Systèmes de reconnaissance à grands vocabulaires : Progrès et défis

Jean-Luc Gauvain

Groupe Traitement du Langage Parlé
LIMSI-CNRS, BP 133, 91403 Orsay, France
gauvain@limsi.fr
http://www.limsi.fr/tlp

ABSTRACT

The last decade has witnessed substantial advances in speech recognition technology, which when combined with the increase in computational power and storage capacity, has resulted in a variety of products already or soon to be on the market. This paper is a review of the state-of-the-art in large vocabulary continuous speech recognition, with a view towards highlighting recent advances. It also highlights issues in moving towards applications, discussing system efficiency, portability across languages and tasks, enhancing the system output by adding tags and non-linguistic information. Current performance in speech recognition and outstanding challenges for various applications are discussed.

1. INTRODUCTION

Ces dix dernières années, la reconnaissance automatique de la parole continue à grands vocabulaires a été un des domaines de recherche centraux en RAP, servant de banc d'essai pour évaluer modèles et algorithmes. L'intérêt pour cette technologie va bien au delà des systèmes de dictée de textes. Elle peut par exemple être employée pour l'accès vocal à des bases de données, ou l'indexation par le contenu de documents audiovisuels. Les progrès dans ce domaine profitent également à d'autres technologies, telles que la reconnaissance du locuteur et de la langue, qui utilisent les mêmes modèles. La reconnaissance de la parole concerne principalement la transcription d'un signal vocal en une suite de mots. La plupart des systèmes repose sur une modélisation statistique du processus de génération de la parole. De ce point de vue, le message est produit par un modèle linguistique qui estime $\Pr(w)$ pour toutes les suites de mots w, et le canal acoustique, encodant le message w dans le signal x, est modélisé par une densité de probabilité $f(x|w)$. Le décodage de la parole consiste alors à maximiser la probabilité *a posteriori* de w, ce qui est équivalent à maximiser le produit $\Pr(w)f(x|w)$. Les principes de base sur lesquels la plupart des systèmes sont fondés sont connus depuis de nombreuses années, c.-à-d. l'application de la théorie de l'information à la reconnaissance de la parole [5, 48], la représentation spectrale du signal vocal [26, 27], la programmation dynamique pour le décodage [93, 94], et l'utilisation de modèles acoustiques en contexte [17, 57, 86]. Malgré cela des progrès considérables ont été faits ces dernières années, en particulier pour la modélisation acoustique et le décodage. Ces progrès sont liés à la disponibilité de grands corpus de parole et de textes ainsi qu'aux puissances de calcul accrues qui ont permis le développement de modèles et d'algorithmes toujours plus complexes.

Cet article présente les avancées récentes de l'état de l'art, et explore des domaines d'application rendus possibles par ces progrès technologiques. Une importante avancée est la capacité des systèmes actuels à traiter des données non homogènes, par opposition à des enregistrements soigneusement préparés. Ceci est illustré par le traitement de documents radio ou télédiffusés: avec de nombreux changements de locuteurs, de conditions acoustiques, de thèmes, voire de langues. De nombreux progrès y ont contribué : une analyse acoustique plus robuste, des techniques d'apprentissage tirant profit de très grands corpus audio et textuels, des algorithmes de segmentation du flux audio, l'adaptation non supervisée des modèles acoustiques, des décodeurs plus performants avec des modèles linguistiques d'ordre plus élevés, et la capacité de traiter des vocabulaires beaucoup plus grands que par le passé (65k mots ou plus). Le développement de systèmes dans le cadre d'applications réelles (hors laboratoire) implique de reconsidérer certaines solutions, telles que l'enregistrement du signal, la compensation du bruit et du canal de transmission, et la capacité de rejet, tout en tenant compte des contraintes matérielles [34]. Les techniques mises en avant dans cet article ont été choisies en fonction de résultats expérimentaux obtenus dans différents laboratoires sur des données publiquement disponibles avec des systèmes au niveau de l'état de l'art.

2. MODÉLISATION ACOUSTICO-PHONÉTIQUE

La plupart des systèmes utilisent des modèles de Markov cachés (MMC) pour la modélisation acoustique [6, 23, 28, 35, 47, 61, 62, 67, 75, 77, 81, 97, 100]. D'autres utilisent des modèles segmentaux [41, 70, 105] ou des réseaux neuronaux [1, 11, 45] pour l'estimation des vraisemblances acoustiques, cependant tous les systèmes se servent du cadre des MMC pour combiner l'information linguistique et acoustique dans un seul réseau représentant le langage de l'application. Pour les systèmes fondés sur des MMC, le modèle est une densité de probabilité sur une séquence de vecteurs acoustiques. Les paramètres des vecteurs acoustiques sont choisis afin de réduire la complexité du modèle tout en essayant de garder l'information appropriée, c.-à-d. l'information linguistique. La plupart des systèmes utilisent des cepstres à court terme obtenus par transformée de Fourier ou via un modèle de prédiction linéaire. Les deux jeux de paramètres les plus util-

isés sont des coefficients cepstraux obtenus avec une analyse de type MFCC [19] ou avec une analyse PLP [44]. Dans les deux cas un spectre de puissance à court terme (20 à 30 ms) est estimé sur une échelle MEL, avec une période le plus souvent égale à 10 ms. Les deux jeux de paramètres ont été utilisés avec succès, mais l'analyse PLP s'avère plus robuste en présence de bruit pour certains systèmes [53, 98].

Les modèles de phones en contexte (triphones ou pentaphones) sont aujourd'hui les unités acoustiques les plus répandues. Comparées à des unités plus grandes telles que les diphones, les demisyllabes ou les syllabes, les modèles de phones en contexte offrent un plus large spectre de dépendances contextuelles avec la possibilité d'un mécanisme de repli vers des contextes fréquents. Le choix de l'ensemble des contextes modélisés est habituellement le résultat d'un compromis entre résolution et robustesse, et dépend fortement des données d'apprentissage disponibles. Ce qui est vraiment essentiel c'est d'ajuster le nombre de paramètres du modèle à la quantité de données d'apprentissage. Une technique très efficace pour limiter le nombres de paramètres des modèles sans sacrifier la résolution, consiste à tirer profit de la similitude entre certains états des MMC d'un même phonème en liant les distributions de ces états. Cette idée fondamentale est utilisée dans la plupart des systèmes avec de légères différences dans la mise en œuvre et dans le nom donné à ces groupes d'états (*senones* [46], *genones* [22], *PELs* [9], *tied-states* [103]). Ce partage de paramètres permet bien entendu de réduire la taille du modèle. Il peut être appliqué à tous les niveaux [90, 99] du modèle (allophone, état MMC, et gaussienne) mais plus de flexibilité est disponible au niveau des gaussiennes, où de grandes réductions peuvent être obtenues sans sacrifier les performances.

3. Modélisation lexicale

Le dictionnaire de prononciations est le lien entre le modèle acoustique et les entrées lexicales du modèle de langage, chaque entrée lexicale étant décrite comme une suite d'unités phonémiques. La conception d'un tel dictionnaire nécessite d'une part la sélection des éléments du vocabulaire en minimisant le nombre de mots hors vocabulaire, et d'autre part la détermination des prononciations possibles de chaque mot de ce vocabulaire [54]. La meilleure couverture lexicale peut être obtenue en retenant les mots les plus fréquents dans les données d'apprentissage, ou en ne prenant qu'un sous-ensemble des données (par exemple les données les plus récentes) [15, 37]. En moyenne, chaque mot hors vocabulaire est la cause de 1,5 à 2,0 erreurs [71]. Contrairement à une croyance largement répandue, un plus grand vocabulaire n'implique pas nécessairement un taux d'erreur plus élevé lorsqu'un modèle de langage adéquat est utilisé. Pour la plupart des systèmes les dictionnaires phonétiques utilisent des prononciations "standards" et ne représentent pas explicitement les allophones, laissant aux modèles acoustiques la représentation des variantes observées dans les données d'apprentissage. Plusieurs études ont été effectuées dans le but d'apprendre automatiquement ces prononciations, mais à notre connaissance ces approches quoique prometteuses n'ont pas encore permis d'améliorer significativement les performances des systèmes [82].

4. Modélisation linguistique

Les modèles de langage sont employés pour modéliser les régularités du langage naturel [78]. Les méthodes les plus utilisées sont fondées sur des statistiques n-grammes qui modélisent les contraintes syntaxiques et sémantiques en estimant la probabilité d'un mot dans un texte étant donnés les n-1 mots précédants. L'approche la plus commune pour "lisser" les statistiques des n-grammes rares est un mécanisme de repli utilisant des statistiques d'ordre inférieur lorsque les données d'apprentissage sont insuffisantes [16, 51]. Dans les systèmes actuels, les modèles de langage de type 3-gramme ou 4-gramme peuvent comprendre quelques dizaines de millions de paramètres. Le mécanisme de repli offre l'avantage supplémentaire que la taille du modèle peut être arbitrairement réduite en augmentant le nombre minimum d'observations requises pour inclure un n-gramme dans le modèle. Les modèles 2-gramme et 3-gramme sont les plus largement répandus. De petites améliorations ont été enregistrées avec l'utilisation de contexte plus large (4 ou 5-gramme) [6, 61, 97] ainsi qu'avec l'utilisation de modèles n-grammes de classe de mots [83].

Étant donné un corpus de textes (ou de transcriptions), il est relativement facile de construire un modèle n-gramme en comptant les occurrences de séquences de n mots [18]. Cependant, cela nécessite au préalable un travail important, tant pour la normalisation des textes avant qu'ils puissent être utilisés, que pour le choix du vocabulaire, la définition des mots, et le traitement des mots composés et des sigles. Fréquemment différentes sources de textes sont disponibles en quantités variables et doivent être combinées. Une solution à ce problème consiste à estimer un modèle par source puis de les interpoler. Les poids d'interpolation sont alors directement estimés sur des données de développement au moyen de l'algorithme EM.

5. Adaptation

Un des principaux défis en matière de reconnaissance de la parole est le développement de systèmes robustes, c.-à-d. qui conservent des performances élevées lorsque les conditions acoustiques de test et d'apprentissage sont différentes. Au niveau acoustique, deux classes de techniques pour augmenter la robustesse des systèmes peuvent être identifiées: les techniques de traitement du signal qui essayent de compenser la différence entre le test et l'apprentissage en modifiant le signal à décoder; et les techniques d'adaptation des modèles qui modifient les paramètres modèles pour les rendre plus représentatifs du signal observé.

Les approches fondées sur le traitement du signal comprennent les techniques de normalisation qui réduisent la variabilité du signal, augmentant les performances en conditions mal adaptées mais souvent avec une réduction des performances en conditions normales, et les techniques de compensation qui reposent sur un modèle du bruit et/ou un modèle de la parole. L'adaptation des modèles est une approche beaucoup plus puissante, en particulier quand le traitement du signal repose sur un modèle de la parole. Par conséquent quand les ressources en calcul ne sont pas considérées, l'adaptation des modèles est la solution de prédilection pour compenser les différences aussi minimes soient-elles.

Les techniques les plus généralement utilisées pour l'adaptation des modèles acoustiques, sont la composition de modèles [32, 33], l'adaptation bayésienne [38, 56, 88, 104], et des méthodes de transformation telles que la régression linéaire[59, 24]. La composition de modèles est essentiellement employée pour compenser des bruits additifs tandis que l'adaptation bayésienne et la régression linéaire sont des outils généraux qui peuvent être utilisés pour l'adaptation au locuteur et à l'environnement acoustique. La normalisation de la longueur du conduit vocal [3, 58, 91] est une autre technique qui a été proposée pour réaliser une certaine normalisation du signal vis-à-vis du locuteur.

Bien entendu l'adaptation peut concerner aussi bien le modèle de langage et le dictionnaire de prononciations que les modèles acoustiques. Diverses approches ont été proposées pour adapter le modèle de langage à partir des mots déjà reconnus dans le document à transcrire: un simple modèle de *cache* [49, 79], un modèle *trigger* [80], et un modèle de concordance de thème [87]. Le modèle de *cache* repose sur l'idée que les mots apparaissant dans un document qui vient d'être dicté ont une plus grande probabilité d'apparaître à nouveau. Pour les documents courts l'avantage de ce modèle est bien entendu très réduit. Le modèle *trigger* essaie de résoudre ce problème en augmentant les probabilités des mots qui apparaissent souvent dans les mêmes documents que les mots observés. Pour le modèle de concordance de thème, des mots-clés présents dans le discours traité sont utilisés pour rechercher des documents sur le même thème, documents à partir desquels des modèles de sous-langage sont élaborés puis utilisés pour redecoder le document courant. En dépit de l'intérêt croissant pour les modèles de langage adaptatifs, seules quelques améliorations minimes ont été obtenues par rapport à un modèle statique.

6. DÉCODEUR

L'un des défis posés par la reconnaissance à grands vocabulaire est la conception d'un algorithme de recherche efficace pour décoder l'énorme espace de recherche obtenu en combinant les modèles acoustique et linguistique. À proprement parler, le but du décodeur est de déterminer la suite de mots ayant la probabilité la plus élevée étant donnés le lexique et les modèles acoustique et linguistique. Dans la pratique, cependant, il est commun de rechercher la séquence d'états des MMC la plus probable, c.-à-d. le meilleur chemin dans un graphe (l'espace de recherche), où chaque nœud associe un état de MMC à une trame de signal. Puisqu'il est évidemment prohibitif de rechercher exhaustivement le meilleur chemin, des techniques ont été développées pour réduire le volume des calculs en limitant la recherche à une petite partie de l'espace total. L'approche la plus généralement utilisée pour de petites et moyennes tailles de vocabulaire est une recherche en faisceau trame-synchrone utilisant un algorithme de programmation dynamique [65]. Cette stratégie de base a été étendue pour traiter de grands vocabulaires en ajoutant des dispositifs tels que le *fast-match* [8, 39], les arbres phonétiques dépendants du mot précédant [66], la recherche avant-arrière [4], le réévaluation des N meilleures solutions [85], la recherche progressive [64] et le décodage dynamique en une passe [68]. Une alternative à la recherche trame-synchrone est une recherche asynchrone utilisant l'algorithme A* (*stack decodeur*) [7, 43, 74]. Les décodeurs dynamiques doivent faire appel à des techniques d'élagage très efficaces afin de prendre en compte toute l'information disponible en une seule passe. Ce type de décodeur est très attrayant pour des applications en temps réel. Cependant, beaucoup de systèmes en cours de développement utilisent les décodeurs à plusieurs passes pour réduire les besoins en calcul lorsque le décodage en temps réel n'est pas nécessaire [4, 36, 64, 76, 97].

Les techniques rapides de décodage sont essentielles pour le déploiement d'applications [84]. Pour des systèmes indépendants du locuteur avec des MMC multigaussiens, entre 30 et 50% du temps de décodage peut être utilisé pour évaluer les distributions gaussiennes. Ce temps peut être réduit d'une part en utilisant une méthode de calcul rapide pour les états des MMC, méthode qui bien entendu nécessite quelques approximations [10], et d'autre part en réduisant la taille des modèles avec des techniques de partage de paramètres, méthode qui a l'avantage de réduire également les besoins en mémoire. Un élagage agressif est généralement nécessaire pour effectuer le traitement en temps réel sur les plate-formes actuellement disponibles. C'est inévitablement une source d'erreurs de recherche, de sorte que de nombreuses techniques ont été proposées pour réduire ces erreurs de recherche et pour limiter leurs effets sur les performances des systèmes.

7. AU-DELÀ DES MOTS

En plus des mots prononcés, d'autres attributs peuvent être identifiés dans la signal audio. Cette information additionnelle peut être de nature linguistique (ponctuation, étiquettes sémantiques), ou de nature acoustique (identité du locuteur, environnement acoustique, tour de parole, mesure de confiance, ...).

En ce qui concerne les attributs de nature acoustique, les mêmes techniques de modélisation ont été appliquées avec succès à la reconnaissance du genre et de l'identité du locuteur, ainsi qu'à l'identification des conditions acoustiques. Pour le traitement d'un flux audio continu, il est avantageux de diviser les données en segments acoustiquement homogènes, et d'identifier et retirer les segments sans parole, puis de regrouper les segments de parole par locuteur. Ces informations peuvent être utilisées pour segmenter les transcriptions et ainsi faciliter leur indexation par un système de recherche documentaire.

Pour certaines applications, il peut être particulièrement utile d'estimer l'exactitude des mots et des phrases reconnus [14, 40, 89, 95, 96]. Pour les systèmes à grand vocabulaire, nous sommes essentiellement intéressés par une mesure de confiance au niveau du mot, le but étant d'estimer $\Pr(w_i|x)$ la probabilité *a posteriori* du i-ème mot du texte, ou alternativement $\Pr(w_i|x,\lambda)$ où λ représente les modèles du système. Une estimation de cette dernière probabilité peut être efficacement calculée en appliquant l'algorithme *forward-backward* à un graphe de mot produit par le système de reconnaissance en même temps que l'hypothèse [96]. Cette estimation reposant sur des modèles, bien entendu incorrects, il est commun d'utiliser d'autres caractéristiques du signal tels que les durées du mot et des phonèmes, le débit d'élocution, et le rapport signal/bruit pour obtenir une meilleure estimation de cette probabilité.

8. APPLICATIONS ET PERFORMANCES

La dictée de textes est l'application la plus évidente pour les systèmes de reconnaissance à grands vocabulaire. Elle a depuis 10 ans fait l'objet de développement de produits et il existe aujourd'hui des logiciels peu coûteux disponibles pour une variété de langages et de plateformes matérielles. Sans doute la caractéristique la plus notable de ce type d'application est que la parole à traiter est produite dans le but explicite d'être transcrite par une machine. Cette application a été largement utilisée pour mesurer les progrès en matière de RAP, car il est facile d'évaluer les résultats en comparant la transcription automatique à une transcription de référence. La métrique généralement utilisée est le taux d'erreur sur les mots défini comme suit : *%erreur = %substitutions + %insertions + %élisions*. Sur des données du corpus NAB du LDC (*North American Business News*, textes lus, micro casque), l'état de l'art pour des systèmes indépendants du locuteur se situe autour de 7% d'erreurs. Les mêmes données enregistrées avec un microphone de table dans un environnement bruité (55dBA, S/B de 15dB), le taux d'erreur est environ 14% avec compensation du bruit [71, 72]. Le taux d'erreur pour la dictée spontanée d'articles financiers est de l'ordre de 14% et est supérieur à 20% pour des textes lus au téléphone. En français, sur le corpus BREF de textes lus du journal Le Monde, le taux d'erreur est d'environ 10% [25] (pour des travaux français sur ce problème cf. [2, 13, 31]).

Le second domaine d'applications concerne la transcription et l'indexation des données audio en général, telles que des émissions de radio et télévision, des téléconférences, ou tout autre document audio susceptible d'être indexé [12, 50, 52, 69, 73]. Plusieurs caractéristiques de ce type de données peuvent être identifiées. D'abord, on peut considérer qu'il s'agit de données "trouvées", qui ne sont pas produites dans le but d'être traitées par une machine. En second lieu, il s'agit de flux audio continus, avec de nombreux changements de locuteurs, sans aucune segmentation à priori. Troisièmement, la prise de son et l'environnement acoustique sont beaucoup moins contrôlés que pour les systèmes de dictée. Sur des documents d'information (radio et TV), le taux d'erreur moyen est de l'ordre de 20% pour l'anglais-américain et environ de 25% pour le français et l'allemand. Une section spéciale de la revue CACM a été récemment consacrée à ce sujet [63]. Sur la tâche de DARPA Hub5 [42] adressant la transcription de la parole conversationnelle au téléphone, le taux d'erreur se situe autour de 40% [102].

La troisième classe d'application est celle des systèmes de dialogue [20]. La plupart de ces systèmes visent à offrir un accès à des bases de données. Il y a de plus en plus de systèmes opérationnels mais ils emploient généralement des stratégies de dialogue beaucoup plus contraintes que les prototypes de laboratoire dits à initiative partagée. L'éventail des taux d'erreur de reconnaissance qui ont été publiés pour ces systèmes s'étend de 5% pour des tâches simples (horaires d'avions) avec micro casque à plus de 25% pour des serveurs téléphoniques.

9. DÉFIS ET PERSPECTIVES

La reconnaissance de la parole est loin d'être un problème résolu, comme cela est démontré par la grande différence entre les performances de la machine et celle de l'auditeur humain [29, 92, 60]. Pour combler cette différence nous devons sans aucun doute améliorer nos modèles à tous les niveaux: acoustique, lexical, syntaxe et sémantique.

Pour les systèmes indépendants du locuteur, il est bien connu qu'il peut y avoir une énorme différence (jusqu'à un rapport 20) entre les taux d'erreur de deux locuteurs [30]. Ceci peut être attribué à une variété de facteurs liés au locuteur et à sa vitesse d'élocution [71]. L'adaptation des modèles acoustiques permet de compenser en partie cette différence, mais nécessite au moins quelques minutes de signal pour être vraiment efficace, ce qui limite son champ d'application. Le développment de techniques d'adaptation plus efficaces et plus rapides qui prennent mieux en compte les corrélations entre les paramètres des modèles est donc une nécessité. La réduction de cette différence doit sans doute aussi passer par l'adaptation du lexique de prononciations, en généralisant les variantes observées sur les données déjà produites par le locuteur. Une personne qui prononce un mot d'une façon donnée est susceptible de prononcer les mots semblables de la même manière. Pour les règles de coarticulation entre mots, différents locuteurs appliquent différentes règles phonologiques, et bien que ces règles soient habituellement systématiques pour un même locuteur, à notre connaissance aucun système ne sait tirer parti de cette information.

Côté modélisation linguistique, les techniques explorées pour effectuer les accords à long terme n'ont pas encore été couronnées de succès. Ces techniques seraient particulièrement utile pour traiter les langues fortement flexionnelles pour lesquelles les modèles n-grammes ne sont clairement pas la solution optimale. L'adaptation des modèles linguistiques est un défi pour les systèmes de transcription de documents d'information radio et TV, où il est particulièrement important de maintenir les modèles à jour. De nouveaux thèmes peuvent apparaître soudainement, et demeurer dans l'actualité pendant un temps très variable. L'existence de sources de données contemporaines, tels que les journaux électroniques disponibles sur Internet, devrait nous permettre de mettre à jour automatiquement les modèles de langage [52].

Le développement de systèmes indépendants de l'application est un autre défi majeur. A partir d'un grand corpus de parole transcrite, il est possible de développer des modèles acoustiques pour une variété d'applications, il n'en est pas de même pour les modèles de langage où une bonne couverture du domaine de l'application est essentielle. Avec la technologie actuelle, le portage d'un système vers une nouvelle application ou une autre langue nécessite l'existence de quantités suffisantes de données transcrites. Le développement de techniques d'apprentissage nécessitant peu de supervision est donc un axe de recherche à explorer.

10. CONCLUSION

En dépit des nombreuses avancées de cette dernière décennie, et de la généralisation des systèmes de dictée de textes, la reconnaissance de la parole est loin d'être un problème résolu. Alors qu'il est clair que nos modèles ont besoin d'être améliorés en particulier pour la parole conversationnelle, nous ne savons pas quel est le chaînon le

plus faible entre le modèle acoustique, le modèle de langage et le dictionnaire de prononciations.

Il apparaît cependant qu'une vaste gamme d'applications est maintenant rendue accessible. Les deux domaines les plus prometteurs concernent les serveurs d'informations, et les systèmes d'indexation automatique de documents audio. Les premières expériences en recherche documentaire dans des documents audio ont conduit à des résultats comparables en utilisant des transcriptions manuelles et automatiques. L'énorme quantité d'information diffusée quotidiennement sous formes audio et audiovisuelle nous permet de mesurer l'intérêt de ce résultat.

BIBLIOGRAPHIE

[1] D. Abberley, D. Kirby, S. Renals et T. Robinson, "The THISL Broadcast News Retrieval System," *Proc. ESCA ETRW on Accessing Information in Spoken Audio*, pp. 14-19, Cambridge, U.K., avril 1999.

[2] G. Adda, M. Adda-Decker, J.L. Gauvain, et L. Lamel, "Le système de dictée vocale du LIMSI pour l'évaluation AUPELF'97", *JST FRANCIL*, Avignon, avril 1997.

[3] A. Andreoum T. Kamm et J. Cohen, "Experiments in Vocal Tract Normalisation", *Proc. CAIP Workshop: Frontiers in Speech Recognition II*, 1994.

[4] S. Austin, R. Schwartz et P. Placeway, "The Forward-Backward Search Strategy for Real-Time Speech Recognition," *Proc. IEEE ICASSP-91* pp. 697-700, Toronto, mai 1991.

[5] L.R. Bahl, J.K. Baker, P.S. Cohen, N.R. Dixon, F. Jelinek, R.L. Mercer, et H.F. Silverman,"Preliminary results on the performance of a system for the automatic recognition of continuous speech," *Proc. IEEE ICASSP-76*, Philadelphia, PA, avril 1976.

[6] L.R. Bahl, S. Balakrishnan-Aiyer, M. Franz, P.S. Gopalakrishnan, R. Gopinath, M. Novak, M. Padmanabhan et S. Roukos, "The IBM Large Vocabulary Continuous Speech Recognizer for the ARPA NAB News Task," *Proc. ARPA Spoken Language Systems Technology Workshop*, pp. 121-126, Austin, TX, janvier 1995.

[7] L.R. Bahl, F. Jelinek et R.L. Mercer, "A Maximum Likelihood Approach to Continuous Speech Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, **PAMI-5**(2), pp. 179-190, mars 1983.

[8] L.R. Bahl, P.V. de Souza, P.S. Gopalakrishnan, D. Nahamoo et M. Picheny, "A Fast Match for Continuous Speech Recognition Using Allophonic Models," *Proc. IEEE ICASSP-92*, CA, **1**, pp. 17-21, San Francisco, CA, mars 1992.

[9] J. Baker, J. Baker, P. Bamberg, K. Bishop, L. Gillick, V. Helman, Z. Huang, Y. Ito, S. Lowe, B. Peskin, R. Roth et F. Scattone, "Large Vocabulary Recognition of Wall Street Journal Sentences at Dragon Systems," *Proc. DARPA Speech and Natural Language Workshop*, pp. 387-392, Harriman, NY, février 1992.

[10] E. Bocchieri, "Vector quantization for efficient computation of continuous density likelihoods," *Proc. IEEE ICASSP-93*, **2**, pp. 692-695, Minneapolis, MN, mai 1993.

[11] H. Bourlard et N. Morgan, "Continuous Speech Recognition by Connectionist Statistical Methods," *IEEE Trans. on Neural Networks*, **4**(6), pp. 893-909, 1994.

[12] F. Brugnara, M. Cettolo, M. Federico et D. Giuliani, "A Baseline for the Transcription of Italian Broadcast News," *Proc. IEEE ICASSP-00*, Istanbul, Turkey, juin 2000.

[13] M.J. Caraty, C. Montacié et F. Lefèvre, "Dynamic Lexicon for a Very Large Vocabulary Vocal Dictation System", *Eurospeech*, Rhodes, pp. 2691-2694, 1997.

[14] L. Chase, "Word and acoustic confidence annotation for large vocabulary speech recognition", *Proc. ESCA Eurospeech'97*, pp. 815-818, Rhodes, Greece, septembre 1997.

[15] L. Chase, R. Rosenberg, A. Hauptmann, M. Ravishankar, E. Thayer, P. Placeway, R. Weide et C. Lu, "Improvements in Language, Lexical and Phonetic Modeling in Sphinx-II," *Proc. ARPA Spoken Language Systems Technology Workshop*, pp. 60-65, Austin, TX, janvier 1995.

[16] S.F. Chen et J. Goodman, "An empirical study of smoothing techniques for language modeling," *Computer, Speech and Language*, **13**(4), pp. 359-394, octobre, 1999.

[17] Y.L. Chow, R. Schwartz, S. Roukos, O. Kimball, P. Price, F. Kubala, M.O. Dunham, M. Krasner et J. Makhoul, "The Role of Word-Dependent Coarticulatory Effects in a Phoneme-Based Speech Recognition System", *Proc. IEEE ICASSP-86*, **3**, pp. 1593-1596, Tokyo, Japan, avril 1986.

[18] P. Clarkson et R. Rosenfeld, "Statistical Language modelling using CMU-Cambridge Toolkit," *Proc. ESCA EuroSpeech'97*, pp. 2707-2710, Rhodes, Greece, septembre 1997.

[19] S. Davis et P. Mermelstein, "Comparison of Parametric Representations of Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. Acoustics, Speech, and Signal Processing*, **28**(4), pp. 357-366, 1980.

[20] R. De Mori, "Spoken Dialogues with Computers," Academic Press, 1998.

[21] N. Deshmukh, A. Ganapathiraju, R.J. Duncan et J. Picone, "Human Speech Recognition Performance on the 1995 CSR Hub-3 Corpus" *Proc. ARPA Speech Recognition Workshop*, pp. 129-134, Harriman, NY, février 1996.

[22] V. Digalakis et H. Murveit, "Genones: Optimization the Degree of Tying in a Large Vocabulary HMM-based Speech Recognizer," *Proc. IEEE ICASSP-94*, **1**, pp. 537-540, Adelaide, Australia, avril 1994.

[23] V. Digalakis, M. Weintraub, A. Sankar, H. Franco, L. Neumeyer et H. Murveit, "Continuous Speech Dictation on ARPA's North American Business News Domain," *Proc. ARPA Spoken Language Systems Technology Workshop*, Austin, TX, pp. 88-93, janvier 1995.

[24] V. Digalakis, D. Rtichev et L.G. Neumeyer, "Speaker adaptation using constrained estimation of Gaussian mixtures", *IEEE Trans. on Speech and Audio*, **3**(5), 357-366, septembre 1995.

[25] J.M. Dolmazon, F. Bimbot, G. Adda, M. El Beze, J.C. Caerou, J. Zeiliger et M.A Decker, "ARC B1 - Organisation de la 1e campagne AUPELF pour l'évaluation des systèmes de dictée vocale," *1ères JST FRANCIL*, Avignon, avril 1997.

[26] J. Dreyfus-Graf, "Sonograph and Sound Mechanics," *J. Acoust. Soc. America*, **22**, pp. 731, 1949.

[27] H. Dudley et S. Balashek, "Automatic Recognition of Phonetic Patterns in Speech," *J. Acoust. Soc. America*, **30**, pp. 721, 1958.

[28] C. Dugast, R. Kneser, X. Aubert, S. Ortmanns, K. Beulen et H. Ney, "Continuous Speech Recognition Tests and Results for the NAB'94 Corpus," *Proc. ARPA Spoken Language Systems Technology Workshop*, Austin, TX, pp. 156-161, janvier 1995.

[29] W.J. Ebel et J. Picone, "Human Speech Recognition Performance on the 1994 CSR Spoke 10 Corpus" *Proc. ARPA Spoken Language Systems Technology Workshop*, pp. 53-59, Austin, TX, janvier 1995.

[30] W. Fisher, "Factors Affecting Recognition Error Rate," *Proc. ARPA Speech Recognition Workshop*, pp. 47-52, Harriman, NY, février 1996.

[31] D. Fohr, J.P. Haton, J.F. Mari, K. Smaïli et I. Zitouni, "MAUD : un prototype de machine à dicter vocale", *1ères JST FRANCIL*, Avignon, avril 1997.

[32] M.J.F. Gales et S.J. Young, "An improved approach to hidden Markov model decomposition of speech and noise," *Proc. IEEE ICASSP-92*, pp. 233-236, San Francisco, CA, mars 1992.

[33] M.J.F. Gales et S.J. Young, "Robust Continuous Speech Recognition using Parallel Model Combination," *Computer Speech & Language*, **9**(4), pp. 289-307, octobre 1995.

[34] J.L. Gauvain et L. Lamel, "Large Vocabulary Continuous Speech Recognition: from Laboratory Systems towards Real-World Applications," *Institute of Electronics, Information and Communication Engineers*, J79-D-II:2005–2021, decembre 1996.

[35] J.L. Gauvain, L.F. Lamel, G. Adda et M. Adda-Decker, "Speaker-Independent Continuous Speech Dictation," *Speech Communication*, **15**(1-2), pp. 21-37, octobre 1994.

[36] J.L. Gauvain, L.F. Lamel, G. Adda et M. Adda-Decker, "The LIMSI Nov93 WSJ System," *Proc. ARPA Spoken Language Technology Workshop*, pp. 125-128, Princeton, NJ, mars 1994.

[37] J.L. Gauvain, L.F. Lamel et M. Adda-Decker, "Developments in Continuous Speech Dictation using the ARPA WSJ Task," *Proc. IEEE ICASSP-95*, pp. 65-68, Detroit, MI, mai 1995.

[38] J.L. Gauvain and C.H. Lee, "Maximum *a Posteriori* Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Trans. Speech and Audio Processing*, **2**(2), pp. 291-298, April 1994.

[39] L. Gillick et R. Roth, "A Rapid Match Algorithm for Continuous Speech Recognition," *Proc. DARPA Speech and Natural Language Workshop*, pp. 170-172, Hidden Valley, PA, juin 1990.

[40] L. Gillick, Y. Ito et J. Young, "A Probabilistic Approach to Confidence Measure Estimation and Evaluation", *Proc. IEEE ICASSP-97*, pp. 879-882, Munich, Germany, avril 1997.

[41] J.R. Glass, T.J. Hazen et I. L. Hetherington, "Real-time Telephone-based Speech Recognition in the Jupiter Domain," *Proc. IEEE ICASSP-99*, **1**, pp. 61-64, Phoenix, AZ, mars 1999.

[42] J. Godfrey, E. Holliman et J. McDaniel, "SWITCHBOARD: Telephone Speech Corpus for Research and Development," *Proc. IEEE ICASSP-92*, pp. 517-520, San Francisco, CA, mars 1992.

[43] P.S. Gopalakrishnan, L.R. Bahl et R.L. Mercer, "A tree search strategy for large-vocabulary continuous speech recognition," *Proc. IEEE ICASSP-95*, **1**, pp. 572-575, Detroit, MI, mai 1995.

[44] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. America*, **87**(4), pp. 1738-1752, 1990.

[45] M.M. Hochberg, S.J. Renals, A.J. Robinson et D. Kershaw, "Large vocabulary continuous speech recognition using a hybrid connectionist-HMM system," in *Proc. ICSLP'94*, pp. 1499-1502, Yokohama, Japan, septembre 1994.

[46] M. Hwang et X. Huang, "Subphonetic Modeling with Markov States - Senone," *Proc. IEEE ICASSP-92*, San Francisco, CA, **1**, pp. 33-36, mars 1992.

[47] X. Huang, F. Alleva, M.Y. Hwang et R. Rosenfeld, "An Overview of the SPHINX-II Speech Recognition System," *Proc. ARPA Human Language Technology Workshop*, Princeton, NJ, pp. 81-86, mars 1993.

[48] F. Jelinek, "Continuous Speech Recognition by Statistical Methods," *Proc. of the IEEE*, **64**(4), pp. 532-556, avril 1976.

[49] F. Jelinek, B. Merialdo, S. Roukos et M. Strauss, "A Dynamic Language Model for Speech Recognition," *Proc. DARPA Speech and Natural Language Workshop*, pp. 293-295, Pacific Grove, CA, février 1991.

[50] F. deJong, J.L. Gauvain, J. deb Hartog, K. Netter, "OLIVE: Speech Based Video Retrieval," *Proc. CBMI'99*, Toulouse, octobre 1999.

[51] S.M. Katz, "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer," *IEEE Trans. Acoustics, Speech, and Signal Processing*, **ASSP-35**(3), pp. 400-401, mars 1987.

[52] T. Kemp et A. Waibel, "Unsupervised Training of a Speech Recognizer: Recent Experiments," *Proc. ESCA Eurospeech'99*, Budapest, Hungary, **6** 2725-2728, septembre 1999.

[53] D. Kershaw, A.J. Robinson et S.J. Renals, "The 1995 Abbot hybrid connectionist-HMM large-vocabulary recognition system," *Proc. ARPA Speech Recognition Workshop*, pp. 93-98, Harriman, NY, février 1996.

[54] L.F. Lamel et G. Adda, "On Designing Pronunciation Lexicons for Large Vocabulary, Continuous Speech Recognition," *Proc. ICSLP'96*, **1**, pp. 6-9, Philadelphia, PA, octobre 1996.

[55] L. Lamel, G. Adda et M. Adda-Decker, "Les lexiques de prononciation dans les systèmes de reconnaissance de la parole," *Proc. Séminaire GDR-PRC CHM Lexique et communication parlée*, pp. 1-10, Toulouse, octobre 1996.

[56] C.-H. Lee et Q. Huo, "On Adaptive Decision Rules and Decision Parameter Adaptation for Automatic Speech Recognition," to appear in *Proc. of the IEEE*, special issue, 2000.

[57] K.-F. Lee, *Large-vocabulary speaker-independent continuous speech recognition: The SPHINX system*, PhD Thesis, Carnegie Mellon University, 1988.

[58] L. Lee et R.C. Rose, "Speaker Normalisation Using Efficient Frequency Warping Procedures", *Proc. IEEE ICASSP-96*, **1**, pp. 353-356, Atlanta, GA, mai 1996.

[59] C.J. Leggetter et P.C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," *Computer Speech and Language*, **9**, pp. 171-185, 1995.

[60] R.P. Lippmann, "Speech recognition by machines and humans," *Speech Communication*, **22**(1), pp. 1-15, 1997.

[61] A. Ljolje, M.D. Riley, D.M. Hindle et F. Pereira, "The AT&T 60,000 Word Speech-To-Text System," *Proc. ARPA Spoken Language Systems Technology Workshop*, pp. 162-165, Austin, TX, janvier 1995.

[62] T. Matsuoka, K. Ohtsuki, T. Mori, S. Furui et K. Shirai, "Large-Vocabulary Continuous Speech Recognition using the Japanese Business Newspaper (Nikkei) Task," *Proc. ARPA Speech Recognition Workshop*, Harriman, NY, février 1996.

[63] M. Maybury (ed.), "News on Demand," Special section in the *Communications of the ACM* **43**(2), février 2000.

[64] H. Murveit, J. Butzberger, V. Digalakis et M. Weintraub, "Large-Vocabulary Dictation using SRI's Decipher Speech Recognition System: Progressive Search Techniques," *Proc. IEEE ICASSP-93*, **II**, pp. 319-322, Minneapolis, MN, avril 1993.

[65] H. Ney, "The Use of a One-Stage Dynamic Programming Algorithm for Connected Word Recognition," *IEEE Trans. Acoustics, Speech, and Signal Processing*, **ASSP-32**(2), pp. 263-271, avril 1984.

[66] H. Ney, R. Haeb-Umbach, B.H. Tran et M. Oerder, "Improvements in Beam Search for 10000-Word Continuous Speech Recognition," *Proc. IEEE ICASSP-92*, **I**, pp. 9-12, San Francisco, CA, mars 1992.

[67] L. Nguyen, T. Anastasakos, F. Kubala, C. LaPre, J. Makhoul, R. Schwartz, N. Yuan, G. Zavaliagkos et Y. Zhao, "The 1994 BBN/BYBLOS Speech Recognition System," *Proc. ARPA Spoken Language Systems Technology Workshop*, Austin, TX, pp. 77-81, janvier 1995.

[68] J.J. Odell, V. Valtchev, P.C. Woodland et S.J. Young, "A One Pass Decoder Design for Large Vocabulary Recognition," *Proc. ARPA Human Language Technology Workshop*, pp. 405-410, Princeton, NJ, mars 1994.

[69] K. Ohtsuki, S. Furui, N. Sakurai, A. Iwasaki et Z.P. Zeang, "Recent Advances in Japanese Broadcast News Transcription," *Proc. ESCA Eurospeech'99*, **2**, pp. 671-674, Budapest, Hungary, septembre 1999.

[70] M. Ostendorf, A. Kannan, O. Kimball et J. R. Rohlicek, "Continuous Word Recognition Based on the Stochastic Segment Model," *Proc. ARPA Workshop on Continuous Speech Recognition*, pp. 53-58, Stanford, CA, septembre 1992.

[71] D.S. Pallett, J.G. Fiscus, W.M. Fisher, J.S. Garofolo, B.A. Lund, A.F. Martin et M.A. Przybocki, "1994 Benchmark Tests for the ARPA Spoken Language Program," *Proc. ARPA Spoken Language Systems Technology Workshop*, pp. 5-36, Austin, TX, janvier 1995.

[72] D.S. Pallett, J.G. Fiscus, W.M. Fisher, J.S. Garofolo, A.F. Martin et M.A. Przybocki, "1995 Hub-3 Multiple Microphone Corpus Benchmark Tests," *Proc. ARPA Speech Recognition Workshop*, pp. 27-46, Harriman, NY, février 1996.

[73] D.S. Pallett, A.F. Martin et M.A. Przybocki, "1998 Broadcast News Benchmark Test Results: English and Non-English Word Error Rate Performance Measures," *Proc. DARPA Broadcast News Workshop* pp. 5-12, Herndon, VA, février 1999.

[74] D.B. Paul, "An efficient A* stack decoder algorithm for continuous speech recognition with a stochastic language model," *Proc. IEEE ICASSP-92*, pp. 405-409, San Francisco, CA, mars 1992.

[75] D.B. Paul, "New Developments in the Lincoln Stack-Decoder Based Large Vocabulary CSR System," *Proc. ARPA Spoken Language Systems Technology Workshop*, Austin, TX, pp. 143-147, janvier 1995.

[76] F. Richardson, M. Ostendorf et J.R. Rohlicek, "Lattice-Based Search Strategies for Large Vocabulary Recognition," *Proc. IEEE ICASSP-95*, **1**, pp. 576-579, Detroit, MI, 1995.

[77] I. Rogina et A. Waibel, "The JANUS Speech Recognizer," *Proc. ARPA Spoken Language Systems Technology Workshop*, Austin, TX, pp. 166-169, janvier 1995.

[78] R. Rosenfeld, "Adaptive Statistical Language Modeling," to appear in *Proc. of the IEEE*, special issue, 2000.

[79] R. Rosenfeld, *Adaptive Statistical Language Modeling*, PhD Thesis, Carnegie Mellon University, 1994. (also *Tech. rep. CMU-CS-94-138*)

[80] R. Rosenfeld et X. Huang, "Improvements in Stochastic Language Modeling," *Proc. DARPA Workshop on Speech and Natural Language*, pp. 107-111, Harriman, NY, février 1992.

[81] R. Roth, L. Gillick, J. Orloff, F. Scattone, G. Gao, S. Wegmann et J. Baker, "Dragon Systems' 1994 Large Vocabulary Continuous Speech Recognizer, *Proc. ARPA Spoken Language Systems Technology Workshop*, Austin, TX, pp. 116-120, janvier 1995.

[82] M.D. Riley, W. Byrne, M. Finke, S. Khudanpu, A. Ljojle, J. McDonough, H. Nock, M. Saraclar, C. Wooters et G. Zavaliagkos, "Stochastic pronunciation modelling from hand-labelled phonetic corpora," *Automatic Speech and Speaker Recognition, Speech Communication* **29**(2-4), pp. 209-224, novembre 1999.

[83] A. Sankar, A. Stolke, T. Chung, L. Neumeyer, M. Weintraub, H. Franco et F. Beaufays, "Noise-Resistant Feature Extraction and Model Training for Robust Speech Recognition," *Proc. ARPA Speech*

Recognition Workshop, pp. 117-122, Harriman, NY, février 1996.

[84] M. Schuster, "Memory-efficient LVCSR search using a one-pass stack decoder," *Computer, Speech and Language*, **14**(1), pp. 47-77, janvier 2000.

[85] R. Schwartz, S. Austin, F. Kubala et J. Makhoul,"New uses for N-Best Sentence Hypothesis, within the BYBLOS Speech Recognition System," *Proc. IEEE ICASSP-92*, **I**, pp. 1-4, San Francisco, CA, mars 1992.

[86] R. Schwartz, Y. Chow, S. Roucos, M. Krasner et J. Makhoul, "Improved Hidden Markov Modeling of Phonemes for Continuous Speech Recognition," *Proc. IEEE ICASSP-84*, **3**, pp. 35.6.1-35.6.4, San Diego, CA, mars 1984.

[87] S. Sekine et R. Grishman, "NYU Language Modeling Experiments for the 1995 CSR Evaluation," *Proc. ARPA Speech Recognition Workshop*, pp. 123-128, Harriman, NY, février 1996.

[88] B. Shahshahani, "A Markov Random Field Approach to Bayesian Speaker Adaptation," *Proc. IEEE ICASSP-95*, pp. 697-700, Detroit, MI, mai 1995.

[89] M. Siu et H. Gish, "Evaluation of word confidence for speech recognition systems", *Computer Speecd & Language*, **13**(4), pp. 299-318, octobre 1999.

[90] S. Takahashi et S. Sagayama, "Four-level Tied Structure for Efficient Representation of Acoustic Modeling," *Proc. IEEE ICASSP-95*, pp. 520-523, Detroit, MI, mai 1995.

[91] L.F. Uebel et P.C. Woodland, "An Investigation into Vocal Tract Length Normalisation", *Proc. ESCA Eurospeech'99*, pp. 2527-2530, Budapest, Hungary, septembre 1999.

[92] D.A. van Leeuwen, L.G. van den Berg, H.J.M. Steeneken, "Human Benchmarks for Speaker Independent Large Vocabulary Recognition Performance," *Proc. ESCA Eurospeech'95*, pp. 1461-1464, Madrid, Spain, septembre 1995.

[93] T.K. Vintsyuk, "Speech discrimination by dynamic programming," *Kibnernetika*, **4**, p. 81, 1968.

[94] T.K. Vintsyuk, "Elements-wise recognition of continuous speech composed of words from a specified dictionary," *Cybernetics*, **7**, pp. 133-143, mars-avril 1971.

[95] M. Weintraub, F. Beaufays, Z. Rivlin, Y. Konig et A. Stolcke, "Neural-Network based Measures of Confidence for Word Recognition," *Proc. ICASSP-97*, pp. 887-890, Munich, Germnay, avril 1997.

[96] F. Wessel, K. Macherey et R. Schlüter, "Using word probabilities as confidence measures," *Proc. IEEE ICASSP-98*, pp. 225-228, Seattle, WA, mai 1998.

[97] P.C. Woodland, C.J. Leggetter, J.J. Odell, V. Valtchev et S.J. Young, "The development of the 1994 HTK large vocabulary speech recognition system," *Proc. ARPA Spoken Language Systems Technology Workshop*, pp. 104-109, Austin, TX, janvier 1995.

[98] P.C. Woodland, M.J.F. Gales, D. Pye et V. Valtchev, "The HTK large vocabulary recognition system for the 1995 ARPA H3 task," *Proc. ARPA Speech Recognition Workshop*, pp. 99-104, Harriman, NY, février 1996.

[99] S.J. Young, "The General Use of Tying in Phoneme-Based HMM Speech Recognisers," *Proc. IEEE ICASSP-92*, San Francisco, CA, pp. 569-572, mars 1992.

[100] S.J. Young, "A Review of Large-Vocabulary Continuous Speech Recognition," *IEEE Signal Processing Magazine*, **13**(5), pp. 45-57, septembre 1996.

[101] S.J. Young, M. Adda-Decker, X. Aubert, C. Dugast, J.L. Gauvain, D.J. Kershaw, L. Lamel, D.A. Leeuwen, D. Pye, H.J.M. Steeneken A.J. Robinson et P.C. Woodland, "Multilingual large vocabulary speech recognition: the European SQALE project," *Computer Speech and Language*, **11**(1):73-89, janvier 1997.

[102] S.J. Young et L. Chase, "Speech recognition evaluation: a review of the U.S. CSR and LVCSR programmes," *Computer Speech and Language*, **12**(4), pp. 263-279, octobre 1998.

[103] S.J. Young et P.C. Woodland, "The Use of State Tying in Continuous Speech Recognition," *Proc. ESCA Eurospeech'93*, **3**, pp. 2203-2206, Berlin, Germany, septembre 1993.

[104] G. Zavaliagkos, R. Schwartz et J. McDonough, "Maximum *a Posteriori* Adaptation for Large Scale HMM Recognizers," *Proc. IEEE ICASSP-95*, pp. 725-728, Detroit, MI, mai 1995.

[105] V. Zue. J. Glass, M. Phillips et S. Seneff, "The MIT SUMMIT Speech Recognition System: A Progress Report", *Proc. DARPA Speech and Natural Language Workshop*, pp. 179-189, Philadelphia, PA, février 1989.