René Lanciné Doumbouya

Reconnaissance Automatique de la Parole

IFT 6010

Université de Montréal

Reconnaissance, Séparation, Autonomie : le triptyque de la Parole

Définition

- Speech Recognition (is also known as Automatic Speech Recognition (ASR), or computer speech recognition) is the process of converting a speech signal to a sequence of words, by means of an algorithm implemented as a computer program.[1]

Historique

Une évolution rapide:

- 1952 : reconnaissance des 10 chiffres, pour un monolocuteur , par un dispositif électronique câblé
- 1960 : utilisation des méthodes numériques
- 1965 : reconnaissance de phonèmes en parole continue
- 1971 : lancement du projet ARPA aux USA (15 millions de dollars) pour tester la faisabilité de la compréhension automatique de la parole continue avec des contraintes raisonnables
- 1972 : premier appareil commercialisé de reconnaissance de mots
- 1978 : commercialisation d'un système de reconnaissance à microprocesseurs sur une carte de circuits imprimés
- 1983 : première mondiale de commande vocale à bord d'un avion de chasse en France
- 1985 : commercialisation des premiers systèmes de reconnaissance de plusieurs milliers de mots
- 1986 : lancement du projet japonais ATR de téléphone avec traduction automatique en temps réel
- 1989 : recrudescence des modèles connexionnistes neuromimétiques
- 1990 : premières véritables applications de dialogue oral homme-machine
- 1994 : IBM lance son premier système de reconnaissance vocale sur PC
- 1997 : lancement de la dictée vocale en continu par IBM

Principe de fonctionnement

- Vous parlez, votre voix est projetée vers un microphone, qui convertit les ondes sonores de votre voix en un signal électrique, qu'un ordinateur ou un mobile peut alors utiliser. La partie difficile est d'analyser votre voix.

Probématiques

- La compréhension humaine de la parole
- Le langage parlé est différent du langage écrit: e.g. a l'oral il y a des répétions, des glissements de langues,changement de sujets au milieu d'un énoncé etc.
- Le bruit: toute information non désirée dans le signal de parole est connu comme bruit.
- Les dialectes,etc…
- Nous avons examiné certaines des difficultés de reconnaissance de la parole mais pas toutes. La question la plus problématique est leur forte variabilité.

Modele standard de Reconnaissance vocale

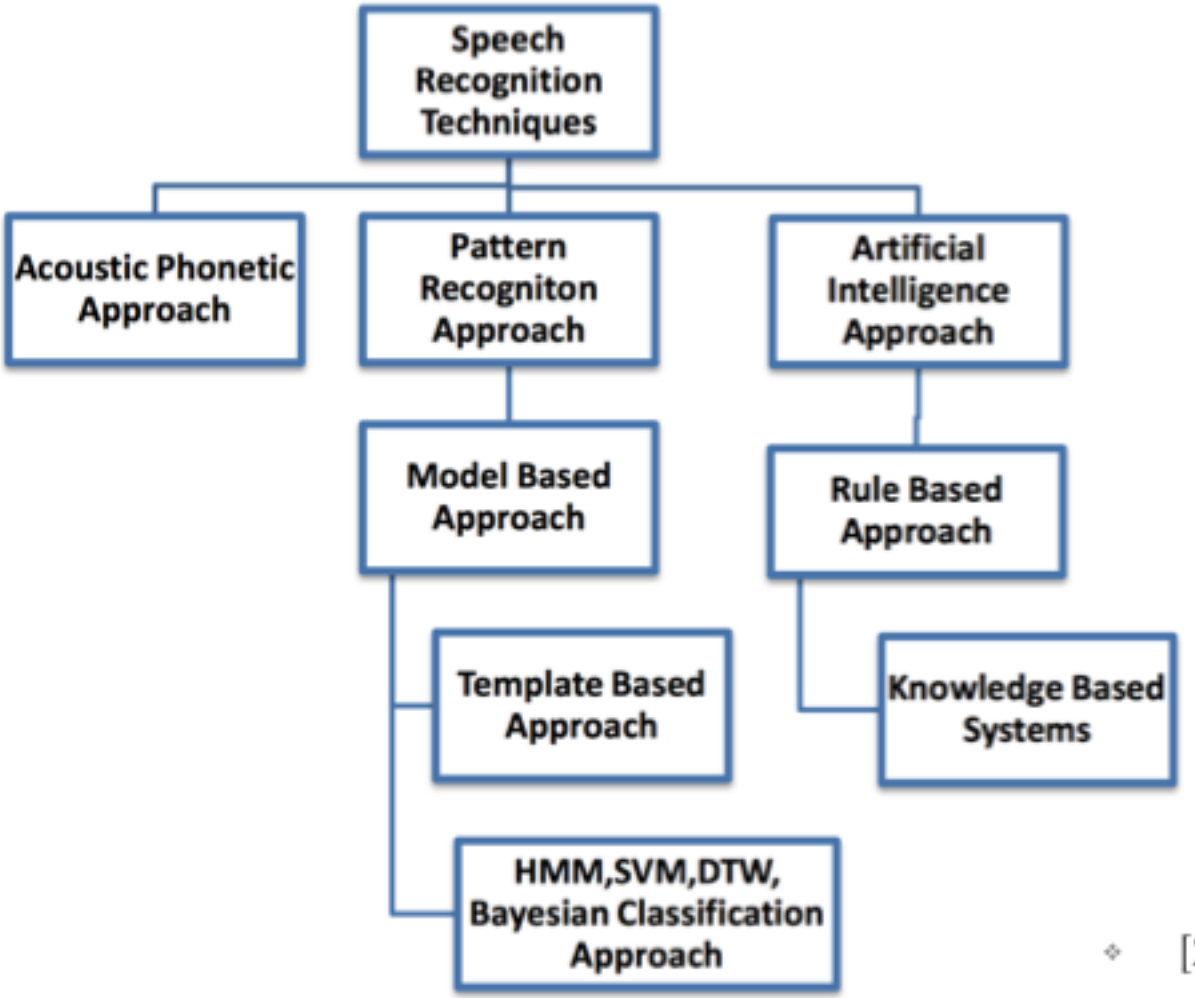
- L'objectif est de décoder la chaîne de mots, sur la base de la séquence d'observation acoustique, de sorte que la chaîne décodée a un maximum de probabilité a posteriori (MAP).En suivant l'approche Bayesienne appliquée aux systèmes RAP[4] on a:
- $P(W \mid A) = \arg\max_w P(W \mid A)$ (1)
- $P(W \mid A) = P(A \mid W).P(W) / P(A)$ (Théorème de Bayes) (2)
- $W = \text{argmax}_w P(A \mid W)P(W)$ (3)

Types de ‘’speech recognizers’’

- Mots isolés: ces systèmes ont des états"Listen/Non-Listen", qui nécessitent un temps d’attente entre les énoncés du locuteur.
- Mots connectés: les systèmes de mots connectés sont similaires aux mots isolés, mais permet aux énoncés distincts d'être ‘exécuter-ensemble’ avec une pause minimale entre eux.
- Parole en continue:ces systèmes permettent aux utilisateurs de parler presque naturellement, alors que l'ordinateur détermine le contenu.
- Parole spontanée: Un tel système devrait être capable de gérer une variété de caractéristiques de parole naturelle comme étant mots mis ensemble,et même de légers bégaiement.

Techniques de reconnaissance vocale:

- Fondamentalement, il existe trois approches de reconnaissance de la parole.



- [2].

L'approche Acoustique-Phonétique

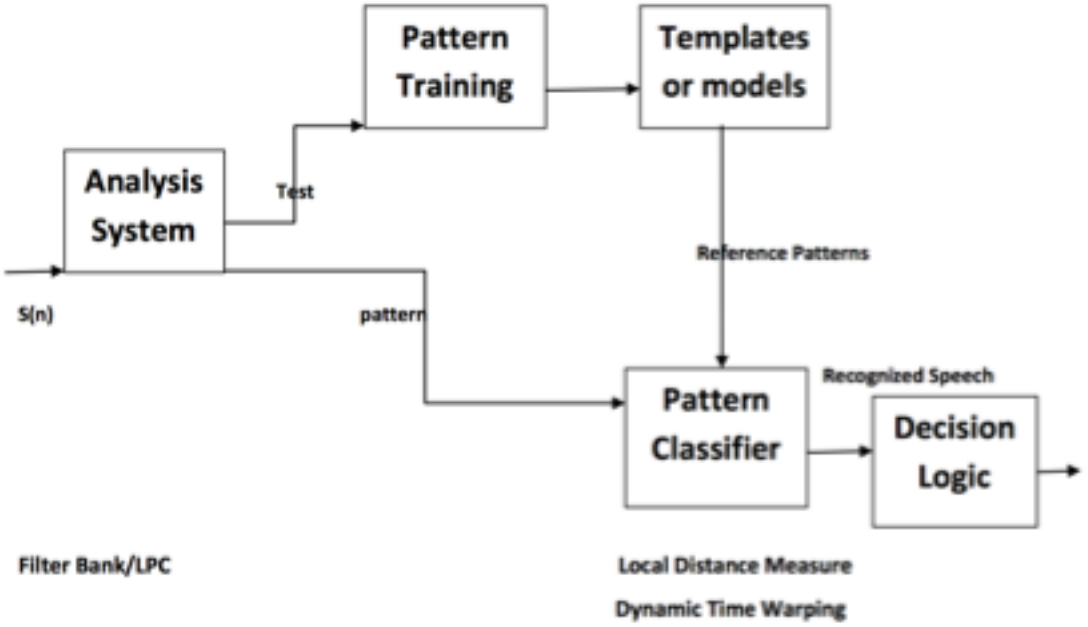
- Cette approche selon Hemdal et Hughes[3], postule qu'il existe des unités phonétiques distinctives dans la langue parlée et ces unités sont caractérisées par un ensemble de propriétés acoustiques.
- Cette approche est mise en œuvre dans l'ordre suivant:
    - l'analyse spectrale,
    - l'extraction des caractéristiques,
    - segmentation et étiquetage,
    - reconnaissance du mot(ou de la chaine) valide
- Cette approche n'a pas été largement utilisée dans les applications commerciales.

L’approche par reconnaissance de modèle[4]

- Cette approche implique deux étapes essentielles :
  - L’entrainement de modèle
  - La comparaison de modèle
- La caractéristique essentielle de cette approche est qu'elle utilise un cadre mathématique bien formulé et établit des représentations cohérentes de la parole.
- Cette approche est devenue la méthode de choix pour la reconnaissance vocale au cours des six dernières décennies.

L'approche par reconnaissance de modèle

- Il existe 2 méthodes:
  - La méthode basée sur les templates
  - La méthode stochastique

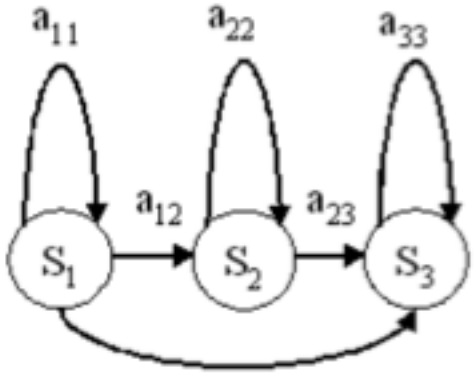


L'approche par reconnaissance de modèle

- La méthode basée sur les templates[5]
  - Idée: une collection de motifs de parole prototypes sont stockées comme des modèles de référence représentant le dictionnaire des mots candidats. La reconnaissance s'effectue ensuite en faisant correspondre un énoncé inconnu pour chacun de ces modèles de référence et en sélectionnant la catégorie du meilleur motif.
- Mais elle a l'inconvénient que les variations de la parole peuvent être modélisées en utilisant de nombreux modèles par mots, qui devient finalement impossible.

L'approche par reconnaissance de modèle

- La méthode stochastique[5]
  - Idée: l'utilisation des modèles probabilistes pour faire face aux informations incomplètes
- L'approche stochastique la plus utilisée a ce jour est la modélisation des modèles de Markov cachés (HMM), celle ci est plus générale et possède une base mathématique plus solide par rapport à la méthode basée sur les templates.



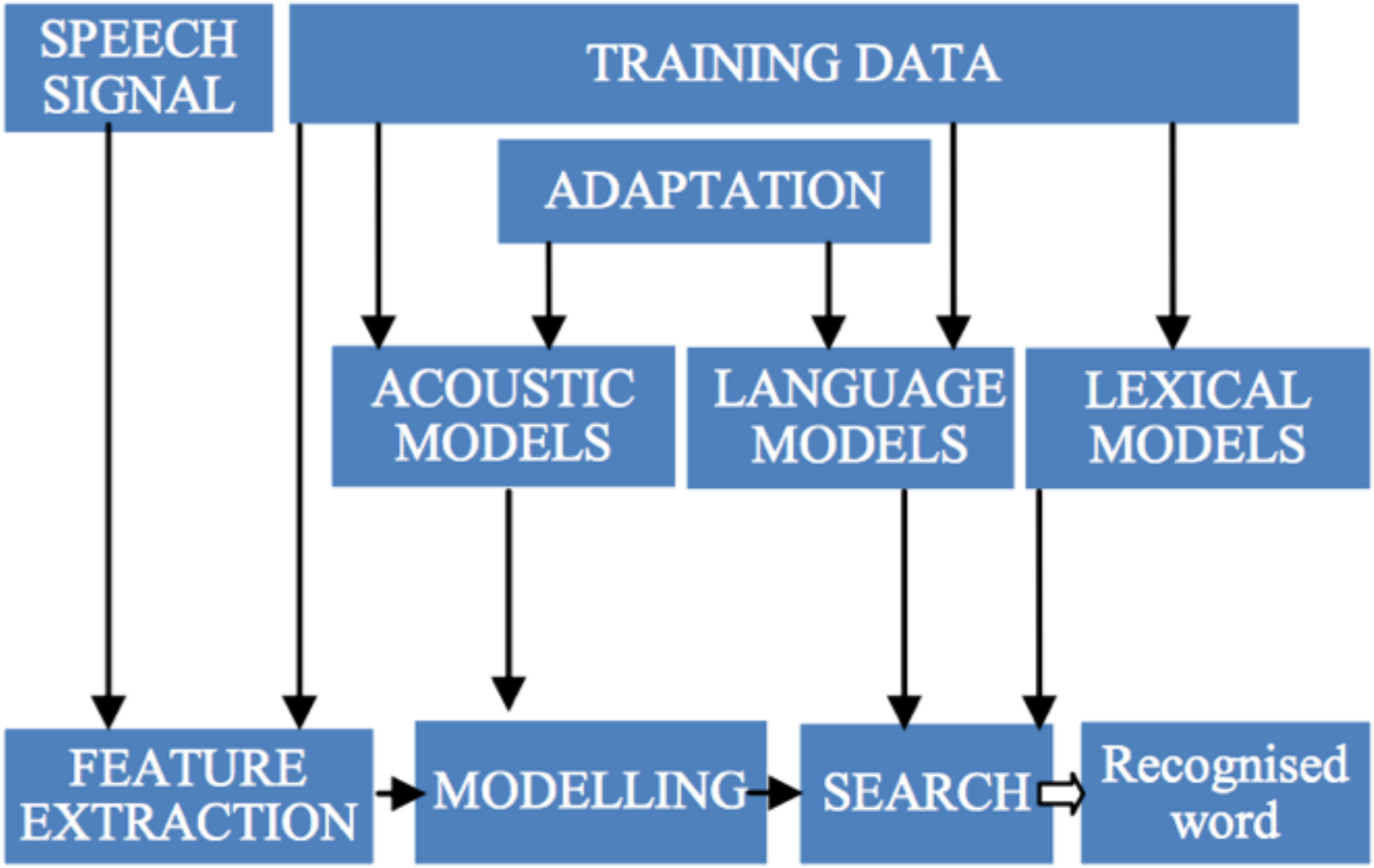
L’approche par l’Intelligence artificielle[5]

- Cette approche est une fusion entre l’approche acoustique-phonétique et celle par reconnaissance de modèle.
- Les connaissances phonétiques-acoustiques permettent d'élaborer des règles de classification pour les sons.
- D'autre part, la littérature linguistique et phonétique fournit des indications sur le traitement de la parole humaine.
- Cependant, cette approche n’a eu un succès que partiel en raison de la complexité d'en quantifier les connaissances d'experts. Un autre problème est l'intégration des niveaux de la connaissance humaine i.e: la phonétique,l’analyse syntaxique et lexicale, sémantique et pragmatique.

Phases

- Un système de reconnaissance automatique de la parole comporte deux phases:
  - la phase d'entrainement.
  - la phase de reconnaissance
- Un système RAP ne peut reconnaitre que ce qu'il a appris a l'entrainement.

Modules


SPEECH SIGNAL
TRAINING DATA
ADAPTATION
ACOUSTIC MODELS
LANGUAGE MODELS
LEXICAL MODELS
FEATURE EXTRACTION
MODELLING
SEARCH
Recognised word

Modules

- L'extraction de caractéristiques(Feature Extraction)
- Différentes techniques pour l'extraction de caractéristiques sont LPC, MFCC, RAS, DAS, ΔMFCC,AMFCC, etc…
- Cette étape conduit à la suppression des principales composantes de bruit.

Modules

- Le modèle acoustique(Acoustic model):
- cette couche représente la majeure partie de la charge de calcul et des performances du système.
- Le modèle acoustique est développé pour détecter le phonème prononcé. Sa création implique l'utilisation d'enregistrements audios et de leurs scripts textuel, puis les compile dans une représentation statistique des sons qui composent les mots.

Modules

Modules

- Le modèle lexical(Lexical model):
- Lexicon est développé, Pour fournir la prononciation de chaque mot dans une langue donnée. Diverses combinaisons de sons sont définis grâce à un modèle de lexique pour donner des mots valides pour la reconnaissance.
- Les réseaux de neurones ont contribué à développer ces modèles.

Modules

Modules

- Modèle de Langue(Language model):
- Les systèmes RAP utilisent les modèles de langue *n-gram* pour rechercher des séquences de mots corrects en prédisant la probabilité du *n-ième* mot sur la base des *n-1* mots précédents.
- $P(W) = P(w_1, w_2, \ldots\ldots\ldots\ldots, w_{m-1}, w_m) = P(w_1).\ P(w_2 \mid w_1).\ P(w_3 \mid w_1 w_2) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ P(w_m \mid w_1 w_2 w_3 \ldots\ldots\ldots\ldots\ldots w_{m-1})$.

Base de données de Parole

Base de données de Parole

- Pour développer une BDD de parole la méthodologie générale suivante est adoptée:
  - Corpus de texte:
    - Génération d'ensemble optimal de phrases textuelles est la première étape.(Collection du corpus de texte - Ecriture en signaux phonétiques le corpus du texte- Optimiser la collection)
  - Collection des données de parole:
    - Le corpus de texte est utilisée finalement pour enregistrer les mots/ phrases par un ou plusieurs locuteur(s) en fonction des besoins(classes, prononciation).(Choix du locuteur - Données statistiques - Correction de la transcription)

Performance

Performance

- La performance de ces systèmes est exprimée en terme de précision et de vitesse:
- La précision est mesurée en terme connu comme le taux d'erreur de mot (Word Error Rate), alors que la vitesse est mesurée avec le facteur temps réel.

Performance

Performance

- Précision:
- Word Error Rate(WER): est une mesure commune de la performance de reconnaissance vocale. Les séquences de mots reconnues ont une longueur différente des séquences de mots de référence, il est difficile de mesurer le rendement.
- WER = (S+D+I) / N
- Avec S le nombre de substitutions,D le nombre de 'Deletions',I le nombre d'insertions et N le nombre de mots dans la référence.
- Parfois taux de reconnaissance des mots (Word Recognize Rate) est utilisé à la place du WER.
- WRR = 1- WER

Performance

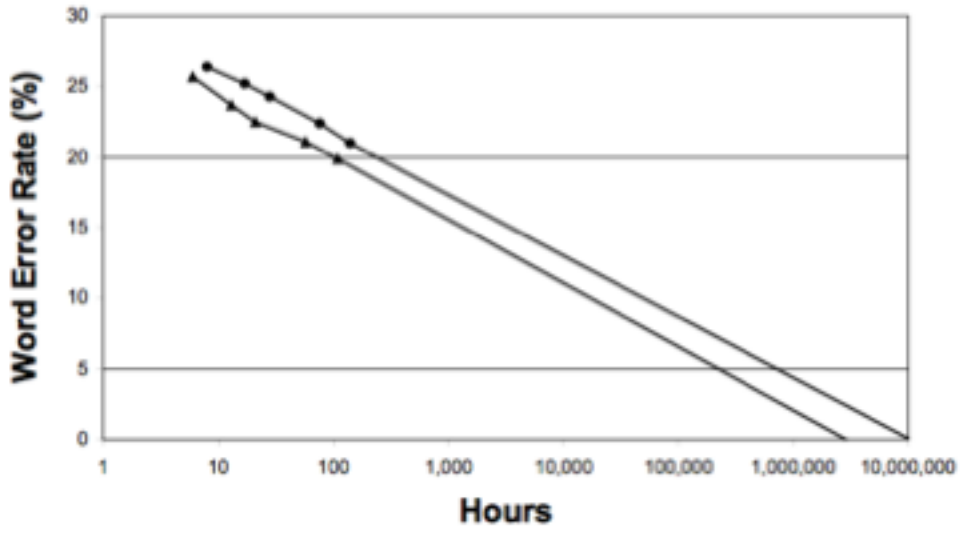
Performance

- Vitesse :
- Elle est mesurée par le facteur de temps réel. Si cela prend du temps T pour traiter une entrée de durée D .
- RTF = T / D
- RTF ≤1 ~ traitement en temps réel.
- E.g: Le facteur temps réel est 2 si cela prend 6 heures de calcul pour traiter un enregistrement d'une durée de 3 heures.

Performance

Performance

UNFILTERED		FILTERED	
Hours	WER	Hours	WER
8	26.4	6	25.7
17	25.2	13	23.7
28	24.3	21	22.5
76	22.4	57	21.1
140	21.0	108	19.9



- LVCSR [6]

Applications

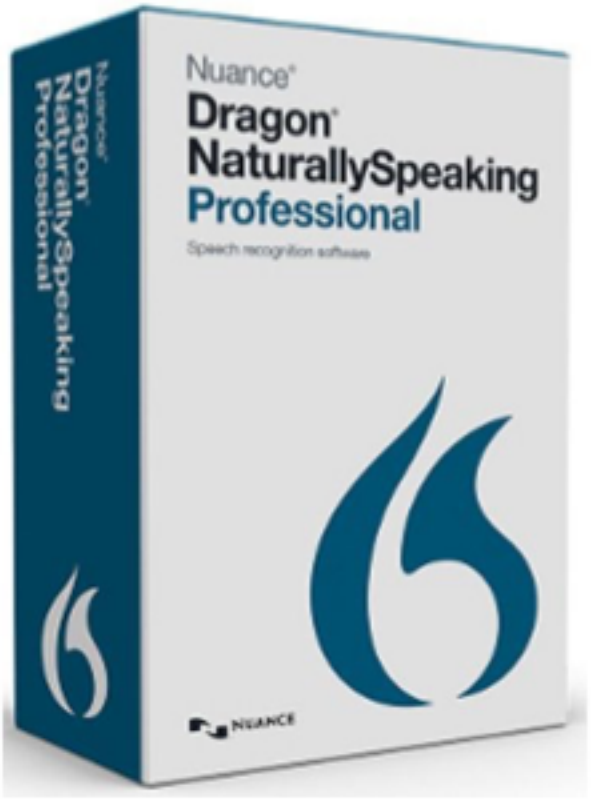
Applications

- La dictée vocale
- Les serveurs d'informations par téléphone
- La recherche d'informations
- La sécurité possible grâce à la signature vocale
- La possibilité de commande et de contrôle d'appareils à distance.
- Accès rapide et mains-libre pour les médecins en pleine opération …

Exemple d'application

Exemple d'application

- Dragon Professional - Nuance



Conclusion

Conclusion

- Il s'agit la d'un domaine complexe mais très prometteur et proposant beaucoup de défis.
- L'utilisation de la reconnaissance automatique de la parole devient courante et devrait très bientôt apparaître dans la plupart des domaines d'activités et la plupart des applications futures.

Références

References

- [1]S.K.Katti, M. A. A. (2009). "Speech Recognition by Machine: A Review." (IJCSIS) International Journal of Computer Science and Information Security, **6**(3).
- [2]Singh, S. J. A. R. P. (2012). "Automatic Speech Recognition: A Review. » International Journal of Computer Applications **60**(9).
- [3]Hemdal, J. F., and G. W. Hughes (1967). "A feature based computer recognition program for the modeling of vowel perception." Models for the Perception of speech and visual form.
- [4]Lawrence Rabiner, B.-H. J. (1993). Fundamentals of speech recognition, Prentice-Hall International, Inc.
- [5]Moore, R. K. (1994). "Twenty things we still don't know about speech." Progress and Prospects of speech Research an Technology.
- [6]Lamel, L., Gauvain, J.-L. and Adda, G. (2000). "Lightly Supervised Acoustic Model Training." Proc. ISCA Workshop on Automatic Speech Recognition,: 150-154.