Reconnaissance automatique de la parole en présence de bruit pour une application grand public

Olivier Siohan, Kamel Smaïli, Jean-Francois Mari

Séminaire IRMA 95

Reconnaissance automatique de la parole en présence de bruit pour une application grand public

Olivier Siohan, Kamel Smaïli et Jean-François Mari
CRIN – CNRS & INRIA Lorraine
BP 239, 54506 Vandœuvre–lès–Nancy

19 septembre 1995

1 Pourquoi et comment utiliser la reconnaissance de la parole ?

La reconnaissance de la parole s'est développée dés 1948-1950 avec comme première motivation la recherche d'un mode de transmission de la voix à faible bande. Actuellement le but essentiel est la réalisation de systèmes de communication homme-machine sur support vocal permettant une amélioration de certains postes de travail et l'accès aux banques de données par l'intermédiaire du réseau téléphonique. L'usage de la voix comme moyen de communication présente de fait des avantages certains qui peuvent justifier les recherches pratiques dans ce domaine :

- Vitesse, simplicité et fiabilité d'une communication orale en comparaison avec d'autres modes : clavier, écriture, ...
- Simultanéité possible avec d'autres tâches permettant une communication en parallèle et préservant une liberté de mouvement du locuteur.
- Accès à distance (téléphone, radio-communication)

Le champ des applications potentielles est donc très vaste. Néanmoins, certains domaines nécessitent encore des travaux de recherche. La machine à dictée est sur le point d'apparaître et les systèmes de dialogue sont pour l'horizon 2000. Les applications actuelles relèvent de la reconnaissance de mots isolés ou enchaînés et concernent essentiellement les domaines militaires et industriels. Ces applications peuvent être classées en trois grandes catégories :

- Interaction avec des systèmes complexes. Le rôle du dialogue est prépondérant. L'exemple de la *souris vocale* de *Mosaic* qui permet une navigation plus aisée sur Internet en s'aidant de la voix pour désigner des points d'ancrage sur des serveurs HTML est une première ébauche d'un tel système.
- Saisie de donnée comme par exemple un système pour aider un garçon de café à prendre une commande.
- Commande de machines à l'aide d'ordres simples par mots isolés.

L'apparition récente de systèmes multi-locuteurs présentant de bonnes performances à travers le réseau téléphonique commuté ouvre de nouveaux champs d'applications : serveurs vocaux d'informations, réservations, autorisations bancaires.

La parole permet aussi à un handicapé moteur un contrôle efficace de son environnement. Les mêmes techniques de reconnaissance sont utilisées pour l'aide à la rééducation vocale des mal-entendants ou l'apprentissage des langues étrangères.

Cet article présente deux aspects d'un système de reconnaissance de la parole dans une application grand public : La robustesse du système dans le bruit et la modélisation du langage naturel oral. Ces deux aspects sont abordés dans l'équipe Rfia du Crin. Un autre aspect important du problème n'est pas abordé. C'est celui du dialogue oral Homme/Machine qui est abordé dans une autre équipe du Crin.

2 La robustesse des systèmes de reconnaissance dans le bruit

Les performances des systèmes actuels de reconnaissance automatique de la parole (RAP) sont satisfaisantes lorsque les systèmes sont évalués sous des conditions contrôlées de laboratoire, où les configurations d'apprentissage et de test sont « similaires ». Ainsi, sur une tâche de reconnaissance de parole continue, en mode dépendant du locuteur, pour un vocabulaire de 2 000 mots, [Gong, 1994] obtient moins de 1% d'erreur de reconnaissance. En reconnaissance de parole continue

indépendante du locuteur, pour un vocabulaire de 20 000 mots, sur la tâche du *Wall Street Journal*, environ 5% d'erreurs sont obtenues [Aubert *et al.*, 1994].

Cependant, ces systèmes sont généralement peu robustes, c'est-à-dire que des variations du signal entre les conditions de test et d'apprentissage peuvent provoquer une dégradation significative des taux de reconnaissance, même si ces variations semblent minimes à l'oreille. En effet, on se retrouve alors dans des configurations de test où l'on cherche à reconnaître des formes qui ne « correspondent » plus à celles apprises.

Les principales sources de variabilité du signal, qui rendent difficile la conception de systèmes de RAP robustes, peuvent être classées selon leur provenance, qu'il s'agisse de l'environnement acoustique, de l'équipement d'acquisition du signal, ou encore du locuteur. Le signal est alors perturbé par le bruit ambiant (stationnaire ou non), les distorsions (linéaires ou non) provenant du canal de communication, et les habitudes articulatoires du locuteur. Notons que les séparations entre les différentes classes ne sont pas toujours nettes, l'environnement pouvant par exemple influencer le mode de production de parole. Le tableau 1 résume ces différentes sources de variabilité.

Environnement	– Bruit corrélé à la parole : réverbération, réflexion – Bruit non corrélé à la parole : bruit additif (stationnaire, non stationnaire)
Locuteur	– Attributs du locuteur : sexe, âge, dialecte. – Mode d'expression : soufflement, bruit des lèvres, stress, effet Lombard, rythme d'élocution, puissance sonore, fréquence fondamentale, locuteur coopératif.
Conditions d'enregistrement	– Microphone – Distance au micro – Filtrage – Matériel de transmission : distorsion, bruit, écho – Matériel d'enregistrement

TAB. 1 - *Principales causes de variabilité du signal de parole (d'après [Furui, 1992])*

L'environnement perturbe le signal de parole sous la forme d'un bruit acoustique, que l'on suppose généralement additif. Cette hypothèse est souvent utilisée, à la fois pour sa simplicité, mais aussi car elle permet de couvrir un grand nombre de situations pratiques. Le signal enregistré est donc considéré comme la somme du signal de parole produit par le locuteur et du bruit ambiant. [Dautrich *et al.*, 1983a] sont parmi les premiers à constater la chute des performances d'un système de RAP entraîné dans des conditions calmes et testé dans le bruit : le taux d'erreur de reconnaissance du système, entraîné sur de la parole propre (SNR[1] > 40 dB) est multiplié par dix lors d'un test sur de la parole bruitée (SNR = 18 dB). Depuis, la littérature fournit pléthore d'observations analogues, et à titre d'exemple nous indiquons Figure 1, l'évolution des taux de reconnaissance en fonction du niveau de bruit présent lors du test, pour un système de reconnaissance de parole continue entraîné à partir de parole propre. Une telle évolution s'avère caractéristique du problème de la reconnaissance de la parole dans le bruit.

En plus des perturbations apportées par le bruit ambiant, le signal de parole est soumis à des distorsions spectrales provoquées par le canal d'acquisition. Dans le meilleur des cas, ces distorsions sont simplement linéaires, mais on rencontre également des distorsions non linéaires, beaucoup plus pénalisantes. Un changement de microphone ou de sa position, entre l'apprentissage et le test d'un système peut affecter de façon significative le spectre du signal, et dégrader ainsi les performances du système. Ainsi par exemple, [Acero et Stern, 1990] constatent que le taux de reconnaissance d'un système de RAP grand vocabulaire, valant initialement 85% s'effondre à 19% lors d'un changement de microphone entre l'apprentissage et le test.

Les variations de nature intra-locuteur sont en général moins pénalisantes que celles provoquées par le bruit ambiant ou le canal d'enregistrement. Cependant, en présence d'un bruit acoustique élevé, le locuteur modifie de façon réflexe son mode d'élocution (effet Lombard [Lombard, 1911]) pour que ses propos restent intelligibles. Cet effet provoque des distorsions importantes du signal de parole, et un système entraîné avec de la parole propre verra ses performances chuter en reconnaissance de parole Lombard [Junqua, 1989].

[1] *Signal-to-Noise Ratio*, rapport signal-à-bruit

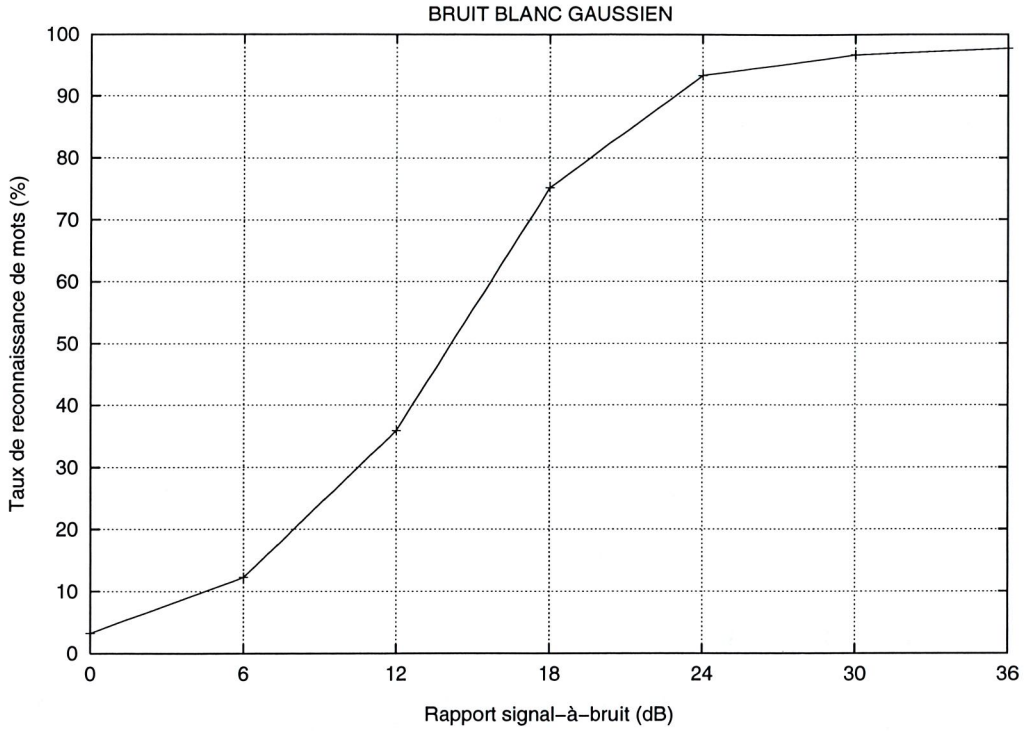

FIG. 1 - *Évolution du taux de reconnaissance d'un système de RAP entraîné en milieu calme (SNR >40 dB), en fonction du rapport signal-à-bruit lors du test [Siohan* et al., *1995].*

Si autrefois les concepteurs de systèmes de RAP travaillaient sous l'hypothèse d'une ambiance acoustique idéale, les problèmes de robustesse doivent aujourd'hui être considérés. En effet, l'utilisateur d'une interface orale homme-machine sera naturellement réticent s'il doit parler d'une façon contrainte, si le système fonctionne mal le jour où il est enrhumé ou fatigué, ou bien encore si les performances s'effondrent en présence d'un bruit de fond inhabituel. Or, les systèmes de communication orale sont généralement destinés à des utilisations en environnements bruités : usines, lieux publics, habitacle de voiture, d'avion, parole téléphonique, etc. Plus les systèmes de RAP seront robustes, plus le nombre de leurs applications potentielles augmentera, et l'absence de robustesse au bruit apparaît comme le principal obstacle au développement commercial de telles applications. L'amélioration de la robustesse des systèmes est donc un thème majeur de recherche, faisant appel à des connaissances pluridisciplinaires (traitement du signal, reconnaissance des formes, intelligence artificielle), essentiel pour permettre le développement d'applications en environnements réels. Dans cet article, nous souhaitons mettons en exergue les problèmes spécifiques à la reconnaissance de parole dans le bruit, ainsi que les principales directions de recherches qui peuvent être développées pour améliorer la robustesse des systèmes de RAP.

3 Reconnaissance automatique de la parole dans le bruit

3.1 Introduction

Un système de RAP est utilisé de façon optimale lorsque ses conditions de test et d'apprentissage sont semblables. On peut alors penser que pour reconnaître de la parole dans un environnement bruité, il suffit d'entraîner le système dans cet environnement. Cette démarche, qui conduit à l'obtention de taux de reconnaissance satisfaisants, n'est cependant jamais applicable en pratique. En effet, la plupart des corpus de parole utilisés pour construire des systèmes de reconnaissance sont constitués de parole de bonne qualité, sans bruit (parole « propre »), et il est très coûteux d'enregistrer un corpus d'apprentissage dans le bruit. De plus, un tel corpus s'avérerait insuffisant car il est difficile de prévoir lors de l'apprentissage quelles seront les conditions de bruit réellement présentes lors de l'utilisation du système. Dans un tel cas, il devient également délicat d'autoriser une variation du SNR ou du type de bruit lors du test. Le problème à aborder est donc le suivant : étant donné un système de reconnaissance de parole continue entraîné à partir de parole propre, quelles méthodes et techniques peut-on mettre en œuvre pour améliorer la robustesse au bruit du système, c'est-à-dire pour que le système reconnaisse correctement de la parole prononcée en environnement réel, *a priori* inconnu ?

Comme nous l'avons noté, l'insuffisance de robustesse des systèmes de RAP provient des variations statistiquement

significatives entre le signal de parole utilisé lors de l'apprentissage, et le signal de parole de test. Par conséquent, l'ensemble des approches pour la reconnaissance de la parole dans des environnements difficiles se focalise sur la réduction des différences entre les conditions d'apprentissage et de test. Cette réduction des différences entre conditions de test et d'apprentissage peut être effectuée selon trois grandes familles d'approches, qui, bien que se différenciant par leurs principes de base, conduisent au développement d'idées assez semblables.

On peut considérer d'une part que la configuration du système de reconnaissance est figée, et que par conséquent, la réduction des différentes distorsions s'effectue en traitant le signal de test. Ces traitements peuvent avoir pour objectif de filtrer le bruit, de compenser les variations de microphone entre test et apprentissage, ou encore de réduire les distorsions provoquées par l'effet Lombard[2]. Ces méthodes sont brièvement présentées au paragraphe 3.2, où l'accent est mis sur la suppression d'un bruit de fond additif.

D'autre part, il est possible de développer une approche duale, qui consiste à autoriser la présence d'un bruit ou d'une distorsion lors du test, en modifiant la configuration du système de reconnaissance. Le système évolue donc selon ses conditions d'utilisation. Le paragraphe 3.3 résume cette famille de travaux.

Enfin, on peut reporter l'amélioration de la robustesse sur la recherche d'un paramétrage du signal ainsi que de mesures de distances associées robustes. Ici encore, la robustesse peut se focaliser sur le bruit, les variations du canal d'enregistrement ou l'effet Lombard. Le système de reconnaissance est alors utilisé quelque soit l'environnement, sans modification de sa configuration. Ces approches sont synthétisées paragraphe 3.4.

3.2 Transformation de la parole

Étant donné un système entraîné à partir d'un corpus de parole propre, l'objectif du filtrage de bruit est de prétraiter le signal bruité de test, afin de pouvoir l'utiliser comme entrée du système de reconnaissance.

La difficulté du filtrage de bruit sur un signal de parole perturbé par un bruit additif large bande réside dans deux points principaux. Tout d'abord, le bruit recouvre le signal à la fois dans le domaine temporel et fréquentiel, et ses caractéristiques sont *a priori* inconnues dans chacun des deux domaines. D'autre part, le signal de parole est non stationnaire ; les distorsions provoquées par le bruit varient au cours du temps et selon les fréquences. Un niveau de bruit constant large bande perturbe donc plus les zones de faible énergie du signal (sons non voisés, transitions) que les zones de fortes énergies (sons voisés). Ainsi, les sons les plus perturbés sont ceux ayant le moins de redondance à exploiter pour définir un filtrage.

Les objectifs du filtrage de bruit sont multiples : d'une part l'amélioration des aspect perceptifs du signal (qualité[3] et intelligibilité[4]), d'autre part l'augmentation de la robustesse au bruit des systèmes de RAP. Qualité et intelligibilité étant évaluées par des tests d'écoute et ne s'exprimant donc pas sous forme mathématique, il est difficile d'évaluer et de concevoir des méthodes permettant d'améliorer ces deux propriétés. De plus, l'amélioration de la qualité se traduit souvent par une dégradation de l'intelligibilité, et les méthodes destinées à l'amélioration de la qualité et l'intelligibilité n'améliorent pas systématiquement la robustesse des systèmes de RAP. Dans les paragraphes qui suivent, nous présentons brièvement les principales méthodes de traitement du signal de test.

3.2.1 Soustraction spectrale

La soustraction spectrale fait partie des méthodes de filtrage du signal de type *Overlap and Add*. La soustraction spectrale consiste à retrancher une estimation de la densité spectrale de puissance du bruit de la densité spectrale de puissance du signal bruité [Weiss *et al.*, 1974; Lim, 1978; Boll, 1979], pour obtenir une estimation du signal de parole débruité. La soustraction spectrale provoque l'apparition d'un artéfact, appelé bruit « musical », qui en dépit d'une énergie très faible par rapport à l'énergie initiale du bruit, dégrade fortement l'intelligibilité du signal, et les performances des systèmes de RAP. Les améliorations de la soustraction spectrale visent essentiellement à limiter cet effet [Weiss *et al.*, 1974; Lim, 1978; Berouti *et al.*, 1979; Boll, 1979]. Leurs formes les plus abouties correspondent à la soustraction spectrale non linéaire de [Lockwood et Boudy, 1991] et à la minimisation des distorsions dans le domaine du logarithme du spectre.

3.2.2 Annulation adaptative de bruit

Contrairement à la soustraction spectrale qui permet de filtrer le bruit d'un signal enregistré avec un seul microphone, l'annulation adaptative de bruit [Widrow *et al.*, 1975] nécessite l'utilisation de deux microphones pour supprimer

2. En présence d'un bruit de fond important, un locuteur modifie de façon réflexe son effort articulatoire pour que ses propos restent intelligibles. Ce réflexe, appelé effet Lombard, provoque des modifications spectrales et temporelles importantes du signal de parole, par rapport à une élocution en environnement calme

3. La qualité est une mesure subjective, caractérisant l'aspect agréable de l'écoute d'un signal.

4. L'intelligibilité est une mesure objective de la quantité d'informations que peut extraire un auditeur lors de l'écoute d'un signal, indépendamment de sa qualité. Un signal peut donc être de mauvaise qualité, tout en ayant une bonne intelligibilité, ou l'inverse.

le bruit [Widrow *et al.*, 1975]. Son grand intérêt est qu'elle ne nécessite pas de connaître *a priori* les statistiques du bruit ou de la parole. De plus, et contrairement à la soustraction spectrale, cette méthode est applicable à des bruits non stationnaires. Malheureusement, les performances de l'ANC sont fortement affectées par les choix d'implantation, en particulier les positions respectives des microphones. Bien souvent, les performances obtenues restent médiocres : [Dal Degan et Prati, 1988] constatent que l'ANC échoue dans un environnement de voiture, à cause de la faible cohérence du bruit entre l'entrée primaire et de référence. [Lecomte *et al.*, 1989] rapportent des observations analogues et concluent que l'ANC n'est pas applicable dans un tel environnement. Dans les environnements fortement bruités où l'estimation *a priori* des caractéristiques de la parole et du bruit sont peu précises, l'ANC fournit cependant une suppression effective du bruit [Boll et Pulsipher, 1980]. Une tendance actuelle semble être l'introduction de contraintes auditives dans le processus de filtrage, ce qui permet d'améliorer la qualité des traitements [Nandkumar et Hansen, 1994].

3.2.3 Transformation d'espace

L'objectif de la transformation d'espace est de définir une transformation permettant de recouvrer la parole propre dans le domaine temporel ou dans un espace de paramètres, à partir de la parole bruitée. Contrairement aux approches fondées sur le filtrage de bruit, la transformation est déterminée sans présumer de la nature de la combinaison entre parole et bruit, c.-à-d. sans connaître la nature des différences entre l'espace de parole de référence et l'espace de parole de test. Ces méthodes permettent donc de prendre en compte des différences entre conditions de test et d'apprentissage, que ces différences proviennent d'un changement de locuteur, de bruit d'environnement ou bien de microphone. La transformation, c.-à-d. la correspondance entre les espaces de référence et de test, est généralement établie à partir de l'observation d'une même séquence de parole dans les deux environnements. L'ensemble des données utilisées pour déterminer la transformation est appelé corpus d'adaptation. Bien évidemment, on recherche en général à utiliser un corpus d'adaptation le plus réduit possible. Les méthodes les plus simples de transformation d'espace sont celles qui établissent une correspondance une à une entre les vecteurs du répertoire de prototypes (*codebook*[5]) de l'espace de référence et ceux de l'espace de test [Shikano *et al.*, 1986; Nakamura et Shikano, 1989]. Dans une seconde famille de méthodes, des transformations analytiques explicites peuvent être définies selon un critère objectif pour projeter un espace sur un autre [Gu et Mason, 1989; Mokbel *et al.*, 1992]. Enfin, les transformations d'espaces peuvent être mises en œuvre par des réseaux de neurones [Tamura, 1989; Ohkura et Sugiyama, 1991; Trompf, 1992].

3.2.4 Exploitation de la structure harmonique du signal de parole

Certaines méthodes de filtrage exploitent la structure harmonique du signal de parole, et en particulier la périodicité des sons voisés [Malah et Cox, 1982; Cox et Malah, 1981; Ramalho et Mammone, 1994; Erell et Weintraub, 1994]. De telles méthodes sont donc limitées et ont généralement un comportement médiocre sur les zones non voisées du signal. De plus, l'estimation et le suivi de la fréquence fondamentale du signal devient délicate en présence de bruit. En présence d'un bruit large bande, la distorsion affectant les zones non voisées de faible énergie sera forte par rapport à celle affectant les zones voisées, et ces méthodes n'améliorent généralement ni la qualité ni l'intelligibilité du signal. Par contre, ces techniques sont plus utiles pour filtrer un bruit additif à bande étroite, ou bien des bruits de parole [O'Shaughnessy, 1989].

3.2.5 Masquage de bruit

En présence d'un bruit de fond large bande, les régions du spectre de faible énergie sont plus affectées que les zones de forte énergie. Il est donc difficile de définir une mesure de distance entre spectres, les zones les plus perturbées étant celles où la mesure de distance est la moins fiable. [Klatt, 1976] introduit alors le masquage de bruit en sortie d'un banc de filtres. Avant masquage, les zones de faible énergie porteuses de peu d'informations intervenaient dans les calculs des distances entre spectres ; après masquage, les zones masquées aux mêmes fréquences sur deux spectres différents ont la même valeur et n'interviennent donc plus dans le calcul de la distance entre ces spectres. On supprime ainsi l'influence des régions du spectre portant peu d'informations. Le masquage de Klatt est reformulé dans le cadre d'un système de RAP à base de HMM dans les travaux de [Varga *et al.*, 1988; Varga et Ponting, 1989], et sous un cadre probabiliste dans [Nadas *et al.*, 1989; Van Compernolle, 1989].

3.2.6 Débruitage à base de modèles

Le signal de parole peut être représenté à court terme par un modèle auto-régressif (AR). À partir du signal de parole bruitée, [Lim et Oppenheim, 1978; Lim et Oppenheim, 1979; Lim et Oppenheim, 1983] cherchent à estimer les paramètres du modèle AR de parole propre ainsi que le signal de parole propre, en utilisant une méthode d'estimation du maximum *a*

5. Un *codebook* est un ensemble fini de vecteurs, ou prototypes, représentatif d'un espace vectoriel donné.

posteriori (MAP). Ces approches de filtrage du signal à base de modèles connaissent de nombreux développement, comme l'introduction de contraintes spectrales inter et intra-trames [Hansen et Clements, 1987; Hansen et Clements, 1991]qui imposent la stabilité du modèle AR, la mise en œuvre de traitement spécifiques à des classes phonétiques grossières [Arslan et Hansen, 1994]. Dans [Ephraim, 1992], le filtrage à base de modèles est présenté sous un cadre unifié pour l'amélioration de la qualité, la reconnaissance et le codage d'un signal de parole bruitée.

3.2.7 Compensation de l'effet Lombard

Différentes techniques *ad-hoc* ont été proposées pour réduire les distorsions provoquées par l'effet Lombard. [Takizawa et Hamada, 1990] proposent de compenser un vecteur de cepstre par un biais rendant compte du déplacement des formants de la parole Lombard, par rapport à la parole normale. [Hansen et Clements, 1989] développent des algorithmes pour compenser de la parole prononcée sous diverses conditions de stress (parole Lombard, parole criée, prononciation lente, rapide, etc.) en utilisant des tables de compensations spécifiques à chaque type de parole, et déterminées après analyse des corpus de parole entre conditions neutres et de stress. Dans [Chen, 1988], le cepstre du signal de test est compensé par une composante additive qui décroît de façon exponentielle en fonction de l'indice des composantes. La compensation de Chen est étendue par [Hansen et Bria, 1990], qui dissocient la compensation des zones voisées de celle des zones non voisées.

3.2.8 Conclusion

Les approches de prétraitement du signal, qui permettent de minimiser les variabilités provoquées par le bruit, peuvent être à l'heure actuelle considérées comme viables pour la reconnaissance de la parole dans le bruit, ce qui n'était pas le cas il y a quelques années (p.ex. très mauvais comportement de la soustraction spectrale classique [Van Compernolle, 1989] ou du filtrage de Kalman [Mokbel, 1992]).

Les progrès proviennent d'une part de la mise en œuvre de traitements spécifiques à des classes de sons, de la minimisation des distorsions provoquées par les techniques classiques de filtrage, et de la prise en compte du rapport signal-à-bruit instantané, éléments qui se retrouvent dans de nombreuses approches de filtrage à base de modèles, de transformation d'espace, ou encore de soustraction spectrale non linéaire. De tels traitements sont justifiés car les différentes zones d'un signal de parole ne sont pas modifiées de façon consistante par un bruit stationnaire, et peuvent alors être traitées spécifiquement en fonction de leurs caractéristiques.

La qualité des traitements est également améliorée par l'exploitation de modèles *a priori* du signal de parole et des corrélations spectrales du signal, qui permettent de prendre en considération les redondances et spécificités du signal de parole.

L'introduction de connaissances et contraintes psychoacoustiques dans les processus de filtrage contribue aussi à l'amélioration des performances, particulièrement en présence d'un niveau de bruit important.

Enfin, les approches de filtrage fondées sur l'utilisation de plusieurs microphones, non présentées ici semblent très prometteuses, en particulier pour limiter les effets des bruits non stationnaires et traiter les problèmes d'échos, même si de tels traitements sont encore peu répandus en reconnaissance de parole.

3.3 Transformation des systèmes de reconnaissance

Dans ces familles de méthodes, le système de reconnaissance de parole est modifié afin de tenir compte de la présence d'un bruit lors de la reconnaissance. La modification peut s'effectuer d'une part au niveau du processus de décodage, pour autoriser la présence d'un signal concurrent perturbant la parole ; d'autre part, il est également possible d'introduire une étape de filtrage dans le processus de décodage, un filtre pouvant être associé à chaque modèle, ou à chaque état d'un modèle stochastique. Des modèles spécifiques aux nouvelles conditions de test peuvent aussi être déterminés à partir des modèles initiaux de parole propre, en utilisant ou non des connaissances sur la nature de la perturbation . L'utilisation de critères discriminants d'apprentissage permet de lutter contre la source de variabilité introduite par le bruit. Enfin, effectuer un apprentissage dans différentes conditions prédéfinies de bruit reste une solution efficace, bien que difficilement réalisable en pratique.

3.3.1 Composition/décomposition de modèles

L'observation d'un signal de parole bruitée correspond à l'observation simultanée de deux signaux (parole et bruit) se combinant selon une certaine relation. Les signaux de parole et de bruit peuvent tous deux être représentés par des modèles stochastiques. À chaque instant, le signal de parole bruitée correspond donc à la combinaison d'une observation de parole associée à un modèle de parole propre, avec une observation de bruit associée à un modèle de bruit. Il est alors possible de déterminer la vraisemblance d'une observation de parole bruitée, et par conséquent de cheminer simultanément

dans les modèles de parole propre et de bruit [Varga et Moore, 1990; Varga et Moore, 1991; Kadirkamanathan, 1992]. Des idées proches de celles de Varga et Moore sont développées dans [Young, 1992], avec la combinaison parallèle de modèles (PMC[6]), qui consiste à construire un modèle de parole bruitée à partir d'un modèle de parole propre et d'un modèle de bruit. De tels travaux ont été développés dans le cadre de la combinaison de modèles de Markov cachés [Gales et Young, 1994; Nolazco Flores et Young, 1994]. Toutes ont en commun le problème de la détermination de la fonction de probabilité de la parole bruitée, connaissant la fonction de probabilité du signal propre et du bruit, ce qui nécessite d'introduire des approximations pour obtenir sous forme close l'expression des fonctions de probabilités [Nadas *et al.*, 1989; Varga et Moore, 1990; Gales et Young, 1992; Kadirkamanathan, 1992; Rose *et al.*, 1994].

3.3.2 Filtrage par état

Le filtrage de Wiener s'avère efficace lorsqu'on l'applique sur un signal stationnaire corrompu par un bruit additif, mais reste limité par la non-stationnarité sous-jacente du signal de parole. Or, les HMMs découpent automatiquement la parole en segments quasi-stationnaires, correspondant aux états des modèles [Beattie et Young, 1991]. Il est alors possible d'associer à chaque état d'un HMM un filtre de Wiener, et d'appliquer un filtrage lors de la reconnaissance de la parole bruitée. Différentes variantes ont été proposées par [Beattie et Young, 1992; Vaseghi et Milner, 1993; Pai et Wang, 1992].

3.3.3 Adaptation des modèles

Un grand nombre de méthodes ont été développées pour permettre de reconnaître de la parole bruitée à partir d'un HMM initialement entraîné sur de la parole propre. Bien souvent, ces approches sont dérivées des travaux sur l'adaptation au locuteur des systèmes de reconnaissance, et consistent à transformer une fois pour toute les modèles de parole propre, la ou les transformations étant déterminées à partir d'un corpus d'adaptation. Une fois les modèles adaptés, ces méthodes ont l'avantage de ne nécessiter aucun calcul supplémentaire. Les méthodes d'adaptation de modèles peuvent être considérées comme des problèmes de réentraînement de modèles à partir d'un faible volume de données [Stern et Lasry, 1987; Gauvain et Lee, 1994]. Beaucoup de méthodes sont dérivées des travaux de transformation d'espace [Mokbel *et al.*, 1992; Ohkura *et al.*, 1992]. Les méthodes mises en œuvre sont souvent applicables à la compensation des variations de nature diverses (locuteur, ligne de transmission, bruit, effet Lombard) [Takahashi et Sagayama, 1994; Suzuki *et al.*, 1994; Leggetter et Woodland, 1994; Siohan *et al.*, 1995].

3.3.4 Apprentissage discriminant

Bien souvent, l'apprentissage de modèles stochastiques est basé sur le critère du maximum de la vraisemblance (MLE[7]), et n'optimise donc pas la discrimination entre les classes à identifier. Ce critère d'estimation garantit cependant l'optimalité de l'apprentissage si les modèles correspondent aux données. En présence de bruit, la discrimination entre les différentes unités de parole devient délicate et l'utilisation d'un critère d'apprentissage discriminant semble donc judicieuse. [Mizuta et Nakajima, 1992] effectuent un apprentissage discriminant en utilisant de la parole bruitée, pour corriger des HMMs initialement entraînés par MLE sur de la parole propre. [Frangoulis et Gaganelis, 1992] adaptent les vecteurs moyennes d'un HMM continu entraîné à partir de parole propre, en utilisant un corpus d'adaptation de parole bruitée. Le critère d'apprentissage du minimum d'erreur de classification (ME[8]) [Chou *et al.*, 1992] semble plus efficace que MLE pour la reconnaissance de mots isolés dans le bruit [Ohkura *et al.*, 1993].

3.3.5 Apprentissage multiréférences

Une stratégie possible pour la reconnaissance de la parole dans le bruit consiste à entraîner les systèmes dans le bruit [Dautrich *et al.*, 1983a; Dautrich *et al.*, 1983b]. Cette approche peut être considérée comme une forme de compensation de modèle, qui supprime totalement les différences entre conditions de test et d'apprentissage. [Morii *et al.*, 1990] ajoutent une estimation du bruit aux vecteurs de références d'un système de reconnaissance à base de programmation dynamique. [Das *et al.*, 1993; Das *et al.*, 1994] superposent également un bruit de fond aux références. Il est cependant difficile de prévoir quelles seront les conditions de bruit lors du test, et le système entraîné dans le bruit devient inefficace pour reconnaître la parole propre. De plus, un système entraîné dans le bruit est très sensible aux variations du niveau et du type de bruit [Kitamura *et al.*, 1992]. L'apprentissage multi-style utilisant différents types et niveaux de bruits est possible, mais provoque une diminution sensible des taux de reconnaissance en parole propre [Kitamura *et al.*, 1992].Afin de prendre en compte les variations provenant du mode d'expression du locuteur, il est possible d'entraîner un système dans différentes conditions. [Lippmann *et al.*, 1987]

6. *Parallel Model Combination*
7. *Maximum Likelihood Estimation*
8. *Minimum Error*

effectuent un apprentissage multi-style qui consiste à entraîner un système avec de la parole propre, Lombard, criée, etc. Bien qu'efficace, cette méthode nécessite de disposer d'un corpus d'apprentissage enregistré sous différentes conditions de stress, dont la collecte est délicate et coûteuse

3.3.6 Conclusion

Comme dans les approches de filtrage de bruit, les méthodes de transformation des systèmes de reconnaissance tirent avantage de la mise en œuvre de compensations associées aux différents sons, c.-à-d. spécifiques aux modèles (combinaisons de modèles, filtrages par états), qui utilisent une connaissance *a priori* du signal de parole propre fournie par les modèles utilisés pour la reconnaissance. La détermination précise des statistiques du bruit est généralement nécessaire, comme dans certaines approches de filtrage du signal, et conditionne fortement la qualité des résultats obtenus. Les méthodes de combinaisons de modèles permettent de prendre en compte des événements concurrents (parole et bruit), en particulier des bruits non stationnaires, tout en ne nécessitant que l'utilisation d'un seul microphone.

Un avantage de certaines approches de transformations de modèles (p. ex. adaptation des paramètres des modèles) réside dans leur caractère généraliste, qui permet d'éviter de se focaliser sur une perturbation particulière comme le bruit additif, pour être potentiellement exploitable pour des applications d'adaptation au locuteur ou à la ligne de transmission. Malheureusement, dans de nombreuses approches d'adaptation de modèles, un corpus d'adaptation de plusieurs minutes de parole est souvent nécessaire afin de déterminer les compensations à appliquer (p.ex. réestimation Bayésienne).

Enfin, il faut noter que certaines méthodes d'adaptation garantissent que leur mise en œuvre ne va pas perturber les performances du système en présence d'un bruit d'environnement faible, ce qui constitue une propriété indispensable pour un système de reconnaissance destiné à fonctionner dans des conditions d'environnement variables.

3.4 Paramétrages et mesures de similarité robustes

La minimisation des différences entre l'environnement de test et de référence d'un système de RAP peut s'effectuer en recherchant un paramétrage du signal de parole et une mesure de similarité associée robustes aux variations des conditions d'environnement. On s'intéresse donc ici plus aux effets du bruit et à la façon de définir un paramétrage insensible à ces effets, qu'à la façon de supprimer ou d'atténuer ce bruit. La représentation du signal de parole étant supposée indépendante du bruit, un système entraîné sur de la parole propre peut alors être utilisé dans un environnement calme ou bruyant, sans modification de sa configuration.

L'avantage des méthodes mises en œuvre est qu'elles ne nécessitent en général que peu de connaissances sur le bruit perturbateur. En particulier, il est inutile de disposer des statistiques du bruit. Cet avantage peut s'avérer être un inconvénient, dans la mesure où on ne tire aucun parti des caractéristiques spécifiques à un bruit.

3.4.1 Représentations acoustiques et distances robustes

L'idée principale pour définir un paramétrage du signal et une mesure associée robustes au bruit consiste à privilégier de façon automatique les zones du spectre les moins perturbées par le bruit, au détriment des zones fortement affectées. La compensation des effets du bruit s'effectue ainsi de façon implicite, par la définition d'une mesure de distance robuste. La distorsion d'Itakura-Saito (IS) [Itakura et Saito, 1968; Itakura, 1975] et une de ses variantes, le rapport de vraisemblance (LR[9]), sont les éléments centraux des mesures de similarité entre signaux de parole, et se basent sur la différence entre spectres dans le domaine logarithmique. Des pondérations spectrales (WLR[10]) ont ensuite été introduites pour améliorer les performances des systèmes de RAP [Shikano et Sugiyama, 1982]. Ces pondérations ont enfin été étendues pour privilégier l'influence des pointes spectrales, plus robustes au bruit que les vallées spectrales. Parmi les distances et paramétrages les plus populaires dans le bruit, on citera la distance RPS[11] [Schroeder, 1981] qui consiste à pondérer les coefficients cepstraux par leurs indices, la représentation SMC[12] qui exploite la cohérence entre segments adjacents du signal de parole, l'analyse homomorphique en racine [Alexandre *et al.*, 1993] qui permet de réduire la sensibilité aux zones du spectre de faible énergie, la mesure de projection cepstrale [Mansour et Juang, 1989] qui exploite le fait que le bruit affecte moins l'angle entre les vecteurs que leurs normes.

9. *Likelihood Ratio*
10. *Weighted Likelihood Ratio*
11. *Root Power Sums*
12. *Short-time Modified Coherence*

3.4.2 Mesures de distorsion et paramétrages discriminants

L'augmentation de la discrimination entre vecteurs de paramètres de parole peut s'effectuer soit en définissant des paramètres discriminants, soit en utilisant des mesures de distorsion discriminantes. L'analyse discriminante linéaire (ADL) peut être utilisée pour projeter un espace de paramètres de parole sur un espace plus discriminant éventuellement de dimension plus réduite, où un critère de séparation de classes est maximisé [Haeb-Umbach *et al.*, 1993]. L'ADL peut être utilisée en reconnaissance de parole dans le bruit. [Hunt et Lefèbvre, 1988] appliquent une analyse linéaire discriminante pour combiner les sorties d'un modèle auditif. Leur travail est étendu dans [Hunt et Lefèbvre, 1989] et conduit à la définition de l'analyse IMELDA [13]. [Trompf *et al.*, 1993; Sorensen et Hartmann, 1993; Siohan, 1995] utilisent également l'ADL sur des applications de RAP dans le bruit .L'augmentation de la discrimination peut être introduite au niveau de la mesure de distance. Étant donné que les zones du signal de forte énergie sont les moins perturbées par le bruit, [Kobatake et Matsunoo, 1994] définissent une distance dans le cadre d'un système de reconnaissance à base d'alignement dynamique temporel, qui privilégie les chemins où le rapport signal-à-bruit instantané est important. [Anglade *et al.*, 1993] effectuent une reconnaissance de mots difficiles dans le bruit, et focalisent la mesure de distance sur une zone discriminante des mots.

3.4.3 Paramétrage à base de modèles auditifs

Le système auditif humain est particulièrement résistant aux bruits perturbant le signal de parole. Aussi, dans beaucoup de travaux, des connaissances sur les mécanismes de l'audition sont utilisées pour analyser le signal. L'utilisation d'une échelle psychoacoustique des fréquences comme l'échelle Mel [Zwicker et Feldtkeller, 1981] est courante en reconnaissance de parole, propre ou bruitée. [Hermansky *et al.*, 1985] développent l'analyse par prédiction linéaire perceptive (PLP [14]), qui consiste à effectuer un filtrage en bandes critiques du signal de parole, suivi d'une pré-accentuation à partir de la courbe d'*isosonie* [15] et d'une compression spectrale pour passer de l'intensité à la *sonie* [16]. Le spectre ainsi obtenu est finalement représenté par un modèle tout-pôle, et des coefficients cepstraux peuvent être calculés. Les principaux modèles auditifs utilisés en RAP sont ceux dérivés des modèles de Lyon, Ghitza et Seneff. Ces modèles se caractérisent par leur grande résolution spectrale, mais les analyses temps-fréquence mises en œuvre conduisent à des charges de calcul importantes. Des phénomènes psychoacoustiques plus complexes sont également pris en compte pour améliorer la robustesse des systèmes de RAP. [Cheng et O'Shaughnessy, 1991] tiennent compte du phénomène d'*inhibition latérale* [17] [Shamma, 1985] ; [Cohen, 1985] utilise le modèle d'*adaptation à court terme* [18] de [Schroeder et Hall, 1974]; dans [Aikawa et Saito, 1994], le masquage proactif est incorporé dans un paramétrage cepstral et s'avère efficace à la fois en reconnaissance de parole propre et bruitée.

3.4.4 Suppression des variations lentes

La plupart des bruits additifs et des distorsions liées au canal d'enregistrement varient lentement par rapport aux variations du signal de parole. Il est donc possible d'utiliser cette propriété pour définir différents paramétrages insensibles aux variations lentes du signal. Cette opération s'apparente à un filtrage des paramètres, et peut être mise en œuvre dans différents espaces de représentation du signal de parole. [Hermansky *et al.*, 1991] proposent l'analyse RASTA [19], qui consiste à supprimer les variations lentes du signal, par filtrage de l'enveloppe log-spectrale. Cette méthode peut être incorporée à l'analyse PLP [Hermansky *et al.*, 1985; Hermansky, 1990]. Dans ses premières formulations, RASTA permettait uniquement de compenser des variations de microphone, mais ne luttait pas contre les perturbations provoquées par un bruit additif. Une version plus récente, J-RASTA, améliore également la robustesse au bruit [Hermansky *et al.*, 1993; Koehler *et al.*, 1994]. [Hirsch *et al.*, 1991] proposent différents filtrages passe-haut d'enveloppes spectrales dans différentes bandes de fréquences, ce qui correspond à une approche semblable à RASTA. La soustraction cepstrale, ou CMN [20] [Atal, 1974], a pour objectif de soustraire à chaque vecteur de cepstre une estimation du cepstre moyen à long terme (calculé sur toute une phrase par exemple). Ce traitement effectue une normalisation de la distribution spectrale de la phrase, ce qui permet de minimiser les variabilités intra-locuteurs [Furui, 1981]. Un tel traitement peut être interprété comme un filtrage inverse qui normalise la réponse en fréquence du canal d'enregistrement, et est donc largement utilisé pour compenser les

13. *Integrated Mel-scale Linear Discriminant Analysis*
14. *Perceptually based Linear Prediction*
15. Les courbes d'*isosonie* relient les niveaux de pression acoustique et fréquence des sons purs qui donnent à l'oreille humaine une égale sensation d'intensité.
16. On appelle *sonie* l'intensité subjective des sons.
17. Le phénomène d'*inhibition latérale* caractérise le fait que la réponse d'une fibre nerveuse à une excitation peut être affectée par la réponse des fibres nerveuses adjacentes.
18. L'*adaptation à court terme* caractérise les variations des réponses à un son des fibres nerveuses, en fonction des caractéristiques du son précédent, et permet ainsi une prise en compte du contexte.
19. *RelAtive SpecTrAl*
20. *Cepstral Mean Normalisation*

variations de microphone entre les conditions de test et d'apprentissage des systèmes de RAP. Sur des tâches de compensation des variations de la ligne de transmission du signal, la supériorité de RASTA sur CMN n'est à l'heure actuelle pas clairement démontrée. Enfin, les paramètres dynamiques (comme Δcepstre) et d'accélération (comme $\Delta\Delta$cepstre) s'avèrent efficaces pour la RAP dans le bruit [Hanson et Applebaum, 1990; Applebaum et Hanson, 1991].

4 Les modèles de langage en reconnaissance de la parole

L'acquisition d'une langue ne se fait pas simplement en apprenant des mots, mais aussi en s'entraînant à construire et à comprendre des phrases. La constitution de ces phrases n'est pas une simple combinaison de mots pris dans n'importe quel ordre, mais un mécanisme de construction très précis obéissant à des règles et ayant des exceptions. Lorsque le langage naturel (écrit ou oral) est utilisé comme média dans un système de communication homme/machine, il faut disposer d'un modèle permettant de ne pas laisser l'utilisateur saisir ou dicter n'importe quoi. Ce modèle est communément appelé modèle de langage. En reconnaissance de la parole, le modèle de langage a un rôle déterminant. Il permet, en effet, de participer à tout instant au choix du prochain mot à reconnaître et à éliminer des hypothèses non valides. Ce qui rend plus difficile la tâche de ce modèle est la multiplication des hypothèses aux niveaux inférieurs du processus de décodage du signal de parole. La modélisation du langage en reconnaissance de la parole peut se faire sous deux aspects. Le premier concerne les applications de reconnaissance ayant un domaine d'application très restreint. Dans ce cas, tous les phénomènes linguistiques de l'application sont modélisables et peuvent être engendrés par une grammaire simple. En revanche, lorsque l'on s'intéresse aux applications pour lesquelles aucune restriction linguistique n'est imposée, le problème devient beaucoup plus ardu. Les linguistes travaillent depuis longtemps sur le développement de formalismes permettant de modéliser tel ou tel aspect de la langue, mais malheureusement, nous ne disposons pas aujourd'hui de grammaires formelles utilisables dans un système de reconnaissance de la parole continue.

La liste des grammaires et des théories sous-jacentes est longue, mais aucune n'est vraiment satisfaisante principalement parce que le langage oral ou spontané possède des particularités. Cela ne remet pas en cause la validité de la théorie chomskyenne qui sert de base à la plupart des constructions syntaxiques.

Un système de reconnaissance de la parole nécessite un modèle de langage opérationnel lui permettant de filtrer les hypothèses fournies par les niveaux inférieurs de la chaîne de communication. Il faut peut être, ne pas chercher à trouver une grammaire modélisant l'ensemble des phénomènes de la langue, mais plutôt trouver un ensemble de règles grammaticales de taille raisonnable permettant l'automatisation du traitement langagier et dans le cas de la parole, d'effectuer une assez bonne prédiction. Gross souligne que les exemples de modèles linguistiques permettant d'effectuer des prédictions sont extrêmement rares et ceux qui sont censés l'être ne sont que des modèles locaux [Gross, 1975], c'est à dire ne recouvrant qu'un fragment de la syntaxe.

Nous citons dans la suite deux expériences de grammaires locales pouvant être utilisés dans des systèmes de reconnaissance de la parole.

4.1 Implantation informatique du français fondamental

Comme nous l'avons vu, il est difficile de disposer d'une grammaire modélisant tout le langage. Ceci admis, il est plus intéressant à court terme de tenter de construire un modèle qui traite l'ensemble des phénomènes linguistiques les plus usuels du français fondamental qui est une partie de la langue conçu initialement pour servir de base à l'enseignement du français aux étrangers. Lacouture [Lacouture, 1988] a implanté un système permettant d'analyser la grande majorité des phrases dérivées du français fondamental.

4.2 Les modèles stochastiques

Dans la langue, une classe de mots ne peut être précédée ou suivie que par un nombre fini de classes. Pour pouvoir modéliser ce phénomène, on a eu recours à l'utilisation des probabilités. Rigoureusement, la probabilité de production d'un mot dans une phrase dépend conditionnellement des mots précédents. La construction d'un modèle probabiliste est exprimée généralement par une source de Markov. Ces modèles sont ceux qu'on utilise le plus aujourd'hui en raison de leur facilité de mise en oeuvre et de l'aide que la théorie de l'information peut procurer. Ces systèmes sont fondés sur une phase d'apprentissage dans laquelle on apprend à reconnaître des suites de mots de taille fixe ou variable pour les utiliser ensuite à prédire soit les prochains mots à reconnaître, soit les prochaines classes contenant les mots susceptibles d'être reconnus. Ces classes peuvent être purement syntaxiques ou de nature syntaxico-sémantique. En regroupant ainsi des mots ayant la même fonction, on peut estimer des probabilités de co-occurrences de mots ou de classes d'une façon plus robuste.

5 Le modèle de langage de MAUD

Dans le système MAUD (prototype de dictée automatique [Smaili, 1991]), le modèle de langage utilisé est de type positionnel probabiliste augmenté d'un certain nombre de règles grammaticales permettant de pallier les insuffisances du modèle positionnel. Ce modèle est composé de sept modules agissant par affinement linguistique graduel de l'interprétation de la phrase prononcée. A chaque niveau de l'interprétation de la phrase, un module de cette composante est activé pour éliminer de l'ensemble des solutions retenues celles qui n'ont pas de validité syntaxico-sémantique au sens de ce module.

5.1 La nécessité de la classification lexicale

Le calcul de la probabilité de production d'un mot sachant un contexte particulier utilise la fréquence d'apparition de la classe précédent ce mot. Il est donc nécessaire de regrouper les mots d'un même dictionnaire dans plusieurs classes. Cette opération est connue sous le terme de classification lexicale. Chaque classe regroupe des mots ayant des comportements syntaxiques similaires selon un certain nombre de propriétés pré-définies. Nos travaux sur MAUD, nous ont conduit dans un premier temps à développer manuellement un jeu de classification de 201 classes syntaxico-sémantiques. L'inconvénient de cette classification est qu'elle est assez générale, autrement dit elle s'adapte mal à des applications spécifiques. Si l'on s'intéresse par exemple à la reconnaissance de documents techniques, on peut garder cette même classification, mais comme elle n'est pas très représentative du domaine, le système émettra à tout moment des hypothèses sans lien avec le domaine traité donc sans grand intérêt. Pour pallier ce problème, il serait commode de pouvoir changer de classification en fonction du domaine étudié et donc finalement d'adapter le modèle de langage.

5.2 La classification automatique du dictionnaire

Pour adapter la classification à une nouvelle application, on pourrait envisager de refaire la classification, mais il s'agit là d'une opération très coûteuse en temps. L'idée est donc de chercher à automatiser cette opération. Pour ce faire, nous nous sommes orientés vers les techniques de recherche opérationnelle et nous avons opté pour la méthode du recuit simulé. Cet algorithme utilisé initialement en physique a été ensuite adapté aux problèmes de recherche opérationnelle. Il a comme objectif de minimiser l'énergie du système qu'on compare à des températures décroissantes. Dans notre cas, l'énergie à minimiser est la perplexité du langage. La simulation dans notre approche n'est pas complètement aléatoire mais orientée. En effet, on ne teste pas n'importe quel mouvement d'un mot d'une classe à une autre mais seulement ceux qui sont les plus plausibles. Nous avons implanté cette méthode sur de petits corpus et les résultats sont satisfaisants, on arrive ainsi à construire automatiquement des classes sémantiques (couleur, direction, verbe de mouvements, prénom, etc ...) et des classes syntaxiques (verbes à la 3ème personne, noms communs, etc ... [Smaili, 1995].

6 Conclusion

Nous avons présenté un panorama de la majorité des approches permettant d'améliorer la robustesse des systèmes de RAP dans des environnements bruités. Il est cependant difficile de se prononcer de façon définitive sur la supériorité de telle ou telle méthode. En effet, les systèmes de RAP dans le bruit peuvent être évalués sous différentes conditions qui évoluent continûment de situations réelles, difficiles à contrôler, à des situations contrôlées mais artificielles :

- parole prononcée dans le bruit *vs* parole propre mixée avec un bruit ;
- parole spontanée *vs* parole lue ;
- vocabulaire ouvert *vs* vocabulaire spécifique à l'application.

Pour permettre une comparaison objective des différentes approches de RAP dans le bruit, il est nécessaire que les différents laboratoires adoptent des procédures communes d'évaluation de leurs systèmes. Un élément primordial concerne le développement de corpus de parole spécifiques à la RAP dans des environnements difficiles. Deux approches sont possibles : d'une part, la définition de corpus de parole enregistrés dans le bruit, mais qui se heurte au problème de l'explosion combinatoire des configurations en types et niveaux de bruit ; d'autre part, l'utilisation de corpus de parole propre associés à des bruits et des procédures standards pour générer la parole bruitée. Cela passe en particulier par la définition d'une mesure du rapport signal-à-bruit, devant tenir compte des problèmes comme l'utilisation ou non des zones de silence, l'application de pondérations spécifiques à certaines bandes de fréquences, la définition du SNR en présence de bruits non stationnaires. Si l'ajout de bruit reste une approche valide pour la simulation des environnements faiblement bruités, l'utilisation de corpus enregistrés sous des conditions réelles est indispensable dès que le niveau de bruit devient important (présence d'effet Lombard).

Au regard des différents travaux présentés, quelques conclusions peuvent cependant être dégagées sur les intérêts respectifs de différentes approches. Ainsi, il nous semble important de :

- privilégier dans le processus de décision, les zones de forte énergie du spectre qui sont les moins perturbées par le bruit ;
- appliquer des traitements spécifiques à des classes de sons, le spectre n'étant pas perturbé de façon consistante tout le long du signal ;
- exploiter les corrélations des composantes spectrales intra-trames, ainsi que les corrélations inter-trames au court du temps ;
- développer la prise en compte des bruits non stationnaires, à l'aide de combinaisons de modèles de parole et bruit, et de filtrages multi-voies ;
- utiliser des connaissances *a priori* sur la parole et le bruit dans les processus de filtrage ou de compensation de modèles ;
- utiliser des connaissances sur les mécanismes de l'audition, et des critères d'estimation significatifs par rapport à la perception humaine ;
- développer les algorithmes auto-adaptatifs, pour permettre aux systèmes d'évoluer automatiquement en fonction des conditions d'environnement.

En ce qui concerne les modèles de langage, on s'oriente de plus en plus vers l'apprentissage automatique permettant ainsi de s'affranchir de la phase d'écriture d'une grammaire lorsque cela est possible et d'éviter la phase d'apprentissage semi-automatique des modèles stochastiques. La classification automatique est déja un premier pas, cependant il reste quelques problèmes posés par la mesure utilisée pour l'évaluation. En effet, la perplexité est le seul moyen d'évaluer un modèle de langage, néanmoins il n'a pas été établi que ce critère permet de faire une analyse fine du comportement du modèle de langage. D'autres travaux ont vu le jour ces dernieres années sur l'inférence grammaticale appliquée au langage naturel [Vidal, 1993] [Dupont, 1994], certes ces travaux sont appliqués à des domaines restreints mais semblent donner de très bons résultats et nous paraissent être les modèles de langage des systèmes de reconnaissance de demain.

Références

[Acero et Stern, 1990] A. Acero et R. M. Stern. Environmental robustness in automatic speech recognition. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 849–852, Albuquerque, New Mexico, Avril 1990. ICASSP'90.

[Aikawa et Saito, 1994] K. Aikawa et T. Saito. Noise robust speech recognition using a dynamic-cepstrum. Dans *Proc. Int. Conf. on Spoken Language Processing*, volume 3, pages 1579–1582, Yokohama, Japan, Septembre 1994.

[Alexandre *et al.*, 1993] P. Alexandre, J. Boudy, et P. Lockwood. Root homomorphic deconvolution schemes for speech processing in car noise environments. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 99–102, Minneapolis, Minnesota, USA, 1993. ICASSP'93.

[Anglade *et al.*, 1993] Y. Anglade, D. Fohr, et J.-C. Junqua. Speech Discrimination in Adverse Conditions Using Acoustic Knowledge and Selectively Trained Neural Networks. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 279–282, Minneapolis, Minnesota, USA, 1993. ICASSP'93.

[Applebaum et Hanson, 1991] T. H. Applebaum et B. A. Hanson. Regression features for recognition of speech in quiet and in noise. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 985–988, Toronto, Canada, 1991. ICASSP'91.

[Arslan et Hansen, 1994] L. M. Arslan et J. H. L. Hansen. Minimum cost based phoneme class detection for improved iterative speech enhacement. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 45–48, Adelaide, Australia, 1994. ICASSP'94.

[Atal, 1974] B. Atal. Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identifition and Verification. *Journal of the Acoustical Society of America*, 55:1304–1312, 1974.

[Aubert *et al.*, 1994] X. Aubert, C. Dugast, H. Ney, et V. Steinbiss. Large Vocabulary, Continuous Speech Recognition of Wall Street Journal Corpus. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 129–132, Adelaide, Australia, Avril 1994.

[Beattie et Young, 1991] V. L. Beattie et S. Young. Noisy speech recognition using hidden Markov model state-based filtering. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 917–920, Toronto, Canada, 1991. ICASSP'91.

[Beattie et Young, 1992] V. L. Beattie et S. J. Young. Hidden Markov Model State-Based Noise Cancellation. Rapport Technique F-INFENG/TR 92, Cambridge University Engineering Department, Février 1992.

[Berouti *et al.*, 1979] M. Berouti, B. Schwartz, et J. Makhoul. Enhancement of speech corrupted by acoustic noise. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 208–211. ICASSP'79, Avril 1979.

[Boll et Pulsipher, 1980] S. F. Boll et D. C. Pulsipher. Suppression of acoustic noise in speech using two microphone adaptive noise cancellation. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28:752–753, Décembre 1980.

[Boll, 1979] S. F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 27:113–120, 1979.

[Chen, 1988] Y. Chen. Cepstral Domain Talker Stress Compensation for Robust Speech Recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 36(4):433–439, Avril 1988.

[Cheng et O'Shaughnessy, 1991] Y. M. Cheng et D. O'Shaughnessy. Speech enhancement based conceptually on auditory evidence. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 39(9):1943–1954, 1991.

[Chou *et al.*, 1992] W. Chou, B. H. Juang, et C.-H. Lee. Segmental gpd training of hmm based speech recognizer. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 473–476, San Francisco, California, 1992.

[Cohen, 1985] J. Cohen. Application of an adaptive auditory model to speech recognition. *Journal of the Acoustical Society of America*, 78 (supplement)(1):S50(A), 1985.

[Cox et Malah, 1981] R. V. Cox et D. Malah. A Technique for Perceptually Reducing Periodically Structured Noise in Speech. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 1981.

[Dal Degan et Prati, 1988] N. Dal Degan et C. Prati. Acoustic noise analysis and speech enhancement techniques for mobile radio applications. *Signal Processing*, 15:43–56, 1988.

[Das *et al.*, 1993] S. Das, R. Bakis, A. Nadas, D. Nahamoo, et M. Picheny. Influence of background noise and microphone on the performance of the IBM TANGORA speech recognition system. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 71–74, Minneapolis, Minnesota, USA, Avril 1993. ICASSP'93.

[Das *et al.*, 1994] S. Das, A. Nadas, D. Nahamoo, et M. Picheny. Adaptation techniques for ambience and microphone compensation in the IBM Tangora speech recognition system. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 21–24, Adelaide, Australia, 1994. ICASSP'94.

[Dautrich *et al.*, 1983a] B. A. Dautrich, L. R. Rabiner, et T. B. Martin. On the effects of varying filter bank parameters on isolated word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 31:793–806, 1983.

[Dautrich *et al.*, 1983b] B. A. Dautrich, L. R. Rabiner, et T. B. Martin. On the Use of Filter Bank Features for Isolated Word Recognition. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 1061–1064, Boston, USA, 1983.

[Dupont, 1994] P. Dupont. Regular Grammatical Inference from Positive and Negative Samples by Genetic Search: the GIG metho. *Lecture Notes in Artificial Intelligence*, 1994.

[Ephraim, 1992] Y. Ephraim. Statistical-model-based speech enhancement systems. *Proc. of the IEEE*, 80(10):1526–1555, Octobre 1992.

[Erell et Weintraub, 1994] A. Erell et M. Weintraub. Estimation of Noise-Corrupted Speech DFT-Spectrum Using the Pitch Period. *IEEE Transactions on Speech and Audio Processing*, 2(1), 1994.

[Frangoulis et Gaganelis, 1992] E. Frangoulis et D. A. Gaganelis. Adaptation of the HMM distribution: Application to a VQ codebook and to a noisy environment. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 489–492, San Francisco, California, 1992. ICASSP'92.

[Furui, 1981] S. Furui. Cepstral Analysis Technique for Automatic Speaker Verification. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 29(2):254–272, 1981.

[Furui, 1992] S. Furui. Recent advances in speech recognition technology at NTT laboratories. *Speech Communication*, 11(2–3):195–204, Juin 1992.

[Gales et Young, 1992] M. J. F. Gales et S. Young. An improved approach to the hidden Markov model decomposition of speech and noise. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 233–236, San Francisco, California, Avril 1992.

[Gales et Young, 1994] M. J. F. Gales et S. J. Young. Robust continuous speech recognition using parallel model combination. Rapport Technique F-INFENG/TR 172, CUED, Cambridge University Engineering Department, Mars 1994.

[Gauvain et Lee, 1994] J.-L. Gauvain et C.-H. Lee. Maximum a Posteriori Extimation for Multivariate Gaussian Mixture Observations of Markov Chains. *IEEE Transactions on Speech and Audio Processing*, 2(2):291–298, Avril 1994.

[Ghitza, 1986] O. Ghitza. Auditory nerve representation as a front-end for speech recognition in a noisy environment. *Computer Speech and Language*, 1:109–130, 1986.

[Gong, 1994] Y. Gong. Stochastic Trajectory Modeling and Sentence Searching for Continuous Speech Recognition. Rapport de recherche, CRIN – CNRS & INRIA Lorraine, 1994.

[Gross, 1975] M. Gross Méthodes en syntaxe - Régime des constructions complétives Herman, 1975.

[Gu et Mason, 1989] Y. Gu et J. S. Mason. Speaker normalization via a linear transformation on a perceptual feature space and its benefits in ASR adaptation. Dans *Proc. European Conf. on Speech Communication and Technology*, pages 258–261. Eurospeech-89, 1989.

[Haeb-Umbach *et al.*, 1993] R. Haeb-Umbach, D. Geller, et H. Ney. Improvements in Connected Digit Recognition Using Linear Discriminant Analysis and Mixture Densities. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 239–242, Minneapolis, Minnesota, USA, Avril 1993. ICASSP'93.

[Hansen et Bria, 1990] J. H. L. Hansen et O. N. Bria. Lombard effect compensation for robust automatic speech recognition in noise. Dans *Proc. Int. Conf. on Spoken Language Processing*, pages 1125–1128, 1990.

[Hansen et Clements, 1987] J. H. L. Hansen et M. A. Clements. Iterative speech enhancement with spectral constraints. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 189–192, Dallas, Texas, USA, Avril 1987. ICASSP'87.

[Hansen et Clements, 1989] J. H. L. Hansen et M. A. Clements. Stress compensation and noise reduction algorithms for robust speech recognition. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 266–269. ICASSP'89, 1989.

[Hansen et Clements, 1991] J. H. L. Hansen et M. A. Clements. Constrained Iterative Speech Enhancement with Application to Speech Recognition. *IEEE Transactions on Signal Processing*, 39(4):795–805, Avril 1991.

[Hanson et Applebaum, 1990] B. A. Hanson et T. H Applebaum. Robust speaker-independent word recognition using static, dynamic and acceleration features: experiments with Lombard and noisy speech. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 857–860, Albuquerque, New Mexico, Avril 1990. ICASSP'90.

[Hermansky *et al.*, 1985] H. Hermansky, B. A. Hanson, et H. Wakita. Perceptually Based Linear Predictive Analysis of Speech. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 509–512. ICASSP'85, 1985.

[Hermansky *et al.*, 1991] H. Hermansky, N. Morgan, A. Bayya, et P. Kohn. Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP). Dans *Proceedings of European Conference on Speech Communication and Technology*, pages 1367–1370. EUROSPEECH'91, 1991.

[Hermansky *et al.*, 1993] H. Hermansky, N. Morgan, et H.-G. Hirsch. Recognition of speech in additive and convolutional noise based on RASTA spectral processing. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 83–86, Minneapolis, Minnesota, USA, 1993. ICASSP'93.

[Hermansky, 1990] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752, Avril 1990.

[Hirsch *et al.*, 1991] H. G. Hirsch, P. Meyer, et H. W. Ruehl. Improved Speech Recognition using high-pass filtering of subband Envelopes. Dans *Proceedings of European Conference on Speech Communication and Technology*, pages 413–416, Genova, Italy, Septembre 1991. EUROSPEECH'91.

[Hunt et Lefèbvre, 1988] M. J. Hunt et C. Lefèbvre. Speaker Dependent and Independent Speech Recognition Experiments with an Auditory Model. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 215–218, New York, USA, 1988.

[Hunt et Lefèbvre, 1989] M. J. Hunt et C. Lefèbvre. A comparison of several acoustic representations for speech recognition with degraded and undegraded speech. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 262–265. ICASSP'89, 1989.

[Itakura et Saito, 1968] F. Itakura et S. Saito. An Analysis-Synthesis Telephony based on Maximum Likelihood Method. Dans *Proc. Int. Congr. Acoust.*, pages C–5–5, Tokyo, Japan, Août 1968.

[Itakura, 1975] F. Itakura. Minimum Prediction Residual Principle Applied to Speech Recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 23, Février 1975.

[Junqua, 1989] J.-C. Junqua. *Contribution à l'amélioration de la robustesse des systèmes de reconnaissance automatique de mots isolés.* Thèse de doctorat, Université de NANCY 1, 1989.

[Kadirkamanathan, 1992] M. Kadirkamanathan. Hidden Markov Model Decomposition recognition of speech in noise: A comprehensive experimental study. Dans *ESCA Workshop Proceedings of Speech Processing in Adverse Conditions*, pages 187–190, Cannes-Mandelieu, France, 1992.

[Kitamura *et al.*, 1992] T. Kitamura, S. Ando, et E. Hayahara. Speaker-independent spoken digit recognition in noisy environments using dynamic spectral features and neural networks. Dans *Proc. Int. Conf. on Spoken Language Processing*, volume 1, pages 699–702, Banff, Alberta, Canada, Octobre 1992.

[Klatt, 1976] D. H. Klatt. A Digital Filter-bank for Spectral Matching. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 573–576, Philadelphia, USA, 1976. ICASSP'76.

[Kobatake et Matsunoo, 1994] H. Kobatake et Y. Matsunoo. Degraded word recognition based on segmental signal-to-noise ratio weighting. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 425–428, Adelaide, Australia, 1994. ICASSP'94.

[Koehler *et al.*, 1994] J. Koehler, N. Morgan, H. Hermansky, H. G. Hirsch, et G. Tong. Integrating RASTA-PLP into speech recognition. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 421–424, Adelaide, Australia, 1994. ICASSP'94.

[Lacouture, 1988] R. Lacouture, G. Lapalme Une implantation informatique du français fondamental Dans *Technique et Science Informatiques*, Vol 4, N 4, 1988.

[Lecomte *et al.*, 1989] I. Lecomte, M. Lever, M. Boudy, et A. Tassy. Car noise processing for speech input. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 512–515, Glasgow, UK, Mai 1989. ICASSP'89.

[Leggetter et Woodland, 1994] C. J. Leggetter et P. C. Woodland. Speaker Adaptation of Continuous Density HMMs Using Multivariate Linear Regression. Dans *Proc. Int. Conf. on Spoken Language Processing*, volume 1, pages 451–454, Yokohama, Japan, Septembre 1994.

[Lim et Oppenheim, 1978] J. S. Lim et A. V. Oppenheim. All-pole modeling of degraded speech. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26(3):197–210, Juin 1978.

[Lim et Oppenheim, 1979] J. S. Lim et A. V. Oppenheim. Enhancement and Bandwidth compression of noisy speech. *Proc. of the IEEE*, 67:1586–1604, Décembre 1979.

[Lim et Oppenheim, 1983] J. S. Lim et A. V. Oppenheim. All pole modeling of degraded Speech. Dans *Speech Enhancement*, rédacteur J. Lim, pages 101–114. Prentice-Hall, Englewood Cliffs, NJ, 1983.

[Lim, 1978] J. S. Lim. Evaluation of a correlation subtraction method for ennhancing speech degraded by additive white noise. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26:471–472, 1978.

[Lippmann *et al.*, 1987] R. P. Lippmann, E. A. Martin, et D. B. Paul. Multi-style training for robust isolated word speech recognition. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 705–708, Dallas, Texas, USA, Avril 1987. ICASSP'87.

[Lockwood et Boudy, 1991] P. Lockwood et J. Boudy. Experiments with a Non-Linear Spectral Subtractor (NSS), Hidden Markov Models and the projection, for robust speech recognition in cars. Dans *Proceedings of European Conference on Speech Communication and Technology*, pages 79–82. EUROSPEECH'91, 1991.

[Lombard, 1911] E. Lombard. Le Signe de l'Élévation de la Voix. *Ann. Maladies Oreille, Larynx, Nez, Pharynx*, 37:101–119, 1911.

[Lyon, 1982] R. F. Lyon. A computational model of filtering, detection, and compression in rhe cochlea. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 1282–1285, 1982.

[Malah et Cox, 1982] D. Malah et R. V. Cox. A generalized comb filtering technique for speech enhancement. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 160–163, 1982.

[Mansour et Juang, 1989] D. Mansour et B. H. Juang. A family of distortion measures based upon projection operation for robust speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(11):1659–1671, 1989.

[Mizuta et Nakajima, 1992] S. Mizuta et K. Nakajima. Optimal Discriminative Training for HMMs to Recognize Noisy Speech. Dans *Proc. Int. Conf. on Spoken Language Processing*, volume 2, pages 1519–1522, Banff, Alberta, Canada, Octobre 1992.

[Mokbel *et al.*, 1992] C. Mokbel, L. Barbier, et G. Chollet. Adapting a HMM speech recognizer to noisy environments. Dans *Proc. ESCA Workshop, Speech Processing in Adverse Conditions*, pages 211–214, Cannes-Mandelieu, France, Novembre 1992. ESCA.

[Mokbel, 1992] C. Mokbel. *Reconnaissance de la parole dans le bruit: bruitage/débruitage.* Thèse de doctorat, ENST Paris, Juin 1992.

[Morii *et al.*, 1990] S. Morii, T. Morii, et M. Hoshimi. Noise Robustness in Speaker Independent Speech Recognition. Dans *Proc. Int. Conf. on Spoken Language Processing*, pages 1145–1148, Novembre 1990.

[Nadas *et al.*, 1989] A. Nadas, D. Nahamoo, et M. A Picheny. Speech recognition using noise-adaptive prototypes. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(10):1495–1502, Octobre 1989.

[Nakamura et Shikano, 1989] S. Nakamura et K. Shikano. Speaker adaptation applied to HMM and Neural Networks. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 89–92. ICASSP'89, 1989.

[Nandkumar et Hansen, 1994] S. Nandkumar et J. H. L. Hansen. Speech enhancement based on a new set of auditory constrained parameters. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 1–4, Adelaide, Australia, 1994. ICASSP'94.

[Nolazco Flores et Young, 1994] J. A. Nolazco Flores et S. J. Young. Continuous speech recognition in noise using spectral subtraction and HMM adaptation. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 409–412, Adelaide, Australia, 1994. ICASSP'94.

[Ohkura *et al.*, 1992] K. Ohkura, M. Sugihama, et S. Sagayama. Speaker Adaptation Based on Transfer Vector Field Smoothing with Continuous Mixture Density HMMs. Dans *Proc. Int. Conf. on Spoken Language Processing*, volume 1, pages 369–372, Banff, Alberta, Canada, Octobre 1992. ICSLP'92.

[Ohkura *et al.*, 1993] K. Ohkura, D. Rainton, et M. Sugiyama. Noise-Robust HMMs Based on Minimum Error Classification. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 75–78, Minneapolis, Minnesota, USA, Avril 1993. ICASSP'93.

[Ohkura et Sugiyama, 1991] K. Ohkura et M. Sugiyama. Speech recognition in a noisy environment using a noise reduction neural network and a codebook mapping technique. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 929–932, Toronto, Canada, Mai 1991. ICASSP'91.

[O'Shaughnessy, 1989] D. O'Shaughnessy. Enhancing speech Degraded by Additive noise or interfering speakers. *IEEE Communications Magazine*, pages 46–52, 1989.

[Pai et Wang, 1992] H. Pai et H. Wang. A Study of the Two-Dimensional Cepstrum Approach for Speech Recognition. *Computer Speech and Language*, 6:361–375, 1992.

[Ramalho et Mammone, 1994] M. A. Ramalho et R. J. Mammone. A New Speech enhancement technique with application to speaker identification. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 29–32, Adelaide, Australia, Avril 1994. ICASSP'94.

[Rose *et al.*, 1994] R. C. Rose, E. M. Hofstetter, et D. A. Reynolds. Integrated Models of Signal and Background With Application to Speaker Identification in Noise. *IEEE Transactions on Speech and Audio Processing*, 2(2):245–257, Avril 1994.

[Schroeder et Hall, 1974] M. R. Schroeder et J. L. Hall. Model for Mechanical to neural transduction in the auditory receptor. *Journal of the Acoustical Society of America*, 55:1055–1060, 1974.

[Schroeder, 1981] M. R. Schroeder. Direct (nonrecursive) relations between cepstrum and predictor coefficients. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 29(2):297–301, 1981.

[Seneff, 1988] S. Seneff. A joint synchrony/mean-rate model of auditory speech processing. *Journal of Phonetics*, 16(1):55–76, Janvier 1988.

[Shamma, 1985] S. A. Shamma. Speech processing in the auditory system II: Lateral inhibition and the central processing of speech evoked activity in the auditory nerve. *Journal of the Acoustical Society of America*, pages 1622–1632, 1985.

[Shikano *et al.*, 1986] K. Shikano, K. F. Lee, et R. Reddy. Speaker adaptation through Vector Quantization. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Tokyo, Japan, 1986. ICASSP'86.

[Shikano et Sugiyama, 1982] K. Shikano et M. Sugiyama. Evaluation of LPC Spectral Matching Measures for Spoken Word Recognition. *Trans. IECE*, J65-D(5):535–541, Mai 1982.

[Siohan *et al.*, 1995] O. Siohan, Y. Gong, et J.-P. Haton. Noise Adaptation Using Linear Regression for Continuous Noisy Speech Recognition. Dans *Proceedings of European Conference on Speech Communication and Technology*, Madrid, Spain, Septembre 1995. EUROSPEECH'95.

[Siohan, 1995] O. Siohan. On the robustness of Linear Discriminant Analysis as a preprocessing step for noisy speech recognition. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 125–128, Detroit, Michigan, USA, Mai 1995. ICASSP'95.

[Smaili, 1991] K. Smaili Conception et réalisation d'une machine à dicter à entrée vocale destinée aux grands vocabulaires : Le système MAUD Thèse de Doctorat de l'université de Nancy I, 1991.

[Smaili, 1995] K. Smaili et J.P Haton Quelle classification lexicale pour un système de dictée automatique IA 95-Génie linguistques, 15 èmes journées internationales, Montpellier 1995.

[Sorensen et Hartmann, 1993] H. B. D. Sorensen et U. Hartmann. Robust speaker-independent speech recognition using Non-Linear Spectral Subtraction based IMELDA. Dans *Proceedings of European Conference on Speech Communication and Technology*, volume 1, pages 235–238, Berlin, Germany, Septembre 1993. EUROSPEECH'93.

[Stern et Lasry, 1987] R. M. Stern et M. J. Lasry. Dynamic speaker adaptation for feature-based isolated word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35(6):751–762, Juin 1987.

[Suzuki *et al.*, 1994] T. Suzuki, K. Nakajima, et Y. Abe. Isolated Word Recognition Using Models for Acoustics Phonetic Variability by Lombard Effect. Dans *Proc. Int. Conf. on Spoken Language Processing*, volume 3, pages 999–1002, Yokohama, Japan, Septembre 1994.

[Takahashi et Sagayama, 1994] J.-I. Takahashi et S. Sagayama. Telephone Line Characteristic Adaptation Using Vector Field Smoothing Technique. Dans *Proc. Int. Conf. on Spoken Language Processing*, volume 3, pages 991–994, Yokohama, Japan, Septembre 1994.

[Takizawa et Hamada, 1990] Y. Takizawa et M. Hamada. Lombard speech recognition by formant-frequency-shifted LPC cepstrum. Dans *Proc. Int. Conf. on Spoken Language Processing*, pages 293–296, 1990.

[Tamura, 1989] S. Tamura. An analysis of a noise reduction neural network. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 2001–2003, Glasgow, UK, Mai 1989. ICASSP'89.

[Trompf *et al.*, 1993] M. Trompf, R. Richter, H. Eckhardt, et H. Hackbarth. Combination of distortion-robust feature extraction and neural noise reduction for ARS. Dans *Proceedings of European Conference on Speech Communication and Technology*, volume 2, pages 1039–1042, Berlin, Germany, 1993. EUROSPEECH'93.

[Trompf, 1992] M. Trompf. Experiments with Noise Reduction Neural Networks for Robust Speech Recognition. Rapport Technique TR-92-035, International Computer Science Institute, Berkeley, CA 94704, Mai 1992.

[Van Compernolle, 1989] D. Van Compernolle. Noise adaptation in a hidden Markov model speech recognition system. *Computer Speech and Language*, 3(2):151–168, 1989.

[Varga *et al.*, 1988] A. Varga, R. Moore, J. Bridle, K. Ponting, et M. Russel. Noise compensation algorithms for use with Hidden Markov Model based speech recognition. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 481–484, New York, USA, Avril 1988. ICASSP'88.

[Varga et Moore, 1990] A. P. Varga et R. K. Moore. Hidden Markov Model Decomposition of Speech and Noise. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 845–848, Albuquerque, New Mexico, Avril 1990. ICASSP'90.

[Varga et Moore, 1991] A. P. Varga et R. K. Moore. Simultaneous recognition of concurrent speech signals using Hidden Markov Model Decomposition. Dans *Proceedings of European Conference on Speech Communication and Technology*, pages 1175–1178. EUROSPEECH'91, 1991.

[Varga et Ponting, 1989] A. P. Varga et K. M. Ponting. Control Experiments on Noise Compensation in Hidden Markov Model Based Continuous word Recognizers. Dans *Proceedings of European Conference on Speech Communication and Technology*, Paris, France, 1989. EUROSPEECH'89.

[Vaseghi et Milner, 1993] S. V. Vaseghi et B. P. Milner. Noisy Speech Recognition Based on HMM, Wiener Filters and Re-evaluation of Most Likely Candidates. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 103–106, Minneapolis, Minnesota, USA, 1993.

[Vidal, 1993] E. Vidal, R. Pieraccini, E. Levin. Learning Associations Between Grammars: a New Approach to Natural Language Understanding Dans *Eurospeech*, Vol2, Berlin 1993.

[Weiss *et al.*, 1974] M. R. Weiss, E. Aschkenasy, et T. W. Parsons. Study and development of the INTEL technique for improving speech intelligibility. Rapport Technique NSN-FR/4023, Nicolet Scientific Corp., 1974.

[Widrow *et al.*, 1975] B. Widrow, J. R. Glover, J. M. McCool, J. Kaunitz, C. S. Williams, R. H. Hearn, J. R. Zeidler, E. Dong, et R. C. Goodlin. Adaptive noise cancelling: principles and applications. *Proc. of the IEEE*, 63(12):1692–1716, Décembre 1975.

[Young, 1992] S. J. Young. Cepstral Mean Compensation for HMM recognition in noise. Dans *ESCA Workshop Proceedings of Speech Processing in Adverse Conditions*, pages 123–126, Cannes-Mandelieu, France, 1992. ESCA.

[Zwicker et Feldtkeller, 1981] E. Zwicker et R. Feldtkeller. *Psychoacoustique: l'oreille récepteur d'information.* Masson, 1981.