Interaction Vocale
Notes de cours

Eléonore Bartenlian, Adrien Pavao

Janvier 2018

1 Introduction

"La linguistique a un double objet, elle est *science du langage* et *science des langues*. Cette distinction (que l'on ne fait pas toujours) est nécessaire : *le langage* est une faculté humaine, caractéristique universelle et immuable de l'homme et est donc autre chose que *les langues* qui sont toujours particulières et variables, et en lesquelles le langage se réalise." - Benveniste

- **Langage :** faculté spécifiquement humaine, universelle. Système de représentation régi par une grammaire.
- **Langue :** Réalisation particulière du langage. Règles et normes partagées par les membres d'une commmunauté.
- **Parole :** Usage oral de la langue (non écrit).

Complexe par rapport à l'écrit : continuité du signal, coarticulation, distortions temporelles (débit variable), variabilité (inter et intra locuteurs, conditions acoustiques), homophonies.

On distingue phonétique et phonologie :

- **La phonologie** (ex : R), sciences des phonèmes de la langue. Objet : la langue, méthodes linguistiques, on décrit les sons fonction des ressemblances et différences phoniques fonctionnelles dans la langue en question. On étudie le statut linguistique des sons (fonctionnel) à l'intérieur d'une langue, d'un système. Cherche à établir la fonction des sons dans une langue.
- **La phonétique** (ex : R roulé, R grasseyé). Sciences des sons (phones) des langues (production des sons). Objet : l'acte de parole. méthodes des sciences naturelles. Les sons sont des entités physiques et on les décrits peut importe la langue d'où ils viennent. Les sons dans leur matérialité, considérés comme des entités physiques et indépendamment de la langue à laquelle ils appartiennent. Etudie les sons des langues du monde.

voyelles orales			
/i/	p**i**e	/a/	p**a**tte
/e/	**é**t**é**	/ɑ/	p**â**te
/ɛ/	mod**è**le	/o/	**au**diteur
/y/	p**u**ni	/ɔ/	p**o**rte
/ø/	d**eu**x	/u/	p**ou**x
/œ/	p**eu**r	/ə/	p**e**tit**e**

voyelles nasales			
/ɑ̃/	**an**	/œ̃/	br**un**
/ɛ̃/	mat**in**	/ɔ̃/	b**on**

plosives orales	labiales	alvéolaires	vélaires
sourdes	/p/ : **p**oids	/t/ : **t**oit	/k/ : **qu**oi
voisées	/b/ : **b**ois	/d/ : **d**oigt	/g/ : **g**oût

occlusives nasales	labiale	alvéolaire	palatale	
	/m/ : **m**on	/n/ : **n**ous	/ɲ/ : a**gn**eau	/ŋ/ : smoki**ng**

fricatives	dentales	alvéolaires	post-alvéolaires
sourdes	/f/ : **f**eu	/s/ : **s**oir	/ʃ/ : po**ch**e
voisées	/v/ : **v**oix	/z/ : **z**éro	/ʒ/ : **j**eu

liquides	/l/ : **l**ong	/ʁ/ : **r**ond

semi-voyelles	/w/ : **ou**i	/j/ : pi**è**ge	/ɥ/ : **l**ui

Figure 1: Les phonèmes de la langue française

Trois modes d'excitation de la source :

- Vibrations quasi-périodiques des cordes vocales (fréquenc fondamentale F_O, son voisé (voyelles))
- Bruits d'écoulement d'air, constriction (son fricatif, fricatives ou constrictives)
- Occlusions rapides, impulsions (son avec explosion, occlusives ou plosives).

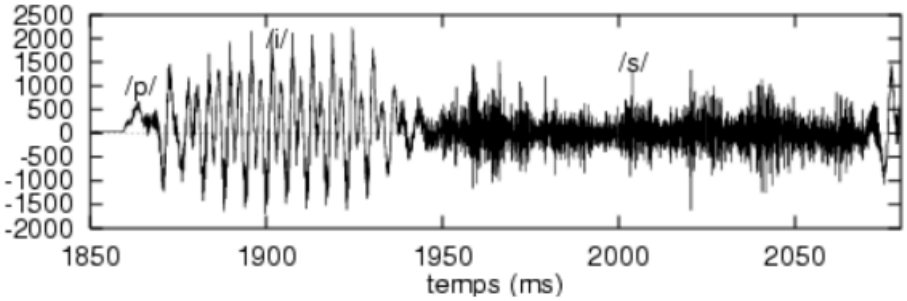


Figure 2: Exemple de son avec explosion, puis son voisé, puis son fricatif

Le signal source produit par l'excitateur se propage dans un volume appelé aussi résonateur. Les composantes fréquentielles de l'excitation sont affaiblies ou renforcées dans le résonateur (dépend du volume de la cavité et de ses ouvertures). Les fréquences de résonances sont appelées formants (caractéristiques du timbre). Les premiers formants sont les plus importants (position des 3 premiers pour caractériser une voyelle. Par ex i : 300, 2200 et 3000Hz).

F_0 et harmoniques

La fréquence de vibration des cordes vocales des sons voisés est appelée fréquence fondamentale ou F_0. Elle correspond à la hauteur "musicale" du son.

Le spectre d'un signal (quasi-)périodique est un spectre de raies aux multiples entiers de la fréquence fondamentale ou harmoniques ($H1 = 2F_0$, $H2 = 3F_0$...). F_0 évolue lentement dans le temps et n'est pas spécifique d'un phonème.

Enveloppe spectrale et formants

En reliant les maxima des raies par une ligne continue on obtient l'enveloppe du spectre. Les fréquences où les maxima de cette enveloppe ont lieu sont les formants du signal F1, F2... Les formants sont spécifiques de chaque phonème.

Formant vs harmonique

Harmonics come from the vocal folds. You can change the harmonics present in the sound by changing the shape of the vocal folds and therefore the pitch being created. More closure in the vocal folds will create stronger, higher harmonics. Harmonics are considered the source of the sound.

Formants come from the vocal tract. The air inside the vocal tract vibrates at different pitches depending on its size and shape of opening. We call these pitches formants. You can change the formants in the sound by changing the size and shape of the vocal tract. Formants filter the original sound source. After harmonics go through the vocal tract some become louder and some become softer.

Applications du traitement automatique de la parole

- Codage (télécommunications)
- Synthèse vocale à partir du texte
- Reconnaissance de la parole

Que reconnaître dans la parole ?

Beaucoup d'informations sont présentes dans un signal de parole :

- Reconnaissance du locuteur
- Transcription
- Identification de la langue
- Reconnaissance des émotions

Aspects non-verbaux de la voix :

- Le timbre, la qualité vocale, les disfluences...
- La prosodie : rythme, intensité, mélodie.

2 Signal

Signal déterministe : $x(t) = Acos(2\pi f_0 t)$

Fourier

Tout signal périodique $x(t)$ de période T peut être décomposé sous la forme d'une série de Fourier :

$$x(t) = \sum_{-\infty}^{\infty} X_n e^{2j\pi nt/T}$$

$$X_n = \frac{1}{T} \int_{-T/2}^{T/2} x(t) e^{-2j\pi nt/T} dt$$

Formule de Parseval

Soient $x(t)$ et $y(t)$ deux signaux périodiques de période T, soit $z(t) = x(t).y^*(t)$ alos $Z_n = \sum_{-\infty}^{\infty} X_k Y^*_{k-n}$.

En faisant *n=0*, on obtient

$$\sum_{k=-\infty}^{\infty} X_k Y^*_{k-n} = \frac{1}{T}\int_{-T/2}^{T/2} x(t)y^*(t)dt$$

En faisant *x(t) = y(t)* on obtient

$$P = \sum_{k=-\infty}^{\infty} |X_k|^2 = \frac{1}{T}\int_{-T/2}^{T/2} |x(t)|^2 dt$$

Interprétation: La puissance d'un signal est égale à la somme des puissances élémentaires de chacune de ses composantes.

Composante = signal « sinusoidal » $X_n e^{2j\pi nt/T}$

Après tout plein d'autres trucs chimico chimiques

Fenetre d'analyse : TODO

Large bande vs bande étroite : TODO

Echelle Mel

Correspond à une approximation psychologie de hauteur d'un son (Tonie). TODO mieux expliquer

Représentation cepstrale

TODO

Paramétrisation MFCC

Une implémentation classique: 13 Coefficients (sans C0), Filtres Mels espaces de 150 Mel (largeur de bandes 300 Mels), Utilisation des dérivées premières et secondes, Soit des vecteurs de 39 paramètres acoustiques.

3 Modélisation statistique de la langue

La prosodie : 3 dimensions

- L'énergie (l'intensité)
- le timbre
- Le rythme

4 Modèles statistiques du langage

4.1 Grammaire formelle

Deux notions importantes :

- **Analyseur de langage :** Le but est de déterminer, à l'aide d'un algorithme déterministe (temps fini), si une phrase donnée appartient à un langage donné. Par exemple l'analyse lexicale et syntaxique d'un compilateur.
- **Générateur de langage :** Un ensemble de règles pour générer toutes les phrases valides possibles du langage. Par exemple une grammaire générative.

Une grammaire formelle est définie par un quadruplet $G = \{N, T, R, S\}$.

- **N** l'ensemble de symboles non-terminaux.
- **T** l'ensemble de symboles terminaux.
- **R** l'ensemble de règles d'écriture.
- **S** le symbole de départ.

4.2 Modèle N-gram

L'objectif d'un modèle N-gram est d'être capable de prédire un mot n à partir des $n-1$ mots. Par exemple, prédire le 4ème mot d'une séquence à partir des trois premiers.

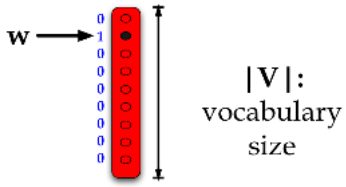
Autrement dit,

$$P(w_i|w_1, ..., w_{i-2}, w_{i-1})$$

4.3 Modèle du langage par réseau neuronal

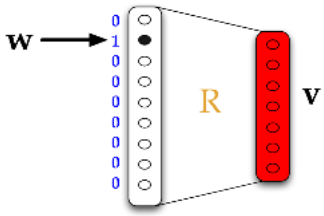
Projeter une séquence de mot dans un espace continu :

- Le vocabulaire est une couche de neurones.



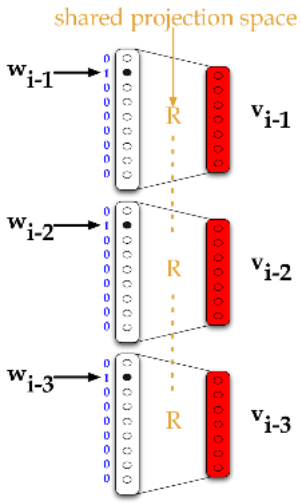
  Cette couche représente un vecteur de valeurs. On a un neurone par valeur.

- On projette le mot dans l'espace continu en ajoutant une deuxième couche *fully-connected*.



  v est un vecteur continu.

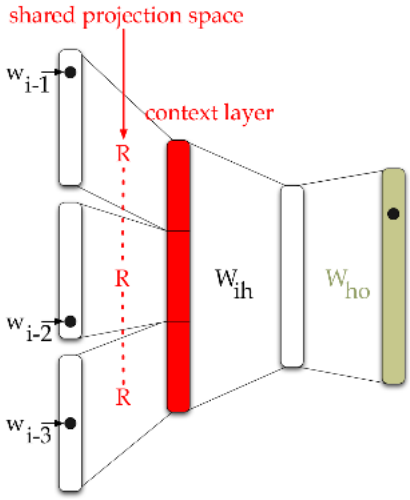
- Par exemple, pour un 4-gram, l'historique est une séquence de 3 mots.



On fusionne ces trois vecteurs afin d'en dériver un vecteur unique pour représentant l'historique.

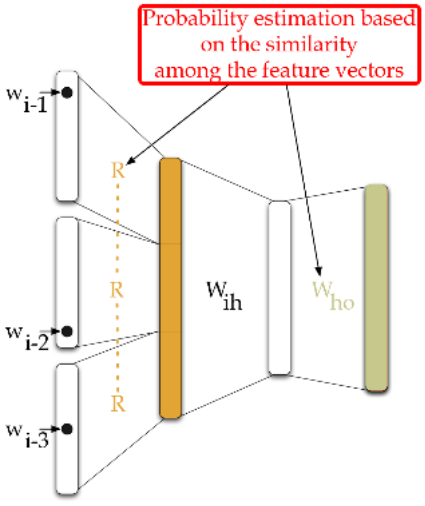
Estimer la probabilité de n-gram :

- On donne l'historique comme un vecteur de features. On crée un vecteur de features pour le mot à prédire dans **l'espace de prédiction**.



On estime les probabilités pour tous les mots à l'aide de l'historique. Tous les paramètres sont appris (les poids des différentes couches).

- On apprend donc simultanément la projection et la prédiction.



4.4 Estimation robuste (smoothing)

La modélisation n-gram présentent de nombreux problèmes. La gestion des mots hors-vocabulaire, la quantité de données d'apprentissage toujours trop faible pour estimer toutes les probabilités, l'hypothèse markovienne insuffisante en pratique, etc.

Pour résoudre cela, on pourra utiliser des méthodes de lissage, s'inspirer de la loi de Zipf.