

Traitement Automatique de la Parole

N7 3SN-M

Jérôme Farinas
jerome.farinas@irit.fr

Institut de Recherche en Informatique de Toulouse
Université de Toulouse

10 janvier 2025



UNIVERSITÉ
DE TOULOUSE

Recherche d'informations audio I

- Hype curve de Gartner pour l'Intelligence Artificielle en 2019¹ :



- 1. New Technology is the name we give to stuff that doesn't work yet (Douglas Adams)

Présentation du cours

L'objectif de ce cours est de vous apporter des connaissances sur les traitements automatiques sur le signal audio, et plus particulièrement la transcription de la parole.

Détail des parties :

- A la découverte d'un signal aléatoire : la parole n'est pas un signal facile à traiter, la connaissance de sa production et perception par l'humain sera utile pour comprendre les traitements qui seront mis en place par les ordinateurs.
- Paramétrisation du signal audio : comment le signal audio arrive-t-il dans l'ordinateur et quels sont les traitements mis en place pour le traiter.
- Traitements automatiques de la parole : reconnaissance du locuteur, reconnaissance de la langue, transcription de la parole.
- Méthodes d'évaluation et implémentations : comment mettre en place un système de traitement automatique de la parole

Organisation de l'enseignement

- 3 séances de Cours-TD
- 3 séances de TP :
 - séance 1 : découverte du signal
 - séance 2 : paramétrisation
 - séance 3 : mise en place d'une modélisation de la parole

Contrôle de connaissances

- UE Audionumérique (N9EN16A) = XX% (N9EN16) comptant X ECTS du parcours Multimedia
- Coefficients :
 - examen écrit 1h30 (notes et support de cours autorisées) = 60%
 - TPs (notebook complété et annoté) = 40%

Données audiovisuelles

- En France, il existe un organisme qui est chargé d'archives radio et télévision : l'Institut National de l'Audiovisuel (INA).

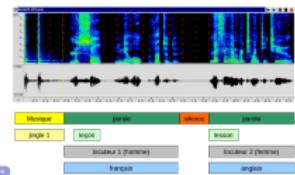


- Jusqu'avant les années 2000 l'indexation du contenu était réalisé de manière manuelle : les métadonnées (sujets, personnes, résumé, etc.) étaient enregistrées en se basant sur les documents réalisés par les producteurs ou bien suite au visionnage du programme. On assiste maintenant à une multiplication de chaînes, le traitement manuel augmenterait de manière exponentielle les moyens financiers nécessaires. L'INA consacre une bonne partie de sa recherche et développement sur l'indexation automatique du contenu audiovisuel.

Indexation automatique du contenu

- L'analyse des documents audiovisuels permet d'extraire de l'information à différents niveaux :
 - segmentation en zones de parole/non parole, musique/non musique, bruits
 - segmentation et identification des locuteurs
 - transcription de la parole
 - détection d'entités nommées (personnes identifiées, lieux, monuments...)
 - segmentation en émissions, détection du thème
 - production de résumés automatiques
- L'analyse peut se faire exclusivement sur le canal audio, ou bien porter sur des analyses combinées audio et vidéo

Exemple de traitements automatiques



<http://www.ina.fr/>

Voxalead : un exemple d'application

http://voxaleadnews.labs.exalead.com/ (HS)

grèves

Première partie I

A la découverte d'un signal aléatoire

Plan

➊ Production et perception de la parole

- ➏ Production de la parole
- ➏ Perception de la parole

➋ Description acoustique des sons

Plan

➊ Production et perception de la parole

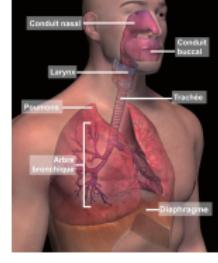
- ➏ Production de la parole
- ➏ Perception de la parole

➋ Description acoustique des sons

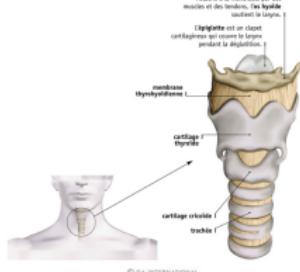
Production de la parole

- ➊ un système : entrée → filtre → sortie
- ➌ parole : poumons, larynx → conduit buccal, nasal, lèvres → onde acoustique
- ➌ musique (violon) : corde + arche → corps du violon → onde acoustique
- ➌ musique (piano) : corde + frappe → type de piano → onde acoustique
- ➌ Phénomène de résonnance fréquentielle
 - ➏ Fréquence fondamentale (fréquence d'entrée)
 $f_0 = \frac{1}{T_0}$ (signal périodique)
 - ➏ Fréquence de résonnance de l'outil (formants)
F1, F2, F3...

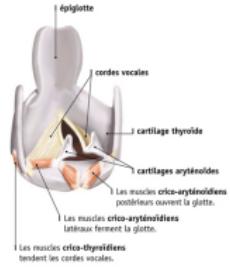
Appareil phonatoire



Larynx



Cordes vocales



Prosodie

$$\begin{aligned} \text{Prosodie} &= f_0 + \text{énergie} + \text{durée} \quad (1) \\ &= \text{hauteur} + \text{intensité} + \text{longueur} \quad (2) \\ &= \text{mélodie} + \text{rythme} \\ &\quad + \text{accentuation} \quad (3) \end{aligned}$$

- ➊ grandeurs acoustiques
- ➋ grandeurs perçues
- ➌ structures

Prosodie : fonctions I

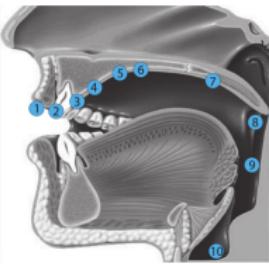
- ➊ accentuation
 - ▶ distinction entre homonymes en anglais : **contrast** (un contre) / **contrast** (contraster)
 - ▶ emphase et focalisation
 - Je vais terminer (par opposition à quelqu'un d'autre)
 - Je vais terminer (par opposition à une action déjà accomplie)
 - Je vais terminer (par opposition à une autre action)
- ➋ structuration de l'énoncé
 - un **vieux armagnac**
 - un **vieux maniaque**
- ➌ mode de la phrase
 - déclaratif
 - interrogatif
 - injonctif
 - exclamatif
- ➍ expressive

Prosodie : fonctions II

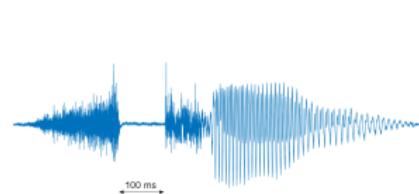
- attitude vis à vis de l'énoncé
 - adhésion plus ou moins forte
 - conviction
 - doute
 - accord ou désaccord
 - approbation ou désapprobation
 - invitation
 - incitation
- gestion des tours de parole
- état psychologique du locuteur
 - calme ou énervé
 - triste ou gai
 - enthousiaste
 - surpris
 - ...

Lieux articulatoires

- ➊ labial
- ➋ labiodental
- ➌ interdental
- ➍ alvéolaire
- ➎ rétroflexe
- ➏ palatal
- ➐ vélaire
- ➑ uvulaire
- ➒ pharyngal
- ➓ glottal



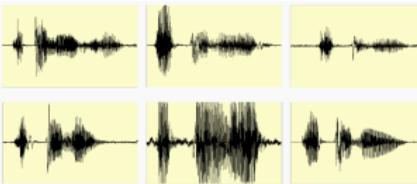
Signal vocal



Variabilité du signal de parole

- ➊ Facteurs intra-locuteurs
 - ▶ variabilité articulatoire
 - ▶ coarticulation (loi du moindre effort)
 - ▶ fréquence de vibration des cordes vocales (f_0) : intonation, hauteur des sons
 - ▶ force d'expiration : intensité des sons
 - ▶ vitesse d'articulation
 - ▶ émotion, stress, rythme...
- ➋ Variabilité para-linguistique
 - ▶ voix chuchotée, normale, criée
 - ▶ débit : réduit, normal, rapide
- ➌ Facteurs inter-locuteurs
 - ▶ variabilité physiologique
 - ▶ variabilité socio-culturelle
- ➍ Environnement
 - ▶ bruits, microphone, canal de transmission
 - ▶ superposition de voix, musique

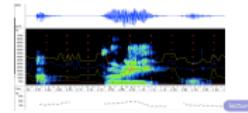
Illustration variabilité de la parole



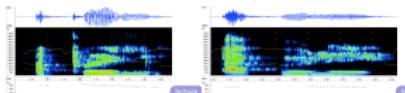
➏ 6 locuteurs différents prononçant le même mot : « appeler »

Illustration variabilité de la parole II

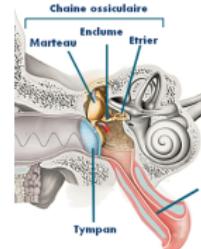
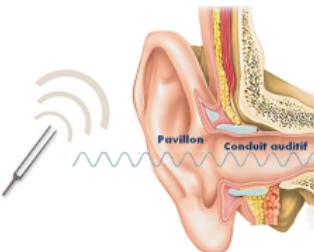
- ➊ parler régional



- ➋ homme/femme/enfant

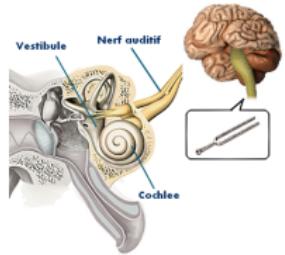


- ① Production et perception de la parole
- Production de la parole
 - Perception de la parole

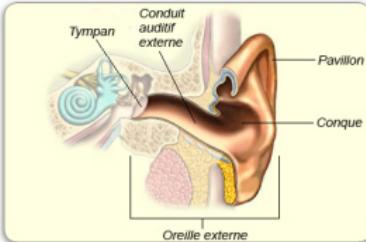


- ② Description acoustique des sons

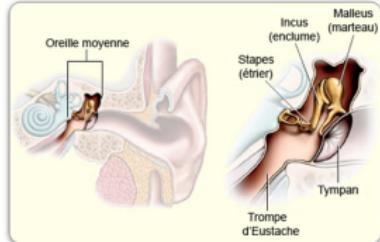
Système auditif humain III



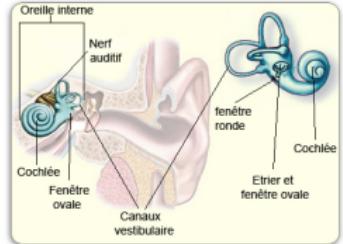
Oreille externe



Oreille moyenne



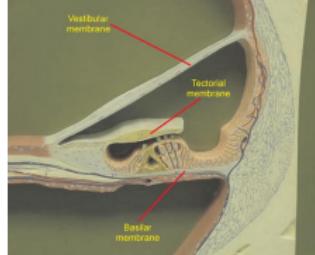
Oreille interne



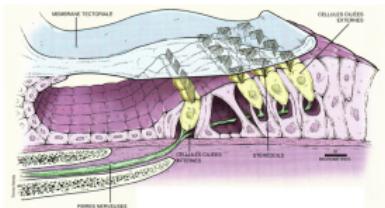
Coupe cochlée



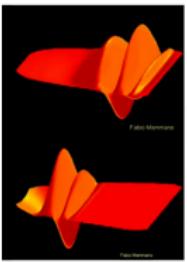
Membranes de la cochlée



Cellules ciliées



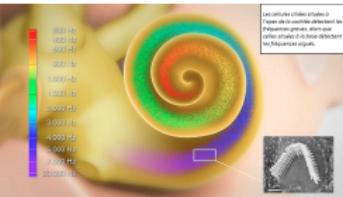
Propagation des sons sur membrane basilaire



Ci-contre, le mouvement de la membrane basilaire (plancher de la cochlée) au passage d'un son grave (fréquence basse) au niveau de l'apex.

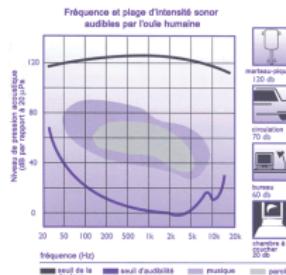
Ci-contre, le mouvement de la membrane basilaire au passage d'un son aigu (fréquence élevée) au niveau de l'apex.

Sensibilité aux sons de la cochlée

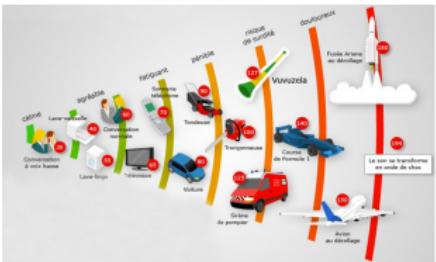


Tous les crêtes ciliaires situées à l'apex de la cochlée détectent les sons aigus. Les crêtes ciliaires situées à la base détectent les fréquences graves.

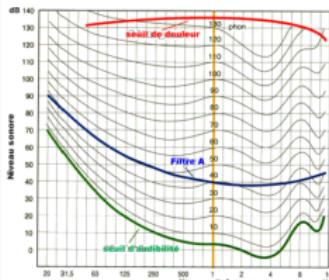
Champ auditif humain



Perception intensité sonore



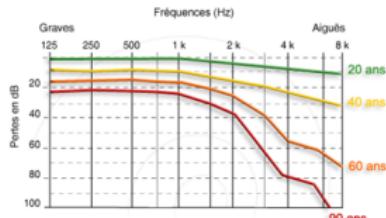
Courbes isosoniques



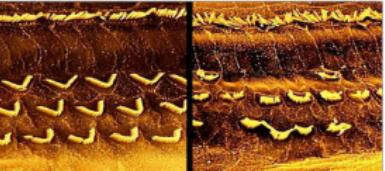
Courbes isosoniques

- Une ligne isosonique, ou courbe isosonique ou courbe d'isosonie, est le lieu des points de même sonie (définis par leur niveau de pression acoustique en dB SPL en fonction de la fréquence), c'est-à-dire provoquant la même sensation d'intensité sonore pour l'oreille humaine.
- Les courbes isosoniques rendent compte de la sensibilité du système auditif humain limité à des fréquences allant de 20 Hz à un maximum d'environ 20 000 Hz. Dans cette gamme de fréquences, la sensibilité est supérieure entre 1 et 5 kHz. Cela est dû principalement à la résonance du canal auditif et à la fonction de transfert des osselets dans l'oreille moyenne.
- Le principe de la mesure consiste à faire entendre aux sujets des sons purs (sinusoïdaux) à différentes fréquences et par incrément de 10 dB. On fait également entendre aux sujets un son de référence à 1 000 Hz. On ajuste l'intensité de ce dernier jusqu'à ce qu'il soit perçu au même niveau sonore que celui en test. Une moyenne des mesures sur les différents sujets est effectuée.

Effet du vieillissement sur la perception des fréquences



Effet du vieillissement sur les cellules ciliées



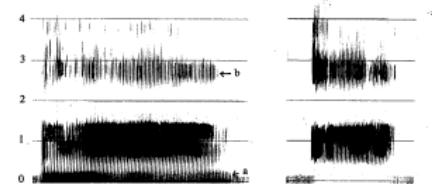
Phonèmes du français²

| | | | | | | |
|-----------|-----|-------|-----|------|-----|--------|
| Consonnes | [p] | paie | [t] | taie | [k] | quai |
| | [b] | baie | [d] | dais | [g] | gai |
| | [m] | mais | [n] | nez | [ɲ] | gagner |
| | [ʃ] | fait | [s] | sait | [ʒ] | chez |
| | [v] | vais | [z] | zéro | [ʒ] | geai |
| | [w] | ouais | [ɥ] | hue | [ɥ] | yéyé |
| | | | [l] | lait | [ʁ] | raie |

| | | | | | | |
|----------|------|------|------|------|------|-------|
| Voyelles | [i] | lit | [y] | lu | [u] | loup |
| | [ɛ] | les | [ø] | leu | [ɔ] | lot |
| | [œ] | lait | [œ̃] | leur | [ɔ̃] | lotte |
| | [ɑ̃] | la | [ɑ̃] | le | [ɔ̃] | brun |
| | [ɛ̃] | lin | [ɛ̃] | lent | [ɔ̃] | long |

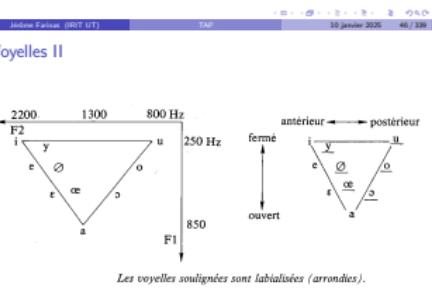
2. Les distinctions vocaliques [i]-[ɪ], [ø]-[œ] et [ɔ̃]-[ɔ̃] ne sont pas faites dans tous les contextes et par tous les locuteurs du français. Par contre certains locuteurs font aussi des distinctions entre patte et pâtre ([ø]-[œ̃]) ainsi qu'entre brin et brun ([ɔ̃]-[ɔ̃]).

Voyelles I



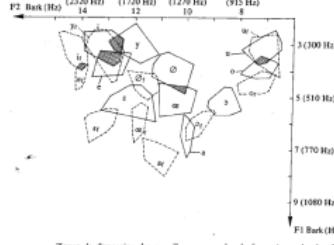
- ➊ Production et perception de la parole
 - ➌ Production de la parole
 - ➍ Perception de la parole

>Description acoustique des sons



Voyelles II

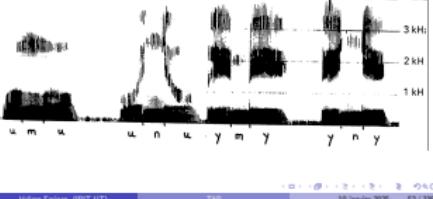
Voyelles III.



Sujets masculins : —
Sujets féminins : —
Les hachures délimitent les zones de recouvrement pour un même sexe.

Consonnes nasales

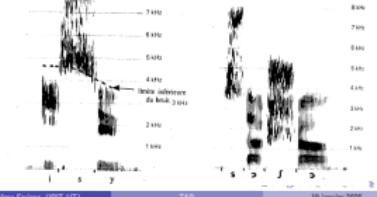
- ➌ couplage conduit oral et conduit nasal
- ➌ occlusion labiale [m], alvéolaire [n], palatale [ɲ]
- ➌ les voyelles s'enchaînent en ignorant [m]



Consonnes fricatives

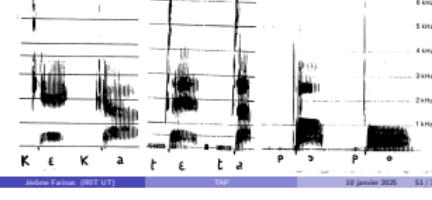
turbulence aérodynamique due à un fort rétrécissement du conduit oral (non voisé ou voisé)

- ➌ labio-dental ([f] et [v])
- ➌ dental ([s] et [z])
- ➌ palatal ([ʃ] et [ʒ])



Consonnes occlusives

- ➌ non voisées ([p],[t],[k]), voisées ([b],[d],[g])
- ➌ occlusion complète du conduit vocal :
 - labiale : [p], [b]
 - dentale : [t], [d]
 - palato-vélaire ou vélaire : [k], [g]
- ➌ suite d'événements
 - ➌ silence ou voisement
 - ➌ barre d'explosion
 - ➌ bruit de friction



Consonnes liquides

- ➌ [l] articulation alvéolaire, passage latéral de l'air. F1=300Hz. F2 très variable, dépend du contexte
- ➌ [ʁ] articulation dorso-vélaire. Dépend du contexte de manière dramatique. Peut-être voisé ou non voisé.



Voyelles dans le monde

La majorité des langues du monde utilisent 5 voyelles
 [i] [e] [a] [o] [u]
 i é a o ou

Le français utilise en moyenne 15 voyelles !

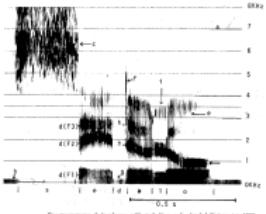
[i] [y] [ɛ] [œ] [ə] [ɔ] [ɑ] [u] [ø] [œ̃] [ɛ̃] [ɔ̃] [ɑ̃] [œ̃̃] [ɛ̃̃]
 du nez mère vué cœur le la cou pal or pain on en un un

Consonnes dans le monde

| | MÉTHODES D'ARTICULATION | | | | | | | | | | | |
|-------------------|-------------------------|-------|-------|-------|-----------|-----------|------------|-------|--------------------|-----------|------------|-------|
| | plissant (oral) | | nasal | | aériques | | fricatives | | lateral fricatives | | satellites | |
| | voisé | voisé | voisé | voisé | non voisé | non voisé | voisé | voisé | non voisé | non voisé | voisé | voisé |
| ❶ IRRÉAL | b | p | m | bv | pf | β | f | | v | | B | |
| ❷ LABIORÉTROGR. | | mj | | v | f | | | | v | | | |
| ❸ INTERDENTAL | | | | θ | θ | | | | | | | |
| ❹ ALVEOLAIRES | d | t | n | dz | ts | Z | s | ç | t | r | r̄ | j |
| ❺ POSTALVÉOLAIRES | | | | dʒ | tʃ | ʒ | ʃ | | | | | j̄ |
| ❻ RÉTRORÉTROGR. | d̄ | l̄ | η̄ | | | k̄ | s̄ | | r̄ | | ç̄ | l̄ |
| ❼ PALATAL | j | c | p̄ | | | χ̄ | s̄ | | | | ç̄ | χ̄ |
| ❽ VELAIRE | g | k | ḡ | | | x̄ | x̄ | | | | ɥ̄ | l̄ |
| ❾ UNGUAIRES | G | q | N | | | χ̄ | χ̄ | | | | | |
| ❿ PHARYNGIALE | | | | | | χ̄ | χ̄ | | | | | |
| ⓫ GLOTTAL | | | | | | B | h | | | | | |
| ⓬ LABIO-PALATAL | | | | | | | | | | | ɥ̄ | |
| ⓭ LABIO-VELAIRE | | | | | | | | | | | w̄ | |
| �� | | | | | | | | | | | | |

Description des consonnes dans l'Alphabet phonétique international (API).

Lecture sur un spectrogramme



— Spectrogramme de la phrase « C'est de l'eau » joué à la fréquence d'0 kHz

0,5 s 0,6 s 0,7 s 0,8 s

— Axe de modulations

0,6 s — Axe de base + périodicité de l'onde sonore

0,7 s — Axe de fréquence de l'onde sonore de la onde sonore

0,8 s — Axe de modulation de l'énergie (modulation d'amplitude)

0,6 s — Axe de modulation de la fréquence fondamentale (modulation de la fréquence fondamentale)

0,7 s — Axe de modulation de l'énergie et de la fréquence fondamentale (modulation d'amplitude et de la fréquence fondamentale)

0,8 s — Axe de modulation de l'énergie et de la fréquence fondamentale et de la périodicité de l'onde sonore

0,6 s — Axe de modulation de l'énergie et de la périodicité de l'onde sonore et de la fréquence fondamentale

0,7 s — Axe de modulation de l'énergie et de la périodicité de l'onde sonore et de la fréquence fondamentale et de la modulation de l'énergie

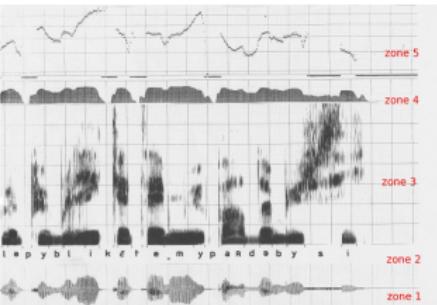
0,8 s — Axe de modulation de l'énergie et de la périodicité de l'onde sonore et de la fréquence fondamentale et de la modulation de l'énergie et de la périodicité de l'onde sonore

Exercices

❶ Écrivez votre prénom en phonétique.

- ❷ Traduisez en phonétique les mots suivants : françois, trois, quatre, cinq, six.
- ❸ Décrivez les différentes zones qui sont représentées sur la planche suivante. Indiquez les unités sur les axes.

Jérôme Faucon (007 UT) TAP 10 janvier 2025 58 / 339



❶ Écrivez votre prénom en phonétique.

- ❷ Traduisez en phonétique les mots suivants : françois, trois, quatre, cinq, six.
- ❸ Décrivez les différentes zones qui sont représentées sur la planche suivante. Indiquez les unités sur les axes.
- ❹ Décdecodez la séquence phonétique sur la planche A10 suivante

Jérôme Faucon (007 UT) TAP 10 janvier 2025 60 / 339

Jérôme Faucon (007 UT) TAP 10 janvier 2025 58 / 339

