Reconnaissance de la parole en français dans un système de dictée automatique pour traducteurs : le projet TransTalk

Julie Brousseau‡, Caroline Drouin‡, George Foster†, Pierre Isabelle†, Roland Kuhn‡, Yves Normandin‡, et Pierre Plamondon†

† Centre d'innovation en technologies de l'information (CITI)
1575, boul. Chomedey, Laval, Québec, Canada, H7V 2X2

‡ Centre de recherche informatique de Montréal (CRIM)
1801, av. McGill College, Montréal, Québec, Canada, H3A 2NA

Résumé

Nous décrivons un système permettant aux traducteurs professionnels de dicter leurs traductions. Puisque ce système accède non seulement au signal acoustique produit par le traducteur, mais aussi au texte source, il est en mesure d'exploiter un modèle statistique de traduction comme outil d'aide à la reconnaissance. Il existe de nombreuses façons de le faire. Laquelle est la meilleure? Nous présentons les expériences qui ont mené à l'intégration du modèle de traduction d'une façon qui améliore à la fois la vitesse et la performance.

1 Introduction

Le projet TransTalk tente d'intégrer la reconnaissance de la parole et la traduction automatique de manière à faire une utilisation maximale de leurs forces complémentaires. Les traducteurs professionnels dictent souvent leurs traductions, qu'ils font ensuite transcrire. Si la dictée est saisie par un système de reconnaissance de la parole qui, en plus, a accès au texte en langue source, des modèles probabilistes de traduction peuvent aider à la reconnaissance. Par exemple, si le système de reconnaissance de la parole doit choisir entre les mots *cheveux* et *chevaux*, qui sont acoustiquement semblables, la présence du mot *horses* (chevaux) dans le texte source orientera le choix.

Nous avons mis au point un prototype du système TransTalk qui saisit en entrée un texte anglais et une traduction dictée en français, et qui produit en sortie un texte français. Dans un document précédent [1], nous nous attardions principalement aux aspects de TransTalk relatifs à la traduction automatique. Depuis la publication de cet article, nous avons apporté des améliorations aux deux composantes du prototype par exemple, la composante de reconnaissance de la parole en français accepte maintenant la prononciation continue. Le présent document traite des problèmes particuliers de la reconnaissance de la parole en langue française pour de grands vocabulaires [2], expose les caractéristiques de notre système visant à résoudre ces problèmes et présente certains résultats expérimentaux.

2 Le Système Transtalk

Nous avons expérimenté trois versions très différentes de notre prototype de TransTalk. La version 1, décrite dans [1], n'acceptait que la dictée par mots isolés. Les deux versions les plus récentes permettent toutes deux la dictée en mode continu, pour un vocabulaire français de 20 000 formes de mots et expressions. Dans la version 2, pour chaque phrase française prononcée, le système produit une liste des 200 hypothèses de séquences de mots les plus probables, qui sont transmises au module de traduction. Le module de traduction utilise sa connaissance de la phrase anglaise correspondante pour choisir l'une des 200 hypothèses de séquences de mots françaises comme produit final du système complet. Dans la version 3, le module de traduction intervient beaucoup plus tôt, en fait avant que la reconnaissance ne s'amorce. À partir de chaque phrase source anglaise, le module de traduction produit un *vocabulaire dynamique* de mots français susceptibles de figurer dans la traduction. Au moment de la reconnaissance de la phrase française dictée, le moteur de reconnaissance n'est autorisé à prendre en considération que les mots appartenant au vocabulaire dynamique courant. Ces différentes façons d'intégrer le module de traduction à le moteur de reconnaissance sont décrites de façon plus détaillée à la section 5 (ci-dessous).

En passant de la version 2 à la version 3, nous avons aussi apporté à le moteur de reconnaissance des changements non liés à la façon dont le module de traduction y est intégré. Dans le moteur de reconnaissance de la version 2 (décrit en détail dans [4]), la modélisation acoustique est effectuée par des modèles markoviens cachés (MMC) à triphones avec distributions de sortie à livres de codes partagés (distributions gaussiennes multimodes avec un livre de codes de densités gaussiennes par phone). La reconnaissance s'effectue en plusieurs passages. Les deux premiers passages (vers l'avant et vers l'arrière) utilisent un modèle de langue à bigrammes et des triphones intra-mot, mais non inter-mots, pour produire un graphe de paires de mots bouclé. Un algorithme de recherche des N meilleurs candidats produit ensuite les 200 énonciations les plus probables, dont les scores sont réévalués et qui sont réordonnées à l'aide de triphones inter-mots avant d'être soumises au module de traduction. Tous ces calculs prennent environ 92,6 fois le temps réel - avant même que les scores ne soient réévalués par le modèle de traduction.

Le moteur de reconnaissance de la version 3 n'utilise plus les triphones comme unité acoustique de base. L'unité de base est maintenant l'*état-arbre*, c'est-à-dire que chaque état du MMC, pour un phone donné, est représenté par un arbre de décision lié au contexte [5]. Jusqu'ici, notre expérience des états-arbres pour la reconnaissance de la parole tant en anglais qu'en français a été extrêmement positive. Comparativement aux triphones, les états-arbres utilisent beaucoup moins de mémoire et accélèrent la reconnaissance, tout en donnant de meilleurs résultats. L'autre changement important apporté à le moteur de reconnaissance a consisté à utiliser des distributions de sortie discrètes, plutôt que des distributions à livres de codes partagés comme dans la version 2. Le but recherché, dans ce cas, était uniquement un gain de vitesse (en général, l'utilisation de distributions discrètes accélère le système, mais réduit la performance).

L'effet net de ces changements les deux augmentant la vitesse, l'un tendant à améliorer la performance, l'autre tendant à la diminuer était une version 3 du système de base (c.-à-d. sans module de traduction) fonctionnant à 15,8 fois le temps réel et dont la performance sur le plan de la reconnaissance de la parole était légèrement inférieure à celle de la version 2. Nous allons voir que l'intégration du module de traduction a permis, dans la version 3, d'obtenir un système complet considérablement plus rapide et plus performant que le système de base.

3 Reconnaissance de la Parole en Français

Dans le lexique français de 20 000 mots du système, chaque entrée a une représentation phonétique basée sur 44 phones comprenant 20 voyelles et 24 consonnes. Pour améliorer le rendement du module de traduction, certaines expressions fréquentes comme *de la* ou *tout de suite* sont entrées comme des unités. Les prononciations de base ont été obtenues automatiquement, à l'aide d'un ensemble de règles de graphème-à-phonème qui tiennent compte des particularités phonétiques du français parlé au Québec (notamment l'assibilation et le relâchement des voyelles), puis ont été vérifiées manuellement.

D'autres particularités du français doivent être prises en compte dans la construction d'un système de reconnaissance de la parole. L'une d'elles est l'explosion du dictionnaire attribuable à l'élision et aux homophones. L'élision signifie l'omission de la dernière voyelle de certains mots-outils (de, la, ne, que) quand le mot suivant commence par une voyelle. Notre système traite ces unités contractées comme des mots distincts.

Les homophones sont plus fréquents en français qu'en anglais et sont souvent la source de mauvais choix de mots dans la reconnaissance de la parole en français. Dans notre vocabulaire de 20 000 mots, il y en a 6 020 pour lesquels il existe au moins un homophone. Autrement dit, 30 % de notre lexique est formé d'homophones, comparativement à 5 % pour le lexique de 19 977 mots du Wall Street Journal [4]. Si un seul membre d'un ensemble d'homophones est prédit par la phrase source anglaise, le module de traduction se révélera particulièrement utile (p. ex. *children* invitera à choisir *enfants* plutôt que *enfant*).

Un des principaux obstacles à surmonter dans la reconnaissance de la parole en français est la liaison. La liaison survient, dans le discours continu, quand une consonne terminant un mot n'est pas prononcée si le mot est énoncé isolément, mais est prononcée si le mot suivant commence par une voyelle. Il est difficile de prédire si une telle consonne est prononcée - cela dépend d'une interaction complexe entre l'orthographe, la syntaxe, la sémantique et d'autres facteurs.

Nous avons essayé plusieurs méthodes de traitement des liaisons pendant l'entraînement et l'expérimentation acoustique. Actuellement, les liaisons possibles sont dérivées automatiquement dans le lexique par un ensemble de règles simples. Par exemple, un mot se terminant par "s" (p. ex. les) engendrera le phone de liaison /z/. Le phone de liaison est représenté comme un champ distinct dans le lexique. Durant l'entraînement et la reconnaissance, ce *champ de liaison* sera actif si le mot suivant commence par une voyelle. Toutefois, même quand le champ est actif, la réalisation acoustique de la liaison est optionnelle; si elle survient, une pénalité d'insertion est imposée. Sept consonnes peuvent participer à une liaison : /z,t,n,r,p,v,k/.

Notons que nos règles de production de liaisons ne sont pas parfaites. Par exemple, elles autoriseront à tort une liaison après un nom propre (comme dans *Gervais /z/ avait deux chiens*) ou même après la conjonction *et* (insertion erronée d'un /t/). Un travail additionnel s'impose dans ce domaine. Nos modèles de langue du français ne sont pas parfaits non plus. Par exemple, les accords à longue portée entre les mots sont une caractéristique importante du français que les modèles classiques à N-grammes arrivent mal á traiter.

4 Description des Corpus

La plupart des phrases du corpus d'entraînement acoustique provenaient de La Presse, un quotidien de langue française de Montréal : 88 personnes différentes ont lu chacune, en moyenne, 256 phrases de La Presse, ce qui a donné 22 528 phrases. Dix-neuf locuteurs ont lu chacun 25 phrases françaises du Journal des débats de la Chambre des communes du Canada (Hansard), pour un total de 475 phrases. Enfin, 29 locuteurs ont lu 43 phrases françaises phonétiquement équilibrées [3], pour un total de 1 247 phrases. Au total, donc, 24 250 phrases ont été lues en mode continu, avec ponctuation énoncée à haute voix.

Pour le modèle de la langue française à bigrammes, nous avons utilisé comme corpus d'entraînement 2,2 millions de mots français tirés de 40 fichiers du Journal des débats. Pour le modèle de traduction, l'entraînement a été fait à l'aide d'un corpus bilingue de 45,6 millions de mots anglais et de 47,0 millions de mots français provenant de 943 fichiers du Journal des débats. (En fait, TransTalk n'aurait pu voir le jour sans la disponibilité d'un vaste corpus bilingue comme le Journal des débats!)

Le système TransTalk a été testé au moyen de 300 phrases du Journal des débats (6 639 mots) qui n'avaient pas servi à l'entraînement acoustique, ni à l'entraînement du modèle de langue ou du modèle de traduction. Ces phrases répondent à deux critères : elles contiennent 40 unités ou moins (y compris la ponctuation), et elles sont entièrement couvertes par le lexique de 20 000 mots de le moteur de reconnaissance. Elles ont été dictées en mode continu, avec ponctuation énoncée à haute voix, sur un Sennheiser HMD 410, par 10 locuteurs (5 femmes et 5 hommes) qui n'avaient pas participé à l'entraînement.

5 Application des Modèles de Traduction

En utilisant comme information une phrase source anglaise et une traduction française dictée, notre système doit tenter de produire la transcription de la phrase française. Divers modèles de traduction, qui exigent tous des données d'entraînement bilingues, peuvent produire, pour une phrase source anglaise donnée, des estimations probabilistes des mots français correspondants [1, 2]. Les modèles se distinguent principalement par la mesure dans laquelle ils tiennent compte de la position de mots donnés dans la phrase anglaise et la phrase française. Pour les besoins du présent article, mentionnons simplement que certains de ces modèles sont élémentaires et peu coûteux en temps de calcul, tandis que d'autres sont complexes et coûteux. Quels modèles sont préférables, et comment les probabilités qu'ils produisent peuvent-elles orienter la reconnaissance de la parole?

Dans la version 1 de TransTalk (décrite en [1]), un modèle combiné traduction-langue (MTL) simple est appliqué au produit d'un moteur de reconnaissance de mots isolés. Pour chaque unité acoustique, le moteur de reconnaissance produit n hypothèses de mots; le MTL est appliqué durant une recherche Viterbi visant à produire une phrase française. Le MTL a une composante qui prédit la catégorie grammaticale française suivante compte tenu des deux catégories précédentes, comme dans un modèle de langue triPOS standard (POS = *part of speech*). Toutefois, la probabilité conditionelle d'un mot, étant donnés sa catégorie et la phrase anglaise correspondante, est fournie par un modèle de traduction simple.

La version 2 utilise la méthode des N meilleurs candidats : le moteur de reconnaissance produit les N hypothèses les plus probables (d'après les données acoustiques et un modèle de langue à bigrammes), et le modèle de traduction est utilisé pour réévaluer les scores de ces hypothèses. L'avantage de cette méthode est que les modèles de traduction les plus complexes et les plus coûteux en temps de calcul peuvent établir en peu de temps les scores d'hypothèses de phrases françaises complètes (pour toute valeur raisonnable de N). L'inconvénient est que ces modèles de traduction ne sont mis à contribution qu'une fois que le moteur de reconnaissance a pris des décisions irréversibles sur le choix des mots.

Dans la version 3, l'ordre est inversé : le même modèle de traduction que dans la version 2 est appliqué avant que la reconnaissance de la phrase française ne s'amorce, et produit à partir de la phrase anglaise un *vocabulaire dynamique* de taille restreinte par rapport au lexique complet de le moteur de reconnaissance. La reconnaissance se fait normalement (uniquement à l'aide du modèle de langue à bigrammes), sauf que le moteur de reconnaissance n'est autorisé à prendre en considération que les mots appartenant au vocabulaire dynamique. Avant de mettre en oeuvre la version 3, nous savions que cette méthode aurait au moins l'avantage d'accélérer la reconnaissance, en réduisant l'espace de recherche. Nous ne pouvions savoir d'avance si le fait d'utiliser le vocabulaire dynamique plutôt

Model	Mots - % exacts	Phrases - % exactes	Rang de la meilleure phrase
ER	80.7%	4.0%	54/200
trigramme	84.5%	8.7%	19/200
tri + MT	86.0%	12.7%	11/200
meilleure	90.0%	22.0%	1/200

Tableau 1: Réévaluation des scores des N meilleures hypothèses (N=200)

que le vocabulaire complet entraînerait une augmentation du nombre d'erreurs de reconnaissance (due à l'élimination erronée, par le modèle de traduction, de mots qui auraient dû être pris en considération) ou une diminution de ces erreurs (grâce à l'élimination, par le modèle de traduction, de mots incorrects qui sont acoustiquement semblables aux mots exacts).

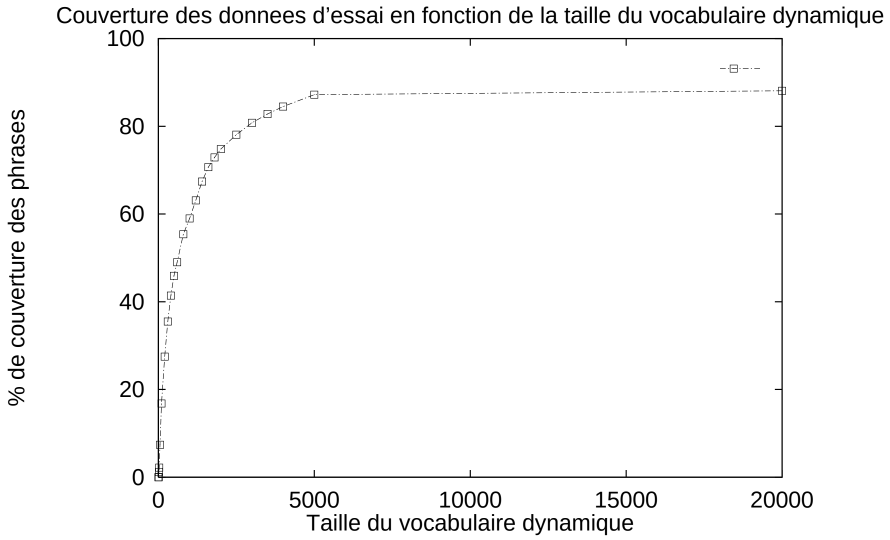
6 Résultats

Les résultats obtenus avec la version 1 du système sont présentés dans [1]. Le tableau 1 présente des chiffres relatifs à la version 2. La ligne *ER* donne les résultats pour la meilleure des 200 hypothèses produites par le moteur de reconnaissance. Le rang moyen de la meilleure hypothèse (celle qui est le plus près de la vérité) est 54. La ligne *trigramme* donne les résultats obtenus après réévaluation des scores des hypothèses au moyen d'un modèle de langue à trigrammes (le moteur de reconnaissance utilise un ML à bigrammes). La ligne *tri + MT* montre les nouveaux scores obtenus avec une interpolation optimisée entre le ML à trigrammes et le modèle de traduction. Enfin, la ligne *meilleure* montre les statistiques propres à la meilleure des 200 hypothèses. Ces résultats montrent qu'on obtient seulement une légère amélioration en réévaluant les scores à l'aide de la combinaison *modèle à trigrammes + modèle de traduction* par rapport aux scores obtenus à l'aide du modèle à trigrammes seulement. Rappelons, par ailleurs, que la version 2 fonctionne à 92,6 fois le temps réel une vitesse peu plausible pour un système opérationnel!

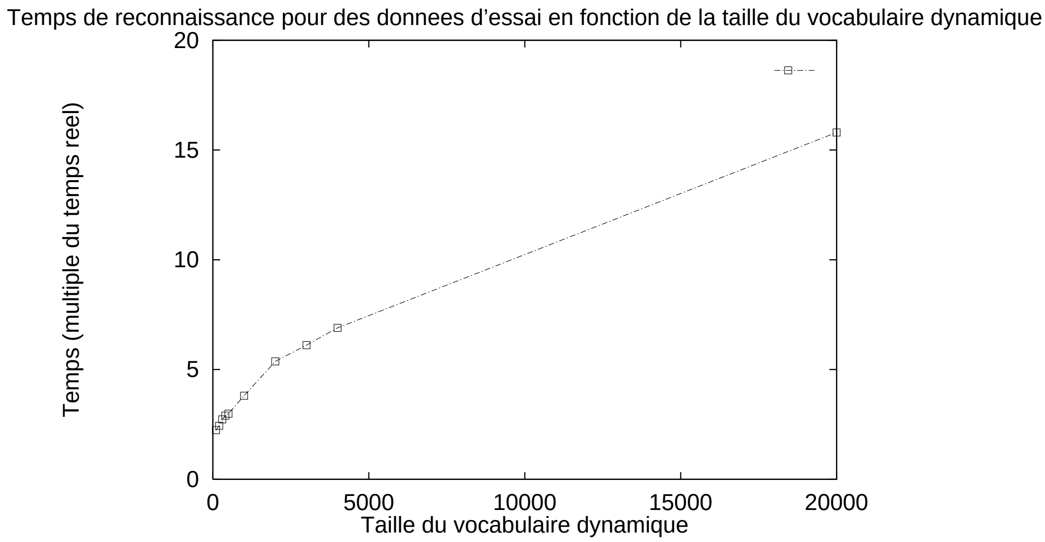
Le système de base de la version 3 (sans module de traduction) utilise le lexique complet de 20 000 mots, fonctionne à 15,8 fois le temps réel et donne une proportion de 75,7 % mots exacts. Le module de traduction examine une phrase anglaise et produit une liste de mots français en ordre décroissant de probabilité. Il y a plusieurs façons d'utiliser cette liste pour définir un *vocabulaire dynamique* :

- choisir tous les mots ayant une probabilité non nulle;
- choisir les M premiers mots, pour un nombre fixe M;
- choisir tous les mots dont la probabilité est supérieure à une probabilité fixe P.

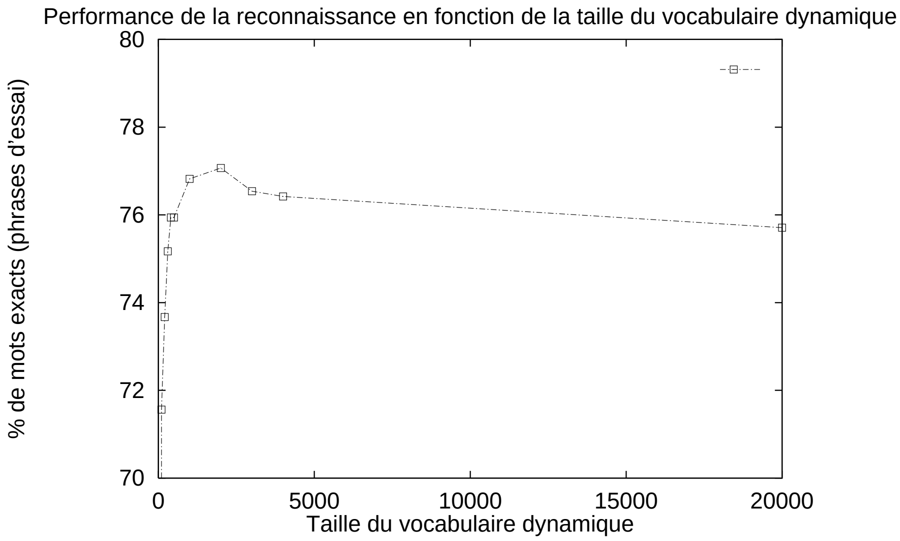
Pour nos 300 phrases d'essai, la première méthode a donné un vocabulaire dynamique comprenant environ 5 000 mots en moyenne, avec un temps



Graphique 1



Graphique 2



Graphique 3

d'exécution de 7,8 fois le temps réel et 76,2 % de mots exacts. Les résultats de l'application de la deuxième méthode sont présentés dans les graphiques cidessus. Le graphique 1 montre la couverture des mots des phrases d'essai par le vocabulaire dynamique. Le graphique 2 montre le temps d'exécution en fonction de la taille du vocabulaire dynamique. Enfin, le graphique 3 montre que la performance s'accroît avec la réduction de la taille du vocabulaire dynamique, le niveau optimal étant atteint autour de 2 000 mots (77,1 % de mots exacts, à 5,4 fois le temps réel). Si nous trichons en fixant la taille du vocabulaire dynamique à tout juste ce qu'il faut pour couvrir tous les mots de la phrase française, la taille moyenne du vocabulaire est de 614 mots, la proportion de mots exacts est de 80,1 % et le temps d'exécution moyen est de 3,4 fois le temps réel. Nous n'avons pas encore de résultats pour l'utilisation d'un seuil de probabilité.

7 Analyse et Travaux Futurs

La version 2 de TransTalk a un temps d'exécution d'environ 93 fois le temps réel. Si nous fixons le vocabulaire dynamique de la version 3 à sa taille maximale (qui est d'environ 5 000 mots en moyenne), la vitesse est d'environ 16 fois le temps réel. Fait remarquable, nous pouvons améliorer à la fois la vitesse et la performance en réduisant la taille du vocabulaire dynamique à sa valeur optimale d'environ 2 000 mots; la vitesse est alors d'environ 5 fois le temps réel. Il est rare, en reconnaissance de la parole, qu'on puisse améliorer simultanément la vitesse et la performance! De plus, nous n'avons apporté que des modifications mineures aux structures de données de notre moteur de reconnaissance pour produire la version 3. Des changements plus profonds nous rapprocheraient davantage du temps réel, tout comme le ferait une règle heuristique (fondée, par exemple, sur la longueur de la phrase anglaise) permettant une variation de la taille du vocabulaire (la taille optimale, pour nos données d'essai, est de 614 mots).

Dans nos prochaines expériences, nous allons faire intervenir le modèle de traduction pendant la reconnaissance, en combinant les probabilités qu'il produit avec celles d'un modèle de langue conventionnel. Notons que puisque le texte source (anglais) est disponible avant que la dictée ne commence, nous pourrions calculer à l'avance ces probabilités combinées, pour chaque phrase, au cours d'un passage hors ligne, de façon à gagner du temps lors de la reconnaissance.

On pourrait envisager un système qui applique des modèles de traduction avant, pendant et après la reconnaissance de la parole :

- un modèle de traduction simple serait utilisé avant la reconnaissance pour déterminer le vocabulaire dynamique, ce qui contribuerait à accélérer la reconnaissance;

- un autre modèle de traduction produirait des probabilités qui seraient combinées aux probabilités du modèle de langue durant la recherche, pour faire en sorte que les mots recevant des probabilités élevées du modèle de traduction et du modèle acoustique, et de faibles probabilités du modèle de langue, ne soient pas perdus;

- un modèle de traduction complexe et coûteux en temps de calcul serait appliqué après la reconnaissance, pour réévaluer les scores d'un ensemble de grilles de N meilleurs candidats ou de graphes de mots produits par le moteur de reconnaissance.

Références

[1] Marc Dymetman, Julie Brousseau, George Foster, Pierre Isabelle, Yves Normandin, et Pierre Plamondon. Towards an automatic dictation system for translators: the TransTalk project. Dans *Proceedings, ICSLP 94*, volume 2, pages 691–694, Yokohama, Japan, September 1994.

[2] J.L. Gauvain and L.F. Lamel et al. Speaker-independent continuous speech dictation. *Speech Communication*, 15: 21–37, 1994.

[3] M. Lennig. 3 listes de 10 phrases phonétiquement équilibrées. *Revue d'acoustique*, 1(56): 39–42, 1981.

[4] Y. Normandin and D. Bowness et al. CRIM's november 94 continuous speech recognition system. Dans *Proceedings of the Spoken Language Systems Technology Workshop*, Austin, Texas, January 1995.

[5] S.J. Young, J. Odell, and P. Woodland. Tree-based state tying for high accuracy acoustic modeling. Dans *Proceedings of the ARPA Workshop on Human Language Technology*, pages 286–291, March 1994.