions, conduire à de meilleurs résultats.
- Enfin, l'approche arbres multiples (Exp.3) doit être accompagnée d'une stratégie de choix du bon arbre qui permettrait une réduction drastique de la taille de la liste de mots candidats.

Perspectives mots isolés, parole continue:

- Le traitement des mots isolés peut se faire comme celui des syllabes isolées. En phase de décodage, on peut avoir une bonne idée de la ou des longueurs possibles du mot en utilisant un système de détection des silences. Il serait peut-être intéressant d'envisager, non pas l'histogramme du mot tout entier, mais des histogrammes par morceaux. En phase de décodage, le problème de la détection des frontières de morceaux dans le mot renvoie à celui de la détection des frontières de mots en parole continue. Pour le résoudre, il faudra des méthodes adaptées. On pourra, par exemple, tester la sensibilité de l'algorithme en essayant quelques longueurs standards.
- Rappelons que le choix de la nature des données sur lesquelles va s'exercer l'analyse discriminante est prépondérant. On peut se poser la question de la qualité des données ici utilisées et essayer de trouver des améliorations possibles. Deux critiques peuvent être faites sur les données que nous avons utilisées. L'une concerne le choix de l'histogramme comme représentant des segments acoustiques: il entraîne la perte de l'information séquentielle sur l'ensemble du mot. En fait, dans le cas idéal d'une parfaite correspondance entre prototypes et phonèmes, ce choix conduirait à rassembler dans une même feuille les anagrammes phonétiques. Ce qui ne constitue pas un inconvénient majeur dans la mesure où l'étape du préfiltre acoustique n'est qu'un filtre intermédiaire. Le choix définitif du mot ne lui appartient pas. Cette remarque conduit à la seconde critique qui porte sur la quantification vectorielle. La quantification vectorielle est ici pratiquée à l'aide de l'algorithme des K-moyennes. Le choix des prototypes initiaux est fait au hasard, leur nombre est fixé d'avance. Leur calcul se fait sans contrainte de correspondance avec les phonèmes de la langue. Deux attitudes sont possibles: soit on cherche à améliorer cette quantification vectorielle, soit on s'en dispense. Si on remet en cause la quantification

vectorielle en décidant de travailler directement sur la représentation spectrale (ou cepstrale) du signal, comme le fait Kitazawa dans [46] pour les plosives, il faudra trouver, à l'échelle du mot, une solution au problème de la normalisation des données pour le calcul de la matrice de covariance. Sinon, on peut chercher à améliorer la quantification vectorielle, par exemple en s'inspirant des travaux de Bahl [12][16] dans lesquels le calcul des prototypes est supervisé de sorte à ce qu'ils représentent chacun un phonème particulier.

CONCLUSION

Le passage d'un mode d'élocution isolé à un mode d'élocution continu pour des systèmes probabilistes de reconnaissance de la parole, est essentiellement un problème acoustique. Une des difficultés réside en l'accroissement de la variabilité contextuelle. Nous avons présenté deux méthodes qui cherchent à modéliser cette variabilité.

La première méthode considère le problème d'un point de vue phonétique. Elle consiste à construire un modèle acoustique phonétique adapté au locuteur. Un processus de sélection choisit parmi les variantes phonétiques proposées par un phonétiseur, la variante plus probable étant donné ce qui à été prononcé. Cette sélection permet l'apprentissage d'un modèle acoustique adapté au locuteur. Sélection et apprentissage sont itérés jusqu'à ce que l'on n'observe plus d'améliorations. Les résultats d'un décodage acoustique montrent une légère amélioration des performances. Trois paramètres conduisent à une amélioration notable des performances:

- la qualité des probabilités diphones
- la recherche optimale de la position du silence
- l'utilisation de phonèmes contextuels, aussi simples soient-ils.

La seconde approche considère le problème d'un point de vue allophonique. Elle cherche à modéliser des variations plus fines que celles prises en compte par la première approche. On utilise des arbres de décision dont le but est de regrouper les réalisations acoustiques de chaque phonème en classes d'équivalences à partir desquelles seront construits des automates allophoniques. Les premiers résultats de décodage en mots isolés pour lesquels les données d'apprentissage des arbres ont été prononcées en syllabes isolées sont relativement décevants. Sur une nouvelle base de données mieux adaptée, le taux d'erreur diminue de 12,7 % par rapport à l'approche fénonique.

- L'efficacité de cette méthode est subordonnée à la cohérence et à la quantité des données disponibles pour la construction des arbres.

Nous avons présenté, d'autre part, une nouvelle méthode permettant d'obtenir rapidement une liste de mots candidats. Le coût du décodage en terme de quantité d'opérations à effectuer croît avec la taille du vocabulaire. Dans le cadre d'une reconnaissance sur un grand vocabulaire, le décodage acoustique doit se faire en deux étapes. Une première étape, grossière, détermine une liste de mots candidats. Une seconde étape, plus fine, sélectionne les meilleurs candidats de la liste. La méthode proposée repose sur des arbres de décision. Le critère de séparation et les questions acoustiques sont obtenus par une analyse discriminante. Les premiers résultats sont satisfaisants. Ils pourront être améliorés par

- une quantification vectorielle supervisée
- l'utilisation de transcriptions phonétiques contextuelles

ANNEXES

ALPHABET PHONETIQUE

[ ]	{ }	/ /	Exemple	[ ]	{ }	/ /	Exemple
a	a	a	**a**b absurdo	p	p	op bp	**p**achyderme
u	u	u	**ou**blier	t	t	ot bt	**t**abac
i	i	i	**i**bériste	k	k	ok bk	**k**abyle
y	y	y	**u**biquiste	b	b	b	**b**abeurre
e	e	e	**é**bahi	d	d	d	**d**actylo
ɛ	ɛ	ɛ	**ai**me	g	g	g	**g**abardine
o	o	o	**au**bade	f	f	f	**f**able
ɔ	ɔ	ɔ	**o**bjecter	s	s	s	**s**able
ø	ø	ø	**eu**calyptus	ʃ	ʃ	ʃ	**ch**aleur
œ	œ	œ	**oe**il	v	v	v	**v**alse
ɑ̃	ɑ̃	ɑ̃	**an**cestral	z	z	z	**z**èbre
ɛ̃	ɛ̃	ɛ̃	**in**ca	ʒ	ʒ	ʒ	**j**aune
œ̃	œ̃	œ̃	**un**tel	l	l	l	**l**ame
ɔ̃	ɔ̃	ɔ̃	**on**ction	r	r	r	**r**ime
j	j	j	**h**iatus	m	m	m	**m**andater
w	w	w	**oi**e	n	n	n	**n**uance
ɥ	ɥ	ɥ	**lu**i	ɲ			li**gn**e
ə̇			e atone	ŋ			par**k**ing
ø̇			ø atone	ə			ø caduc
ȯ			o atone	*silence*		.. .c .v .g	

[ ]: Alphabet phonétique international

{ }: Réalisations acoustiques

/ /: Automate phonétique

CLASSIFICATION DES PHONEMES

CLASSIFICATION ARTICULATOIRE DES CONSONNES

liquide	l	r												
obstruante		m	n	p	t	k	b	d	g	f	s	ʃ	v	z	ʒ
nasale		m	n												
obs. sourde				p	t	k				f	s	ʃ			
obs. sonore							b	d	g				v	z	ʒ
occlusives				p	t	k	b	d	g						
fricative										f	s	ʃ	v	z	ʒ
occ. sourde				p	t	k									
occ. sonore							b	d	g						
fricative sourde										f	s	ʃ			
fricative sonore													v	z	ʒ
obs. bilabiale		m		p			b								
obs. apico-dent.			n		t			d							
obs. dorso-pala.						k			g						
fric. labio-dent.										f			v		
fric. alvéolaire											s			z	
fric. palatale												ʃ			ʒ

CLASSIFICATION ARTICULATOIRE DES VOYELLES

	antérieure	antérieure arrondie	postérieure
fermée	i	y	u
Semi fermée	e	ø	o
semi ouverte	ɛ (ɛ̃)	œ (œ̃)	ɔ (ɔ̃)
ouverte	a (ɑ̃)		

Les voyelles nasales sont mises entre parenthèses. Les voyelles antérieures sont aussi appelées: palatales, les postérieures: vélaires, et les palatales arrondies, labio-dentales.

LES SEMI CONSONNES OU SEMI VOYELLES

semi consonne	j w ɥ
semi consonne palatale	j ɥ
semi consonne labiale	ɥ w

ENSEMBLE DES QUESTIONS

voyelles	a i y u e ɛ ø œ o ɔ ɑ̃ ɛ̃ œ̃ ɔ̃
voyelles orales	a i y u e ɛ ø œ o ɔ
voyelles nasales	ɑ̃ ɛ̃ œ̃ ɔ̃
voyelles antérieures	i e ɛ ɛ̃ a ɑ̃
voyelles ant. arrondies	y ø œ œ̃
voyelles postérieures	u o ɔ ɔ̃ w
voyelles ouvertes	a ɑ̃
voyelles fermées	i y ɥ
voyelles semi ouvertes	ɛ œ ɔ ɛ̃ œ̃ ɔ̃
voyelles semi fermées	e ø o
voyelles antérieures fermées	i y e ø ɥ j
voyelles antérieures ouvertes	a ɛ œ ɑ̃ ɛ̃ œ̃
semi consonnes	j ɥ w
semi consonnes palatales	j ɥ
semi consonnes labiales	ɥ w
consonnes	bp bt bk b d g f s ʃ v z ʒ m n r l ɥ j w
obstruantes	bp bt bk b d g f s ʃ v z ʒ m n
obstruantes sourdes	bp bt bk f s ʃ
obstruantes sonores	b d g v z ʒ
occlusives	bp bt bk b d g
fricatives	f s ʃ v z ʒ
occlusives sourdes	bp bt bk
occlusives sonores	b d g
fricatives sourdes	f s ʃ
fricatives sonores	v z ʒ
consonnes nasales	m n
consonnes liquides	r l
obstruantes bilabiales	bp b m
obstruantes apico-dentales	bt d n
obstruantes dorso-palatales	bk g j
fricatives labio-dentales	f v
fricatives alvéolaires	s z
fricatives palatales	ʃ ʒ

silences	.. op ot ok .c .v .g
occlusion devant [p], [t] et [k]	op ot ok
Questions singleton	a
	u
	●
	●
	●
	..
inconnue	!

BIBLIOGRAPHIE

[1] Aho A., Sethi R. Ullman J. "*Compilers, principles, techniques and tools.*" Addison-Wesley 1986.

[2] Atal B., Hanauer S. "*Speech analysis and synthesis by linear prediction of speech wave.*" JASA n° 50. 1971.

[3] Bahl L., Brown P., De Souza P., Picheny M., Mercer R. "*Acoustic Markov model used in Tangora speech recognition system.*" ICASSP 1988.

[4] Bahl L., Raimo B., De Souza P., Mercer R. "*Polling: a quick way to obtain a short list of candidate words in speech recognition.*" ICASSP. 1988.

[5] Bahl L., Gopalakrishnan P., Kanevsky D., Nahamoo D. "*Matrix Fast-Match: a fast method for identifying a short list of cadidate words for decoding.*" ICASSP 1989.

[6] Bahl L., DeGennaro S., Gopalakrishnan P., Mercer R. "*A fast approximate acoustic match for large vocabulary speech recognition.*" IEEE Trans. on speech and Audio processing, Vol.1. n°1. Janv. 1993.

[7] Bahl L., Brown P., De Souza P., Mercer R. "*A tree-based language model for natural language speech recognition.*" ICASSP 1989.

[8] Bahl L., Gopalakrishnan P., Nahamoo D., Picheny M. "*Context-dependent modeling of phones in continuous speech using decision trees.*" ICASSP 1991.

[9] Bahl L., Das S., De Souza P., Epstein M., Mercer R., Merialdo B., Nahamoo D., Picheny M., Powell J. "*Automatic phonetic baseforms determination.*" ICASSP 1991.

[10] Bahl L., De Souza P., Gopalakrishnan P., Nadas A., Nahamoo D., Picheny M. "*Splitting rules for phonological decision trees.*" IBM Research Report. 1992.

[11] Bahl L., De Souza P., Gopalakrishnan P., Nahamoo D., Picheny M. "*A fast match for continuous speech recognition using allophonic models.*" ICASSP 1992.

[12] Bahl L., De Souza P., Gopalakrishnan P, Picheny M. "*Context dependent vector quantization for continuous speech recognition.*" ICASSP 1993.

[13] Baker J. "*The DRAGON System: an Overview.*" Proc. IEEE Trans. Acoust., Speech and Signal Proc. Vol.23. 1975.

[14] Baker J. "*DragonDictate™ -30K natural language speech recognition statistical methods.*" Eurospeech 1989.

[15] Baum L. "*An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov process.*" Academic Press, Inequalities, Vol.3. 1972.

[16] Bellegarda J., De Souza P., Nahamoo D., Picheny M., Bahl L. "*A supervised approach to the construction of context-sensitive acoustic prototypes.*" ICASSP 1993.

[17] Breiman L., Friedman J., Olshen R., Stone C. "*Classification and regression trees.*" Wadsworth Inc. 1984.

[18] Calliope "*La parole et son traitement automatique.*" CNET-ENST Masson. 1989.

[19] Carapiperis M.R. "*Système de conversion graphème-phonème du français.*" DESS systèmes et communication Homme-Machine. LIMSI 1986.

[20] Childers D., Skinner D. "*The Cepstrum: a guide to processing.*" Proc. IEEE 1977

[21] Chow Y., Schwartz R., Roucos S., Kimball O., Price P., Kubala F., Dunham M., Krasner M., Makhoul J. "*The role of word-dependent coarticulation effects in a phoneme-based speech recognition system.*" ICASSP 1986.

[22] Cohen M., Bernstein J., Murveit H. "*Pronunciation variation within and across speakers.*" Paper P9, presented at the 113th meeting of the Acoustic Society of America, Indianapolis. 1987.

[23] Combescure P. "*Phonetically balanced sentences.*" Recherches en acoustique. Vol.6 CNET Lannion. 1979-1980.

[24] Deng L., Lennig M., Seitz F., Mermelstein P. "*Large vocabulary word recognition using context-dependent allophonic hidden Markov models.*" Computer Speech and Language. Vol.4. 1990.

[25] Derouault A.M., Merialdo B. "*Language modeling at the syntactic level.*" ICPR 1984.

[26] Derouault A.M., Merialdo B. "*Natural language modeling for phoneme-to-text transcription.*" IEEE Trans. on PAMI. n°6. Novembre 1986.

[27] Derouault A.M. "*Context-dependent phonetic Markov models for large vocabulary speech recognition.*" NATO Advanced Institute on Pattern Recognition. Juil. 1987.

[28] Derouault A.M. "*Construction d'un corpus d'apprentissage rendant compte des contextes phonétiques.*" Etude de service du Centre Scientifique d'IBM France. 1988.

[29] Derouault A.M., El-Bèze M. "*A morphological model for large vocabulary speech recognition.*" ICASSP 1990.

[30] D'Orta P, Ferretti M., Scari S. "*Phoneme classification for real time speech recognition in Italian.*" ICASSP 1987.

[31] El-Bèze M. "*Caractérisation des sons stables et assistance automatique pour améliorer leur production.*" Mémoire de fin d'Etude. CNAM 1984.

[32] El-Bèze M. "*Refutation based recognition to help vowel articulation.*" ICASSP 1986.

[33] El-Bèze M. "*Choix d'unités appropriées et introduction de connaissances dans des modèles probabilistes pour la reconnaissance de parole.*" Thèse de Doctorat. Université Paris 7. 1990.

[34] Emerard F. "*Synthèse par diphones et traitement de la prosodie.*" Thèse de 3e cycle, Univ. des Langues et Lettres. Grenoble. 1977.

[35] Giachin E.P, Rosenberg A.E., Lee C.H. "*Word juncture modeling using phonological rules for HMM-based continuous speech recognition.*" Computer Speech and Language. Vol.5. 1991.

[36] Grevisse M."*Le Bon Usage, grammaire française.*" Duculot 1986.

[37] Guédon Y. "*Review of several stochastic speech unit models.*" Computer Speech and Language. Vol.6. 1992.

[38] Gupta V. Cornilliat B. "*Multiple-branch tied modeling for automatic speech recognition.*" en préparation.

[39] Haton J.P., Pierrel J.M. "*Organization and operation of a connected speech understanding system, at lexical, syntactical and semantical levels.*" ICASSP 1976.

[40] Haton J.P., Pierrel J.M., Perennou G., Caelen J., Gauvain J.L. "*Reconnaissance automatique de la parole.*" Afcet-Dunod informatique. 1991.

[41] Jelinek F. "*Continuous speech recognition by statistical methods.*" Proc. of IEEE Vol.64. 1976,

[42] Jelinek F., Mercer R.L. "*Interpolated estimation of Markov source parameters from sparses data.*" Pattern Recognition in practice, E. Gelsema and L.N. Kanal. 1980

[43] Jelinek F., Mercer R.L., Bahl L. "*Continuous speech recognition: statistical methods.*" Handbook of statistics, Vol.2, Classification, pattern recognition and reduction of dimensionality. Edité par P.R Krishnaiah et L.N. Kanal. North Holland. 1982.

[44] Jouvet D, Gagnoulet C. "*Reconnaissance de parole et modélisation statistique, l'expérience du CNET.*" Revue Traitement du Signal, Numéro spécial Reconnaissance Automatique de la Parole. 1990.

[45] Juang B., Rabiner L. "*A probabilistic distance measure for hidden Markov models.*" Bell Syst. Tech. J. Vol.64 1985.

[46] Kitazawa S. et Doshita S. "*Discrimination of isolated vowels and stop consonants using features extracted from spectrum.*" ICASSP 1986.

[47] Kitazawa S. "*Statistical discrimination of french initial stops and nasals.*" Rapport ENST 87D007. Août 1987.

[48] Klatt D. "*The Klattalk text-to-speech conversion system.*" ICASSP 1982.

[49] Kurematsu A. "*Speech and langage processing for automatic telephone interpretation.*" IEICE Trans. Vol.75, n°10. 1992.

[50] Lacheret-Dujour A. "*Contribution à l'analyse de la variabilité phonologique pour le traitement automatique de la parole continue multilocuteur.*" Thèse de Doctorat. Paris 7. 1990.

[51] Laporte E. "*Méthodes algorithmiques et lexicales de phonétisation de textes. Applications au français.*" Thèse de doctorat en informatique. Paris 7. 1988.

[52] Lee C.H., Soong F., Juang B. "*A segment model approach to speech recognition.*" ICASSP 1988.

[53] Lee K.F. "*An overview of the Sphinx speech recognition.*" IEEE Trans on ASSP. Janvier 1990.

[54] Lee K.F. "*Context dependent phonetic HMM for speaker independent continuous speech recognition.*" IEEE Trans on ASSP. Avril 1990.

[55] Lee K.F., Hayamizu S., Huang H.W.H.C., Swartz J., Weide R. "*Allophone clustering for continuous speech recognition.*" ICASSP 1990.

[56] Lennig M., Gupta V., Mermelstein P. "*A language model for very large-vocabulary.*" Computer Speech and Language. Vol.6. 1992.

[57] Lesser W.R. et al. "*Organization of the Hearsay II speech understanding system.*" IEEE Transactions ASSP. Vol.23. 1975.

[58] Liénard J.S., Mlouka M. "*Segmentation automatique de la parole en phonatones.*" III^{e} JEP, GALF, Lannion. 1972.

[59] Lin-Shan Lee, Chiu-yu Tseng, Hung-yan Gu, Lui F.H., Chang C.H., Hsieh S.H., Chen C.H. "*A real-time Mandarin dictation machine for chinese language with unlimited texts and very large vocabulary.*" ICASSP 1990.

[60] Lowerre B.T. et al. "*The Harpy speech recognition system.*" Technical report. Carnegie Mellon Univ. 1976.

[61] Lucassen J., Mercer R. "*An information theoretic approach to the automatic determination of phonetic baseforms.*" ICASSP 1984.

[62] Malmberg B. "*La phonétique.*" Que sais-je? Presses Universtaires de France. 1re édition: 1954, 4^{e} édition corrigée 1987.

[63] Mariani J., Lienard J.S. "*Esope 0: un programme de compréhension automatique de la parole procédant par prédiction-vérification aux niveaux phonétique, lexical et syntaxique.*" Congrés AFCET RFIA 1978.

[64] Mariani J., Gauvain J.L. "*Evaluation of time compression for connected word recognition.*" ICASSP 1984.

[65] Meloni H. "*Etude et réalisation d'un système de reconnaissance automatique de la parole continue.*" Thèse de docteur d'état, Aix-Marseille II. 1982.

[66] Meloni H., Gilles P. "*Décodage acoustico-phonétique ascendant.*" Traitement du Signal, Vol 8. n°2. 1991.

[67] Merialdo B. "*Speech recognition with very large size dictionary.*" ICASSP 1987.

[68] Merialdo B. "*Multi level decoding for very-large-size-dictionary speech recognition.*" ICASSP 1988.

[69] Mercier G. "*The Keal speech understanding system.*" Spoken Language Generation and Undertsanding J.C. Simon (Ed) D. Reidel, 1980.

[70] Miclet G., Grenier Y., Le Roux J. "*Phonetic recognition by linear prediction experiences at ENST*" Automatic speech analysis and recognition. J.P. Haton editor, Reidel. 1982.

[71] Nadas A., Nahamoo D., Picheny M. "*Normalization of speech by adaptative transformation based on vector quantization.*" ICASSP 1988.

[72] Nadas A., Nahamoo D., Picheny M. "*Speech recognition using noise-adaptative prototypes.*" ICASSP 1988.

[73] Nadas A., Nahamoo D., Picheny M., Powell J. "*An iterative "Flip-Flop" approximation of the most informative split in the construction of decision trees.*" ICASSP 1991.

[74] Pierrel J.M. "*Utilisation des contraintes linguistiques en compréhension de la parole continue: le système Myrtille II.*" TSI Vol. 1. n°5. 1982.

[75] Prouts B. "*Contribution à la synthèse de la parole à partir de texte, transcription graphème-phonème en temps réel sur microprocesseur.*" Thèse de Docteur Ingénieur, Université de Paris 11. 1980.

[76] Rabiner L., Levinson S., Sondhi M. "*An introduction to the application of the theory of probabiblistic functions of a Markov process to automatic speech recognition.*" B.S.T.J. Vol. 62, n° 4. Avril 1983.

[77] Rabiner L., Wilpon J, Soong F. "*High performance connected digit recognition using hidden Markov models.*" ICASSP 1988.

[78] Randolph M. "*A data-driven method for discovering and predicting allophonic variations.*" ICASSP 1990.

[79] Revuz D. "*Algorithme linéaire de minimisation des automates déterministes cycliques.*" Institut Blaise Pascal. Publication LITP. 1990

[80] Romeder J.M. "*Méthodes et programmes d'analyse discriminante.*" Dunod. Bordas 1973.

[81] Sakoe H., Chiba S. "*Dynamic programming algorithm optimization for spoken word recognition.*" ICASSP 1978.

[82] Saporta G. "*Théories et méthodes de la Statistique.*" Technip 1978.

[83] Sebestyen G.S. "*Decision making process in pattern recognition.*" The Macmillan Company. 1962.

[84] Tubach J.P., Boe L.J. "*Quantative knowledge on word structure, from a phonetic corpus, with application to large vocabularies recognition system.*" ICASSP 1986.

[85] Tubach J.P., Descout R., Isabelle P. "*Reconnaissance en entrée d'un système de traduction automatique, en anglais et en français.*" 18ième JEP. 1990.

[86] Viterbi A. "*Error bounds for convolutional codes and an asymptotically optimum decoding algorithm.*" IEEE Trans. Inform. Theor., Vol.IT-13. 1967.

[87] Vittorelli V., Adda G., Billi R., Boves L., Jack M., Vivalda E. "*Esprit POLYGLOT project: Multilingual speech recognition and synthesis.*" Proc. ICSLP. 1990.

[88] Waast C. "*Phonétiseur du français avec variantes, son intégration dans un système de reconnaissance probabiliste.*"8e congrès AFCET-RFIA. 1991.

[89] Waibel A., Toshiyuzi Hanazawa, Hints G., Kihohiro Shikano, Lang K.Y. "*Phoneme recognition using time delay neural networks.*" IEEE Trans. on Acoustic Speech and Signal Processing, 37, 3. 1989.

[90] Wolf J.J., Woods W.A. "*The HWIN speech understanding system..*" ICASSP 1977.

[91] Zue V., Lamel L. "*An expert spectrogram reader: a knowledge-based approach to speech recognition.*" ICASSP 1986.

[92] Zue V., Glass J., Phillips M, Seneff S. "*Acoustic segmentation and phonetic classification in SUMMIT speech recognition system.*" ICASSP 1989.

TABLE DES MATIERES

Dépôt légal : 1er trimestre 1995
imprimé à l'Ecole Nationale Supérieure des Télécommunications - Paris
ISSN : 0751 - 1353


BIBLIOTHEQUE
UNIVERSITE de MARNE-LA-VALLEE

Contribution à l'élaboration d'un système de reconnaissance de parole continue à grand vocabulaire

Le premier chapitre présente quelques systèmes de reconnaissance vocale ; ils sont classés selon différents critères permettant de distinguer la dictée automatique de la commande vocale ou de la compréhension de la parole.
Cet état de l'art est complété de la description des recherches présentées et de l'exposé de la nature des problèmes abordés.
Le second chapitre présente un phonétiseur de texte avec variantes traitant des problèmes classiques de co-articulation (liaison, e caduc, assimilation, nasalisation, etc.) et un processus de sélection de la meilleure variante qui permet de calculer des modèles acoustiques adaptés aux locuteurs.
Le troisième chapitre présente une méthode de construction d'automates phonétiques contextuels. Elle est fondée sur une classification des réalisations acoustiques des phonèmes au moyen d'arbres de décision.
Le dernier chapitre expose une technique de prÉfiltrage acoustique du lexique fondée sur un arbre de décision.

TELECOM
PARIS

Ecole Nationale Supérieure des Télécommunications
46 rue Barrault 75634 Paris Cedex 13
Téléphone (1) 45 81 77 77 Télécopie (1) 45 89 79 06
Enseignement supérieur des Télécommunications
RCS Paris B 380 129 866