UNIVERSITÉ DE SHERBROOKE
Faculté de génie
Département de génie électrique et de génie informatique

Modèles de Markov cachés à haute précision dynamique

Mémoire de maîtrise
Spécialité : génie électrique

Sébastien GAGNON

Jury : Jean Rouat (directeur)
Ramin Pichevar
Éric Plourde

Sherbrooke (Québec) Canada 05 2016

À ma compagne de vie, Véro, et à mes parents

RÉSUMÉ

La reconnaissance vocale est une technologie sujette à amélioration. Malgré 40 ans de travaux, de nombreuses applications restent néanmoins hors de portée en raison d'une trop faible efficacité. De façon à pallier à ce problème, l'auteur propose une amélioration au cadre conceptuel classique. Plus précisément, une nouvelle méthode d'entraînement des modèles markoviens cachés est exposée de manière à augmenter la précision dynamique des classificateurs. Le présent document décrit en détail le résultat de trois ans de recherche et les contributions scientifiques qui en sont le produit. L'aboutissement final de cet effort est la production d'un article de journal proposant une nouvelle tentative d'approche à la communauté scientifique internationale.

Dans cet article, les auteurs proposent que des topologies finement adaptées de modèles markoviens cachés (HMMs) soient essentielles à une modélisation temporelle de haute précision. Un cadre conceptuel pour l'apprentissage efficace de topologies par élagage de modèles génériques complexes est donc soumis. Des modèles HMM à topologie *gauche-à-droite* sont d'abord entraînés de façon classique. Des modèles complexes à topologie générique sont ensuite obtenus par « *écrasement* » des modèles *gauche-à-droite*. Finalement, un enchaînement successif d'« *élagages* » et d'entraînements Baum-Welch est fait de manière à augmenter la précision temporelle des modèles.

Mots-clés : Reconnaissance vocale automatique, modèles markoviens cachés, structure temporelle fine, élagage, déséquilibre de plages dynamiques de probabilités

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ACRONYMES

Acronyme	**Définition**
ASR	Automatic Speech Recognition
DNN-HMM	Deep Neural Network Hidden Markov Model
GMM-HMM	Gaussian Mixture Model Hidden Markov Model
HMM	Hidden Markov Model
MFCC	Mel Frequencies Cepstral Coefficients
pdf	probability density function

CHAPITRE 1

INTRODUCTION

La reconnaissance vocale automatique est un domaine de recherche consacré à imiter la capacité des humains à comprendre la parole. Le but premier est de pouvoir interagir avec la technologie qui nous entoure à l'aide de notre mode de communication préféré. Les objectifs principaux sont clairs : se faire comprendre rapidement et efficacement. Le développement d'un système remplissant pleinement ces critères marquerait l'avènement d'une nouvelle ère en interactions personne-machine.

Bien que la technologie existe déjà et est présente dans notre vie de tous les jours (triage d'appels téléphoniques automatique, *Siri* de *Apple*, *Dragon NaturallySpeaking* de *Nuance*, etc.), celle-ci n'est qu'une pâle imitation des capacités humaines. Ce fait est non seulement prouvé scientifiquement [Lippmann, 1997], mais est également une connaissance populaire. Par conséquent, de nombreuses applications potentielles sont malheureusement hors de portée et le seront tant et aussi longtemps que la reconnaissance vocale automatique (« Automatic Speech Recognition » ou ASR) n'approchera pas le niveau d'interaction humaine. On pense notamment à l'opération de robots et ordinateurs par la voix, le triage des appels d'urgence, la rédaction automatique sans assistance, etc.

Toutes les technologies ASR modernes font appel à des *modèles acoustiques*. Ceux-ci sont comparés d'une façon ou d'une autre aux signaux vocaux enregistrés par un microphone. C'est de cette comparaison que naît la *classification* des signaux reçus. De façon générale, chaque *classe* (que ce soit un mot, une syllabe ou un son) possède un *modèle acoustique* la caractérisant. La qualité des modèles utilisés a un impact énorme sur l'efficacité de la reconnaissance.

Le présent travail de recherche vise donc à proposer une méthode de conception de modèles acoustiques de meilleure qualité dans le but d'augmenter l'efficacité des systèmes l'employant. De façon plus spécifique, l'auteur s'attaque à une caractéristique peu explorée en ASR : la *précision temporelle*. Celle-ci se définit comme le niveau de détail des différents cheminements temporels que peuvent emprunter les représentations MFCC («*MEL* Frequencies Cepstrum Coefficients» des signaux acoustiques appartenant à la même *classe*.

Les objectifs de ce travail sont de développer une méthode de modélisation capable d'augmenter la précision temporelle et d'en d'évaluer son efficacité.

Le travail présenté fait état d'avancements technologiques qui sont soumis sous la forme d'un article de journal scientifique. L'auteur démontre d'abord que la composante temporelle des modèles acoustiques utilisés en reconnaissance vocale depuis 40 ans n'a qu'un faible niveau de précision et celle-ci ne peut être facilement augmentée sans modifications à la technologie.

En réponse à cette limitation, l'auteur propose une amélioration capable d'augmenter substantiellement la précision temporelle des modèles acoustiques. L'article soumis démontre également comment cette amélioration impacte positivement les performances de reconnaissance de systèmes ASR, en l'occurrence un système de type *Gaussian mixture model hidden Markov model* (GMM-HMM). Finalement, certaines conclusions quant à la différence majeure de contexte entre la reconnaissance par mots (« word-based ASR ») et par phonèmes (« phone-based ASR ») sont tirées.

Le présent travail présente d'abord l'état de l'art et certains concepts clés en reconnaissance de la parole nécessaires à la compréhension de l'article. Celui-ci est ensuite exposé. Ce travail est finalement terminé par une conclusion énumérant les contributions spécifiques et les avenues de recherches envisageables.

CHAPITRE 2

ÉTAT DE L'ART

2.1 Introduction au traitement de la parole

Avant d'analyser les différentes techniques qui entrent dans le fonctionnement d'un algorithme de reconnaissance vocale de l'état de l'art, il convient de prendre en compte certains phénomènes importants de la linguistique, un domaine de recherche consacré à la structure du langage humain.

Premièrement, le langage humain peut être considéré comme l'entrelacement temporel et fréquentiel d'événements acoustiques. Ce processus est considéré comme hautement variable. Parmi les nombreuses sources de variabilité, on dénombre notamment l'identité du locuteur, le contexte (lecture à voix haute ou parole *spontanée*), le stress (émotivité), la syntaxe de la phrase et bien d'autres. Celles-ci peuvent s'observer sous différentes formes, par exemple au niveau des fréquences dominantes (*formants*) d'un son voisé (qui utilise les cordes vocales ; ex : /a/, /o/ et /i/), du rythme d'élocution et de la suppression d'objets acoustiques (omission de sons en fonction de l'élocution). Tout ceci confère à la parole une complexité difficile à gérer pour un algorithme de reconnaissance artificielle, mais qui n'est pas un problème pour l'audition humaine. Pour cette raison, une tendance populaire dans le domaine de la reconnaissance vocale consiste à comprendre et à imiter le traitement fait par l'oreille.

L'oreille modifie notamment la différence entre les fréquences dans sa perception. En effet, la distance perçue entre les différents niveaux de *hauteur tonale* (Caractère subjectif d'un son par lequel on donne une place sur une échelle de perception des fréquences qui est dite échelle de *tonie*), varie de façon non-linéaire avec les fréquences. L'échelle *MEL* est souvent utilisée pour caractériser la relation entre hauteurs tonales (exprimées en MELS) et fréquences. Cette échelle de perception fréquentielle est couramment utilisée dans l'extraction de coefficients caractéristiques à la parole, notamment dans les *coefficients cepstraux MEL* (MFCC).

L'étude de la perception auditive nous enseigne également que l'oreille tend, particulièrement en perception vocale, à accorder plus d'importance à l'enveloppe temporelle du signal qu'à son contenu fréquentiel exact. Les expériences de Shannon [Shannon et al., 1995], par

exemple, ont démontré que même si un signal de parole est synthétisé avec seulement quatre (4) sous-bandes fréquentielles différentes et que l'enveloppe temporelle dans ces sous-bandes est conservée, un auditeur serait généralement en mesure de comprendre correctement une phrase. Ceci semble démontrer que la dynamique du signal acoustique est très importante à l'oreille et est peut-être même plus significative que le contenu fréquentiel exact.

Certaines expériences en reconnaissance automatique de parole *spontanée* ont également révélé que les humains ont une tendance naturelle à prendre des raccourcis (omettre des sons importants) à l'élocution [Greenberg, 1999]. Ceci démontre à la fois la richesse temporelle des signaux de parole et le fait que sa compréhension nécessite l'apprentissage de nombreuses dynamiques acoustiques différentes.

2.2 Introduction à la reconnaissance vocale

2.2.1 Extraction des coefficients

La reconnaissance vocale est divisée en deux processus distincts : l'extraction des coefficients et la classification. La parole enregistrée et non-identifiée est d'abord décomposée en *coefficients caractéristiques*. Ceux-ci sont conçus pour extraire l'information utile des signaux acoustiques et la soumettre au classificateur sous un format prédéfini. Nombreux sont les modes d'extractions des coefficients, mais le plus populaire est sans aucun doute celui des *coefficients cepstraux MEL* (« *MEL Frequencies Cepstral Coefficients* » ou MFCCs) [Fang *et al.*, 2001]. Ceux-ci sont généralement représentés comme une série de points dans un espace à 39 dimensions et échantillonnés à 10 ms d'intervalle. Chacun de ces points est calculé à partir d'une *trame*, soit un très court signal sonore qu'on suppose dépourvu de dynamique.

2.2.2 Classification

La tâche d'un classificateur en reconnaissance vocale se résume généralement à évaluer lequel de ses modèles possède le plus d'affinité avec le signal de parole inconnu présenté à l'entrée. Ceci représente généralement tout un défi. Comme présenté dans l'introduction au traitement de la parole, les différentes *énonciations* (exemples d'une même classe) sont très différentes les unes des autres en raison de variabilité fréquentielle (contenu en fréquences des sons) et temporelle (dynamiques acoustiques différentes). Depuis près de 40 ans, le

modèle de prédilection est le *modèle markovien caché* (« *hidden Markov model* » ou HMM [Rabiner, 1989]).

En reconnaissance vocale par *mots* (« *word-based speech recognition* »), ces modèles représentent différents mots. Ce genre d'approche est généralement réservé aux applications à court *dictionnaire*, c'est-à-dire des implémentations où seuls quelques mots doivent être reconnus. À l'inverse, une application à large dictionnaire requiert généralement la capacité de reconnaître tous les mots d'un dialecte. Ce genre de problème ne peut pas être abordé par une reconnaissance à base de mots en raison de limitations computationnelles. La reconnaissance de *phonèmes* (« *phone-based speech recognition* ») est alors utilisée. Le phonème, ou *monophone*, est l'unité phonétique de base d'un dialecte et le nombre total de variations possibles est généralement assez limité. En Anglais, par exemple, on dénombre 44 phonèmes. En reconnaissance vocale à large dictionnaire, les classificateurs font généralement usage de monophones ou de *triphones* soit des modèles acoustiques de trois monophones consécutifs.

2.2.3 Le modèle markovien caché (HMM)

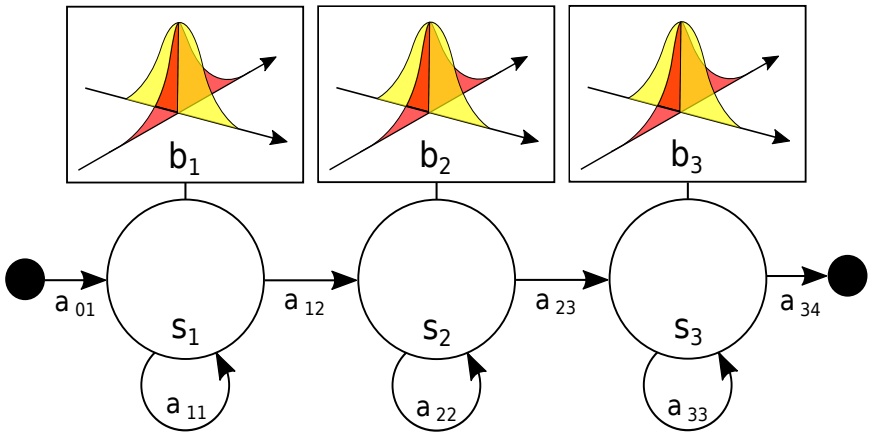
Que ce soit un HMM à réseau de neurones profond (« *Deep Neural Network Hidden Markov Model* » ou *DNN-HMM* [Mohamed et al., 2012]) ou un HMM à mixture de gaussiennes (« *Gaussian Mixture Model Hidden Markov Model* » ou *GMM-HMM*), tous les modèles Markoviens cachés utilisent la même structure de fonctionnement de base. Celui-ci est un modèle qui peut être *entraîné* à imiter un processus et qui peut permettre d'évaluer la *vraisemblance* que celui-ci ait généré une séquence d'observations particulières. Dans le domaine de la reconnaissance vocale, ce processus est la *classe* (mot, syllabe, phonème). Selon le théorème de Bayes, la probabilité à posteriori qu'un signal appartienne à une classe particulière est déterminée par :

$$P(C_i|x) = \frac{P(x|C_i)P(C_i)}{P(x)} \tag{2.1}$$

Où $P(x|C_i)$ est la vraisemblance que x a été généré par C_i et $P(x)$ et $P(C_i)$ sont les probabilités à priori d'observer le signal et la classe, respectivement. Dans ce contexte, on utilise le modèle HMM comme modèle de la vraisemblance $P(x|C_i)$ que le signal ait été généré par la classe C_i.

L'utilisation d'un HMM se fait toujours en deux phases distinctes : l'*entraînement* et le *décodage*. À l'entraînement le modèle est adapté aux signaux qu'il doit modéliser. Au dé-

Figure 2.1 Architecture classique d'un modèle de Markov caché



L'organisation des états et transitions (architecture) d'un modèle de Markov peut varier grandement. Celle qui est exposée dans cette figure est la forme la plus utilisée, que l'on appelle « *left-to-right* ». s_i désigne l'état i et a_{ij} représente la transition de l'état i à j. Finalement, b_i est le modèle d'émission de l'état i, représenté dans la figure comme une distribution de densité de probabilité (« *probability density function* » ou *pdf*) à une seule dimension. La *pdf* illustrée pour b_1, b_2 et b_3 est une *mixture de gaussiennes* (« *Gaussian mixture model* » ou *GMM*) à une (1) composante sur deux (2) dimensions.

codage, celui-ci permet de reconnaître des signaux similaires à ceux qui ont étés présentés à l'entraînement à partir de l'équation (2.1) qui permet de trouver la probabilité à postériori.

Au niveau structurel, le HMM est constitué de plusieurs *états* stationnaires et de *transitions* entre ces états. Chaque transition possède une probabilité statique d'être empruntée. Le fonctionnement est simple : à chaque instant t un seul état est occupé et à chaque moment $t+1$ une transition est faite vers un autre (ou vers le même si une transition le permet). La figure (2.1) illustre un exemple classique de HMM.

À chacun des moments t une *observation* est générée par l'état occupé. Cette génération est aléatoire et gouvernée par un *modèle d'émission* propre à l'état occupé. L'*observation* est un vecteur de *coefficients caractéristiques*, généralement un vecteur MFCC à 39 dimensions. La clé du pouvoir modélisateur du HMM est dans le fait que l'on ne sait jamais quel état est occupé au moment t et donc tous les états peuvent être occupés avec une certaine probabilité (qui peut cependant être de 0% ou 100% dans certain cas). À l'*entraînement*, les probabilités de transitions et les modèles d'émission sont adaptés de façon à représenter correctement le comportement des signaux appartenant à la classe pour laquelle on estime la vraissemblance. Ceci se fait à l'aide de signaux d'exemple et d'un algorithme de

réestimation comme celui de *Baum-Welch* [Rabiner, 1989]. Au *décodage*, un algorithme tel que celui de *Viterbi* [Rabiner, 1989] permet de trouver la plus probable séquence d'états (et sa probabilité cumulative) associés à une série de *trames* non-identifiées (dont on ne connaît pas la classe) obtenue en entrée.

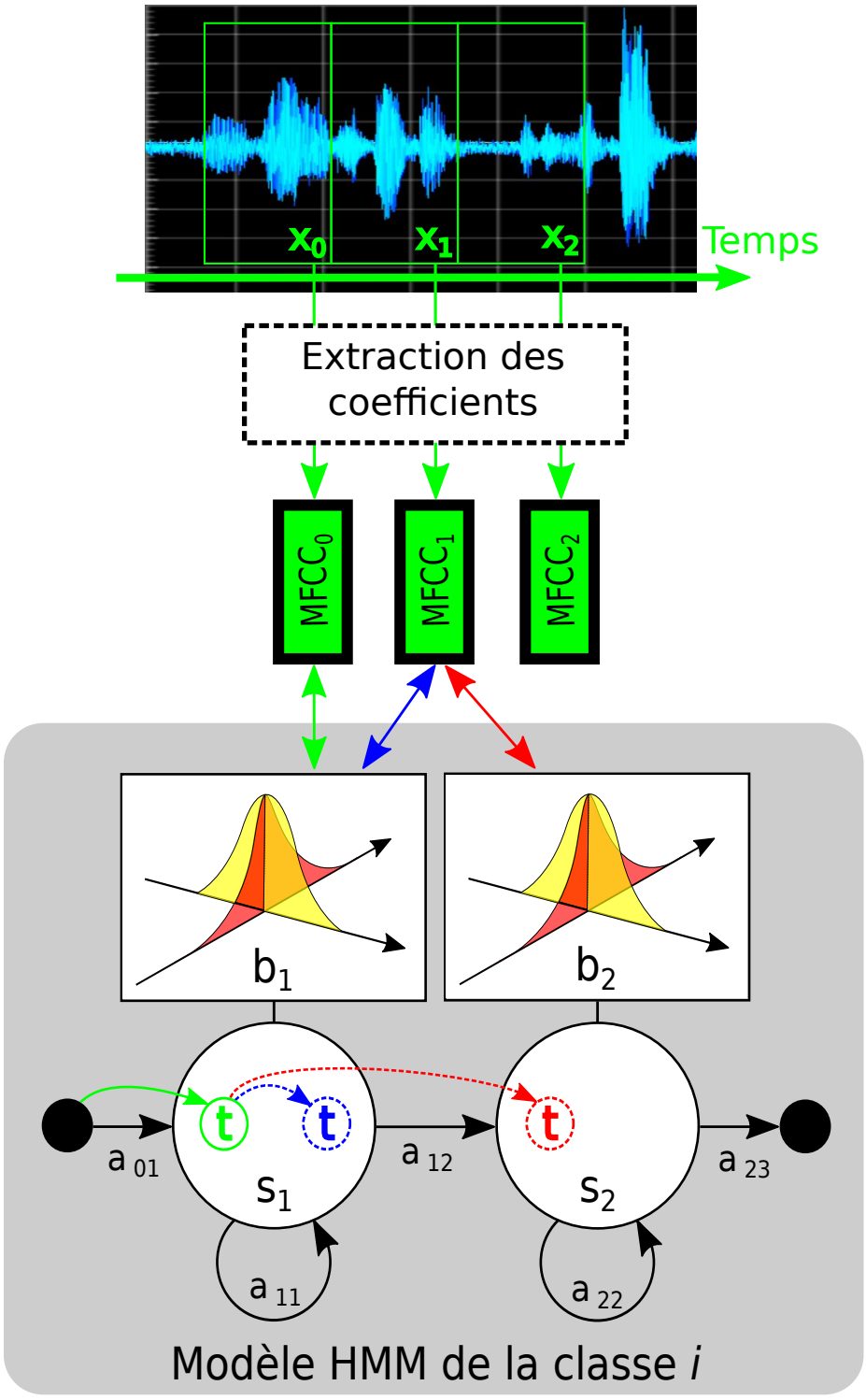
La figure (2.2) représente le fonctionnement haut-niveau du décodage d'un signal acoustique par rapport à un modèle HMM. L'extraction des coefficients est d'abord faite de façon systématique, c'est-à-dire qu'un vecteur MFCC est calculé pour chaque *trame* f_x. Il convient ensuite d'associer ces *observations* avec leur état approprié.

Dans cette association, ou *décodage du chemin emprunté*, deux probabilités distinctes sont considérées : les probabilités de transition et d'émission. Les probabilités d'émission sont calculées à partir des modèles d'émission b_x. Dans le cas classique des GMM-HMM, employant des modèles d'émission composés de mixtures de fonctions gaussiennes (« *Gaussian mixture models* » ou GMMs), ceux-ci sont de simples *fonctions de densités de probabilités* (« *probability density fonction* » ou *pdf*). Ces densités de probabilités, existant dans un monde à 39 dimensions, peuvent fournir une probabilité non-nulle pour tout vecteur MFCC. Ainsi, chaque état d'un GMM-HMM possède une *pdf* unique sensible (c.-à-d. qu'elle donne une haute probabilité) à un ou des sons *stationnaires* (dont la dynamique ne change pas dans le temps).

Les modèles d'émissions ne sont pas restreints à des constructions totalement stochastiques comme les *pdfs*. Dans le cas des HMM à *réseaux de neurones profonds* (« *Deep Neural Networks HMMs* » ou *DNN-HMMs*), l'état de l'art actuel en reconnaissance vocale, les modèles d'émission sont des réseaux de neurones générant une probabilité [Mohamed et al., 2012].

Les probabilités de transition sont complètement invariables au *décodage*. Toutes les transitions *quittant* le même état ont une valeur de probabilité complémentaire, c'est-à-dire leur somme est toujours égale à 100%. Ainsi donc, un état quitté par une seule transition doit toujours être quitté par celle-ci. Le réseau qui comprends des états et des transitions forment ce qu'on appelle la *topologie*, ou l'*architecture* du modèle. Cette *topologie* est absolument immuable au *décodage* et l'est souvent également à l'*entraînement* (à l'exception des valeurs de probabilités). La richesse dynamique des signaux acoustiques modélisés est majoritairement encodée dans cette topologie.

Figure 2.2 Schéma de fonctionnement général du décodage *token passing*



Le signal acoustique est représenté par le graphique du haut sous la forme d'une tension électrique variant dans le temps. Les x_x désignent les trames de signal acoustique. Le modèle HMM décodé, représentant la classe i, est représenté par la zone grise. La variable b_x désigne le modèle d'émission propre à l'état s_x. Les a_x représentent les transitions unidirectionnelles empruntables entre tout moment t et $t+1$. Les points noirs pleins sont les points d'entrée ($t=0$) et de sortie ($t=T$) du HMM. Finalement, le jeton vert (marqué d'un t) représente le seul et unique chemin possible entre $t=0$ et $t=1$. Les jetons bleu et rouge sont les 2 chemins possibles entre $t=1$ et $t=2$.

2.3 Introduction au décodage d'un HMM

En ce qui a trait au décodage, différents algorithmes existent et tous ne fonctionnent qu'à partir des probabilités de transition et d'émission. Il faut également comprendre qu'en reconnaissance vocale tous les problèmes ne sont pas équivalents et donc différents algorithmes sont disponibles. Le plus simple de ces problèmes est celui de la reconnaissance de *classes uniques isolées*. Dans ce cas, on sait à priori que le signal en entrée ne contient qu'une seule classe (mot, syllabe, phonème, etc.) et que celle-ci est entièrement contenue dans ce signal. Ce contexte est propice au décodage par calcul de la *probabilité totale d'émission* (probabilité tenant compte de tous les chemins possibles à travers le HMM) par l'algorithme récursif *forward-backwards* [Rabiner, 1989; Young et al., 1995]. Cet algorithme garantit une solution optimale. Malheureusement, ce type de problème est rare et peu d'applications concrètes s'y rattachent.

Un peu plus difficile est la reconnaissance de *classes multiples isolées*. Contrairement au cas précédent, on sait à priori que l'échantillon acoustique peut contenir un nombre indéterminé de classes à identités inconnues. On sait cependant que les classes sont entièrement contenues dans le signal d'entrée (aucune n'est coupée par le début ou la fin de l'enregistrement). Ce genre de problème est beaucoup plus fréquent ; on pense notamment à des commandes simples vocalisées dans un laps de temps prédéfini. L'efficacité de ce genre de solution est cependant moindre puisque la coupure d'une classe (vocalisation pendant le début ou la fin de l'enregistrement) est souvent fatale à la reconnaissance. Ce genre d'application est généralement réservée au domaine du prototypage. Au décodage, l'algorithme de *Viterbi* [Rabiner, 1989; Young et al., 1995] est généralement utilisé.

L'application la plus commune de reconnaissance vocale a trait à un problème encore plus difficile que celui des *classes multiples isolées*. Il s'agit de la reconnaissance *continue de classes*. Contrairement au cas précédent, des classes peuvent être coupées par la fin de l'enregistrement. Pour répondre à cette contrainte supplémentaire, une modification majeure de l'algorithme de *Viterbi* est nécessaire. Cette modification réduit significativement l'optimalité de la procédure de décodage. Cette implémentation de l'algorithme de *Viterbi* est connue sous le nom de « *token passing* ».

De façon formelle, l'implémentation « *token passing* » de l'algorithme de Viterbi spécifie que l'état $s(t+1)$ occupé au temps $t+1$ est donné par :

$$s(t+1) = \underset{k=1,2,\ldots,N}{\operatorname{argmax}} [a_{s(t)\to k} b_k(O_{t+1})] \tag{2.2}$$

Où N est le nombre total d'états émetteurs dans le modèle HMM, O_{t+1} est le vecteur d'observations obtenu au temps $t+1$, b_k est le modèle d'émission de l'état k et $a_{s(t)\to k}$ est la probabilité de transition de l'état occupé au temps t vers l'état k. La figure (2.2) explique le fonctionnement général de *token passing* dans sa forme la plus simple, c'est-à-dire sans considérer les autres classes du classificateur. En commençant à l'instant $t = 0$, le point d'entrée émet un jeton à tous les états atteignables selon les transitions. Dans la figure (2.2) il s'agit du jeton vert. Ensuite, et pour tout moment successif, chaque jeton existant émet un autre jeton pour chaque transition quittant l'état occupé. Chaque jeton possède une valeur de probabilité. Celle-ci se calcule comme le produit de la probabilité de son jeton « géniteur », de la probabilité de transition et de la probabilité d'émission de l'état atteint par rapport à l'observation au moment $t+1$.

2.4 La topologie d'un HMM

Peu importe le type de modèle d'émission, qu'il soit purement stochastique comme les GMM-HMMs ou hybride (en partie stochastique) comme les DNN-HMMs [Mohamed et al., 2012], la topologie des HMMs dans les systèmes ASR de l'état de l'art n'a pas changé depuis les années 80. La structure utilisée porte le nom de *gauche-à-droite* (« *left-to-right* » ou L2R).

Gauche-à-droite est une topologie très facilement reconnaissable : tout état n'a que 2 transitions le quittant : une allant vers le prochain état et l'autre retournant sur le même état. Le HMM illustré à la figure (2.2) possède cette architecture. Celle-ci est très simple et est uniformément utilisée pour toutes les classes d'un classificateur. La plupart du temps, les différents modèles ont exactement le même nombre d'états. Ce faisant, ceux-ci ont aussi la même structure. Considérant que tous les mots ou phonèmes n'ont pas du tout la même dynamique acoustique, ceci est bien particulier et très contraignant.

En fait, le problème vient du fait que bien qu'un HMM soit entraîné sur des signaux, la topologie est généralement fixée à priori par le concepteur et celle-ci est immuable. Il faut comprendre que trouver la topologie appropriée à une classe est difficile. Les travaux associés à ce genre d'amélioration ne présentent généralement pas d'améliorations de *précision de reconnaissance* (« *recognition accuracy* »). Ainsi donc, il est coutume d'utiliser la topologie *gauche-à-droite* puisqu'elle donne des résultats satisfaisants et que développer des architectures temporelles plus précises ne semble pas avoir d'impact.

2.5 Vers l'utilisation de topologies plus précises

Il est expliqué dans l'article au chapitre suivant comment la topologie *gauche-à-droite* est fonctionnelle puisqu'elle est peu précise et qu'elle est donc très tolérante aux particularités dynamiques de la parole. En ce qui a trait à l'apprentissage de la topologie, ou au développement de topologies plus précises, le manque de résultats concluant semblerait pouvoir être lié à une limitation intrinsèque aux HMMs. Expliqué par Rabiner comme un problème de débalancement entre les plages dynamiques des probabilités de transition et d'émission [Rabiner et Juang, 1992], les auteurs démontrent expérimentalement l'influence de ce phénomène sur le décodage par *token passing*. Ce débalancement peut expliquer, du moins en partie, pourquoi certaines topologies plus complexes sont en fait moins précises que *gauche-à-droite*. Finalement, les auteurs proposent une nouvelle méthode d'entraînement de topologies visant à atteindre une modélisation de plus haute précision temporelle.

CHAPITRE 3

VERS LA MODÉLISATION DE LA PAROLE PAR MODÈLES DE MARKOV CACHÉS À HAUTE PRÉCISION DYNAMIQUE

3.1 Avant-propos

L'article présenté dans cette section expose en détail l'entièreté des travaux effectués.

Premier auteur : Sébastien Gagnon, étudiant

Second auteur : Jean Rouat, Ph. D., professeur titulaire à la Faculté de génie électrique et informatique à l'Université de Sherbrooke. Membre fondateur du groupe de recherche NECOTIS

Date de soumission : 15 juillet 2016

Journal : ELSEVIER COMPUTER SPEECH AND LANGUAGE

3.2 Résumé de l'article

Dans cet article, nous proposons que des topologies finement adaptées de modèles markoviens cachés (HMMs) soient essentielles à une modélisation temporelle de haute précision. Nous soumettons donc un cadre conceptuel pour l'apprentissage efficace de topologies par élagage de modèles génériques complexes. Des modèles HMM à topologie *gauche-à-droite* sont d'abord entraînés de façon classique.

Des modèles complexes à topologie générique sont ensuite obtenus par « *écrasement* » des modèles *gauche-à-droite*. L'« *écrasement* » est une technique introduite par [Zhao et Juang, 2012] permettant de transformer un modèle GMM-HMM simple en sa forme complexe en formalisant les transitions intrinsèques aux modèles d'émission GMM.

Finalement, un enchaînement successif d'« *élagages* » et d'entraînements Baum-Welch est fait de manière à augmenter la précision temporelle des modèles.

Des expériences de reconnaissance vocale menées sur la base de données Aurora-2 démontrent une précision de reconnaissance (« *recognition accuracy* ») accrue en conditions propres (« *clean conditions* », c'est-à-dire sans ajout de bruit) et *réverbérées*. Ces résultats semblent démontrer que la technique proposée permet le développement de modèles à plus haute précision temporelle. Ceci semble d'autant plus vrai considérant que la *réverbération* introduit des dépendances temporelles aux signaux de parole et augmente donc leur complexité dynamique.

3.3 Citation de l'article

3.3.1 abstract

Hidden Markov Model (HMM) is often regarded as the dynamical model of choice in many fields and applications. It is also at the heart of most state-of-the-art speech recognition systems since the 70's. However, from Gaussian mixture models HMMs (GMM-HMM) to deep neural network HMMs (DNN-HMM), the underlying Markovian chain of state-of-the-art models did not changed much. The "left-to-right" topology is mostly always employed because very few other alternatives exist. In this paper, we propose that finely-tuned HMM topologies are essential for precise temporal modelling and that this approach should be investigated in state-of-the-art HMM system. As such, we propose a proof-of-concept framework for learning efficient topologies by pruning down complex generic models. Speech recognition experiments that were conducted indicate that complex time dependencies can be better learned by this approach than with classical "left-to-right" models.

3.3.2 Introduction

The correctness of a hidden Markov model's (HMM) topology can strongly influence the model accuracy of a HMM systems, especially for signals with high dynamic variability. This graphical architecture is usually hand-designed to a simple and generic form (usually shared across all classes), whereas constructing precisely tuned class representations can be challenging.

In the 70's a "left-to-right" topology was first proposed for speech modelling, meaning that feature changes through time always flowed in a specific sequential order [Jelinek, 1976]. It is however simplifying considering that spontaneous speech dynamics are known to be very variable [Greenberg, 1999]. Up to these days, most state-of-the-art ASR systems such

as *deep neural networks*-HMMs (DNN-HMMs, [Mohamed et al., 2012]) are still based on that architecture.

As states of HMMs encode static feature space distributions, simple HMM topologies can only model coarse dynamics. A dynamic process constructed from static events is as detailed as the number of such events. Too little precision results in an *underfitted* model with low discriminative power in classification systems. Too much precision, however, can lead to an *overfitted* model without generalizing power [Cawley et Talbot, 2010], making it unable to recognize anything but the training signals. In model selection, as discussed in [Cawley et Talbot, 2010], the key is to balance precision and generalization for maximum performance.

Robustness, the ability of a system to tolerate recording environment changes, is also to be considered and seems to be strongly related to a model's generalizing power [Xiao *et al.*, 2010]. As such, improving model precision would decrease robustness. *Online adaptation* techniques, however, can easily compensate for such a drawback. Based on a more comprehensive approach, these methods deal with unseen noise by modifying model statistics at testing time [Kalinli *et al.*, 2010; Li *et al.*, 2007; Narayanan et Wang, 2015]. Performance of such systems in noisy conditions are quite remarkable.

Even if "left-to-right" topologies are good for speech signals, some datasets with higher temporal complexity need more precise architectures. The fact, however, is that topologies are usually hand-designed and kept simple, following Occam's razor principle [MacKay, 2003]. According to this heuristic, in a setting where a high number of alternative models exist, simple solutions are better than complex ones since less potentially wrong assumptions are made. Therefore, the main goal of this work is to provide a framework to automatically learn precise HMM topologies from data.

"Left-to-right" *Hierarchical Dirichlet process*-HMM (HDP-HMM, [Torbati *et al.*, 2014]), a recently developed technique, is one example of such a framework ; it is however unbound by Occam's razor principle or any strong underfitting/overfitting criterion. The expected size of the learned topology being dependant on a *concentration* parameter chosen *a priori*, the designer has indirect control over the degree of temporal precision. Furthermore, the model is allowed to both increase or decrease its size according to a Dirichlet process and is thus unconstrained to follow Occam's principle [MacKay, 2003]. On the other hand, this pioneer work allow for the development of HMM speech models without any *a priori* knowledge of the dynamics, something that is not possible with other approaches.

While adept at learning complex topologies from data, “left-to-right" HDP-HMM is oblivious to recognition accuracy and how the architecture influences it. Like most HMM training procedures, this is caused by a mismatch between training and decoding alignement methods, i.e. *forward-backward* vs *Viterbi*. In [Torbati *et al.*, 2014], for example, some learned monophone topologies allow decoded paths to be as short as 2 time frames long whereas in standard monophone systems the shortest path is 3 time frames. According to our preliminary experiments, this tend to generate insertion errors when decoding, enough to significantly lower accuracy as defined by the word error rate (WER) standard. As shown in TIMIT benchmarks listed in [Lopes et Perdigao, 2011], considering insertion errors always significantly decreases the recognition performance of a system. Thus, while useful in approximating the topology needed for each class, “left-to-right" HDP-HMM procedure alone is not ideal for our intended goal.

Conventional approaches to dynamics encoding with HMMs usually substitute topology learning with transition probabilities estimation. In speech recognition, this is the most popular paradigm : generic “left-to-right" architectures are adapted to target signal’s dynamics by tuning a few persistent parameters. In [Zhao et Juang, 2012], such an approach is attempted on a complex generic topology with improved additive noise robustness. However, clean speech performance are not reported, which may suggests that precision has not improved. In fact, improved robustness can be linked to a greater generalizing power, as explored in [Xiao *et al.*, 2010].

These results might be explained by an intrinsic problem of HMMs, the imbalance between the dynamic ranges of the transition and emission probabilities. Exposed by Rabiner and Huang in [Rabiner et Juang, 1992], this phenomenon is at the root of the popular thought that transition probabilities are almost useless. It is even a common practice for designers to implement HMM ASR systems with untrained transitions, because the loss in performance is fairly small. Explained in [Rabiner et Juang, 1992] as a lack of pervasive discriminative power of the transition probabilities in path decoding, we conceptualize its effects as rendering equiprobable all transitions that leave the same state. Thus, tuning transition probabilities cannot be a good substitution to complex topology learning.

In this work, we first analyse the effects of the imbalance phenomenon. We show that all paths leaving the same state are effectively equiprobable in the standard TIMIT monophone recognition experiment. Thus, topology learning is shown essential for precise dynamics modelling, for which we then propose a simple and accessible framework. Assuming that HMM spoken word models in conventional ASR systems are closer to underfitting than overfitting (a reasoning we based on [Greenberg, 1999]), we propose to use model

flattening [Zhao et Juang, 2012] in conjunction with *transitions pruning* to extract precise class topologies. *Flattening* is the process of transforming a simple "left-to-right" *Gaussian mixture model*-HMM (GMM-HMM) into an equivalent complex HMM with single Gaussian emission models. Using *transitions pruning* to reduce the flatten model complexity then reveals a more precise dynamic model while still following Occam's razor principle. We finally demonstrate that with the same number of emission model parameters, our technique clearly outperforms the classic "left-to-right" topology on clean word recognition tasks.

3.3.3 Transition and Emission Probabilities Imbalance

As discussed in [Rabiner et Juang, 1992], transition probabilities may not play a significant role in path decoding (using Viterbi) ; recognition could be entirely independent of them. Then, all transitions leaving the same state could be considered effectively equiprobable during path decoding. To the knowledge of the authors, this phenomenon has not been quantitatively documented for speech recognition. In the *token passing* implementation of the Viterbi algorithm, the state $s(t+1)$ occupied at time $t+1$ is given by [Young *et al.*, 1989] :

$$s(t+1) = \underset{k=1,2,...,N}{\operatorname{argmax}} [a_{s(t)\to k} b_k(O_{t+1})] \tag{3.1}$$

Where N is the total number of emitting states in the model, O_{t+1} is the observation vector obtained at time $t+1$, b_k is the emission model for state k and $a_{s(t)\to k}$ is the transition probability from the state occupied at time t to the state k. Let there be a distinction between zero and non-zero transition probabilities :

$$\forall (i,k) \in [1,...,N]; a_{i\to k} \neq 0 \to k \in \varphi_i \tag{3.2}$$

Were φ_i regroups all non-zero transition probabilities leaving state i. Thus, (3.1) can be reformulated in the following fashion :

$$s(t+1) = \underset{k\in\varphi_{s(t)}}{\operatorname{argmax}} [a_{s(t)\to k} b_k(O_{t+1})] \tag{3.3}$$

Equation (3.3) only takes into account states that are linked by a non-zero transition from $s(t)$. Formula (3.3) can be formulated as follows :

$$s(t+1) = i \;\; if \;\; \frac{b_i(O_{t+1})}{b_k(O_{t+1})} > \frac{a_{s(t)\to k}}{a_{s(t)\to i}}, \forall (i,k) \in \varphi_{s(t)}, k \neq i \tag{3.4}$$

$$\text{Lets define : } \beta = \frac{b_i(O_{t+1})}{b_k(O_{t+1})} \;\; \alpha = \frac{a_{s(t)\to k}}{a_{s(t)\to i}}$$

We defined α and β as ratios of probabilities to isolate the respective transition and emission discriminability forces. The variance of these variables, respectively the transition and emission discriminability coefficients, give a good estimate of their dynamic ranges. To evaluate them, we conducted an experiment on the TIMIT training set with conventional 5-states (3 emitting states) “left-to-right" monophonic models with Gaussian mixture models (GMMs) of 16 components on each state. The procedure is done in 2 steps for each training utterance : first, using the appropriate models (listed in the signal's label) an ideal path is computed with the *forward-backward* algorithm [Rabiner, 1989]. Then, for each transition taken in the decoded path the α and β values are computed. The variances are then estimated across all the training set. The path alignment method used here, forward-backward, is not equivalent to (3.4). In fact, Token passing does not take the backward probability into account and is therefore less optimal. This was done purposefully to favor high emission probabilities in an effort to minimize discriminative power. One must understand that strong model mismatch comes from emission probability values being several orders of magnitude different from one state to another, which far more happens in low probabilities (in mismatched dynamics). In other words, the emission discriminability coefficient calculated is minimized to a level unattainable in practical applications.

$$\sigma^2(ln(\alpha)) = 0.80 \;\; \sigma^2(ln(\beta)) = 193.24$$

Where $\sigma^2(i)$ is the variance of i. In the linear domain, the standard deviation of β is roughly 440,000 times larger than α. Thus, the transition probabilities are in some sense binary variables, i.e. they are, or not, members of $\varphi_{s(t)}$ in (3.4).

This is because emission probabilities have a near-infinite dynamic range, while transition probability do not, for any given $s(t)$. In fact, considering how this problem is exposed in [Rabiner et Juang, 1992], we infer that only in topologies with states of near-infinite branching ratios may this imbalance vanishes. It is therefore safe to assume that all discrete topologies considered in this work are equally affected by this imbalance.

3.3.4 Pruning

Encoding acoustic dynamic properties in a generic HMM model is to change its topology, i.e. by activating or deactivating transitions. A deactivated transition has a probability value of 0 and is therefore not involved in (3.4).

Learning the topology can be done in 3 ways : either “growing" from a simple prototype model (ex. [Jitsuhiro et Nakamura, 2004]), “pruning" from a complex generic model (ex. [Mak et Chan, 2005]) or a mixture of both (ex. [Torbati *et al.*, 2014]). With “growing" techniques, i.e. increasing the model's complexity, an almost groundless guess must be made to determine how the expansion is done. This is very much subject to human error.

On the other hand, “pruning" processes are much more reliable as one removes only the paths that are not often visited. Mak and Chan [Mak et Chan, 2005], for example, have successfully used pruning on a “left-to-right" topology with long range transitions. When compared with an unpruned system, they obtained a significant improvement on the accuracy in a clean word recognition task. Our work follows that line of thinking and implements pruning instead of other alternatives.

3.3.5 Proposed System

Integration of the Pruning Module and Threshold Optimization

A modification of the standard HMM training procedure is proposed for increasing the temporal modelling precision. Fig. 3.1 illustrates the full proposed system. The pruning (step #5) is done by comparing each individual transition probability with a threshold value, ϵ :

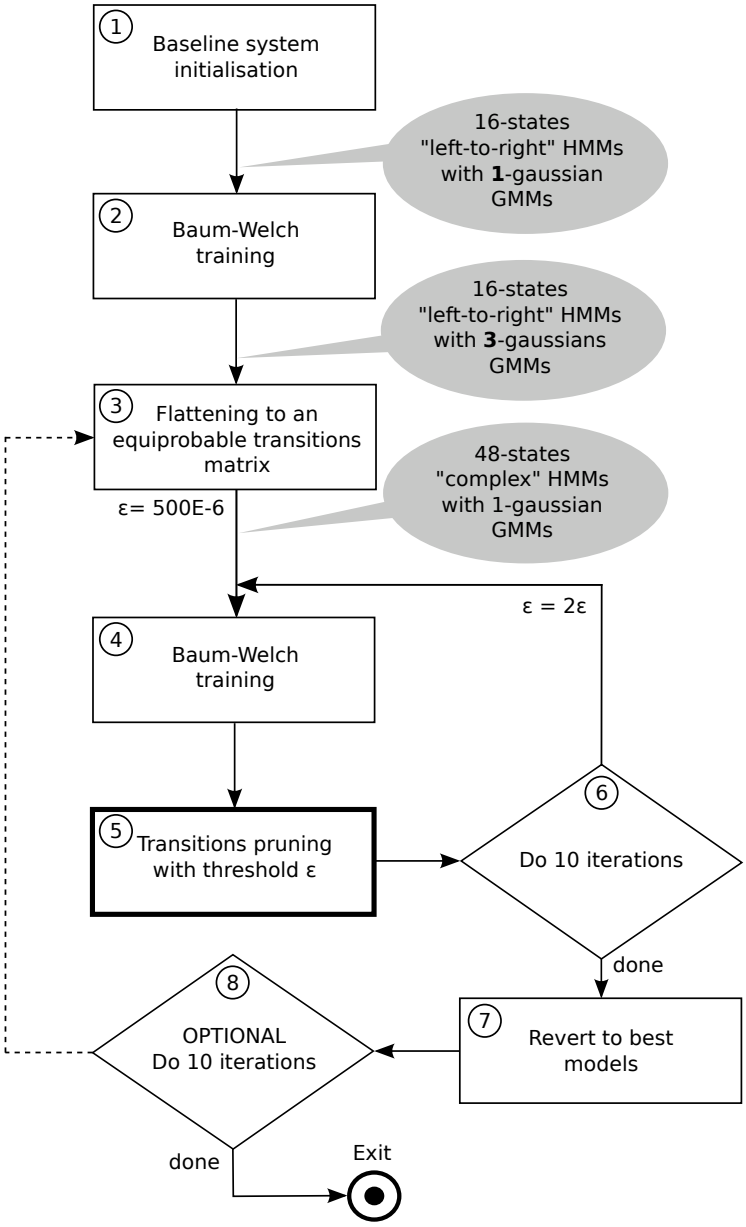
$$\text{if } a_{i \to j} > \epsilon \text{ then keep, else } a_{i \to j} = 0 \tag{3.5}$$

Since the value of ϵ is unknown, optimization steps are required during training to find ϵ (using the loop step #6 in Fig. 1). The very simple ϵ optimization process is designed as such to exploit the steady relation between pruning threshold and performance, thus avoiding unpredictable local minimums.

Initialization by Model Flattening

To maximize the beneficial impact of transition pruning we work on a complex prototype model (high number of states and transitions). However, this can be difficult since the transition parameter space is larger than with simple “left-to-right" topologies. Furthermore, if it is improper, alternate paths tend to die off during training (i.e. very low occupancy

Figure 3.1 Proposed system diagram



1) Flat initialisation ; 2) Conventional Baum-Welch training with GMM mixture splitting ; 3) Flattening process on the baseline GMM-HMM, transition matrix is equiprobable on the allowed transitions ; 4) Baum-Welch training without GMM mixture splitting ; 5) Pruning as per (5) ; 7) The model with highest decoding accuracy on the training set is kept, all the others are discarded ; 8) (Optional) Emission models feedback : emission models of the model kept in 7 are given to their respective states in the flatten model in 3. All pruned transitions are thus reactivated.

probability) to only favour a single path through the topology. This effectively returns the model to a long "left-to-right" chain with mediocre performance. Thus, the initialization of a complex HMM model is an important aspect of this work, for which the *flattening* technique presented in [Zhao et Juang, 2012] is used. Since a multi-gaussian mixture model can be viewed as an HMM of single-gaussian states, the flattening consists in replacing the states of a "left-to-right" GMM-HMM by their respective HMM's form. This effectively flattens the representation to a lattice of mono-gaussian densities, as illustrated in Fig. 2. The final form of a trained "left-to-right" model is already fine-tuned to the acoustic properties of the modelled class, it is therefore an ideal configuration.

Feedback of Emission Models

To further increase the recognition performance of the proposed system, the emission models of the pruned HMMs are fed back to the initialization step (link between steps #8 and #3 in Fig.1) :

$$\forall i \in [1, ..., N]; b_i^{pre}(\cdot) \leftarrow b_i^{post}(\cdot) \tag{3.6}$$

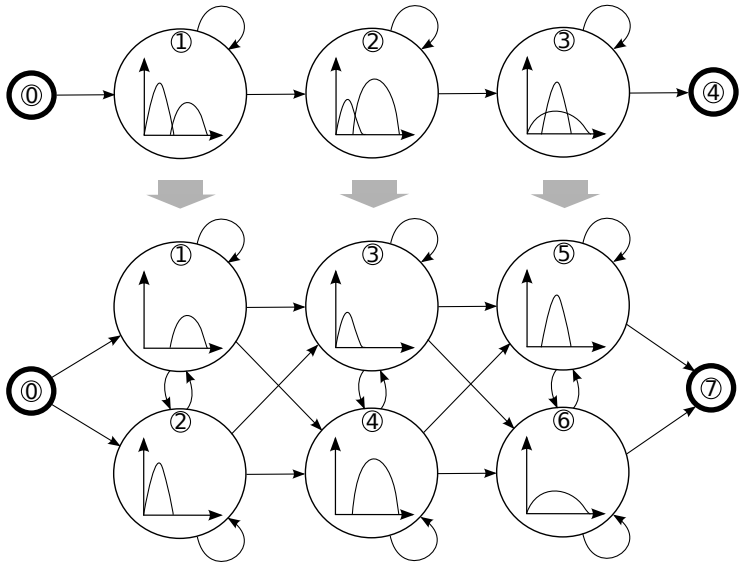
Where $b_i^{pre}(\cdot)$ and $b_i^{post}(\cdot)$ are respectively the emission distributions of the state i at initialization and after training. As complex models require more care to train adequately because of their vast transition parameter space, this step is added to allow slower convergence. We observed experimentally that with about 10 iterations of this "feedback" process (step #8 in Fig. 1) the recognition performance on the training set seems to saturate. While this step is not required to beat baseline accuracies on clean word recognition, on average it increases performance even further (Table 3.1). However, it also sharply increases the computational effort required for training.

3.3.6 Experimental Framework

To evaluate the temporal modelling precision, we chose to view it in terms of recognition accuracy. Our premise is that classical "left-to-right" modelling of speech dynamics is closer to *underfitting* than *overfitting* and therefore improved precision should yield better results. According to [Greenberg, 1999], this is most likely true for spoken word models. Furthermore, clean speech accuracies (the tested signal was recorded in the same conditions as the training signals) are the main focus since, as explored in [Xiao *et al.*, 2010], noisy recognition seems to deal more with generalizing power than precision.

For all conducted tests, we want any increase in recognition accuracy to be entirely attributed to the higher temporal modelling precision. This is done by ensuring that the total amount of Gaussian distributions for each HMM was the same, i.e. the summed

Figure 3.2 Flattening process



The top “left-to-right" model has a 2 Gaussian mixture model for each state; the bottom “complex" model is a flatten version of the top model with only 1 Gaussian per state.

amount of GMM mixture components across each model is identical for both the reference and the proposed systems.

First, large dictionary speech recognition applicability is evaluated with a word recognition task with monophonic models on TIMIT. The baseline “left-to-right" 5-states (3 emitting states) models have 10 Gaussian mixture components per state. The proposed system uses flatten versions of the “left-to-right" models and are thus 30 states long with single-Gaussian GMMs.

Next we tested on the Aurora-2 digits word classification task. Since monophonic models should not posses much temporal structure, by linguistic definition, a word recognition task is thought to be more representative of the difference in temporal modelling precision (word models encode much richer dynamics than monophones). The baseline “left-to-right" HMM models used for the Aurora-2 tasks are 18-states long (16 emitting states) with 3-Gaussian GMMs. The proposed models are 50-states long (48 emitting states) with 1-Gaussian GMM per state. Training is done on the Aurora-2 clean speech TRAIN corpus. Noisy recognition tasks (using standard Aurora-2 additive noises) are also performed to evaluate the robustness of the technique.

Since the clean speech recognition accuracies of the reference “left-to-right" GMM-HMMs on this last dataset are very high (>99%), improvements may not be significant. As a re-

sult, we generated 4 new versions of the Aurora-2 digits dataset, each of them convoluted with a different real world reverberation impulse response (IR) taken from the Openair IR database [Murphy et Shelley, 2016]. The IRs are chosen on the basis of their uniqueness. We selected "Maes Howe", "Falkland Palace Royal Tennis Court", "Purnode's Tunnel" and "Tyndall Bruce Monument". For these experiments, both systems are trained on a convoluted version of the "TRAIN" corpus. Clean and noisy testing datasets are also convoluted in the same fashion, which means the latter is corrupted by IRs with its additive noise.

Word error rates (WER) are computed in the standard fashion, taking into account insertion and deletion errors. For every one of the 3 Aurora-2 sets (A, B, and C), presented WER values are averaged over all additive noise types. Furthermore, noisy tests are averaged over SNRs 20, 15, 10, 5 and 0 dB.

3.3.7 Results and Discussion

Table 3.1 shows significant WER reductions with the proposed approach in clean word recognition. When trained and tested in reverberated environments without additive noise, performance are also significantly improved on average. Since reverberation adds time-dependencies to signals and thus increases their temporal complexity, this demonstrates that temporal precision has indeed been increased. However, the proposed system is also less robust to additive noise. This may be explained by a loss of generalizing power caused by the improved precision [Xiao *et al.*, 2010]. Also noteworthy is the fact that the system perform poorly when baseline performances are very low, as demonstrated by the results on the "Tennis Court" corpus. This makes sense considering that the proposed architecture is built using a "left-to-right" prototype which is, in this case, poorly adapted to the dataset dynamics. To our knowledge, the poor performance with the TIMIT database (shown in Table 3.1) are best explained by the fact that monophonic HMMs have linguistically little to no temporal structure, as the name implies. Hence, a model specially designed to encode complex temporal behaviours is unsuited to this recognition task.

We thus see that monophone-based speech recognition does not seem to profit from models with higher temporal precision while word-based classification does. As discussed in [Greenberg, 1999], this indicates that a high temporal variability exists at the syllable level. Considering this, we suggest that the proposed approach should perform better in large dictionnary systems with classes representing linguistic units of increased length, such as triphones [Lopes et Perdigao, 2011].

Tableau 3.1 Performances measured on the Aurora-2 and TIMIT datasets

		WER (%)			Relative WER reduction vs baseline (%)	
AURORA-2						
Convolutive noise	**Additive noise SNR**	**Baseline**	**Proposed, no feedback**	**Proposed, 10 iterations feedback**	**Proposed, no feedback**	**Proposed, 10 iterations feedback**
Clean	Clean	0.93	0.81	0.70	12.59	24.97
	20dB-0dB	40.33	47.06	45.75	-16.71	-13.46
Maes Howe	Clean	6.54	5.67	5.70	13.43	12.93
	20dB-0dB	55.42	55.32	55.96	0.17	-0.97
Tennis court	Clean	56.88	65.30	60.70	-14.81	-6.71
	20dB-0dB	78.91	91.64	90.25	-16.13	-14.38
Tunnel	Clean	27.18	25.50	20.16	6.19	25.82
	20dB-0dB	69.72	77.54	73.55	-11.22	-5.49
Tyndall Bruce	Clean	18.52	16.32	17.08	11.85	7.77
	20dB-0dB	58.91	63.80	67.35	-8.29	-14.31
			Average	Clean	5.85	12.96
				20dB-0dB	-10.44	-9.72
TIMIT						
Clean	Clean	31.98	-	35.88	-	-12.20

The proposed proof-of-concept framework could be improved in a number of ways for hypothetical increased performances. First, a less trivial optimization algorithm (see Fig.1) could be used in the pruning iterative mechanism. This could yield very precise pruning strengths leading to high recognition accuracy.

Better initialization models and methods could also greatly benefit our solution. As it was discussed earlier, complex and finely-tuned models are very sensitive to their initialization conditions and as such, there is much work to be done in optimizing them. Finally, the proposed framework can also be coupled with complementary state-of-the-art techniques that implement better emission models such as DNN-HMMs [Mohamed et al., 2012] and online model adaptation [Li *et al.*, 2007].

CHAPITRE 4

CONCLUSION

L'augmentation de la précision de reconnaissance sur la base de données Aurora-2 en conditions propres et réverbérées semble démontrer que la technique proposée permet en effet une modélisation temporelle plus précise des modèles de mots. Les résultats sur la base TIMIT semblent aussi confirmer ceci considérant que l'augmentation de la précision dynamique n'impacte que négativement les performances. Ceci s'explique par le fait que les modèles de monophones, tel que leur nom l'indique, contiennent peu d'information dynamique.

Il est démontré dans le travail présenté que le déséquilibre entre les probabilités d'émission et de transition est un phénomène non négligeable en reconnaissance vocale. Son impact premier est de réduire le pouvoir discriminatif des probabilités de transition au décodage. Se faisant, seule la présence de celles-ci (une probabilité non-nulle) est vraiment importante. Cette connaissance en main, je propose une méthode d'entraînement de topologies de modèles markoviens cachés (HMMs). Les modèles générés avec cette méthode sont capables de modéliser de plus complexes dynamiques acoustiques. Les expériences menées en reconnaissance vocale démontrent également que cette précision accrue peut impacter positivement les performances d'un classificateur de mots.

Le cadre conceptuel d'entraînement de topologie proposé est en soit une contribution non négligeable au domaine. Bien que son efficacité n'est démontrée qu'avec les GMM-HMMs, rien n'indique que les DNN-HMMs, l'état de l'art, ne pourraient en profiter, puisque les systèmes ASR de l'état de l'art partagent la même topologie depuis 40 ans. Autre contribution importante est la démonstration que la reconnaissance de mots est fondamentalement différente de la reconnaissance de monophones et ne devrait pas être approchée de la même manière. En effet, les résultats obtenus semblent démontrer que les monophones, comme leur nom l'indique, sont très pauvres en dynamiques acoustiques et ne profitent pas pleinement des capacités de gestion temporelle des HMMs.

Deux avenues de recherche futures sont à considérer : l'entraînement de modèles GMM-HMM triphoniques employant la méthode développée et l'application de celle-ci aux DNN-HMMs. La première nécessite une grande base de données contenant des exemples de chacun des 125 000 triphones anglais. Les systèmes ASR standards à base de GMM-HMMs

triphoniques n'ont pas cette restriction puisque la majorité des triphones sont estimés. Cette estimation est possible en assumant l'absence de richesse dans la dynamique acoustique des triphones, une hypothèse impossible dans le cas du cadre conceptuel proposé.

L'application de la technique d'entraînement développée aux DNN-HMMs nécessite l'implémentation d'un système ASR capable de produire les mêmes performances que les résultats présentés en littérature. À défaut de ceci il est difficile de comparer les techniques. L'outil communément utilisé pour ceci est le système *Kaldi* [Povey et al., 2011], or celui-ci est incompatible avec l'approche proposée puisqu'il ne peut supporter l'entraînement de modèles complexes (dont tous les états ne sont pas forcément utilisés au décodage). Ceci est parce que *Kaldi* n'implémente pas l'algorithme standard *Baum-Welch* mais plutôt l'entraînement *Viterbi* puisque celui-ci a un poids computationnel plus faible. Cette avenue de recherche impliquerait donc l'implémentation d'un système ASR complet utilisant à la fois les DNN-HMMs et l'algorithme d'entraînement *Baum-Welch*, ce qui dépasserait les objectifs visés par ce mémoire de maîtrise.

LISTE DES RÉFÉRENCES

Cawley, G. C. et Talbot, N. L. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, volume 11, p. 2079 – 2107.

Fang, Z., Guoliang, Z. et Zhanjiang, S. (2001). Comparison of different implementations of mfcc. *Journal of Computer Science and Technology (English Language Edition)*, volume 16, numéro 6, p. 582 – 9.

Greenberg, S. (1999). Speaking in shorthand-a syllable-centric perspective for understanding pronunciation variation. *Speech Communication*, volume 29, numéro 2-4, p. 159 – 76.

Jelinek, F. (1976). Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, volume 64, numéro 4, p. 532 – 56.

Jitsuhiro, T. et Nakamura, S. (2004). Automatic generation of non-uniform hmm structures based on variational bayesian approach. Dans *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing.* volume vol.1. p. 805 – 8.

Kalinli, O., Seltzer, M. L., Droppo, J. et Acero, A. (2010). Noise adaptive training for robust automatic speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, volume 18, numéro 8, p. 1889 – 1901.

Li, J., Deng, L., Yu, D., Gong, Y. et Acero, A. (2007). High-performance HMM adaptation with joint compensation of additive and convolutive distortions via vector taylor series. Dans *2007 IEEE Workshop on Automatic Speech Recognition and Understanding.* p. 65 – 70.

Lippmann, R. P. (1997). Speech recognition by machines and humans. *Speech Communication*, volume 22, numéro 1, p. 1 – 15.

Lopes, C. et Perdigao, F. (2011). *Phone Recognition on the TIMIT Database.* Rijeka, Croatia, p. 285 – 302.

MacKay, D. J. (2003). *Model Comparison and Occam's Razor.* Cambridge, UK, p. 345 – 357.

Mak, B. et Chan, K.-W. (2005). Pruning hidden markov models with optimal brain surgeon. *IEEE Transactions on Speech and Audio Processing*, volume 13, numéro 5, p. 993 – 1003.

Mohamed, A. et al. (2012). Acoustic modeling using deep belief networks. *IEEE Transactions on Audio, Speech and Language Processing*, volume 20, numéro 1, p. 14 – 22.

Murphy, D. et Shelley, S. (2016). Open acoustic impulse response (open air) library.

Narayanan, A. et Wang, D. (2015). Improving robustness of deep neural network acoustic models via speech separation and joint adaptive training. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, volume 23, numéro 1, p. 92 – 101.

Povey, D. et al. (2011). The kaldi speech recognition toolkit. Dans *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*.

Rabiner, L. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, volume 77, numéro 2, p. 257 – 86.

Rabiner, L. R. et Juang, B. H. (1992). Hidden markov models for speech recognition - strengths and limitations. *Speech Recognition and Understanding. Recent Advances, Trends and Applications. Proceedings of the NATO Advanced Study Institute*, p. 3 – 29.

Shannon, R. et al. (1995). Speech recognition with primarily temporal cues. *Science*, volume 270, numéro 5234, p. 303 – 4.

Torbati, A., Picone, J. et Sobel, M. (2014). A left-to-right HDP-HMM with HDPM emissions. Dans *2014 48th Annual Conference on Information Sciences and Systems (CISS)*. p. 6 pp.

Xiao, X., Li, J., Chng, E. S., Li, H. et Lee, C.-H. (2010). A study on the generalization capability of acoustic models for robust speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, volume 18, numéro 6, p. 1158 – 69.

Young, S. et al. (1995). *The HTK Book*. Microsoft Corporation.

Young, S., Russell, N. et Thornton, J. (1989). Token passing : a simple conceptual model for connected speech recognition systems.

Zhao, Y. et Juang, B.-H. (2012). Stranded gaussian mixture hidden markov models for robust speech recognition. Dans *Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2012)*. p. 4301 – 4.