Techniques d'apprentissage machine appliquées à la modélisation du timbre d'un individu pour la synthèse de voix réaliste

Pierre-Alexandre Fournier, B. Ing.
Directeur: Jean-Jules Brault, Ph. D.

Département de génie électrique

Soutenance de mémoire, 24 novembre 2005



Introduction

Bienvenue à ma présentation de mémoire. Je vais aujourd'hui vous présenter la méthode que j'ai développée pour compléter le corpus d'un locuteur dans le but de faire une synthèse de voix réaliste.
Construire une base de données pour une voix réaliste est un processus fastidieux, qui peut prendre des mois, voire des années, et dont les résultats ne sont jamais parfaits. Le but de ma recherche était donc de trouver un moyen de compléter le corpus incomplet d'un locuteur pour permettre de retrouver le timbre d'un phonème manquant.

Ma recherche s'est concentrée sur un sous-ensemble de phonèmes, constitué des cinq voyelles a, é, i, o et ou. À partir de ce sous-ensemble, j'ai développé un système qui montre que l'on peut utiliser l'information disponible d'un locuteur pour compléter son corpus.



Sommaire

1. Problématique
2. Le signal vocal
   - Production et perception du signal vocal
   - Les voyelles
   - Analyse harmonique et résiduel
   - Normalisation
3. Prédiction de voyelle
   - Objectifs
   - Noyaux
   - Algorithme génétique
   - Résultats



Problématique

- Nous voulons construire un modèle de la voix d'un locuteur pour avoir un son réaliste.
- Il faut réduire le temps de modélisation, présentement long et coûteux, et dans certains cas impossible.
- Nous voulons pouvoir compléter la base de donnée servant de modèle lorsque certains phonèmes ne sont pas disponibles.



Outline

1 Problématique

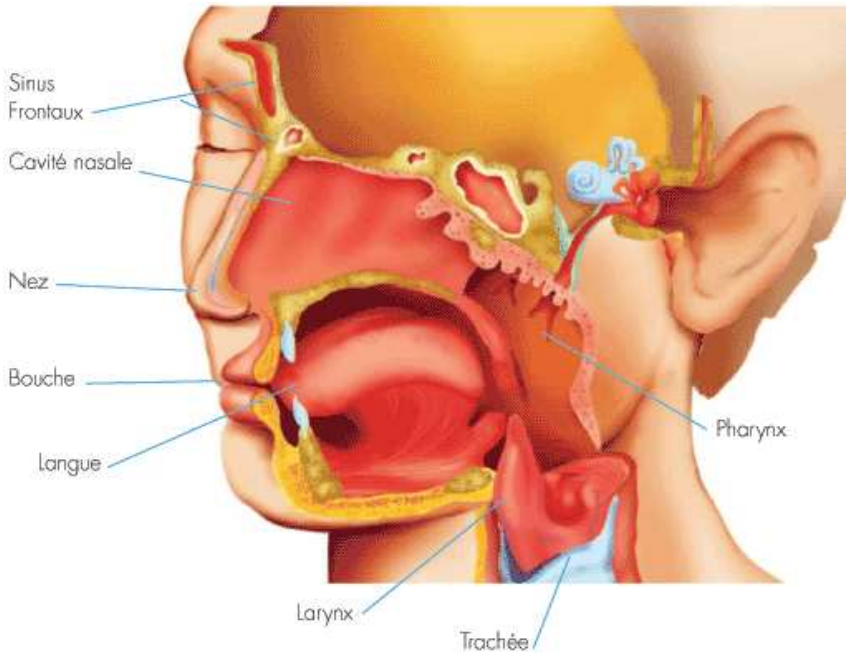
2 Le signal vocal
- Production et perception du signal vocal
- Les voyelles
- Analyse harmonique et résiduel
- Normalisation

3 Prédiction de voyelle
- Objectifs
- Noyaux
- Algorithme génétique
- Résultats



Production du signal vocal

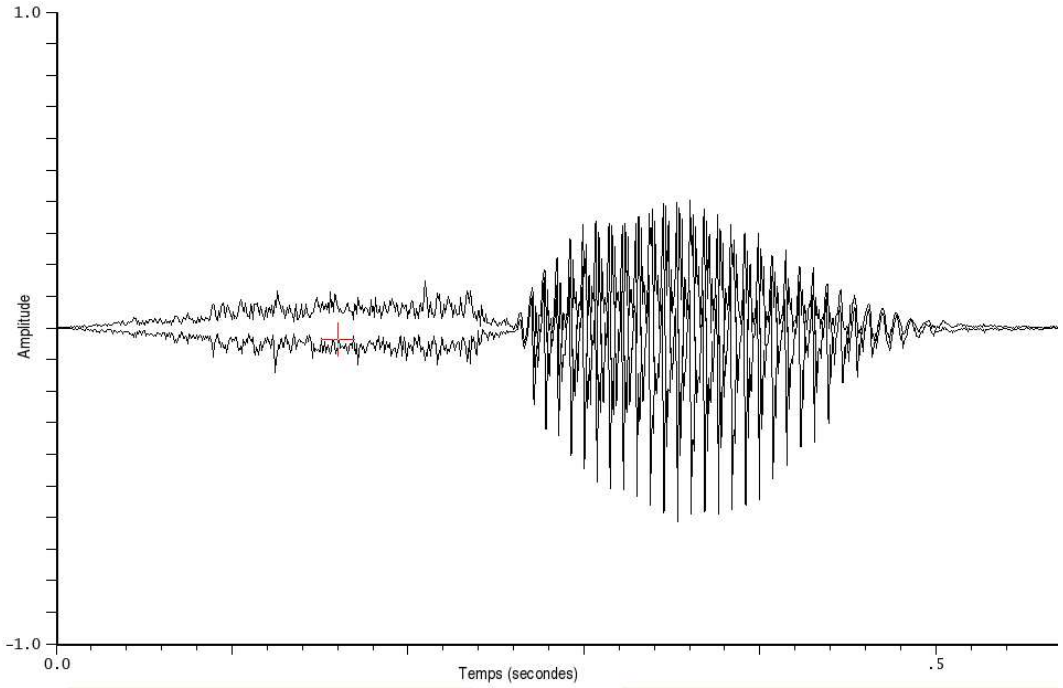
Comment la voix est elle produite?



L'air poussé dans le conduit vocal produit les sons.

Production du signal vocal

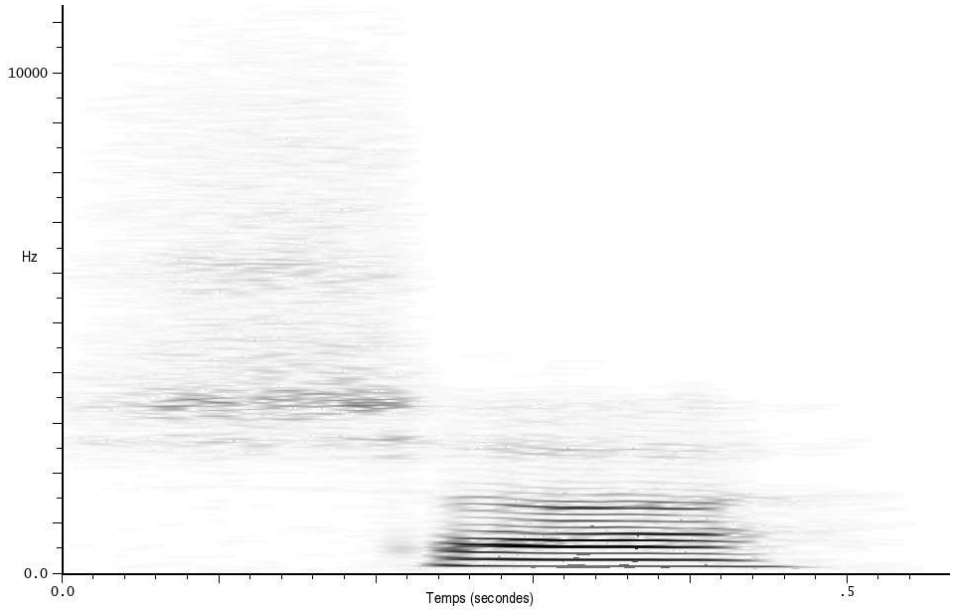
L'exemple du mot "chat":



Lorsque les cordes vocales vibrent, on dit que le son est voisé.

Production du signal vocal

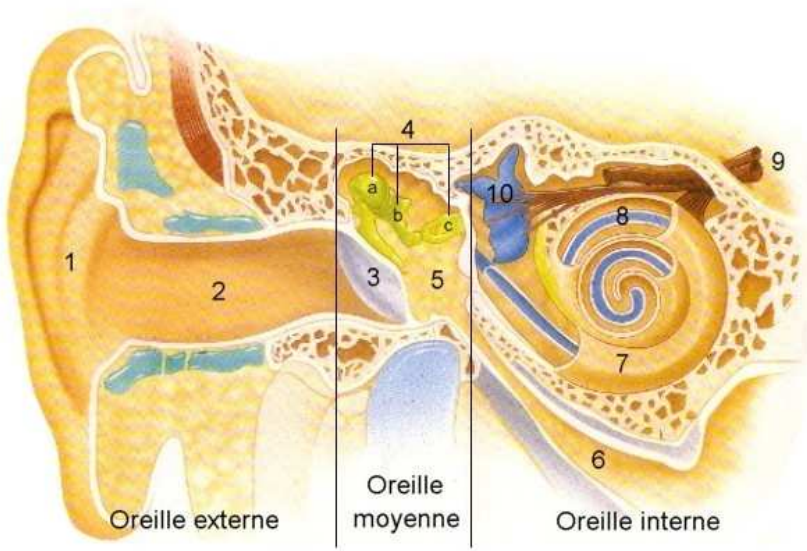
Sonogramme du mot “chat”:



Les sons voisés ont une plus grande cohérence temporelle que les sons non-voisés. Ils ont la caractéristique d’être composés de sinusoïdes ayant une relation harmonique.

Perception du signal vocal

L'oreille:



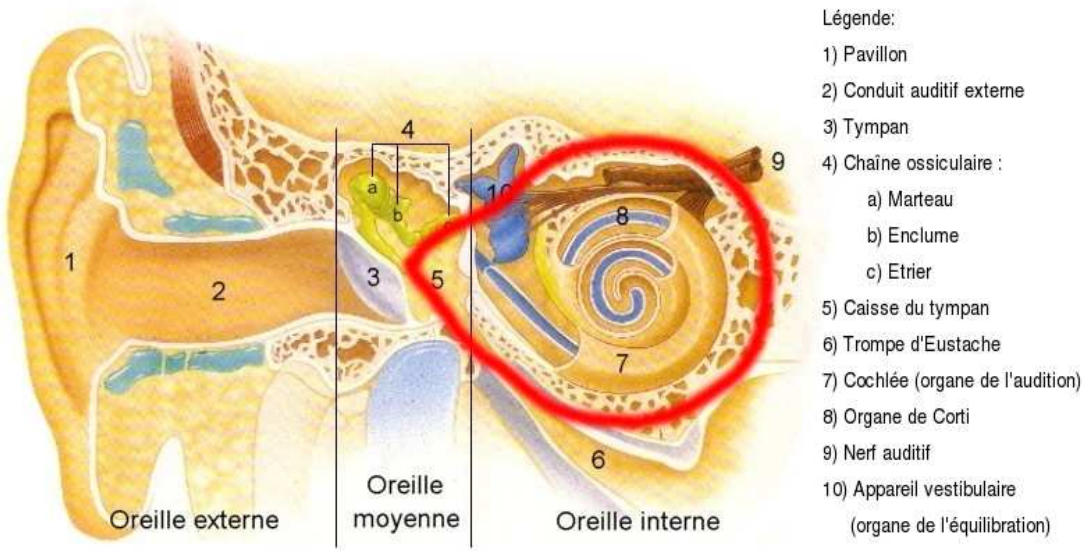
Légende:
1) Pavillon
2) Conduit auditif externe
3) Tympan
4) Chaîne ossiculaire :
a) Marteau
b) Enclume
c) Etrier
5) Caisse du tympan
6) Trompe d'Eustache
7) Cochlée (organe de l'audition)
8) Organe de Corti
9) Nerf auditif
10) Appareil vestibulaire (organe de l'équilibration)

Le signal vocal est perçu par le système auditif.



Perception du signal vocal

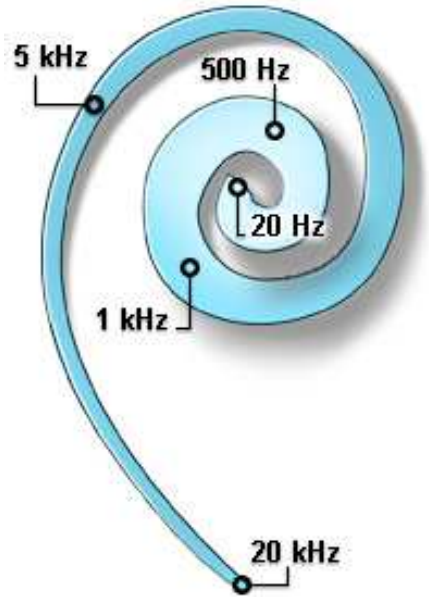
L'oreille:



La cochlée traduit les vibrations mécaniques en excitation neuronale.

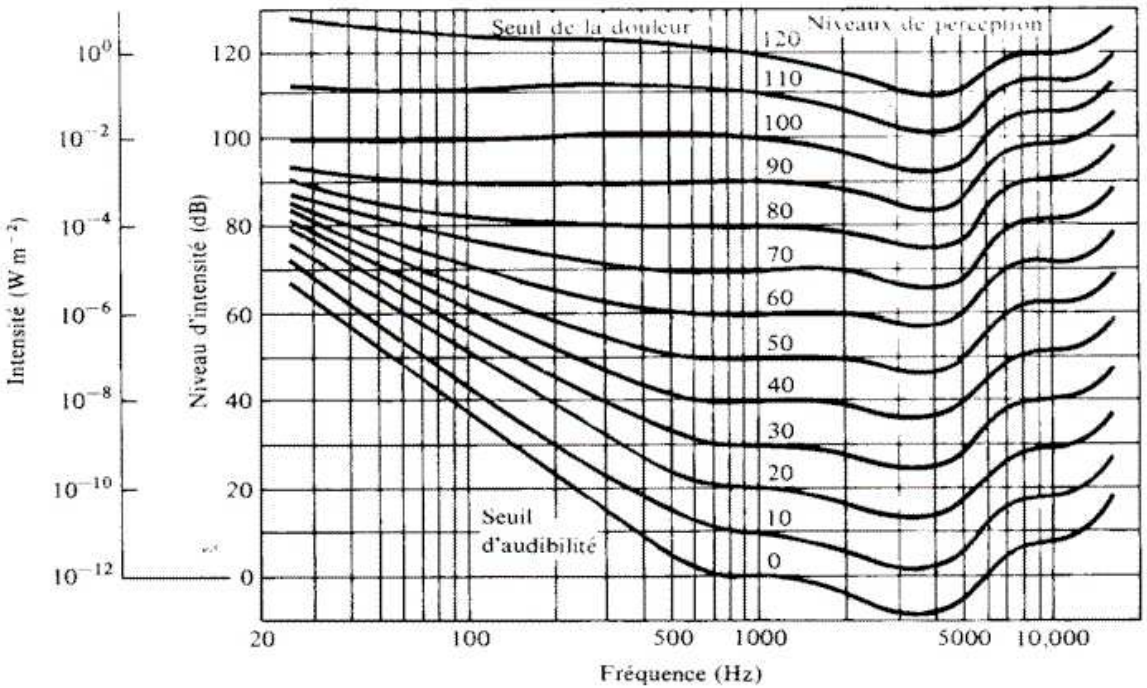
Perception du signal vocal

Les **basses fréquences** sont perçues à l'entrée de la cochlée, les **hautes fréquences** au fond.



La cochlée a une résolution logarithmique en fréquence.

Perception du signal vocal



Les fréquences entre 1000 et 5000 Hz sont perçues plus facilement pour une même puissance.

Outline

1 Problématique

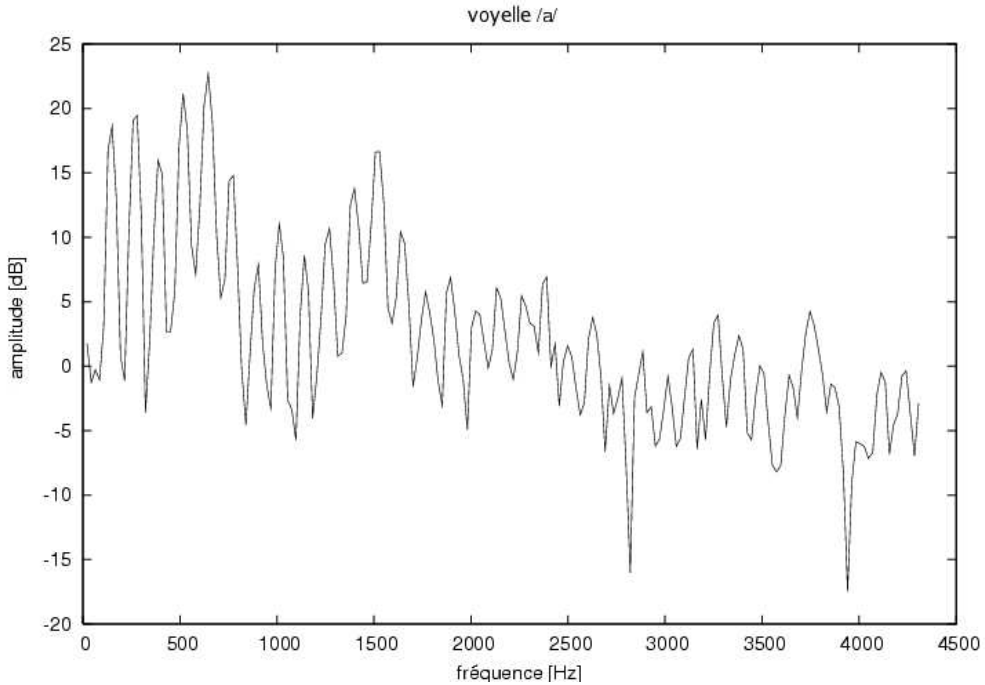
2 Le signal vocal
- Production et perception du signal vocal
- Les voyelles
- Analyse harmonique et résiduel
- Normalisation

3 Prédiction de voyelle
- Objectifs
- Noyaux
- Algorithme génétique
- Résultats



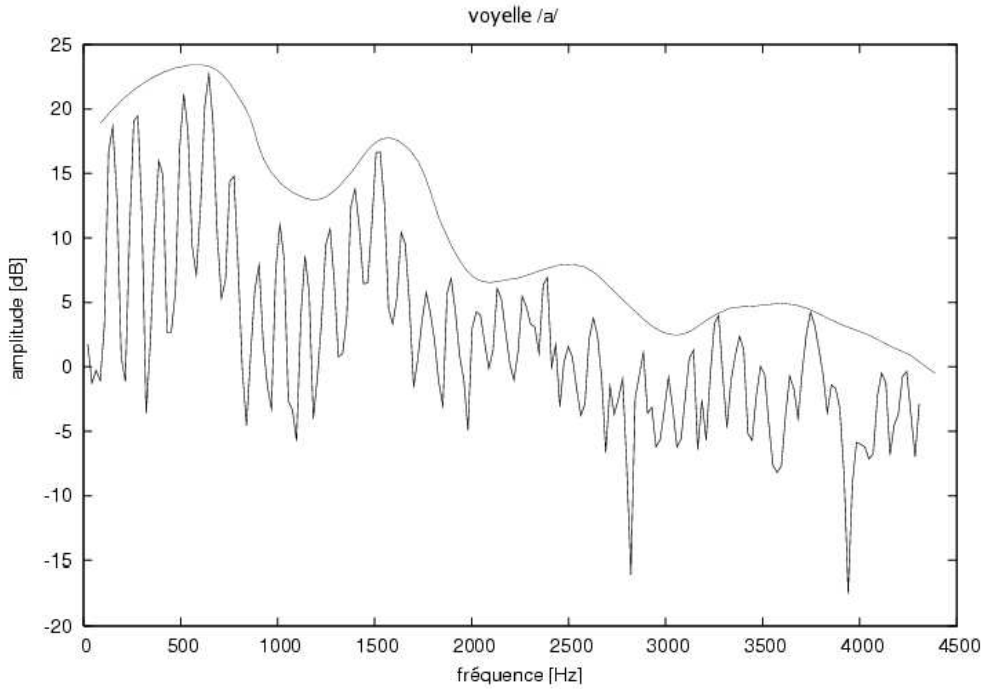
Les voyelles

Les voyelles sont les unités phonétiques d'une personne les plus déterminantes pour l'identification du locuteur.



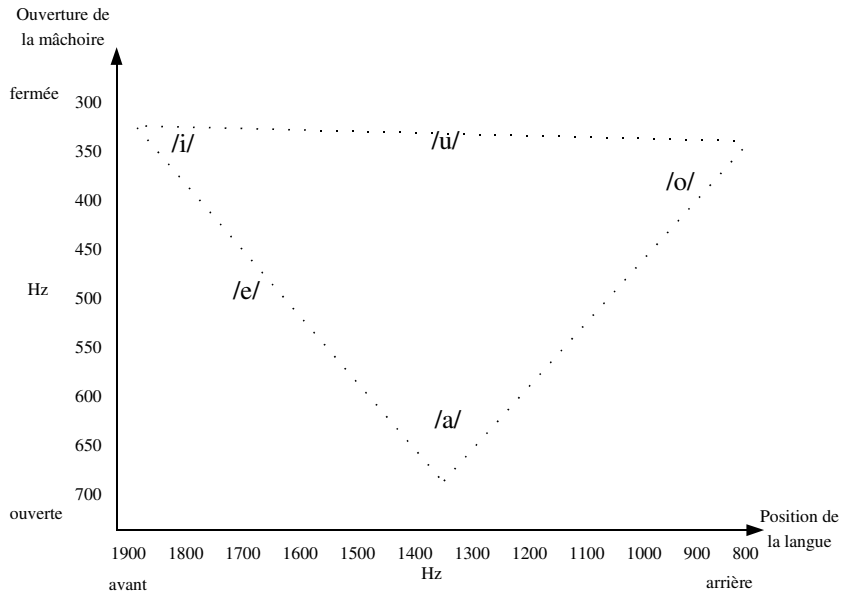
Les voyelles

Les voyelles sont souvent approximées par des sommets spectraux appelés **formants**.



Les voyelles

Le triangle vocalique:



permet de situer approximativement les voyelles selon la position des deux premiers formants.

Outline

1 Problématique

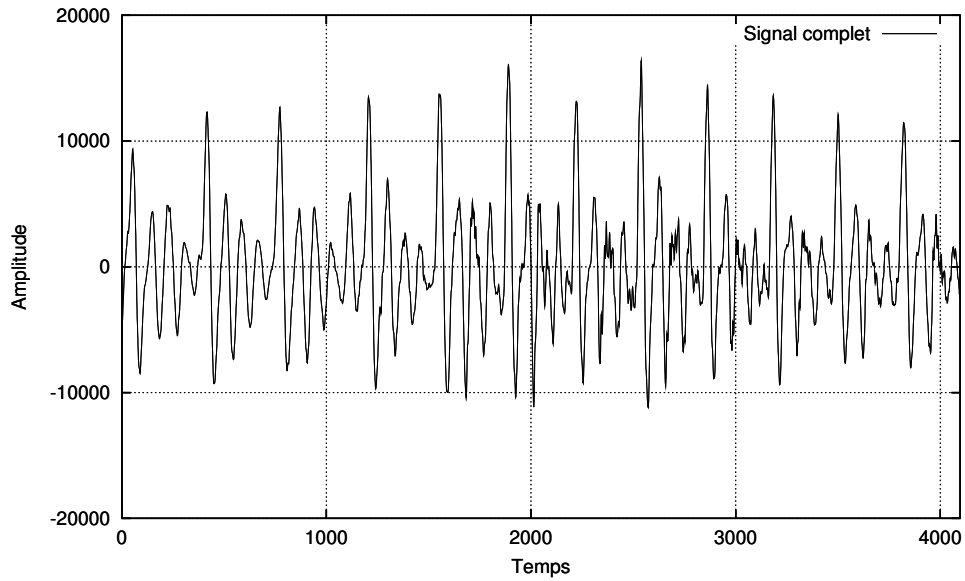
2 Le signal vocal
- Production et perception du signal vocal
- Les voyelles
- Analyse harmonique et résiduel
- Normalisation

3 Prédiction de voyelle
- Objectifs
- Noyaux
- Algorithme génétique
- Résultats

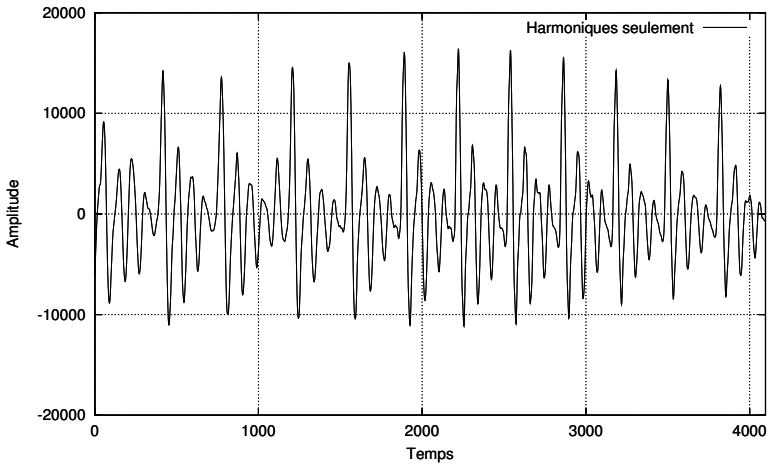


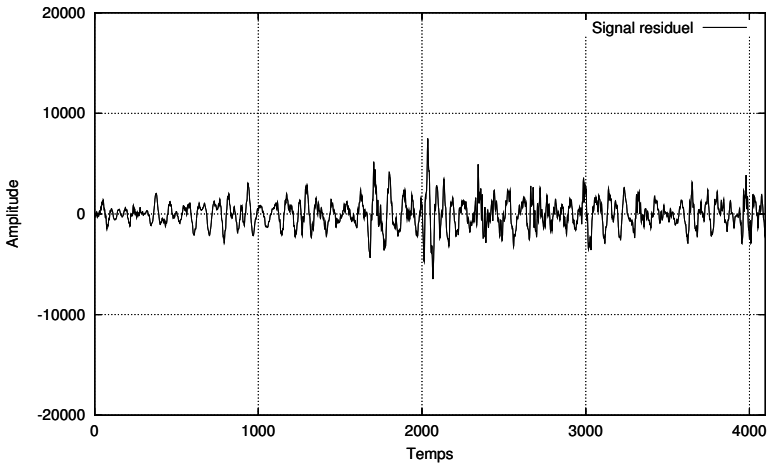
Analyse harmonique et résiduel

Le modèle harmonique et résiduel représente le signal vocal en le séparant en deux parties.



Analyse harmonique et résiduel





La partie harmonique, plus cohérente, et la partie résiduelle, considérée “aléatoire”.



Outline

1 Problématique

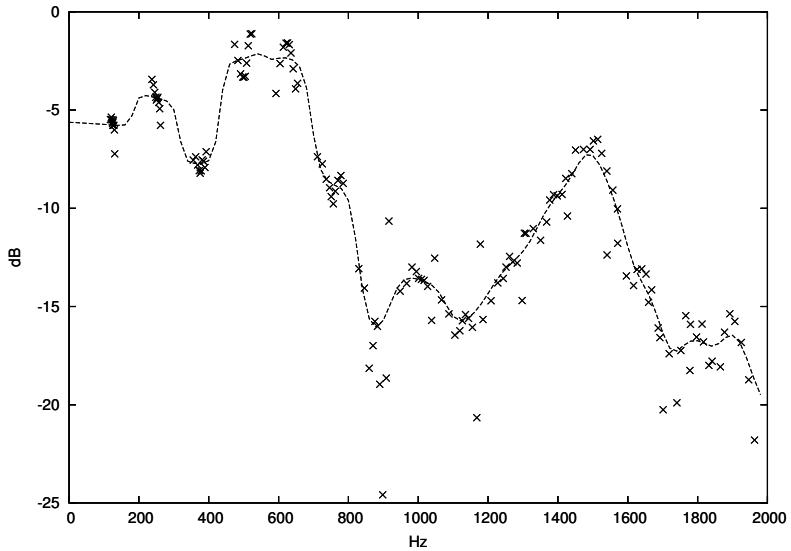
2 Le signal vocal
- Production et perception du signal vocal
- Les voyelles
- Analyse harmonique et résiduel
- Normalisation

3 Prédiction de voyelle
- Objectifs
- Noyaux
- Algorithme génétique
- Résultats



Normalisation des voyelles d'un locuteur

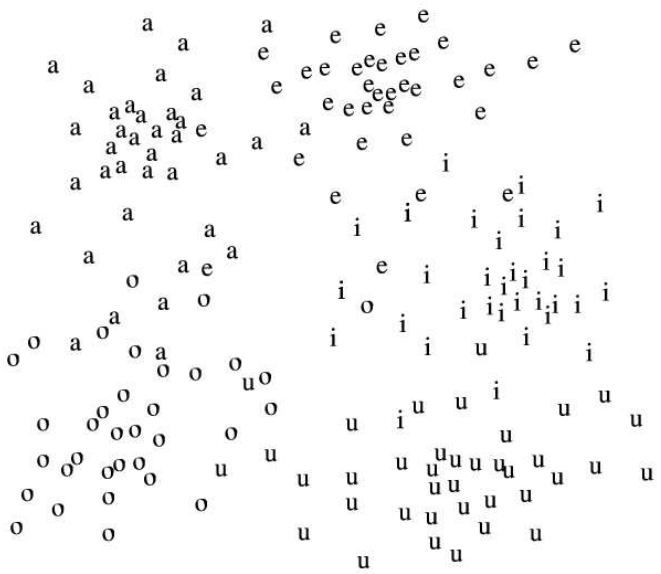
Une voyelle peut-être prononcée de plusieurs manières par un même locuteur. Nous devons donc normaliser les exemples et définir une voyelle typique.



En prenant des données de plusieurs fenêtres d'analyse, nous obtenons une meilleure résolution du spectre de puissance.

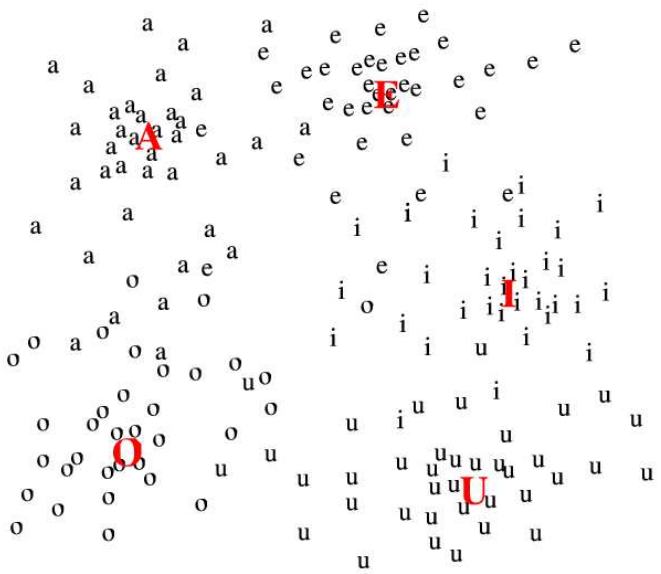
Normalisation des voyelles d'un locuteur

Nous avons pris un **sous-ensemble typique** des fenêtres d'analyse pour faire cette normalisation. Nous espérons de cette manière modéliser le **mode** de la distribution de la voyelle.



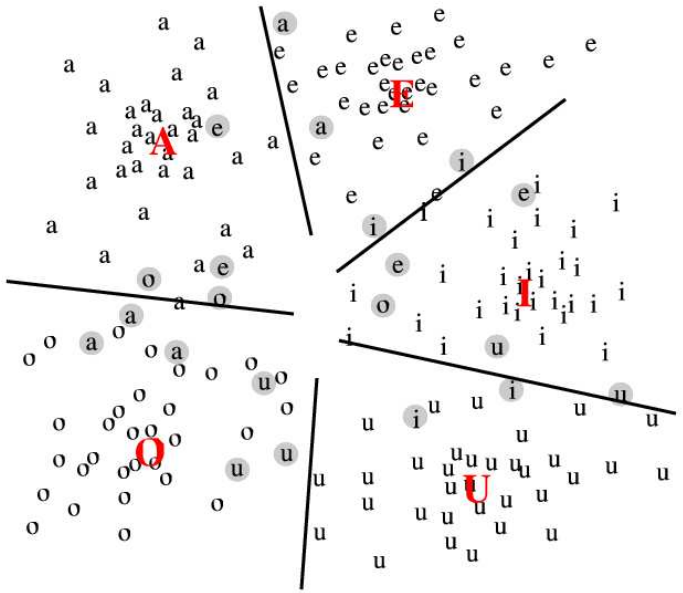
Normalisation des voyelles d'un locuteur

Pour éliminer les échantillons atypiques, nous avons calculé leur distance par rapport à la moyenne de chaque catégorie de voyelle.



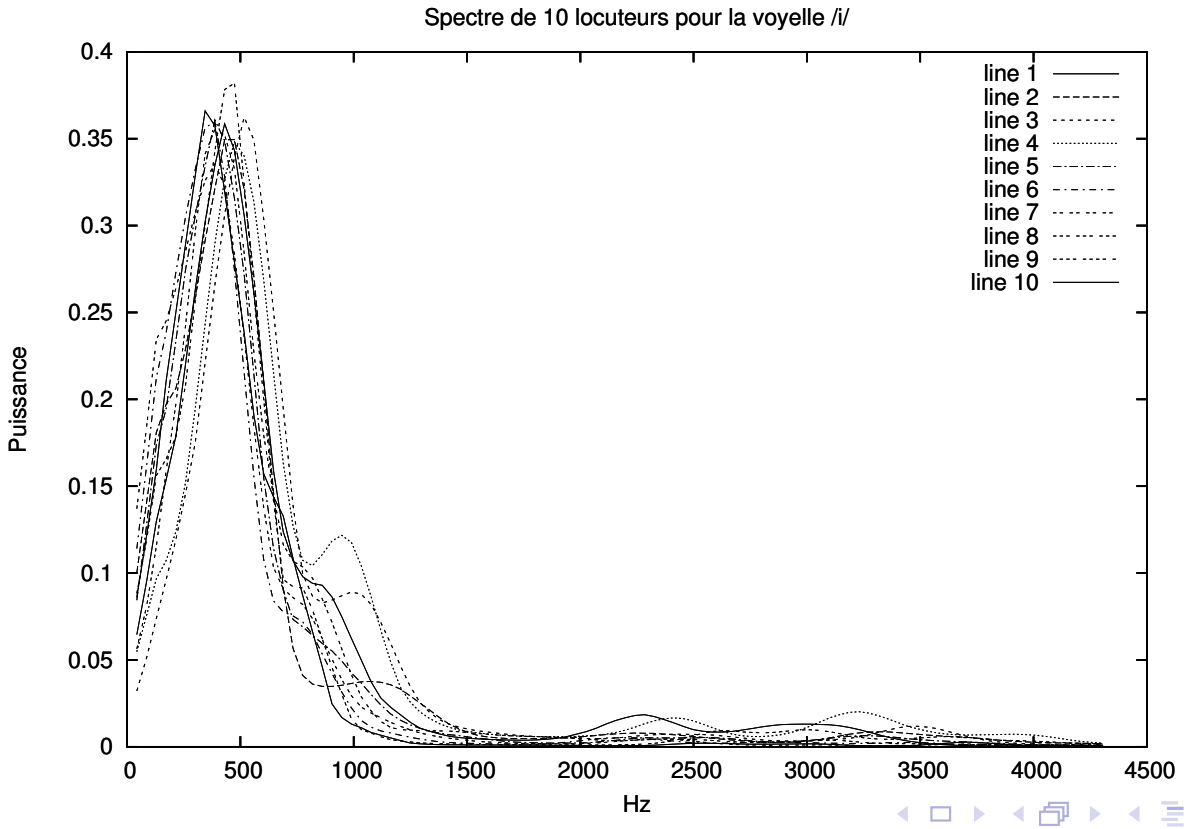
Normalisation des voyelles d'un locuteur

Puis nous avons progressivement éliminé les échantillons situés plus près d'une moyenne d'une autre catégorie que la leur.



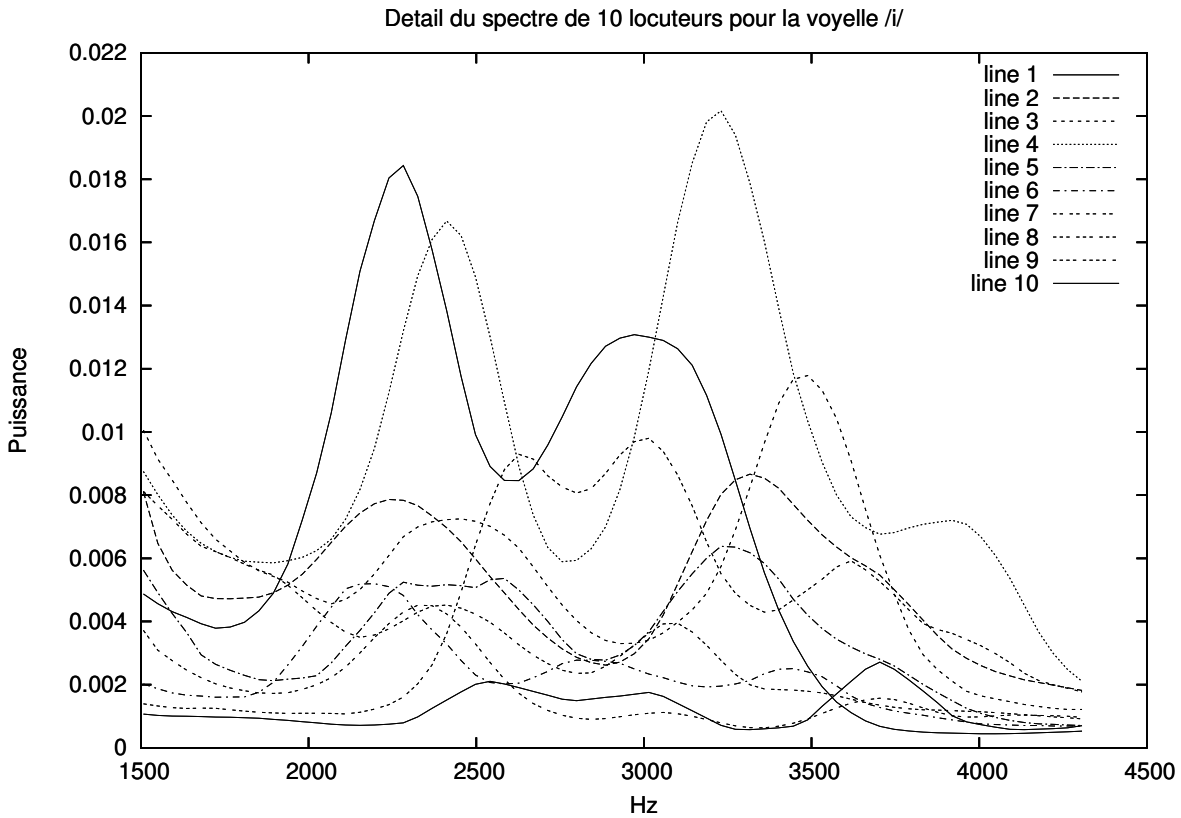
Normalisation des voyelles d'un locuteur

Finalement, nous obtenons les données utilisées pour notre expérience. Voici des spectres typiques pour la voyelle /i/:



Normalisation des voyelles d'un locuteur

Détail des spectres typiques pour la voyelle /i/:



Outline

1 Problématique

2 Le signal vocal
- Production et perception du signal vocal
- Les voyelles
- Analyse harmonique et résiduel
- Normalisation

3 Prédiction de voyelle
- Objectifs
- Noyaux
- Algorithme génétique
- Résultats



Objectifs de la prédiction

Objectifs de la prédiction:

1. Reproduire le spectre de puissance harmonique d'une voyelle d'un locuteur.
2. Le spectre variant selon le contexte, nous voulons repoduire un spectre **typique**.
3. Le modèle doit se servir de l'information disponible du locuteur.



Objectifs de la prédiction

Objectifs de la prédiction:

1. Reproduire le spectre de puissance harmonique d'une voyelle d'un locuteur.
2. Le spectre variant selon le contexte, nous voulons repoduire un spectre **typique**.
3. Le modèle doit se servir de l'information disponible du locuteur.



Objectifs de la prédiction

Objectifs de la prédiction:

1. Reproduire le spectre de puissance harmonique d'une voyelle d'un locuteur.
2. Le spectre variant selon le contexte, nous voulons repoduire un spectre **typique**.
3. Le modèle doit se servir de l'information disponible du locuteur.



Outline

1. Problématique
2. Le signal vocal
   - Production et perception du signal vocal
   - Les voyelles
   - Analyse harmonique et résiduel
   - Normalisation
3. Prédiction de voyelle
   - Objectifs
   - Noyaux
   - Algorithme génétique
   - Résultats



Noyaux

Les noyaux permettent de faire une interpolation basée sur une fonction de distance.

Prédiction:

$$\hat{y}_j(s_j) = \frac{\sum_i K_\lambda(s_j, s_i) y_i}{\sum_i K_\lambda(s_j, s_i)}$$

Noyau:

$$K_\lambda(s_j, s_i) = \frac{1}{\lambda} exp[\frac{-||s_i - s_j||^2}{2\lambda}]$$



Noyaux

Notre fonction de distance:

- Distance euclidienne pondérée (justification psycho-acoustique)

$$d_2(s_i, s_j) = \sqrt{||s_i - s_j||^2}$$

- Mesure d'erreur = carré de la distance



Outline

1. Problématique
2. Le signal vocal
   - Production et perception du signal vocal
   - Les voyelles
   - Analyse harmonique et résiduel
   - Normalisation
3. Prédiction de voyelle
   - Objectifs
   - Noyaux
   - Algorithme génétique
   - Résultats



Algorithme génétique

- Population = 20 individus
- Initialisation près du noyau symétrique optimal
- Taux de survie = 50 %
- Individus mutants = 25 %
- Individus hybrides = 25 %
- Mutations: $\lambda_{i,n+1} = max(0.0001, \lambda_{i,n} + N(0, 0.01))$
- Accouplement = 50 % des paramètres de chaque parent



Outline

1. Problématique
2. Le signal vocal
   - Production et perception du signal vocal
   - Les voyelles
   - Analyse harmonique et résiduel
   - Normalisation
3. Prédiction de voyelle
   - Objectifs
   - Noyaux
   - Algorithme génétique
   - Résultats



Erreur de prédiction

Distorsion harmonique moyenne avec chaque méthode.

Noyau	Distorsion	Gain
Voyelle moyenne	-9.58 dB	N/A
3 plus proches voisins	-10.74 dB	23.5 %
Gaussien ($\lambda = 0.1$)	-10.91 dB	26.4 %
Gaussien ($\lambda = 0.06$)	-11.00 dB	27.9 %
Gaussien (meilleur λ par voyelle)	-11.41 dB	34.4 %
Gaussien (Algorithme génétique)	-11.56 dB	36.6 %



Coefficients optimaux

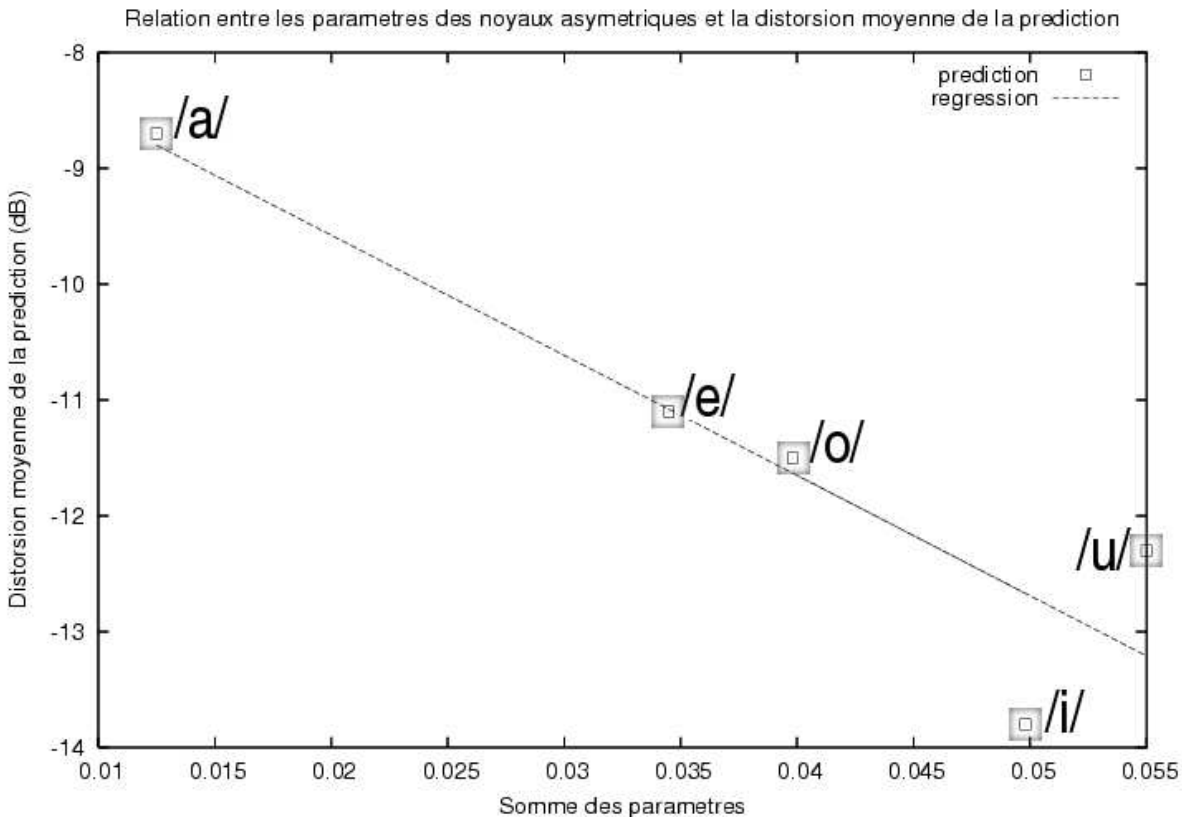
Les coefficients trouvés montrent que certaines voyelles sont plus utiles que d'autres pour la prédiction:

Voyelle à prédire	λ_a	λ_e	λ_i	λ_o	λ_u
/a/	N/A	0.0122	0.0001	0.0001	0.0001
/e/	0.0256	N/A	0.0001	0.0087	0.0001
/i/	0.0340	0.0130	N/A	0.0018	0.0010
/o/	0.0169	0.0218	0.0001	N/A	0.0010
/u/	0.0001	0.0307	0.0001	0.0241	N/A



Prédiction de l'erreur de prédiction

Les coefficients trouvés semblent donner un indice du pouvoir de prédiction.



Résumé

- On peut utiliser les phonèmes connus d'un locuteur pour approximer les phonèmes inconnus avec **plus de précision** que les **phonèmes moyens**.
- Les **noyaux** permettent de faire cette approximation même à partir d'une **petite base de données**.
- Un **algorithme génétique** simple a permis d'optimiser la **fonction de distance** du noyau.
- Certains coefficients sont pratiquement nuls, ce qui suggère que pour des dizaines de phonèmes la matrice de relation serait **creuse**.



Recherche future

- Construire une plus grande base de données.
- Prédire le spectre de phase harmonique.
- Prédire le spectre de puissance de la partie résiduelle.
- Prédire l'évolution temporelle des spectres dans le cas de diphones.



Publication

FOURNIER, P.-A. et BRAULT, J.-J., 2005, “Harmonic Envelope Prediction for Realistic Speech Synthesis Using Kernel Interpolation”, Proceedings of the International Joint Conference on Neural Networks 2005 (IJCNN05), Montréal.



Questions?

?

