



<http://www-clips.imag.fr/geod/>

Le système de reconnaissance de la parole RAPHAEL a été développé pour être utilisé dans des applications comme la réservation touristique ou hôtelière par exemple. Il peut aussi être utilisé pour détecter des mots clefs prononcés par un malade à son domicile et caractérisant un état de détresse.

1 La parole

La parole est un son émis par le **locuteur**, c'est à dire une variation de pression acoustique plus ou moins rapide et plus ou moins forte qui est captée par un microphone placé à proximité. La possibilité de reconnaître la phrase sera donc très dépendante des **conditions d'enregistrement** : qualité du microphone lui-même, distance au locuteur, niveau du bruit environnemental.

La source d'énergie utilisée pour produire les sons est l'air contenu dans les poumons. Le flux d'air sous pression parvient à travers la trachée jusqu'au conduit vocal, aux fosses nasales, aux organes d'articulation (langue, lèvres...) qui vont avoir chacun leur rôle dans la production de la parole. La parole sera donc très dépendante des caractéristiques physiques du locuteur : âge, taille, sexe...

La parole captée par le microphone est caractérisée par un **spectre de fréquences acoustiques** qui couvre des fréquences allant de 50 Hz à 8 kHz environ suivant la personne qui parle. Les lignes téléphoniques classiques qui coupent les fréquences plus basses que 300 Hz et plus hautes que 3,3 kHz modifient la perception dans le cas des personnes ayant une voix très grave ou très aiguë. La partie supérieure de la figure 1 montre l'évolution au cours du temps de la pression acoustique mesurée par le microphone lorsque un locuteur prononce la phrase "Ça va pas du tout".

Sur la partie représentant le sonogramme en figure 1, le spectre est affiché entre 0 Hz et 5.000 Hz (échelle des ordonnées sur la gauche), le premier formant apparaît sous forme d'une ligne (échelle des ordonnées sur la droite).

Les mots constituant une phrase peuvent se décomposer en une succession de **phonèmes** qui constituent eux-mêmes les sons élémentaires des mots. Les phonèmes sont en général différents d'une langue à l'autre. Chaque phonème est influencé par le phonème qui le précède et par celui qui le suit, c'est le phénomène de coarticulation.

Les sons élémentaires de parole peuvent être classés en fonction de trois variables essentielles : le **voisement** (activité des cordes vocales), le **mode d'articulation** (type de mécanisme de production) et le **lieu d'articulation** (endroit de resserrement du conduit vocal). Le français comprend 12 voyelles orales émises seulement par la bouche, 4 voyelles nasales correspondant à la mise en parallèle des cavités nasales avec la bouche. Il comprend 6 consonnes occlusives correspondant à une ouverture brusque du conduit vocal après son obstruction, 6 consonnes fricatives produites par un rétrécissement local du conduit vocal, 2 consonnes liquides, 3 nasales et 4 semi-consonnes.

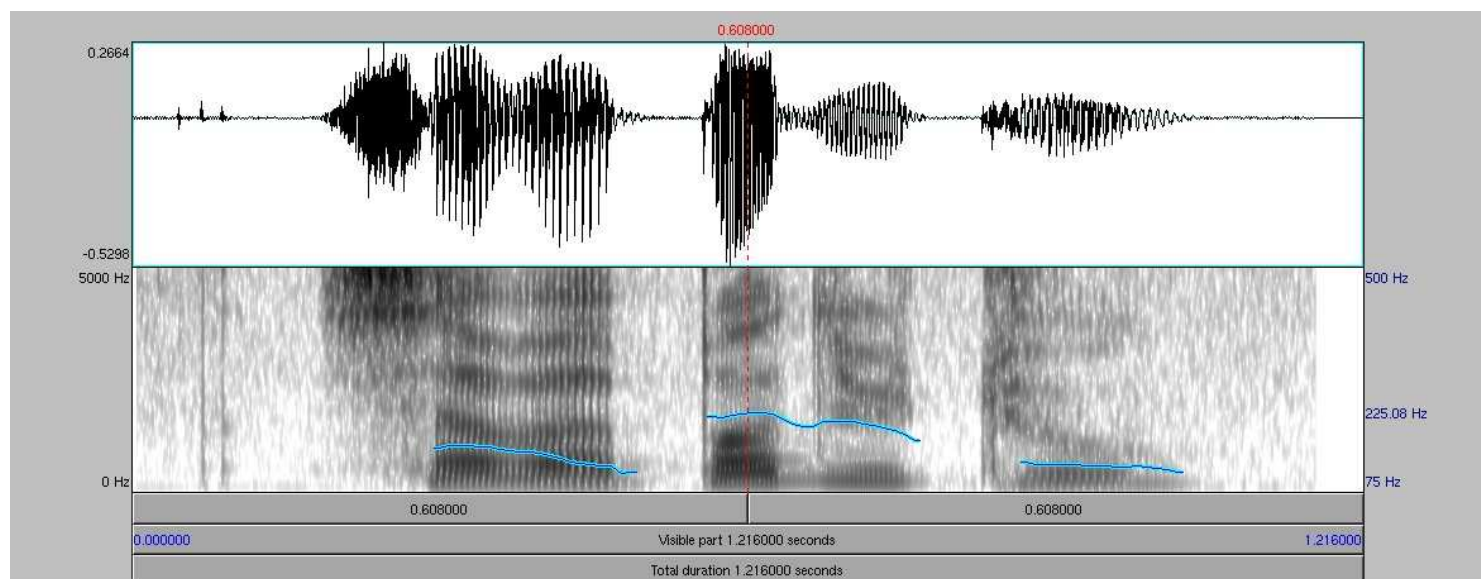


FIG. 1 – Phrase prononcée "Ça va pas du tout" et en dessous le sonogramme correspondant

Il est donc clair que la réalisation d'un système de reconnaissance de parole sera dépendante de la langue parlée et devra prendre en compte la variabilité apportée par les différents locuteurs. Par ailleurs il est possible pour une raison ou pour une autre que le locuteur ne prononce pas correctement la phrase, en cas de stress ou de fatigue par exemple. Il faut aussi tenir compte des différents accents régionaux toujours possibles.

2 Architecture du système RAPHAEL

Le système RAPHAEL est un système basé sur les statistiques, il comprend 4 étages de traitement qui sont représentés (voir figure 2). Les 2 premiers étages ont pour rôle :

- **Interface audio** : de calculer les paramètres **acoustiques** à partir du spectre (analogue à l'affichage d'un sonogramme),
- **Module acoustique** : d'extraire les successions les plus probables de **phonèmes**.

Pour s'affranchir le plus possible de l'influence du **locuteur**, le système a été entraîné sur des bases de données contenant des enregistrements de plus de 200 locuteurs. La méthode statistique utilise les modèles de Markov cachés (HMM en anglais), chaque phonème est représenté par une sorte d'automate à 3 états : établissement, continuation, disparition.

Les 2 étages suivants vont procéder à l'extraction des mots les plus probables puis des phrases les plus probables :

1. **Extraction de mots** à partir d'un dictionnaire faisant la correspondance entre chacun des mots (environ 20.000) et les phonèmes (notons qu'il peut y avoir plusieurs façons de prononcer un même mot),
2. **Extraction de phrases** à partir des probabilités d'apparition de chacun des mots pris séparément, pris 2 à 2 ou 3 à 3 ; ces probabilités sont apprises en utilisant des grands **corpus** de texte comme par exemple le journal "Le Monde".

Pour être capable de reconnaître une phrase quelconque, il faut que le système contienne le plus de mots possible, cela conduit à des probabilités très faibles pour chacun des mots d'où un accroissement du taux d'erreur de reconnaissance. Par ailleurs les corpus écrits contiennent du texte en français grammaticalement correct ce qui est bien moins respecté à l'oral. Les mots employés en parole spontanée ne sont pas non plus les mêmes.

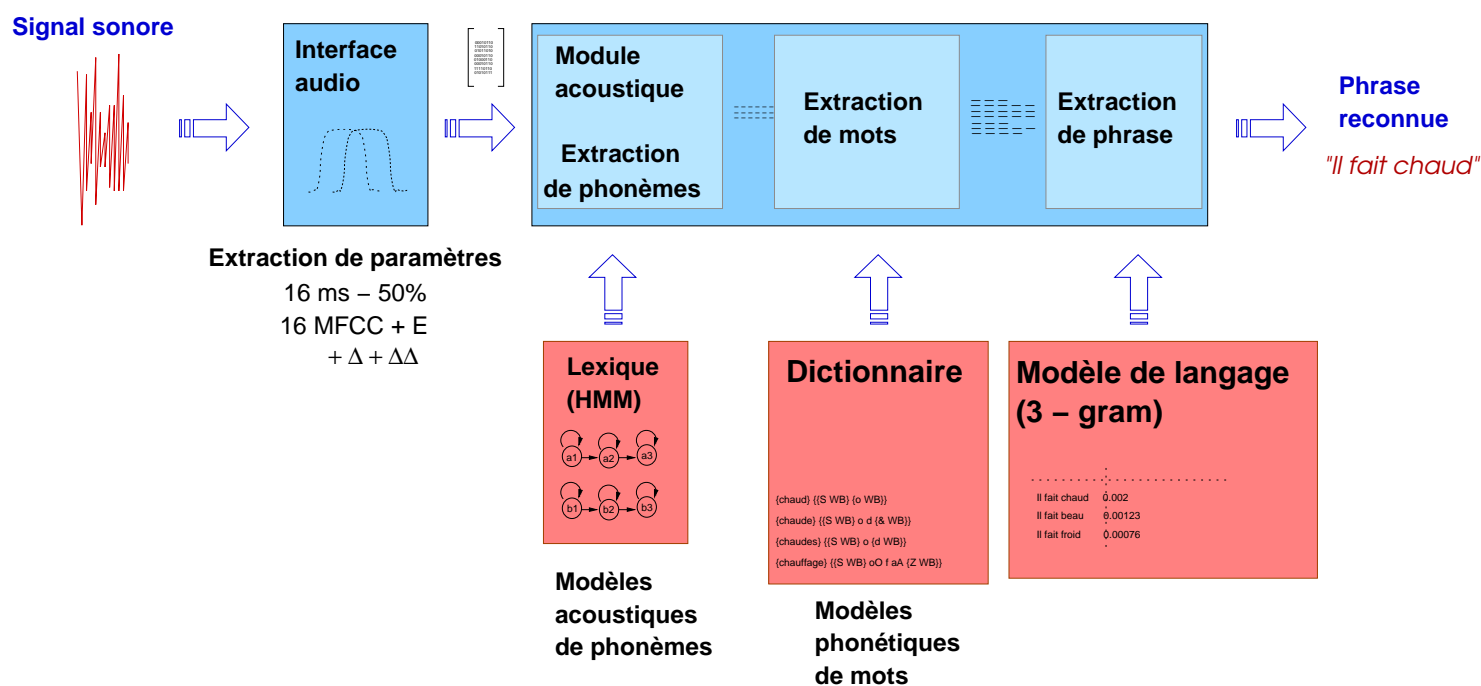


FIG. 2 – Comment fonctionne le système RAPHAEL

3 Encore de nombreux obstacles à lever

Pour obtenir un système robuste il faut s'affranchir des problèmes suivants :

- **bruit** présent lors de l'enregistrement,
- **parole spontanée** différente de phrases lues en studio, les corpus de parole spontanée sont presque inexistantes alors qu'ils sont indispensables pour obtenir les modèles de langage adaptés,
- **vocabulaire** important plus difficile à reconnaître qu'un ensemble de mots restreint et limités à un domaine particulier,
- **reconnaissance multilingue**.

4 Contacts

Laurent.Besacier@imag.fr, Brigitte.Bigi@imag.fr, Jean-Francois.Serignat@imag.fr, Michel.Vacher@imag.fr