La reconnaissance vocale

Théodore APAPOULLE

Thibault CHATIRON

Plan

- Introduction
- Définition
- Applications de reconnaissance vocale
- Problématiques liées à la reconnaissance vocale
- Principe de fonctionnement
- Robustesse des systèmes de reconnaissance vocale
- Conclusion

Traitement automatique de la parole

- Ensemble de 6 grands thèmes:
  - Codage et compression de la parole
  - Synthèse de la parole
  - Reconnaissance et vérification du locuteur
  - Identification de la langue
  - Détermination de l'état émotionnel d'un locuteur
  - Reconnaissance de la parole

La reconnaissance vocale

- Domaine recouvrant tous les aspects liés à l'interprétation, par la machine, du langage humain.
- Domaine de la science ayant toujours eu un grand attrait auprès des chercheurs comme auprès du grand public
- Exemples
  - Piloter son installation domestique à la voix
  - Ne plus avoir à taper pendant des heures sur un clavier pour rédiger un rapport

Applications de la reconnaissance vocale

- Trois grands types de systèmes :
  - Les systèmes de commandes vocales
  - Les systèmes de dictée automatique
  - Les systèmes de compréhension.

- Exemples :
  - Aide à la navigation à bord de voiture
  - Aide au handicap
  - Saisie de données

La parole humaine

- Flux continu constitué d'une suite de mots, eux mêmes étant constitués d'un enchainement de phonèmes et de bruits articulatoires.

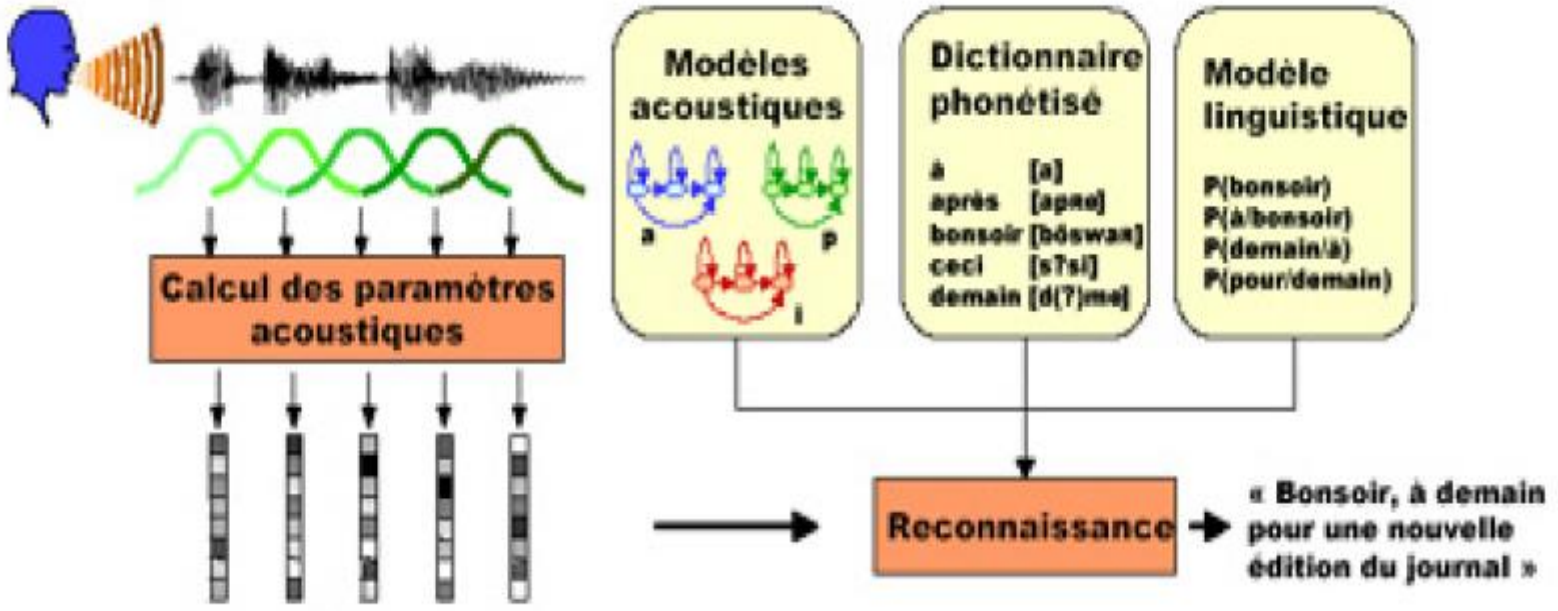
- Phonème : Unité distinctive de prononciation dans une langue.
  - Exemple : /ɛ / et / ɛ: / dans père et paire

- Parole humaine: Très variable puisqu'un même phonème possède de nombreux paramètres qui sont en fonction du locuteur.
  - Intensité de la voix
  - hauteur de la voix
  - type de son émis par le locuteur (chuchotement, chant, parole)
  - émotion dans la voix du locuteur

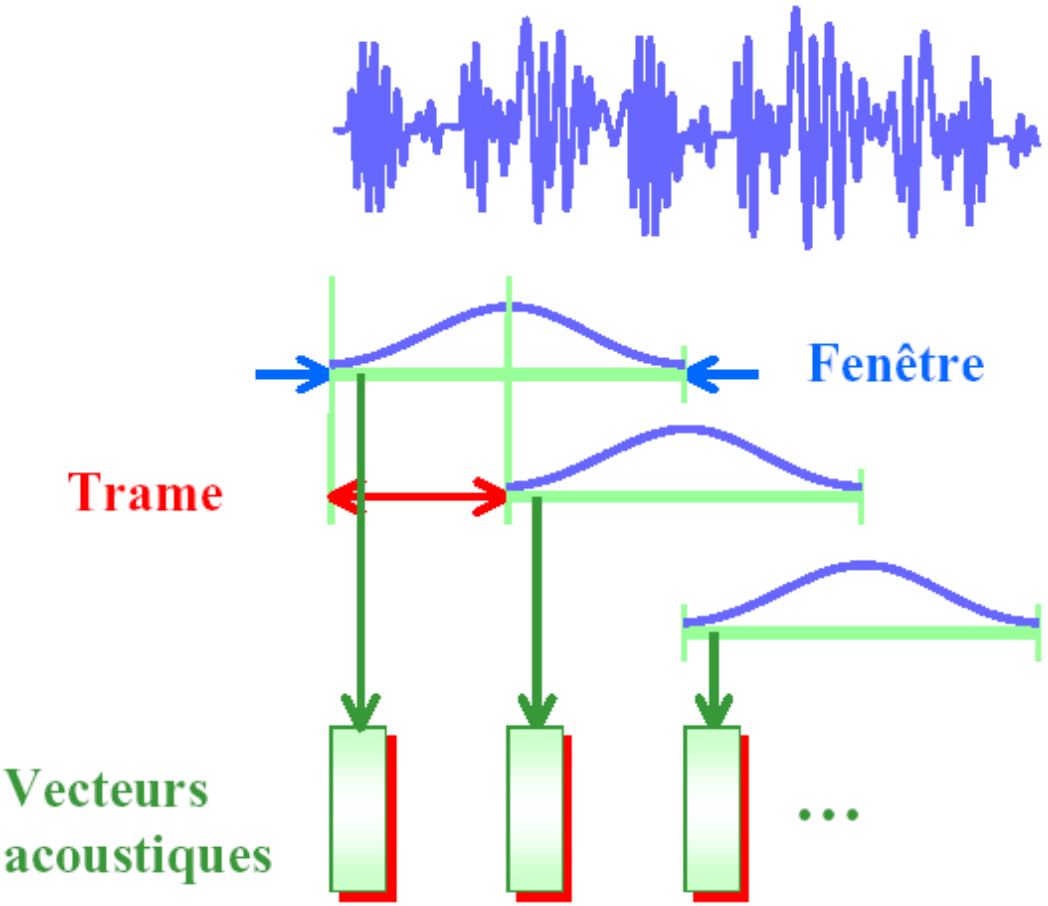
Problématiques

- Plusieurs problèmes font que le reconnaissance de la parole est un domaine difficile
  - Une grande variabilité de la parole
    - Variabilité intralocuteur : voix chantée, criée, murmurée, enrhumée, enrouée, sous stress, bégaiement . ..
    - Variabilité interlocuteur : timbres différents, voix masculines, féminines, voix d'enfants
  - Continuité et coarticulation
    - La production d'un son est fortement influencée par le son qui le précède et qui le suit en raison de l'anticipation du geste articulatoire.

Architecture d'un système de reconnaissance vocale



Analyse acoustique du signal parole



Conversion analogique/numérique

- Onde acoustique de parole captée par le microphone
- Transformation de l'onde acoustique de parole en un signal électrique.
- Filtrage pour éliminer tous les composants du signal en dehors de la bande passante [50 Hz - 8 kHz]
- Conversion analogique-numérique du signal :
  - Echantillonnage : la fréquence d'échantillonnage doit donc au moins 8 kHz pour la parole de qualité téléphonique et de 16 à 20 kHz pour la parole de bonne qualité
  - Quantification

.

Préaccentuation

- Le signal échantillonné est pré-accentué : Ressortir les hautes fréquences avec un filtre numérique à réponse impulsionnelle finie de premier ordre
- Hautes fréquences moins énergétiques que les basses fréquences

Segmentation

- Méthodes du traitement de signal utilisées dans l'analyse du signal opèrent sur des signaux stationnaires
- Parole: un signal non stationnaire.
- Solution : Analyse de ce signal effectuée sur des trames successives de parole, de durée relativement courte sur lesquelles le signal peut en général être considéré comme quasi stationnaire
- Découpage du signal pré accentué en trames de N échantillons de parole.
- En général N est fixé de telle manière à ce que chaque trame corresponde à environ 20 à 30 ms de parole.

Fenêtrage

- La segmentation du signal en trames produit des discontinuités aux frontières des trames (Lobes secondaires).
- Réduction de ces effets en multipliant les échantillons de la trame par une fenêtre de pondération telle que la fenêtre de Hamming

Analyse à court terme

- Analyse à court terme : chaque trame fenêtrée du signal est ensuite convertie en un vecteur acoustique constitué d'un ensemble réduit de paramètres

- Différentes méthodes coexistent pour la transformation d'une trame fenêtrée de signal en un vecteur acoustique
  - Méthodes non paramétriques
  - Méthodes paramétriques
  - Méthodes avec modèles de perception

Analyse à court terme

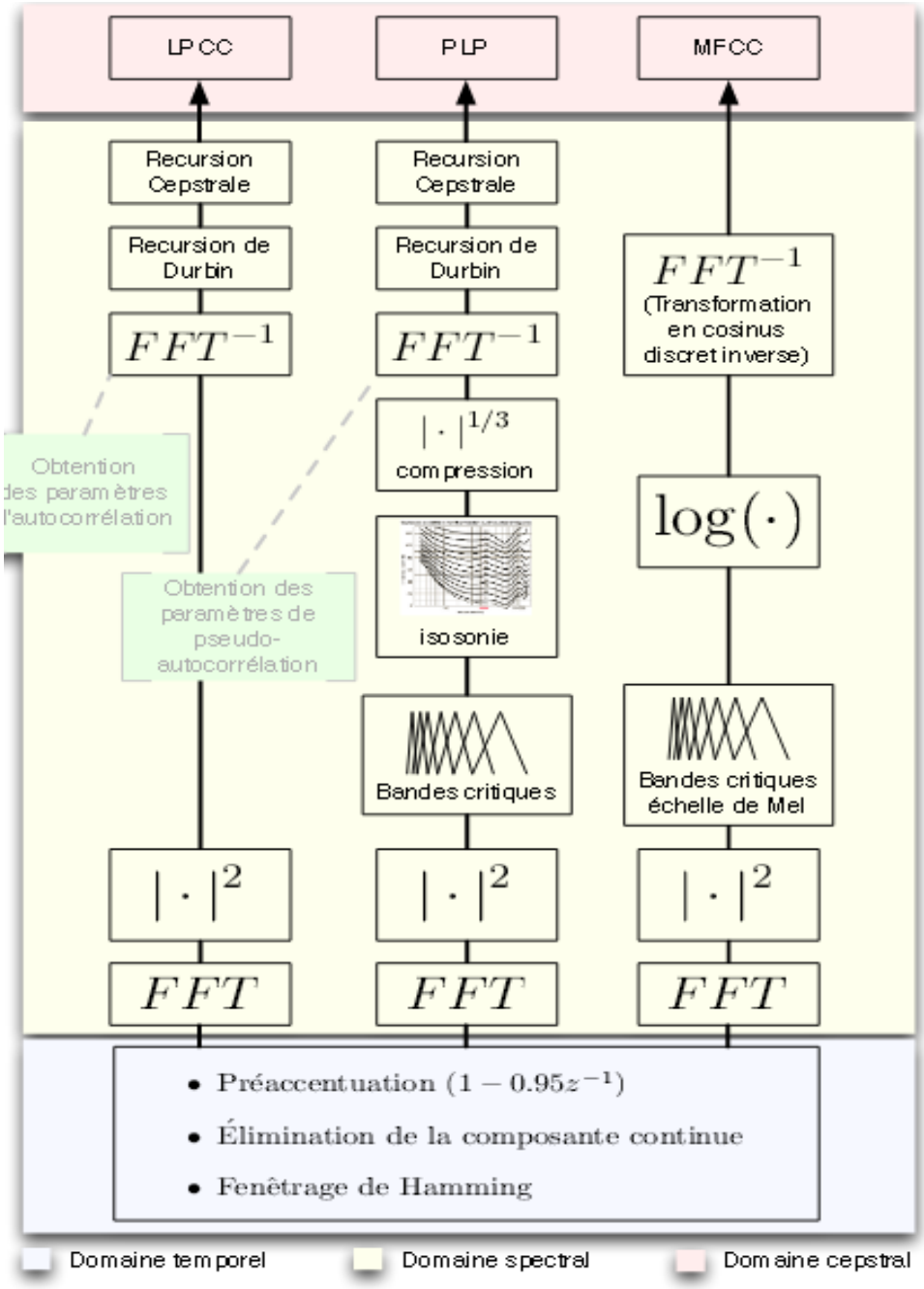
- Les méthodes paramétriques qui se basent sur un modèle de production
  - Codage par prédiction linéaire LPC (Linear Prediction Coding)
  - LPCC (Linear Prediction Cepstral Coefficients).

- Les méthodes non paramétriques
  - le taux de passage par zéro,
  - la fréquence fondamental (pitch),
  - la transformée de Fourier discrète,
  - l'énergie du signal,
  - les sorties d'un banc de filtres numériques
  - la transformée en ondelettes.

- Les méthodes fondées sur un modèle de perception
  - MFCC (Mel Frequency Cepstral Coefficients)
  - PLP(Perceptual Linear Prediction)

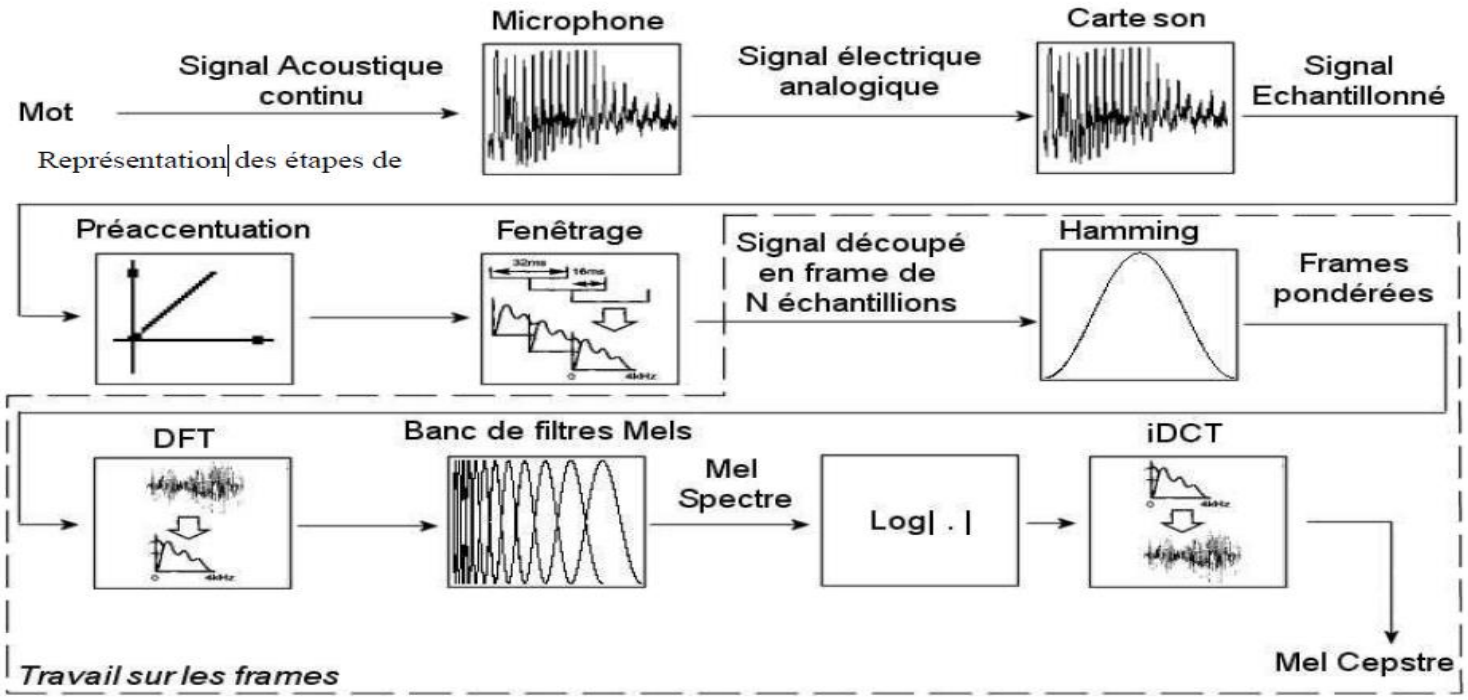
Analyse à court terme

- Les méthodes les plus couramment utilisées:
  - MFCC
  - PLP
  - LPCC



Méthode MFCC

- Exploiter les propriétés du système auditif humain par la transformation de l'échelle linéaire des fréquences en échelle Mel



Représentation des étapes de création des MFCC.

Méthode de reconnaissance vocale

- Décodage acoustico-phonétique
  - Extraire les paramètres choisis pour représenter le signal
  - Décoder le signal d'entrée

Les techniques (1/2)

- Approche globale : le mot
  - Fournir une image acoustique de chaque mots à identifier

- Limite :
  - petits vocabulaires
  - nombre restreint de locuteurs

Les techniques (2/2)

- Approche analytique : la structure des mots
  - Identifier les composantes élémentaires (phonèmes, syllabes, …) => unités de base

- Meilleure approche :
  - Pour reconnaître de grands vocabulaires, il suffit d'enregistrer dans la mémoire de la machine les principales caractéristiques des unités de base.

Les phases (1/2)

- La phase d'apprentissage :
  - un locuteur prononce l'ensemble du vocabulaire, souvent plusieurs fois, pour créer en machine le dictionnaire de références acoustiques. Pour l'approche analytique, l'ordinateur demande à l'utilisateur d'énoncer des phrases souvent dépourvues de toute signification, mais qui présentent l'intérêt de comporter des successions de phonèmes bien particuliers.

Les phases (2/2)

- La phase de reconnaissance :
  - un locuteur prononce un mot du vocabulaire. Ensuite la reconnaissance du mot est un problème typique de reconnaissance de formes.
  - Tout système de reconnaissance des formes comporte toujours les trois parties suivantes:
    - Un capteur permettant d'appréhender le phénomène physique considéré (microphone),
    - Paramétrisation des formes (analyseur spectral),
    - Décision de classer une forme inconnue dans l'une des catégories possibles

Reconnaissance du mot

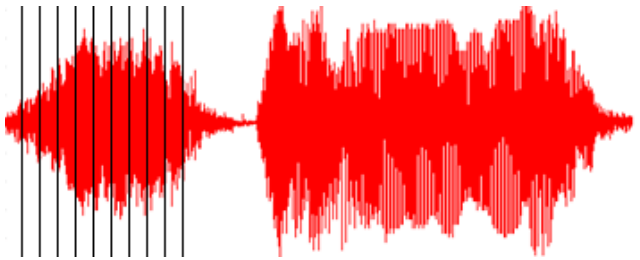
- Signal vocal comparé aux mots du dictionnaire de référence

- L'algorithme de reconnaissance permet de choisir le mot le plus ressemblant, en calculant le taux de similitude entre le mot prononcé et les diverses références.
  - Les modèles de Markov à états cachés (Hidden Markov Model)
    - Modèle acoustique
    - Résultats :
      - Donne la probabilité de correspondance à phonème
      - Associer le phonème le plus probable à la tranche

- Le programme va comparer le mot prononcé par le locuteur avec ceux qui sont en mémoire depuis l'apprentissage
  - Trouver le signal le plus ressemblant.

L'étape de reconnaissance

- Analyser le signal inconnu sous la forme d'une suite de vecteurs acoustiques
- Comparer la suite inconnue à des exemples préalablement enregistrés.
- Le mot «reconnu» sera alors celui dont la suite de vecteurs acoustiques colle le mieux à celle du mot inconnu.



Principe HMM

- Soit A un signal acoustique, le processus de reconnaissance peut être décrit comme le calcul de la probabilité P(W | A) qu'une suite de mots (ou phrase) W corresponde au signal acoustique A, et la détermination de la suite de mots qui maximise cette probabilité.

$$P(W|A) = P(W) \cdot P(A|W)/P(A)$$

et

$$P(\hat{W}|A) = \max_i \frac{P(A|W_i) \cdot P(W_i)}{P(A)}$$

- P(W), probabilité de la suite de mots W
- P(A | W), probabilité du signal acoustique A, étant donné la suite W,
- P(A), probabilité du signal acoustique.

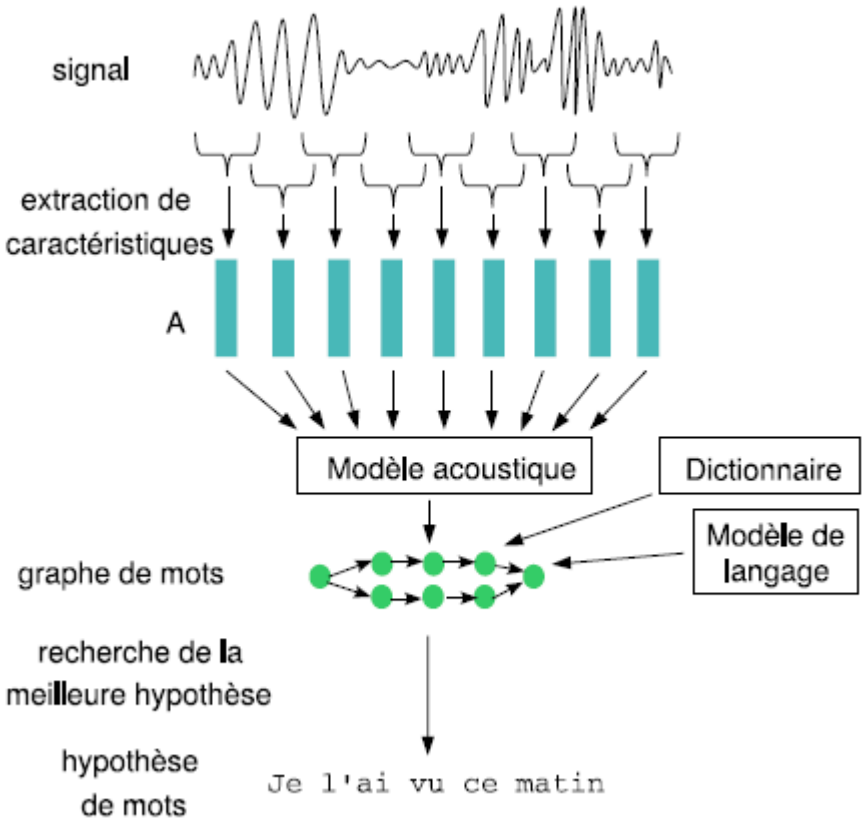
=> Nécessaire de considérer P(A | W) (modèle acoustique) et P(W) (modèle linguistique).

Modèle linguistique

- Travail sur la syntaxe et la sémantique propre à la langue
  - Probabilité qu'une suite de mots existe dans la langue
  - Introduction de la notion d'approximation avec N-grams

- Algorithme N-grams
  - Agrégation en 2 ou 3 mots avec une probabilité associée
  - Approximation de probabilités de séquences plus longues
  - Calcul des probabilités sur ces séquences plutôt que sur des mots
  - Probabilité d'obtenir un mot connaissant les mots précédents

Modèle acoustique + linguistique



Robustesse

- Le système est-il capable de fonctionner dans des conditions difficiles ?
  - Bruits d'environnement (dans une rue, etc…)
  - Déformation de la voix par l'environnement (réverbérations, échos, etc…)
  - Qualité du matériel utilisé (micro, carte son etc…)
  - Bande passante fréquentielle limitée (fréquence limitée d'une ligne téléphonique)
  - Elocution inhabituelle ou altérée (stress, émotions, fatigue, etc…)
- Certains systèmes peuvent être plus robustes que d'autres à l'une ou l'autre de ces perturbations, mais en règle générale, les systèmes de reconnaissance de la parole sont encore sensibles à ces perturbations.

Conclusion

- Aujourd'hui :
  - Systèmes fonctionnels basés sur une approche statistique
  - Logiciels de reconnaissance du langage continu
    - Tailles de vocabulaire allant à 60 000 mots,
    - Dictée à la vitesse de 120 à 160 mots par minute
    - Succès de reconnaissance supérieur à 95%.
- Avenir :
  - Améliorer les modèles acoustiques
  - Améliorer les modèles linguistiques :
    - techniques statistiques et réseaux neuronaux.
  - Rendre les modèles indépendants du locuteur

Réference

- Reconnaissance automatique de la parole : Du signal à son intépretation par *Jean paul Haton, Christophe Cerisara, Dominique Fohr,Yves Laprie, Kamel Smaili*