Reconnaissance des commandes vocales d'un robot mentor dans un environnement bruité à base HMM

Khenfer-koummich fatima, Mesbahi Larbi, Hendel Fatiha

الملخّـص

تعرض هذه الدّراسة نظام التّعرّف الآلي على الكلام المنجز، باستعمال نماذج «ماركوف المخفيّة» لأداء أوامر محدّدة من قبيل التحكّم في الآلة عن طريق الكلام الطّبيعيّ، مثل : دوران، افتح، إلخ.

يأخذ هذا النّظام بعين الاعتبار الضّجيج المنبعث من البيئة، وقد تمّ تطبيق تلك النّماذج على كلمات تمثّل أوامر للتّحكم في الآلة، نُطقت باللغتين العربيّة والفرنسيّة. معدّل التّعرّف المحصّل عليه في كلا اللغتين متقارب في حالة الكلمات دون ضجيج، مع اختلاف طفيف يعود لفائدة اللّغة العربيّة، عندما يضاف الضّجيج الأبيض الغوصيّ، بنسبة إشارة إلى الضّجيج مساوية لـ 30 ديسيبل، حيث يصل معدّل التعرّف إلى 69 % و 80 % بالنسبة للّغتين الفرنسيّة والعربيّة، على التّوالي. يمكن تفسير ذلك بقدرة السّيّاق اللّفظيّ لكلّ لغة عند تعرّضها لتأثير الضّجيج.

الكلمات المفاتيح : نماذج ماركوف المخفيّة، اللغة العربيّة، اللغة الفرنسيّة، التّعرف الآلي على الكلام، نسبة الإشارة إلى الضّجيج.

Reconnaissance des commandes vocales d'un robot mentor dans un environnement bruité à base HMM

Khenfer-Koummich Fatima[1], Mesbahi Larbi[2], Hendel Fatiha[1]

[1] Université des Sciences et de Technologie d'Oran
Mohamed Boudiaf (USTO-MB), Oran, Algérie.
[2] Institut Français des Sciences et Technologies des Transports de
l'Aménagement et des Réseaux (IFSTTAR), Lille, France.
fatimakhanfar@yahoo.fr, mesbahi_99@yahoo.com, fa_hendel@yahoo.fr

Résumé

Cet article présente une approche basée sur les modèles de Markov cachés (HMM -Hidden Markov Model-), appliquée à des mots isolés de type commande robot. Le but c'est de permettre à un opérateur de commander un robot mentor pour exécuter des tâches bien précises de type tourner, monter ou fermer, etc. Cette tâche doit tenir en compte des différents niveaux de bruit d'environnement. Cette approche a été appliquée sur des mots isolés représentants les commandes robot prononcées en deux langues : le Français et l'Arabe. Le taux de reconnaissance obtenu dans les deux langues est comparable dans la parole neutre. Néanmoins, il y a une légère différence au profit de la langue Arabe lorsqu'un bruit blanc gaussien est ajouté, avec un Rapport Signal sur Bruit (RSB) égale à 30dB, on affiche des taux de reconnaissances de 69% et 80% pour le Français et l'Arabe respectivement. Cela peut s'expliquer par la capacité du contexte phonétique de chaque langue à contenir l'influence du bruit.

Mots clés : Commande vocale, HMM, Bruit, HTK, langue arabe.

1. Introduction

La communication est le moyen naturel d'une interaction humaine, actuellement, il est possible de communiquer avec la machine via la parole [1], ceci nécessite la création d'une interface homme-machine réalisée avec un système de reconnaissance vocale. Pour réaliser cette interface, le système doit comprendre deux parties: un encodeur et un décodeur. L'encodeur analyse le signal et extrait un certain nombre de paramètres pertinents. Le décodeur utilise ces paramètres pour formuler la décision qui correspond au signal de parole d'entrée. Différentes recherches dans le domaine de la reconnaissance de la parole sont menées avec différents angles, en passant en premier par l'étape d'analyse et de paramétrisation : mfcc, lpc, cepstres discrets, etc. [2]. Ces choix de paramètres reposent sur la capacité d'interpolation, robustesse au bruit et l'adaptation à la variabilité inter et intra locuteurs [3-4]. En second, le choix des méthodes d'apprentissage (HMM, réseaux de neurones, SVM, GMM, etc.) [5-6] dépend des paramètres d'analyse, la taille des données et la capacité de généralisation. En somme, il faut tenir compte des

méthodes de test et de généralisation pour valider le modèle d'apprentissage.

Vu le progrès technologique rapide, les tendances actuelles s'intéressent aux applications interactives avec des robots manipulateurs [7], des machines ou engins automatisés [8]. Différentes applications sont concernés surtout l'intervention dans des milieux hostiles, aide à l'handicapé [9], ou déminage, etc. A cet effet, le développement de module de reconnaissance de la parole s'avère primordial. Malgré que les taux de reconnaissance soient excellents, le problème de la parole spontanée reste toujours posé, de même que l'accent et les contraintes de bruit d'environnement. Si la tâche dans un milieu bruité s'annonce très critique et importante, le recours au modèle le plus résistant et adapté au bruit devient une nécessité. Dans ce contexte nous proposons un système de reconnaissance à base de HMM appliqué à une taille de vocabulaire limité prononcé en Français et en Arabe, en vue de tester la capacité de réaction du robot exprimé par le taux de reconnaissance. Le but à long terme consiste à élargir d'une part le vocabulaire de commandes et d'étudier d'autres part les caractéristiques phonétiques (effet de l'articulation des consonnes et voyelles) dans des milieux fortement bruités.

2. Modélisation par HMM

Selon le formalisme des modèles de Markov cachés (HMM), le signal de parole est supposé être produit par un automate d'états fini stochastique construit à partir d'un ensemble d'états stationnaire régis par les lois statistiques [10]. Le formalisme des modèles HMM suppose que le signal de parole est formé d'une séquence de segments stationnaires, tous les vecteurs associés à un même segment stationnaire étant supposés avoir été générés par le même état HMM. Chaque état de cet automate est caractérisé par une distribution de probabilité décrivant la probabilité d'observation des différents vecteurs acoustiques. Les transitions entre les états sont instantanées, elles sont caractérisées par une probabilité de transition. Si chaque état du modèle permet de modéliser un segment de parole stationnaire. La séquence d'état permet quant à elle de modéliser la structure temporelle de la parole comme une succession d'états stationnaires. De ce fait, les modèles utilisés en reconnaissance automatique de la parole sont généralement du type gauche-droite ou les transitions possibles sont soit des boucles sur le même état, soit le passage à un état suivant**,** ce type de modèle est présenté (voir section 2.1). L'aspect séquentiel du signal de parole est ainsi modélisé [11]. Les modèles sont dit cachés car la séquence d'états n'est pas directement observable; seule la séquence de vecteurs acoustiques est visible et est considérée comme une fonction statique de la séquence d'états (ayant généré les observations). Chaque unité linguistique est modélisée par plusieurs états stationnaires. Ces unités peuvent être : des mots, ou des phonèmes. Les mots sont ensuite construits suivant une grammaire et un dictionnaire. Un exemple de principe de la connaissance des mots est présenté dans la figure.1.

2.1. Description du HMM

Un Modèle de Markov Caché (MMC) en anglais Hidden Markov Model (HMM) est un modèle statistique dans lequel le système modélisé suit un processus markovien de paramètres inconnus.

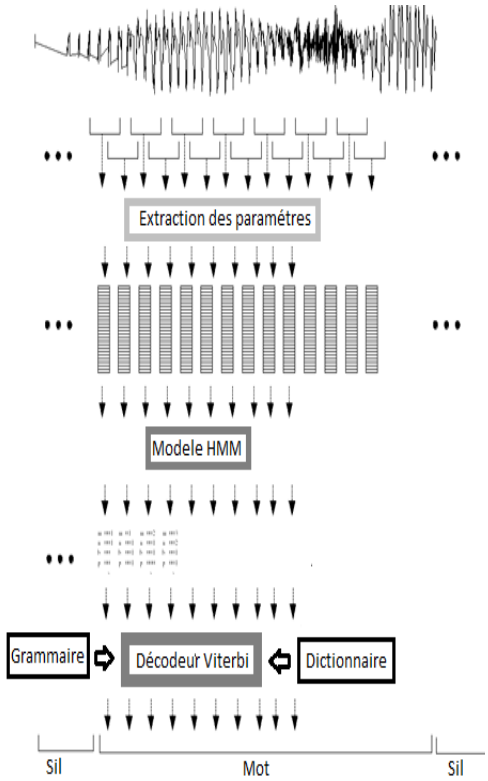

Figure 1: Principe général de la reconnaissance de la parole par HMM.

Les modèles de Markov cachés sont utilisés notamment en reconnaissance de formes, en intelligence artificielle ou encore en traitement automatique du langage naturel. Chaque modèle est constitué de nœuds cachés et de nœuds d'observation (voir la figure .2).

Pour chaque mot, un modèle "gauche-droit" a été défini et présenté par l'ensemble de paramètres de modèle : $\lambda = \{\pi, A, B\}$ ou :

S_i : L'état i ;

π_i : La probabilité que S_i soit l'état initial;

a_{ij} : La probabilité de la transition $S_i \to S_j$;

$b_i(k)$: La probabilité d'émettre le symbole k étant dans l'état S_i ;

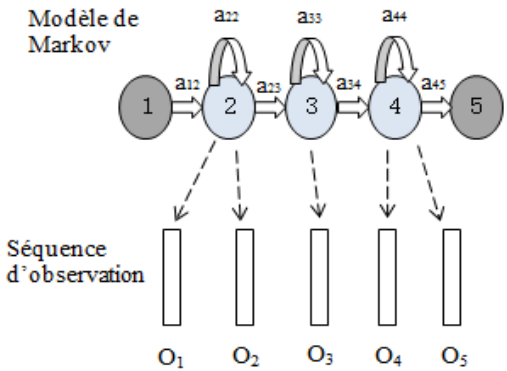

Figure 2: Modèle de Markov à 5 états [12].

Sous contrainte :

- $\sum_i \pi_i = 1$
- $\forall i, \sum_j a_{ij} = 1$ et $a_{ij} \geq 0$ (Si $i > j, a_{ij} = 0$)
- $\forall i, \sum_k b_i(k) = 1$ et $b_i(k) \geq 0$

Chaque mot est représenté par une séquence de vecteurs ou d'observations O défini par :

$$O = o_1, o_2, ..., o_t, ...o_T$$

Où o_t est le vecteur observé à l'instant t et T est le nombre d'observations (nombre de vecteurs).

Le but principal est de déterminer la probabilité d'une séquence d'observations.

2.2. Apprentissage

Pour ré-estimer les modèles HMM, l'algorithme de Baume-Welch est utilisé pour calculer les valeurs optimales des paramètres HMM (probabilités de transition, plus les vecteurs de moyenne et de variance pour des fonctions d'observations) obtenues après un nombre d'itérations.

2.3. Grammaire et Dictionnaire

La figure .3 montre l'architecture de notre système, elle est définie en tenant compte de la grammaire suivante : un silence de départ, suivi par un mot unique (ferme/monte/…), suivi par un silence de fin. Le dictionnaire contient l'ensemble des mots à reconnaitre et aussi les modèles HMM qui sert pour la phase de reconnaissance.

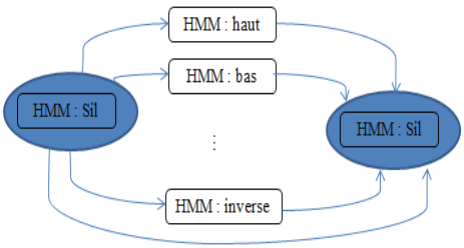

Figure 3: Grammaire pour la reconnaissance des mots isolés.

2.4. Reconnaissance

La reconnaissance revient alors à calculer la vraisemblance de la suite d'observations acoustiques constituant le mot à reconnaître par rapport à chacun des modèles optimisés dans la phase d'apprentissage. Le modèle présentant la plus grande vraisemblance (la probabilité maximale) d'avoir émis cette suite d'observations fournit le mot reconnu.

3. Partie expérimentale

3.1. Description de la base TIMIT

La base TIMIT est une base de données acoustique et phonétique dédiée à la reconnaissance de la parole indépendamment du locuteur [13]. Elle contient les enregistrements de 630 locuteurs américains (438 hommes et 192 femmes), répartis en 8 "dialectes régionaux"("dr1" à "dr8") et prononçant chacun 10 phrases. La base TIMIT contient un total de 6300 phrases, Pour chaque phrase, nous disposons du texte en Anglais, du signal échantillonné à 16 kHz sur 16 bits, de la segmentation phonétique en 61 classes, et de la segmentation en mots, le vocabulaire total de la base est de 6100 mots. Cette base sert pour valider nos résultats obtenus sur la base de commande.

3.2. Description de la base des commandes vocales

Notre base de données est constituée de dix (10) mots (commandes) prononcés séparément en Arabe / Français. Chaque mot a été prononcé trois fois par 10 locuteurs d'âge et de sexe différent (5 hommes et 5 femmes), il y a 300 prononciations pour chaque langue, et ces mots sont utilisés pour l'apprentissage. Une autre base est constituée de 300 mots a été utilisée pour faire les tests. Le signal de parole est enregistré à une fréquence d'échantillonnage de 11025 Hz. Le choix de cette fréquence peut varier de 6000 à 16000 Hz, pour les techniques d'analyse, de synthèse ou de reconnaissance de parole. La quantification s'effectue sur 8 bits dans un environnement ambiant. Cette base a été ensuite segmentée et étiquetée manuellement avec le logiciel Wavesurfer [14].

Un Bruit blanc gaussien avec différents niveaux de Rapport Signal sur Bruit RSB

est ajouté aux bases d'apprentissage et de test pour les deux langues; car les systèmes de reconnaissance sont généralement dédiés à des applications qui doivent tenir compte de l'environnement bruité. Ces commandes permettent au bras manipulateur/robot mentor d'exécuter des actions relatives à chaque commande (Voir Table 1).

Bas : orientation de la pince vers le Bas
Haut : orientation de la pince vers le haut.
Descend : le bras descend d'un pas selon l'axe des z.
Monte : le bras monte d'un pas selon l'axe des z.
Droite : le bras tourne de gauche vers droite.
Gauche : le bras tourne de droite vers gauche.
Tourne : la pince tourne.
Inverse : la pince tourne en sens inverse.
Ouvre : Ouverture de la pince.
Ferme : Fermeture de la pince.

Table 1: Commandes du robot

3.3. Reconnaissance en temps réel

Le rebot mentor ou bras manipulateur à 5 axes est un rebot industriel. Lorsqu'un opérateur veut utiliser ce type de robot, il doit prononcer un mot isolé, ce mot est d'abord transformé en série de vecteurs acoustiques (MFCC), cet observation d'entrée est reconnue par un algorithme de Viterbi, qui aligne les observations avec les HMM, si cette commande est reconnue, l'angle correspondante à cette commande est envoyée via un réseau local au serveur qui envoie à son tour les ordres au rebot (Faire tourner les servomoteurs de rebot). Cette procédure est schématisée dans la figure .4.

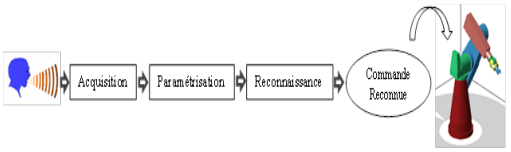

Figure 4: Procédure de la reconnaissance en temps réel

3.4. Implémentation

L'implémentation de notre application a été réalisée sous l'environnement Linux avec les outils de HTK et la programmation en C-Shell.
L'utilisation de la base de données des mots ou des phonèmes nécessite une analyse du signal vocal pour extraire les caractéristiques de celui-ci. Nous avons choisi les paramètres MFCC (*Mel Frequency Cepstral Coding*) grâce à leurs caractéristiques de bonne représentation du signal. D'autres paramètres ont été ajoutés comme la vitesse Δ et l'accélération $\Delta\Delta$ de ces MFCCs (c'est-à-dire les dérivées et dérivées secondes du vecteur) dans le but d'une meilleure représentation de signal vocal [1-2] car ils fournissent des informations supplémentaires.
Le processus d'apprentissage des phonèmes ou mots isolés procède par l'initialisation des paramètres HMM, afin d'obtenir une convergence rapide et précise de l'algorithme d'apprentissage. L'outil HTK permet d'initialiser le modèle avec le nombre d'états et les probabilités de transitions ainsi que les vecteurs de variance et de moyenne. Chaque état du HMM reçoit le même vecteur de variance et de moyenne : ceux-ci sont calculés globalement sur le corpus entier d'entrainement.

4. Résultats et discussion

4.1. Base TIMIT

On est intéressé par la reconnaissance des phonèmes de la base de données TIMIT en utilisant le toolkit HTK. On a évalué la

technique de modélisation par HMM avec des différents paramètres d'apprentissage (nombre d'états, nombre de mixtures, nombre d'itérations, etc.) pour obtenir un meilleur taux de reconnaissance des phonèmes obtenus par HMM. Après plusieurs tests appliqués à cette base, on est arrivé à améliorer les résultats (Voir Table 2), cela nous a permis de choisir les bons paramètres et valider nos résultats obtenus par la base enregistrée.

	Taux de reconnaissance
Base d'apprentissage	78.05%
Base de test	70.24%

Table 2. Taux de reconnaissance des phonèmes.

4.2. Base des commandes vocales

4.2.1. Reconnaissance des mots neutres

Notre système a été évalué sur une base de 30 exemples pour chaque mot. Table 3 indique que les mots d'apprentissage de la base Arabe et Français sont comparables car le taux d'apprentissage obtenu tourne autour de 97%. Concernant la reconnaissance des mots de test, on remarque que les taux des mots du Français et de l'Arabe sont respectivement 94.67% et 93%. Cette légère différence est liée au nombre de paramètres d'entrée et la différence des caractéristiques propres à chaque langue. En ajoutant à cela le taux de confusion qui varie d'une commande à l'autre en tenant en compte les particularités phonétiques de chaque langue.

4.2.2. Reconnaissance des mots bruités

Après un ajout de bruit blanc gaussien de différents niveaux de RSB allant de 0 dB à 65 dB appliqué sur la base d'apprentissage et de test des deux langues.

Mots Français	**Train %**	**Test %**	**Mots Arabes**	**Train %**	**Test %**
Bas	90	87.1	Asfel	100	93.3
Haut	100	93.1	Aala	96.8	86.7
Descend	100	100	Ihbit	100	93.3
Monte	100	100	Isaad	96.6	83.3
Droite	100	96.7	Yamine	100	90
Gauche	96.7	96.7	Yassar	100	100
Tourne	100	100	Dawarane	100	100
Inverse	100	96.7	Iklib	96.7	100
Ouvre	86.7	83.3	Ifteh	96.7	90
Ferme	96.7	93.3	Aghlik	93.3	93.3
Taux	97.01	94.67	Taux	97.99	93

Table 3. Taux de reconnaissance des commandes.

On a obtenu des courbes représentants l'évolution du taux de reconnaissance en fonction du niveau de RSB. D'après les courbes obtenues dans la figure 5 (a et b), on a observé que le système de reconnaissance des mots bruités devient performant à partir d'un seuil de bruit (niveau de RSB) correspondant à 30 dB pour l'Arabe (le taux de test est égal à 80%) et un niveau de RSB de 35 dB pour le Français (le taux de test est égal à 82.33%). On remarque aussi que lorsque le RSB varie dans l'intervalle de [10 dB, 45dB], le taux de reconnaissance des mots Arabes est nettement supérieur au taux de reconnaissance des mots Français que ce soit dans la base d'apprentissage (voir Figure 5.a) ou dans la base de test (voir Figure 5.b). Ces résultats indiquent que la langue arabe résiste mieux au bruit par rapport à la langue française car les mots du Français sont plus consonantiques et les mots d'Arabe sont plus voyellisés, par conséquence la partie consonantique résiste moins à l'influence du bruit, ces résultats sont obtenus sur l'ensemble du corpus qui contient 300 mots pour chaque langue.

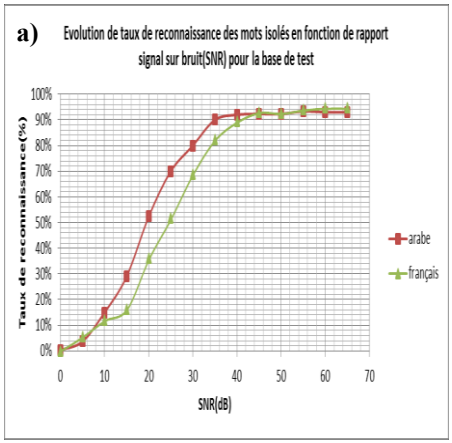

Figure 5: Reconnaissance des 300 mots bruités

La figure 6 représente le taux de reconnaissance de l'ensemble des mots des deux langues mais cette fois-ci avec une base qui contient 1100 mots du Français et 1100 mots de l'Arabe ainsi qu'un mélange des deux langues qui contient 2200 mots. Lorsqu'on mélange les commandes des deux langues on remarque qu'il y a une chute de taux de reconnaissance par rapport à ceux obtenus pour chacune des deux langues séparément. Cela peut être justifié par le fait que le taux de confusion augmente en augmentant l'espace du contexte phonétique.

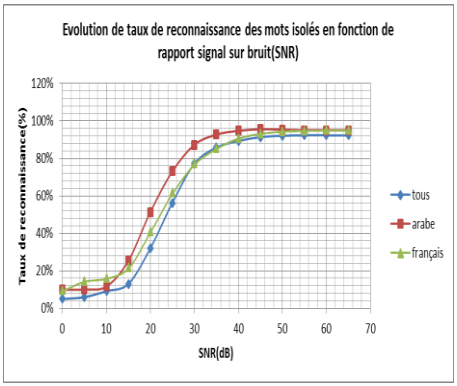

Figure 6: Reconnaissance des 1100 mots bruités

5. Conclusion

Nous avons développé dans ce papier un module de reconnaissance à base HMM pour commander le robot mentor par la voix.

Ces commandes ont été reconnues avec succès dans la majorité des cas, avec un taux de reconnaissance dépassant les 93% pour la base de test et les 97% pour la base d'apprentissage.

Notre contribution consiste à évaluer notre système de reconnaissance dans un environnement bruité avec différents niveaux de RSB allant jusqu'à dominance du bruit sur le signal parole. Le choix des commandes en deux langues Arabe et Français était dans le but de tester laquelle des deux langues résiste mieux au bruit et établir le lien entre le contexte phonétique et le taux de reconnaissance.

Notre futur travail se focalisera sur l'introduction de nouvelles méthodes, en augmentant la taille du vocabulaire et en tenant compte de la distance qui sépare l'opérateur du robot. Ainsi pour garantir la sécurité de la tâche à exécuter, il faudra étudier les paramètres qui permettent

d'obtenir de meilleurs résultats en présence du bruit avec différents niveaux.

6. References

[1] Gallardo-Estrella, L. and Poncela, A., "Human/Robot Interface for Voice Teleoperation of a Robotic Platform", Springer-Verlag Berlin Heidelberg, 2011.

[2] Ben Fredj, I. and Ouni, K., "Optimization of Features Parameters for HMM Phoneme Recognition of TIMIT Corpus", International Conference on Control, Engineering & Information Technology (CEIT'13). Vol.2, pp. 90-94, 2013.

[3] Umesh, S., "Studies on inter-speaker variability in speech and its application in automatic speech recognition", Sadhana Vol. 36, Part 5, October 2011, pp. 853–883.© Indian Academy of Sciences.

[4] Bänziger, T., Klasmeyer, G., Johnstone, T., Kamceva, T. and Scherer, K. R., "Améliorer les systèmes de vérification automatique du locuteur en intégrant la variabilité émotionnelle", XXIIIèmes Journées d'Etude sur la Parole, Aussois, 19-23 juin 2000.

[5] Rabiner, L. and Juang. B.,"Fundamentals of Speech Recognition", PTR Prentice Hall (Signal Processing Series), Englewood Cliffs NJ, 1993, ISBN0-13-015157-2.

[6] Kevin, J., Lang, A. and Waibel, "A Time Delay Neural Network Architecture for Isolated Word Recognition", Neural Networks, Vol. 3, pp. 23-43, 1990.

[7] Heidaria, H. and Gobeeb, S., "Isolated Word Command Recognition for Robot Navigation", International Symposium on Robotics and Intelligent Sensors 2012 (IRIS 2012), sciencedirect, Procedia Engineering 41 (2012) 412 – 419.

[8] Ferre M., Macias-Guarasa, J., Aracil R.and Barrientos A." Voice command generator for teleoperated robot systems." In Proceedings of the IEEE ROMAN 1998, Takamatsu, Japan 1998.

[9] Dragon Naturally Speaking de Nuance. http://www.nuance.fr/Dragon12.

[10] Juang, B. H. and Rabiner, L. R., "Hidden Markov Models for Speech Recognition", Technometrics is currently published by American Statistical Association, Technometrics, Vol. 33, No. 3. (Aug., 1991), pp. 251-272.

[11] Rabiner, L. R., "A Tutorial on Hidden Markov Models and selected applications in speech recognition", Proceedings of IEEE, Vol. 77, N°2, pp: 257-286, Feb. 1989.

[12] Young, S., Evermann, G., Kershaw, D., Moore, D., Odell, J., Ollason, D., Valtchev, V. and Woodland, P. "The HTK Book (for HTK version 3.4)". Cambridge University Engineering Department, December 2006. http://htk.eng.cam.ac.uk.

[13] The DARPA TIMIT Acoustic-phonetic Continuous Speech Recognition Database CDROM, NIST, 1990.

[14] http://www.speech.kth.se/