h e p i a

Haute école du paysage, d'ingénierie
et d'architecture de Genève

Classification de phonèmes pour le jeu sérieux FunSpeech

Yohann Perez

/perezy/phoneme

Plan

- FunSpeech
  - Problématique et algorithme actuel
- Résolution de la problématique
  - Production d'un son
  - Jeu de données
  - Caractéristiques acoustiques & algorithmes de classification
  - Protocole de test
- Résultats
- Conclusion

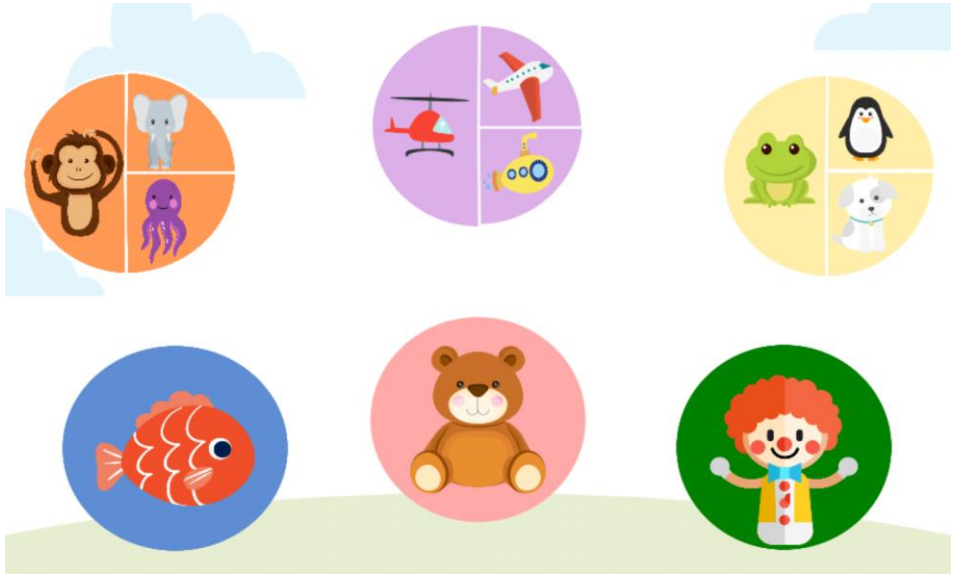
Qu’est ce qu’un phonème ?

- Un phonème est une unité de base formant la parole
- Son distinct
  - Exemples : [a], [o], [i], [ch], [t] ...

FunSpeech

- Application tablette
- Développée à HEPIA
- Destinée aux enfants sourds implantés
  - 2 à 6 ans
- But : Encourager à la production de sons

Source : https://en.wikipedia.org/wiki/Cochlear_implant



/perezy/phoneme

FunSpeech - problématique

- Jeu du clown
- Algorithme de reconnaissance de phonèmes

Investiguer les algorithmes de reconnaissance de sons.
Trouver une combinaison d'algorithmes qui devrait dans le futur remplacer celle actuellement utilisée

Source : FunSpeech

/perezy/phoneme

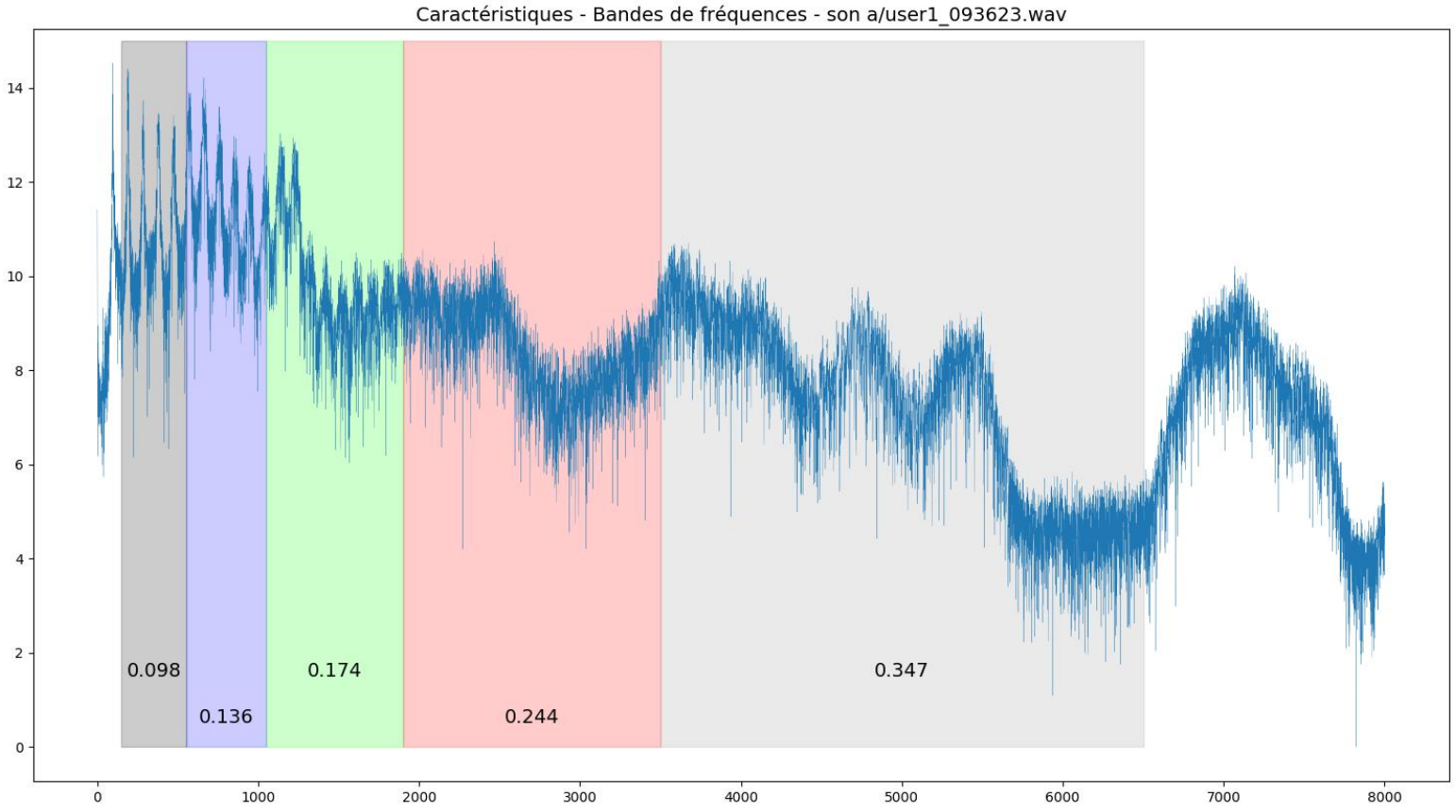
Algorithmes utilisés dans FunSpeech

/perezy/phoneme

Caractériser un phonème
Bandes de fréquences

$$\text{"Énergie" } d'une\ bande = \frac{\sum {x_i}^2}{énergie\ tot.}$$

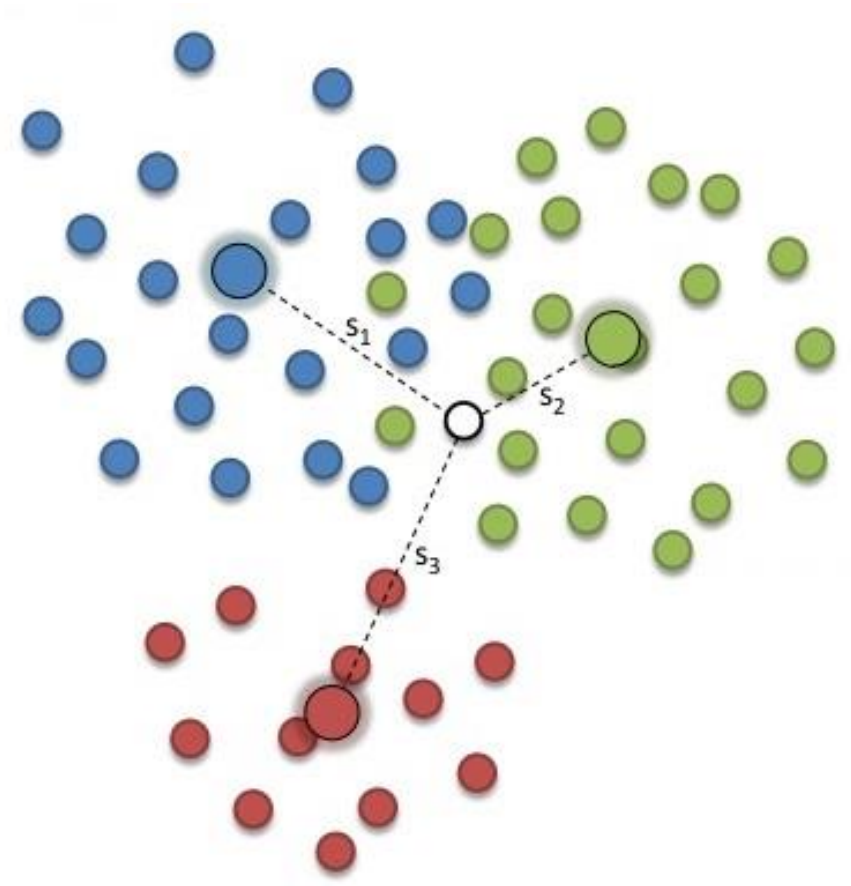
$Caractéristiques\ acoustiques\ de\ ce\ phonème$:
$[0.098, 0.136, 0.174, 0.244, 0.347]$



/perezy/phoneme

Classifier un phonème
Plus proche barycentre

$$Distance\ euclidienne\ (p,q) = \sqrt{\sum_{i=1}^{n=5}(q_i - p_i)^2}$$



/perezy/phoneme

Source: Miha Grčar

Démarche pour résoudre la problématique

- Jeu de données
- Caractéristiques acoustiques
- Algorithmes de classification
- Protocole de test

	File	Group	Information	Length
2	m26er.wav	er	man	0.252
3	m49aw.wav	aw	man	0.293
4	w17aw.wav	aw	woman	0.341
5	g15er.wav	er	girl	0.331
6	w14iy.wav	iy	woman	0.261
7	b18er.wav	er	boy	0.304
8	b09ah.wav	ah	boy	0.285
9	m16er.wav	er	man	0.204

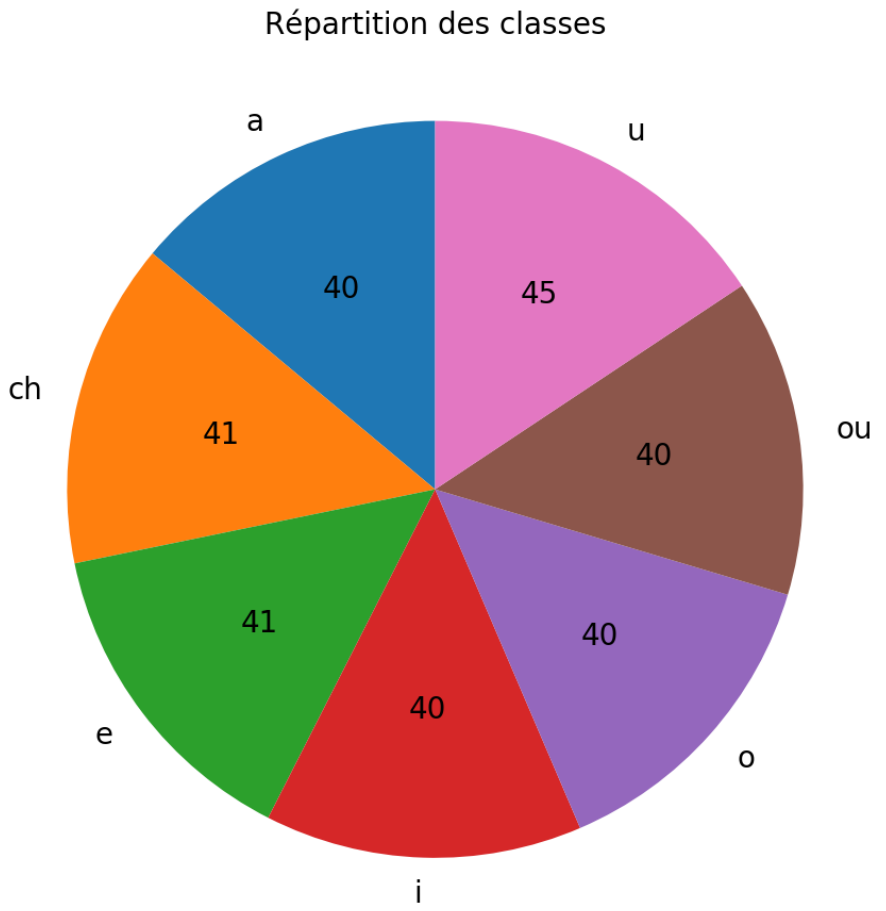
Jeu de données

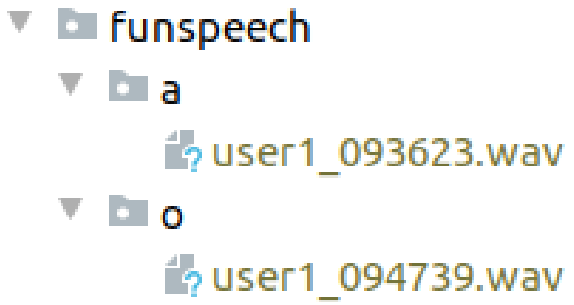
Jeu de données FunSpeech

- 287 données
- 4 locuteurs (hommes)
- 7 phonèmes français

Inconvénient majeur

- Pas de phonèmes d'enfants alors qu'on souhaite reconnaitre des phonèmes d'enfants

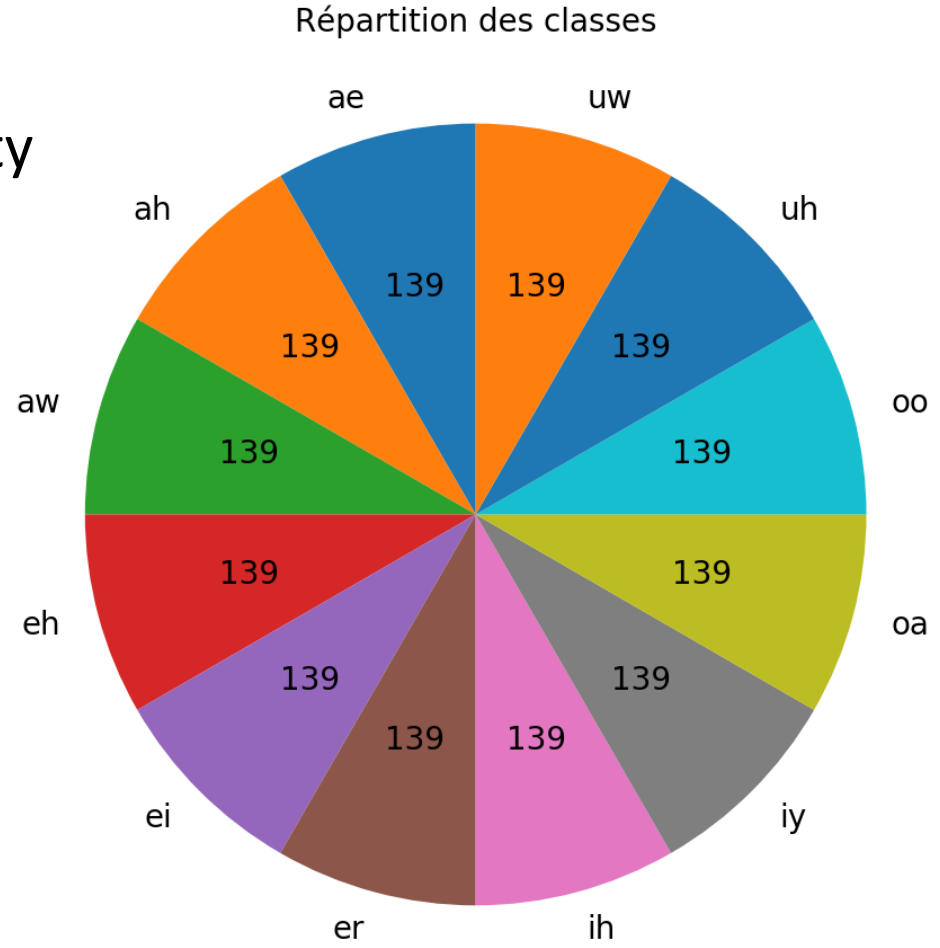


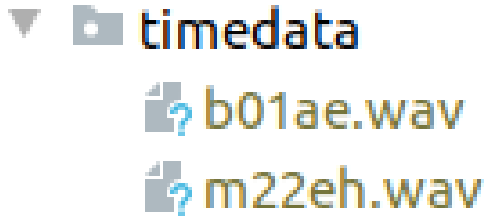


/perezy/phoneme

Jeu de données Timedata

- Western Michigan University
- 1667 données
- 12 phonèmes anglais
- 139 locuteurs
  - 48 femmes et 45 hommes
  - 19 filles et 27 garçons





/perezy/phoneme

dataset_generate_csv.py

FunSpeech

	File	Group	Information	Length
2	ou/user4_095201.wav	ou	user4	2.62
3	ou/user2_095002.wav	ou	user2	2.32
4	ou/user2_094848.wav	ou	user2	1.42
5	ou/user1_095001.wav	ou	user1	2.44

Timedata

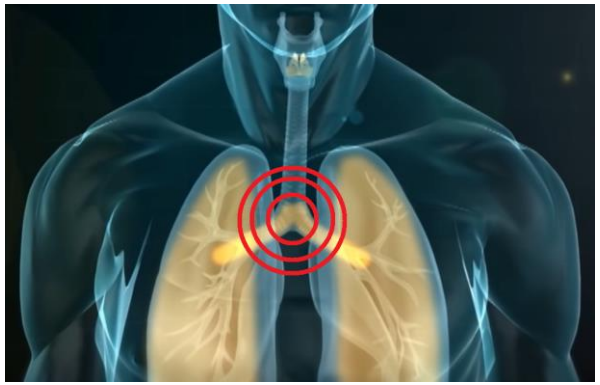
	File	Group	Information	Length
2	m26er.wav	er	man	0.252
3	m49aw.wav	aw	man	0.293
4	w17aw.wav	aw	woman	0.341
5	g15er.wav	er	girl	0.331

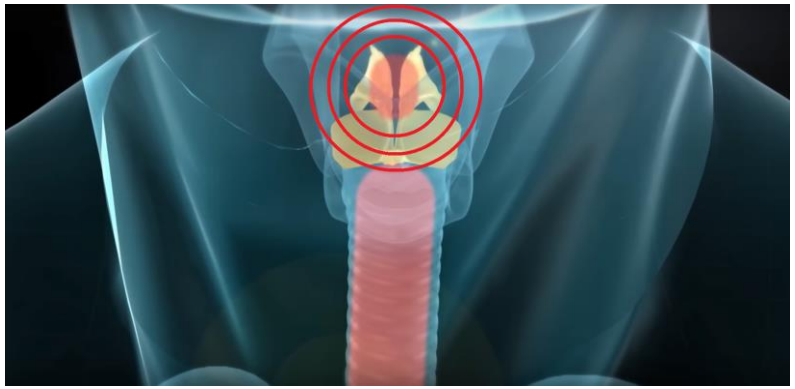
Pandas

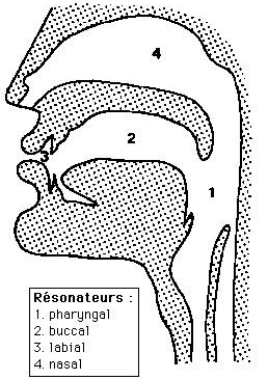
/perezy/phoneme

Caractéristiques acoustiques

Production d’un son







Source : Lucie Cambrai, Youtube

- Expiration
  - **Détermine intensité (faible ou fort)**
- Impulsion glottale
  - **Détermine la hauteur du son (aigu ou grave)**
- Forme du canal vocal
  - **Détermine le phonème qui va être produit**

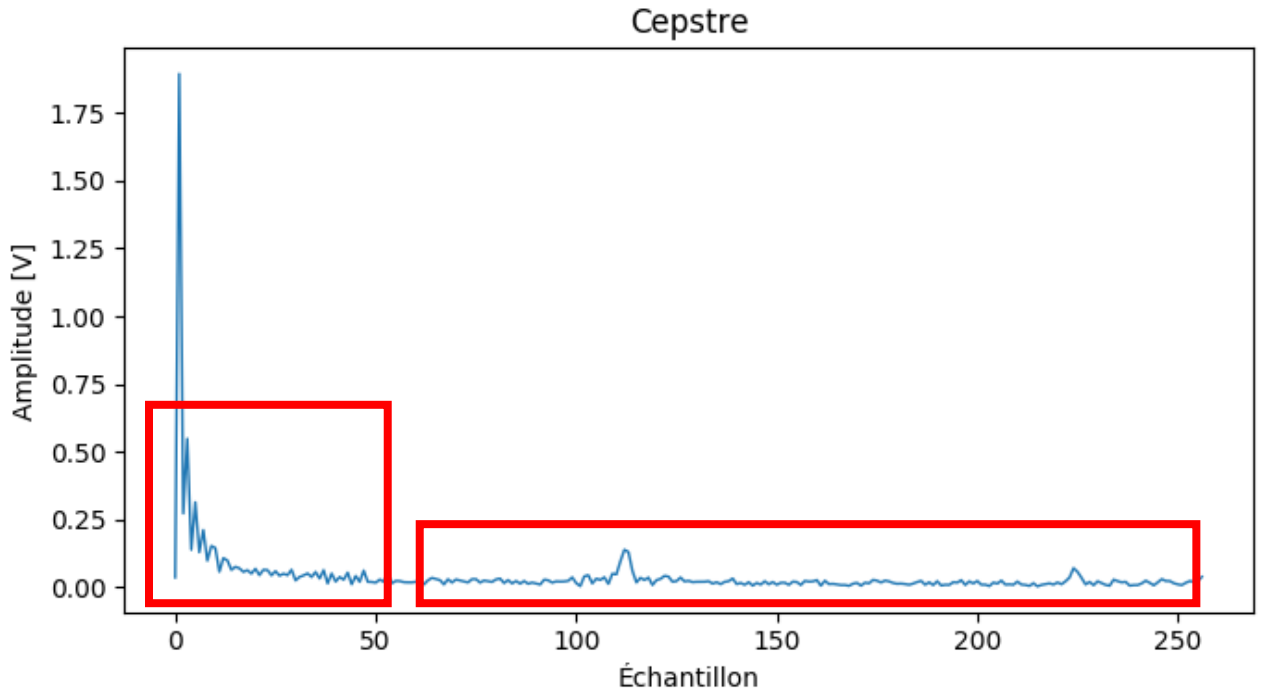
Dans un signal vocal, l’impulsion glottale et la forme du canal vocal sont convoluées

/perezy/phoneme

Domaine cepstral

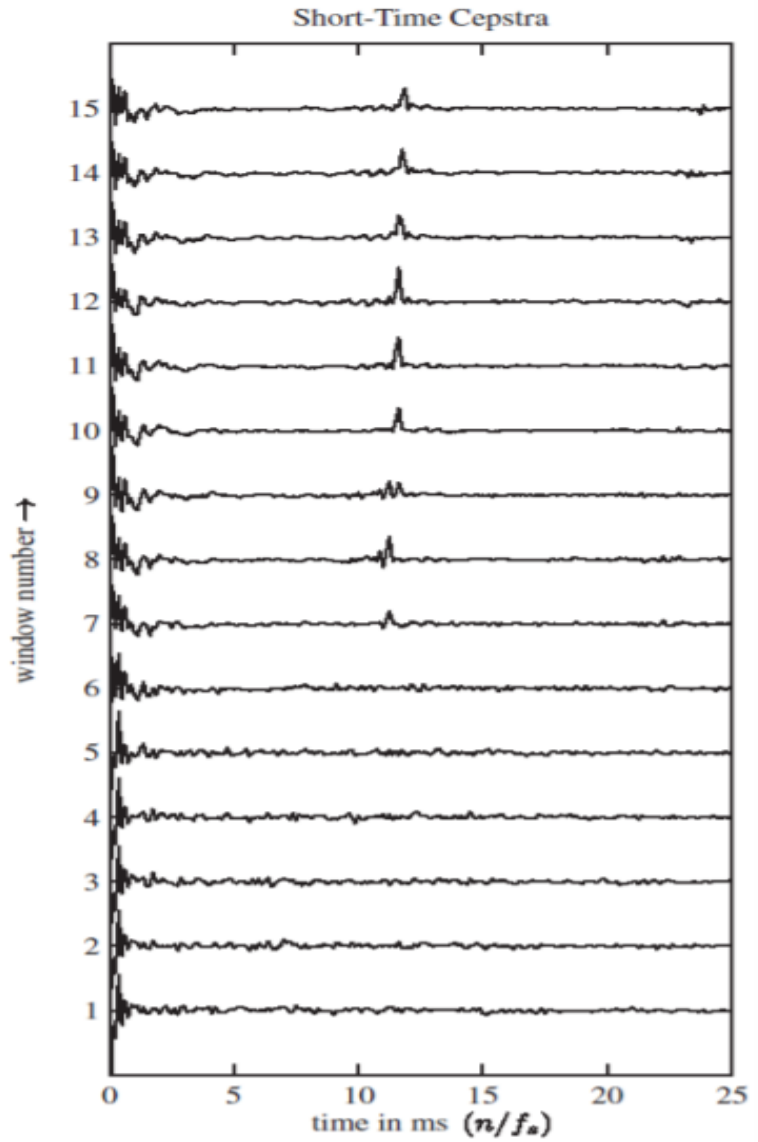
- Distinguer l'impulsion glottale et la forme du canal vocal
  = Analyse cepstrale

$$Cepstre = FFT^{-1}(\ln(|FFT(x)|))$$



Mel Frequency Cepstral Coefficients

- Évolution de l'impulsion glottale et de la forme du canal vocal au cours du temps

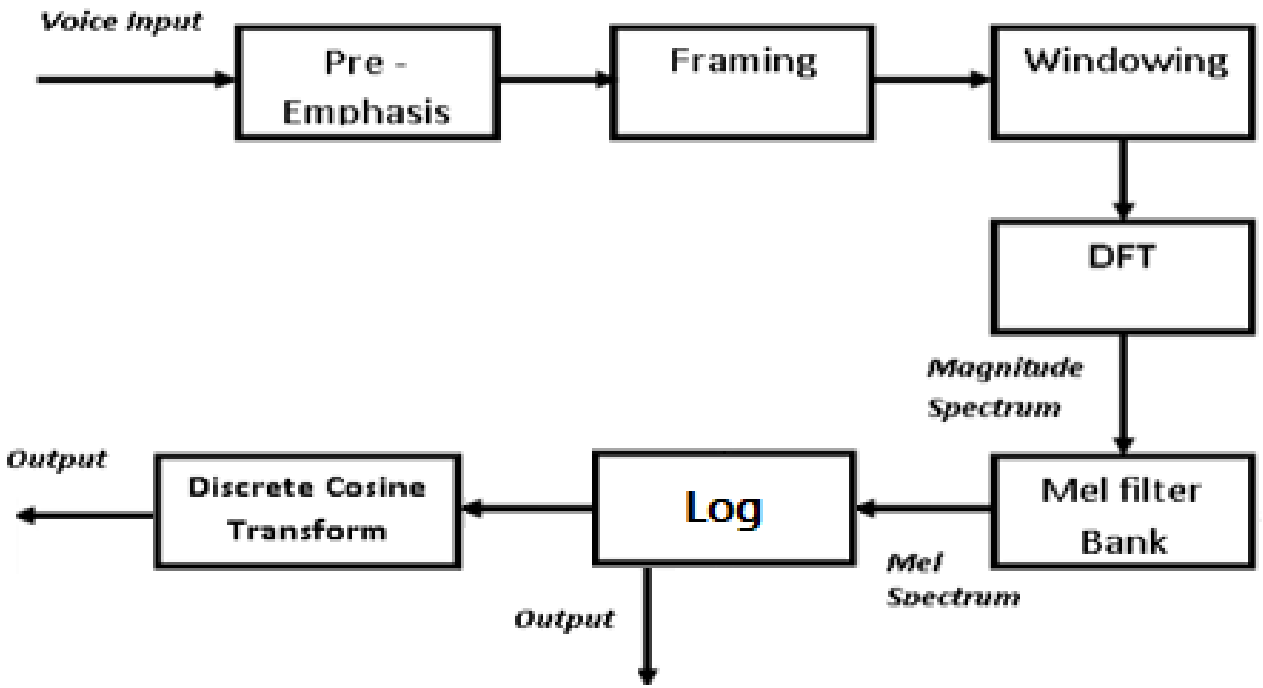


Source :
Rainer & Schäfer, 2007

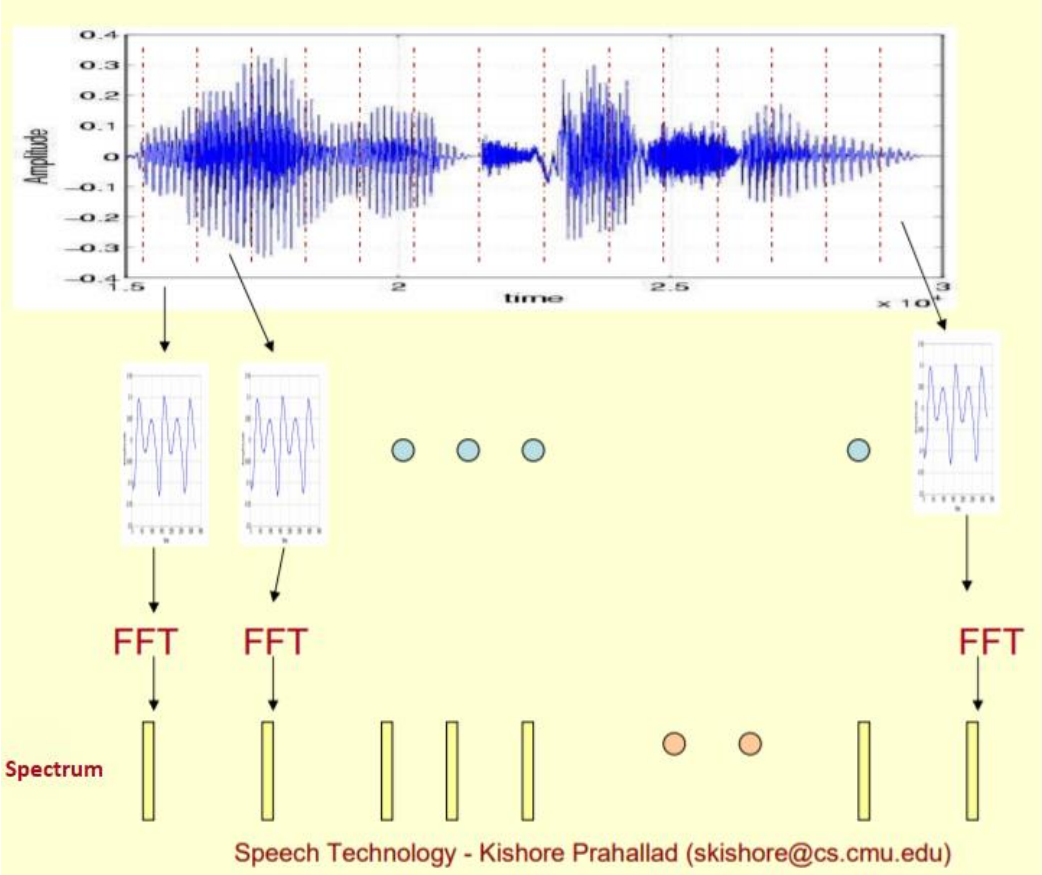
/perezy/phoneme

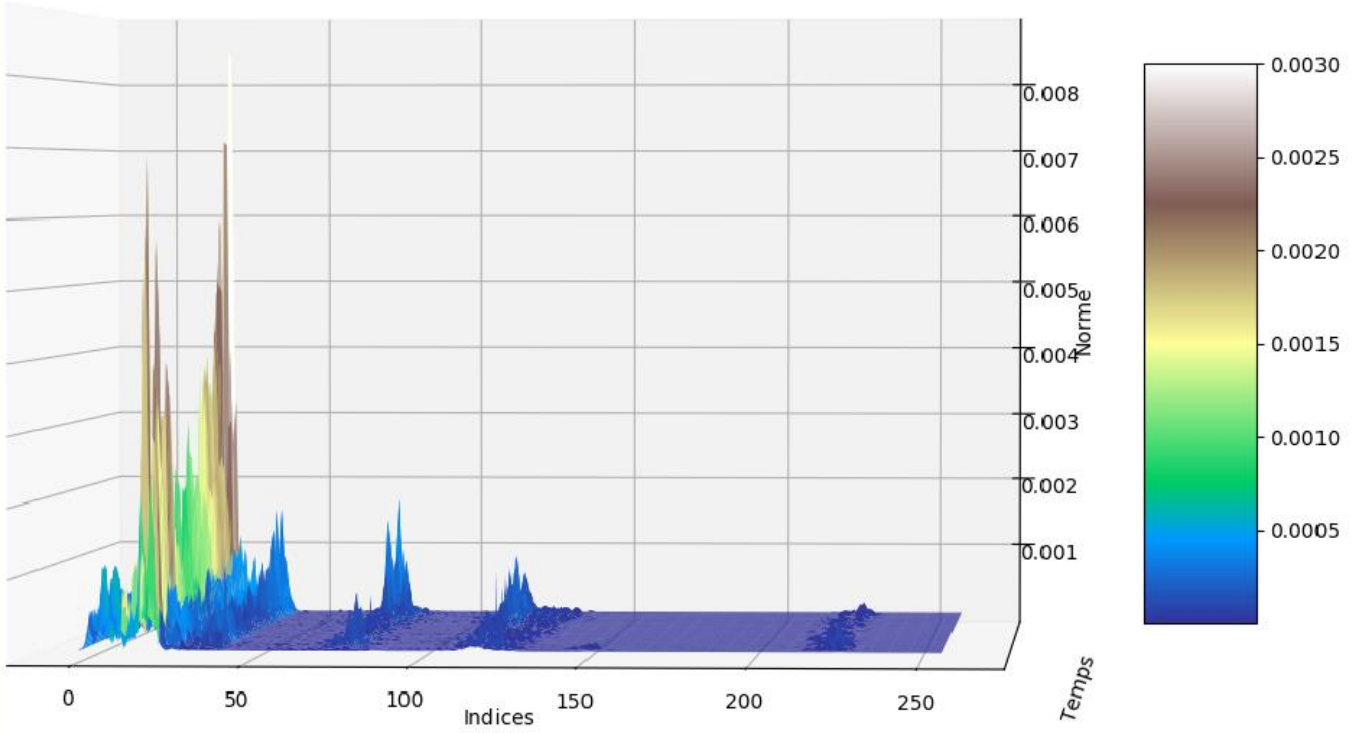
Mel Frequency Cepstral Coefficients

- Analyse par banc de filtres
- Analyse cepstrale



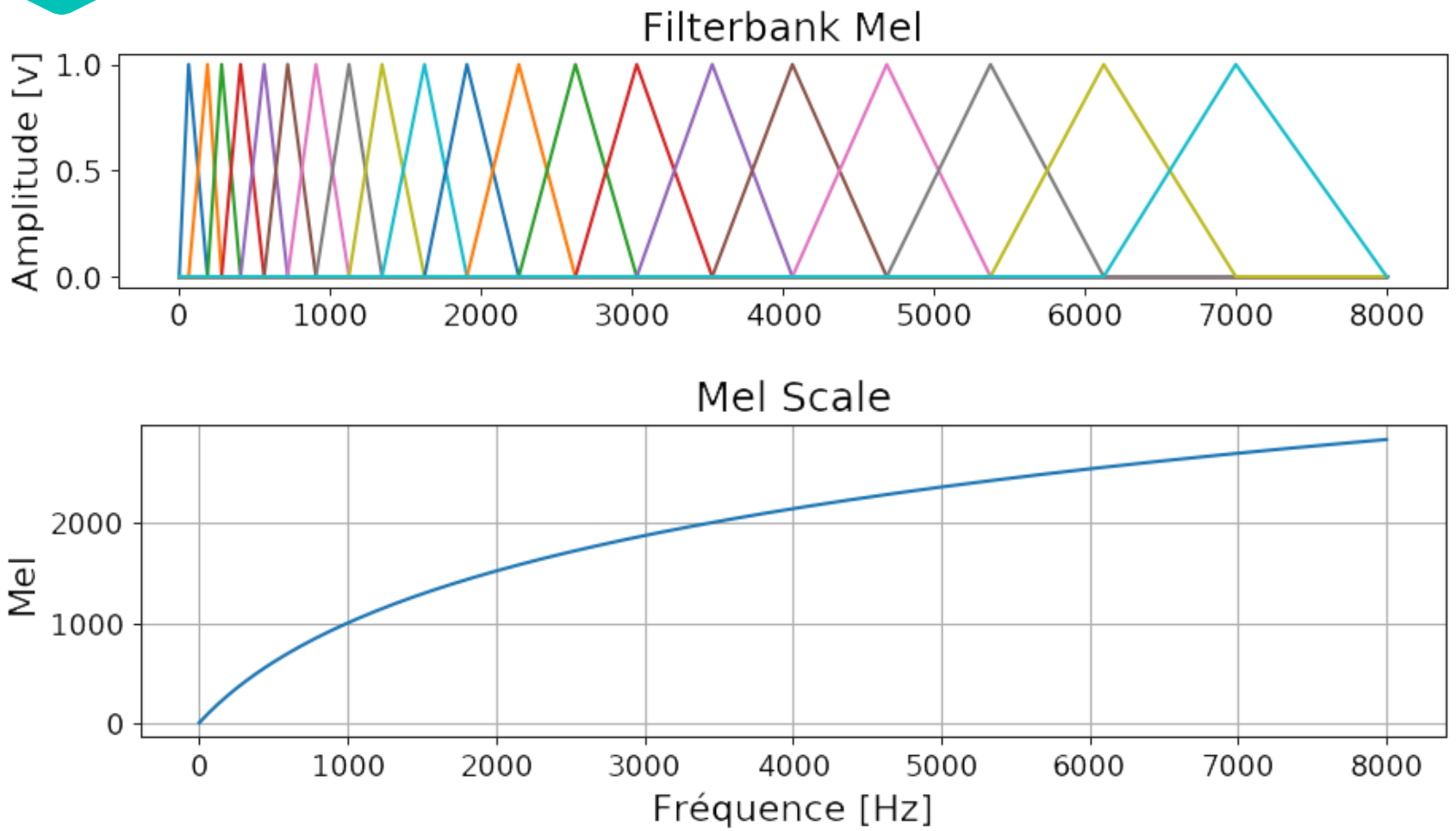
Mel Frequency Cepstral Coefficients





/perezy/phoneme

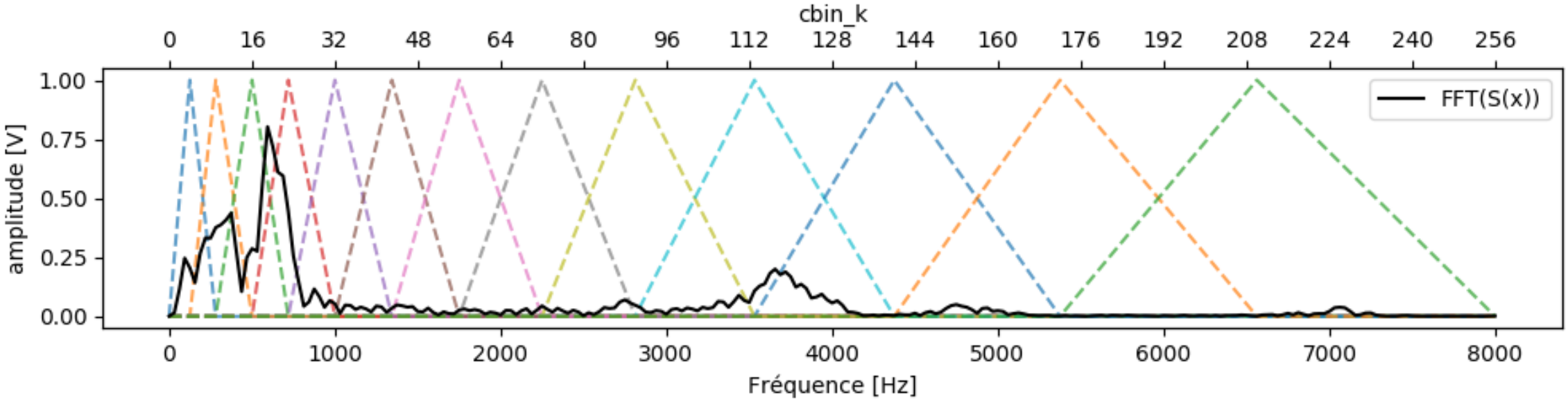
Mel Frequency Cepstral Coefficients



/perezy/phoneme

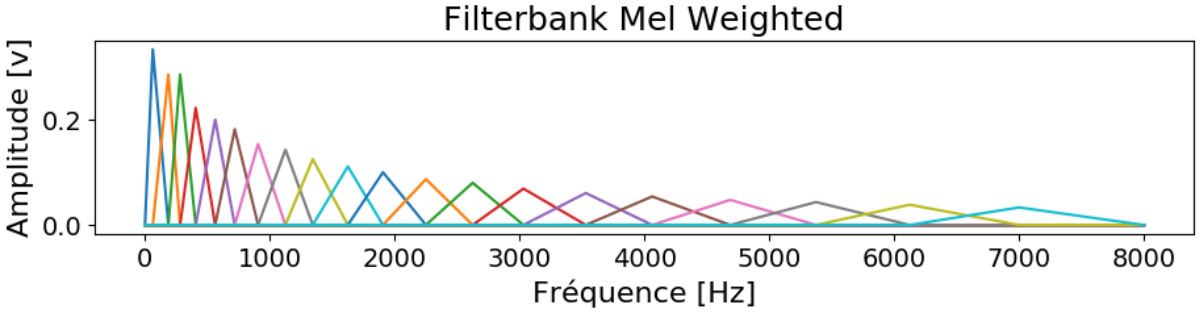
Mel Frequency Cepstral Coefficients

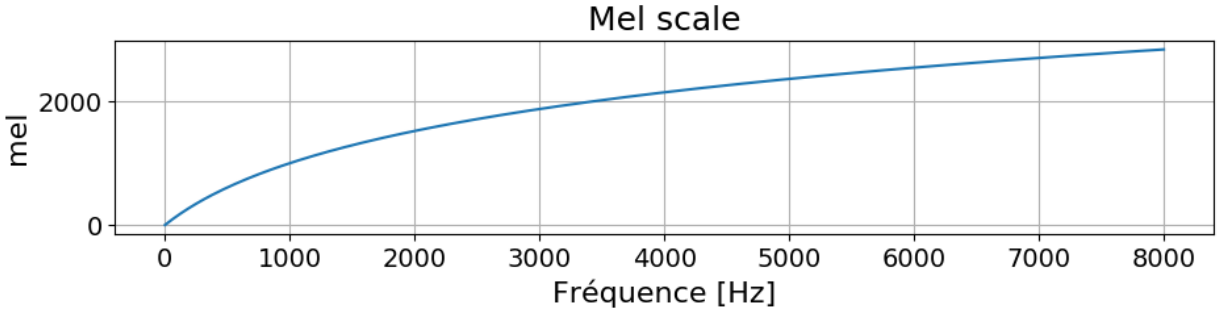
Banc de filtres pour capturer l'enveloppe spectrale

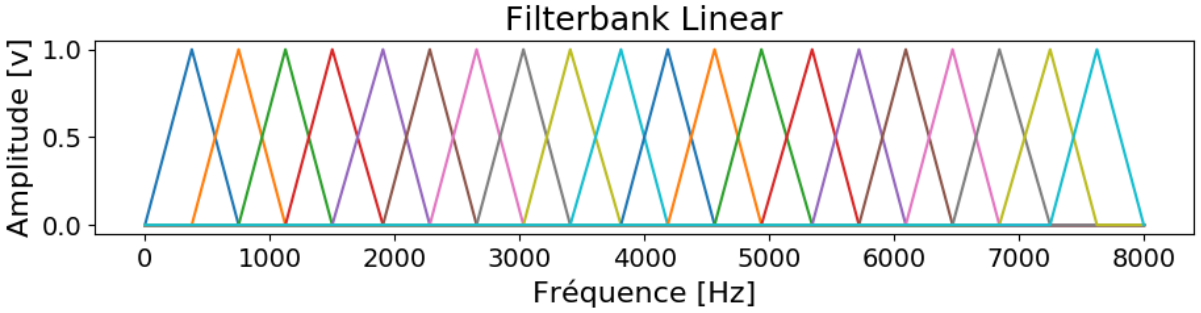


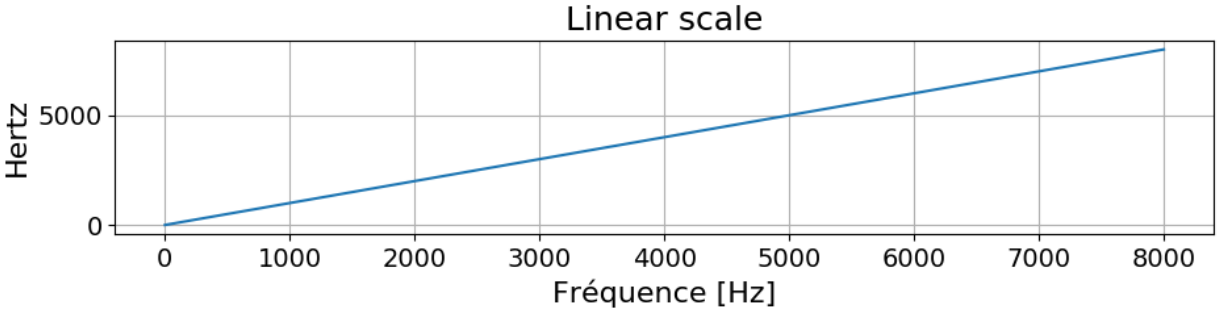
/perezy/phoneme

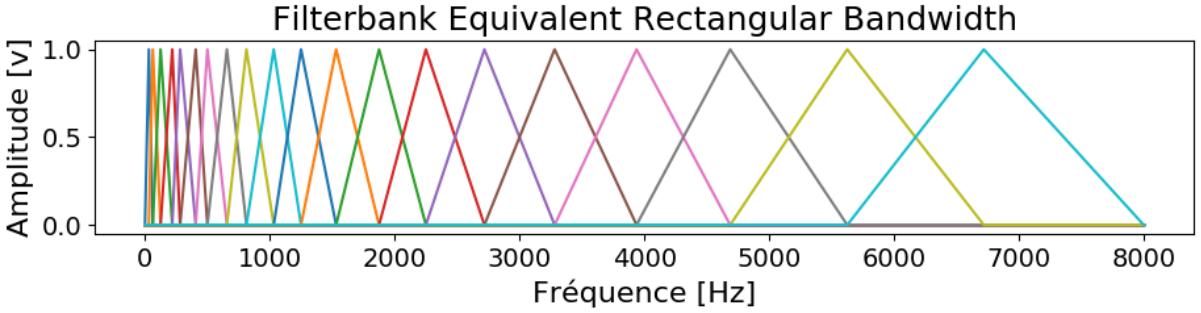
Autres bancs de filtres

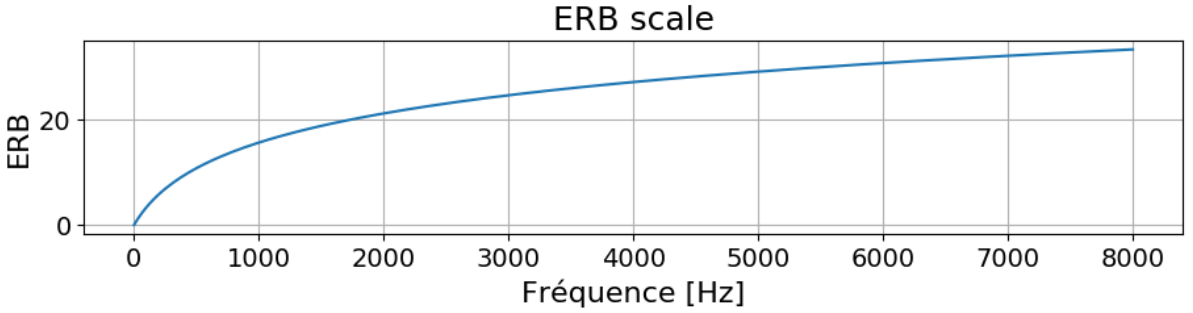






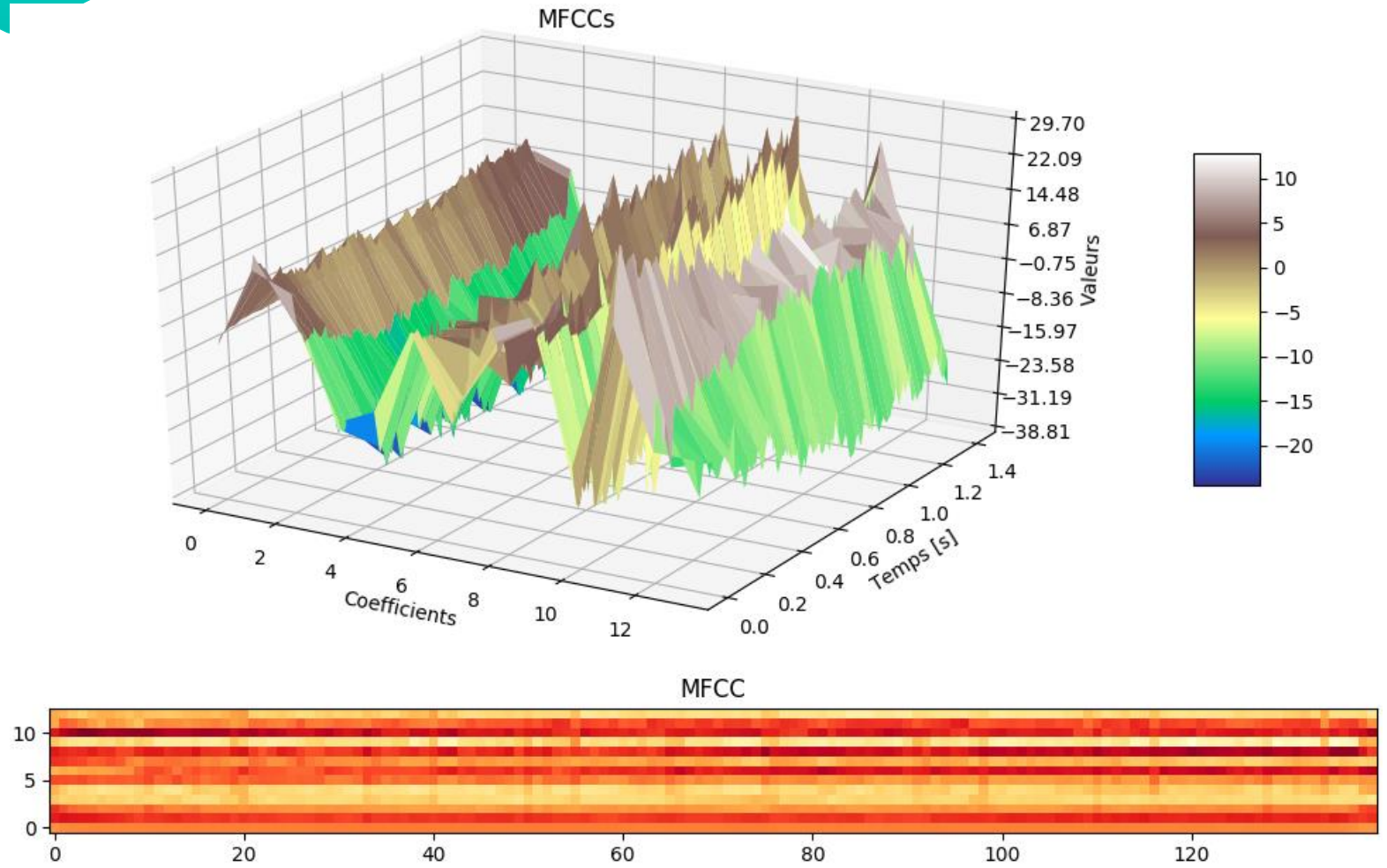






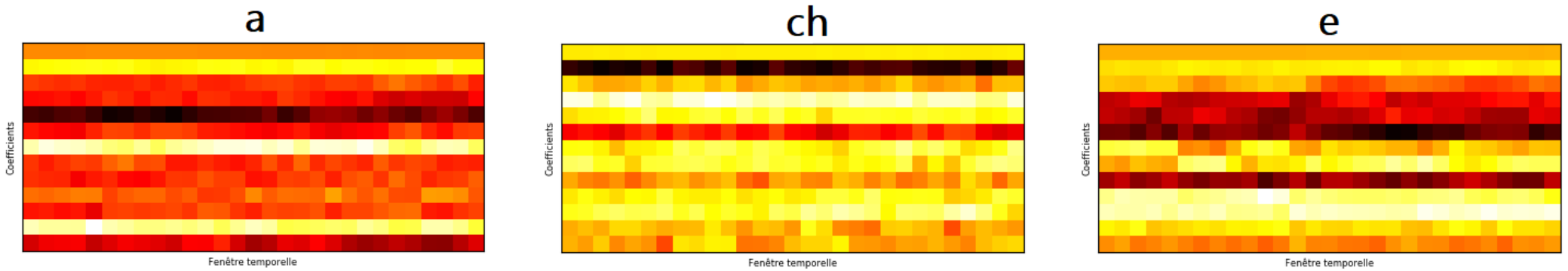
/perezy/phoneme

Mel Frequency Cepstral Coefficients



/perezy/phoneme

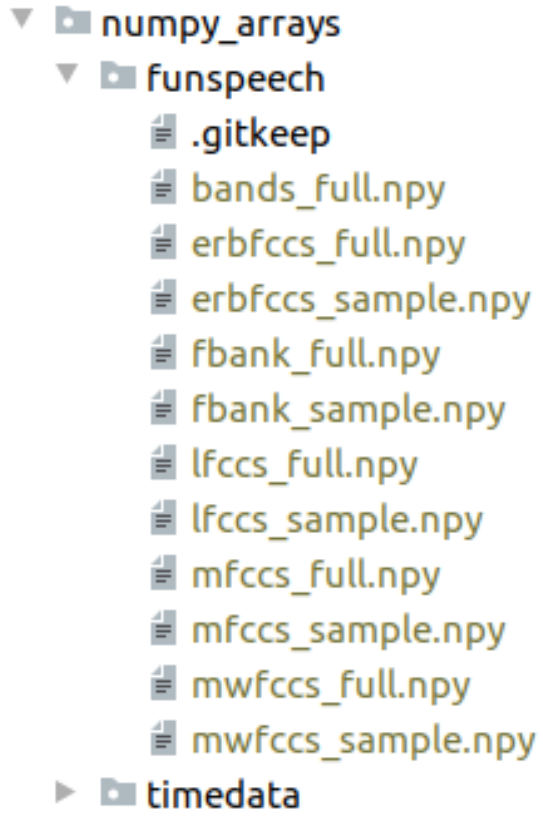
Mel Frequency Cepstral Coefficients



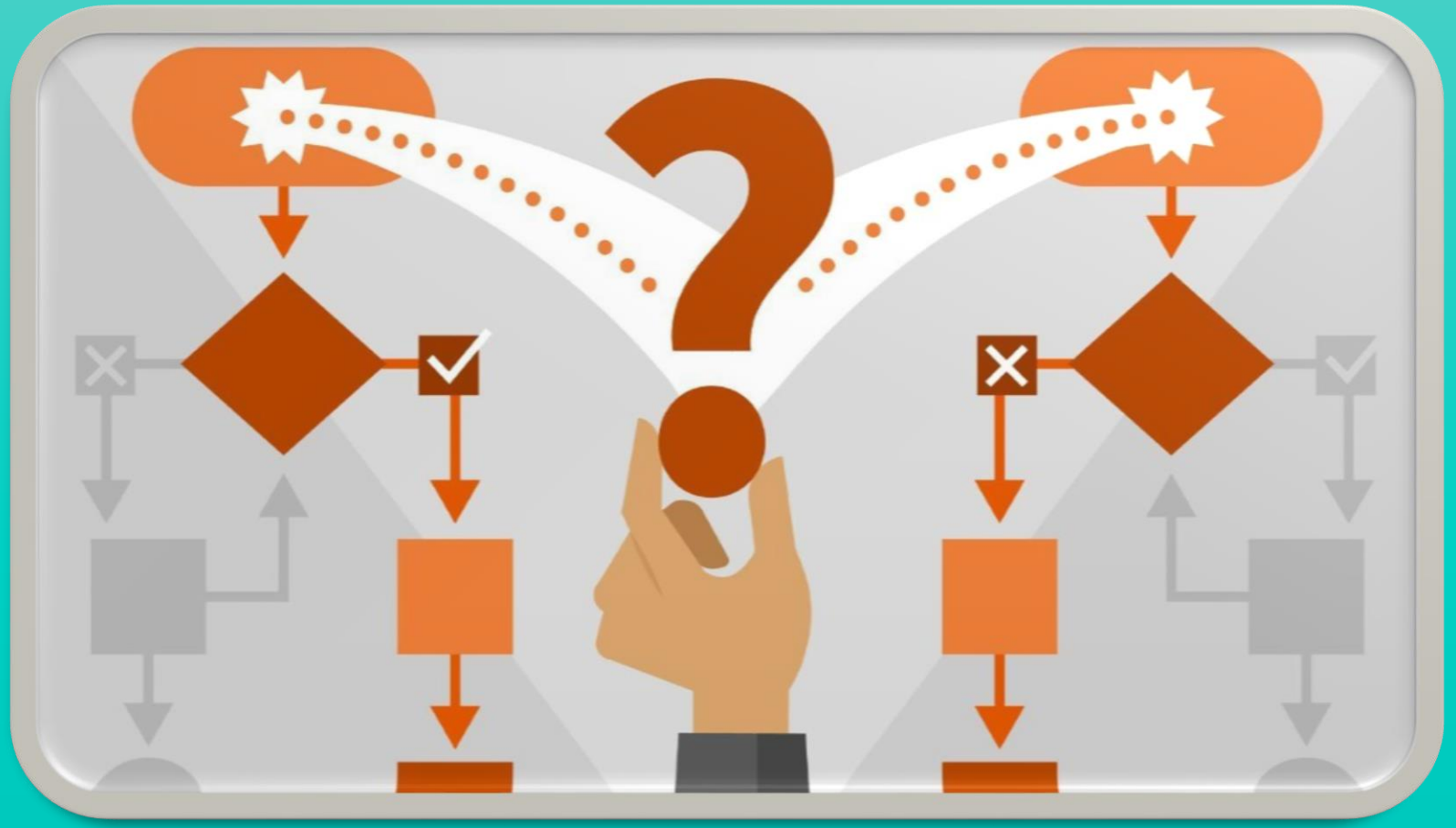
/perezy/phoneme

features_extraction.py

- Extrait les caractéristiques sur
  - Échantillon de son
  - Toute la durée du son
- Stocke les caractéristiques
  - Fichiers contenant des tableaux numpy



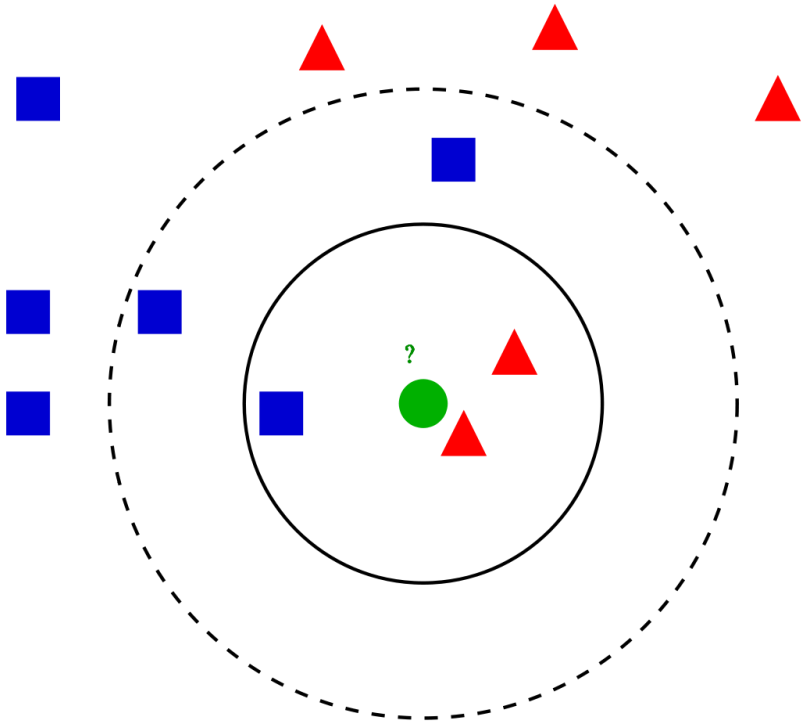
/perezy/phoneme



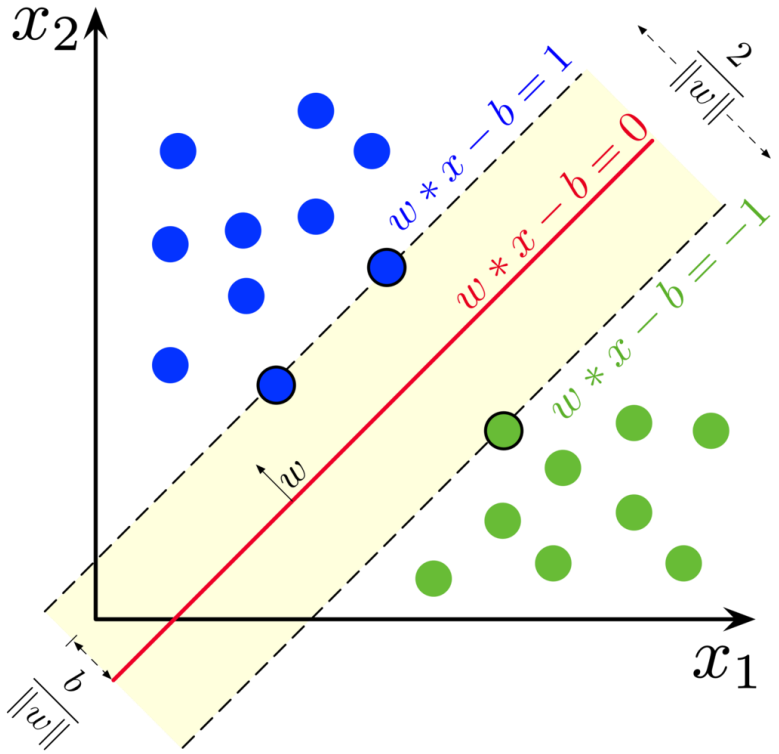
Algorithmes de classification

Algorithmes de classification

K – plus proches voisins



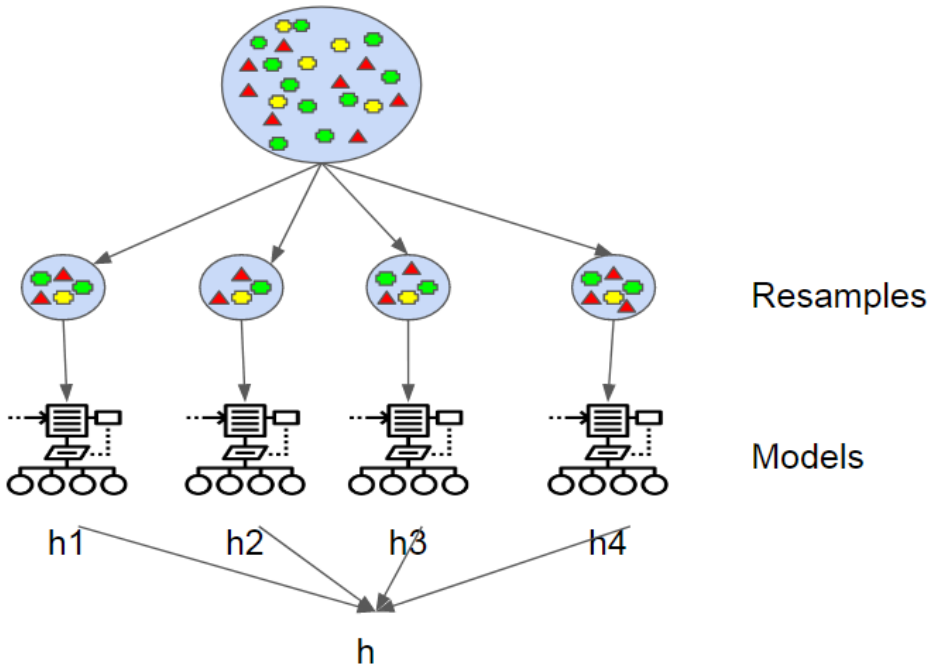
Séparateurs à vastes marges



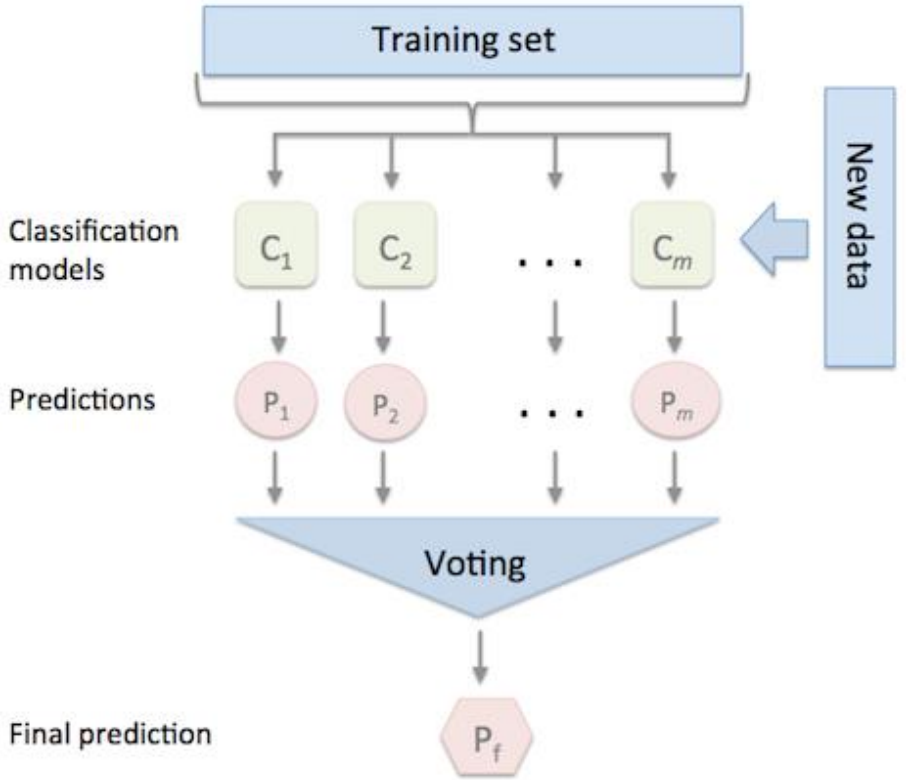
/perezy/phoneme

Algorithmes de classification

Bagging



Combinaison de classificateurs



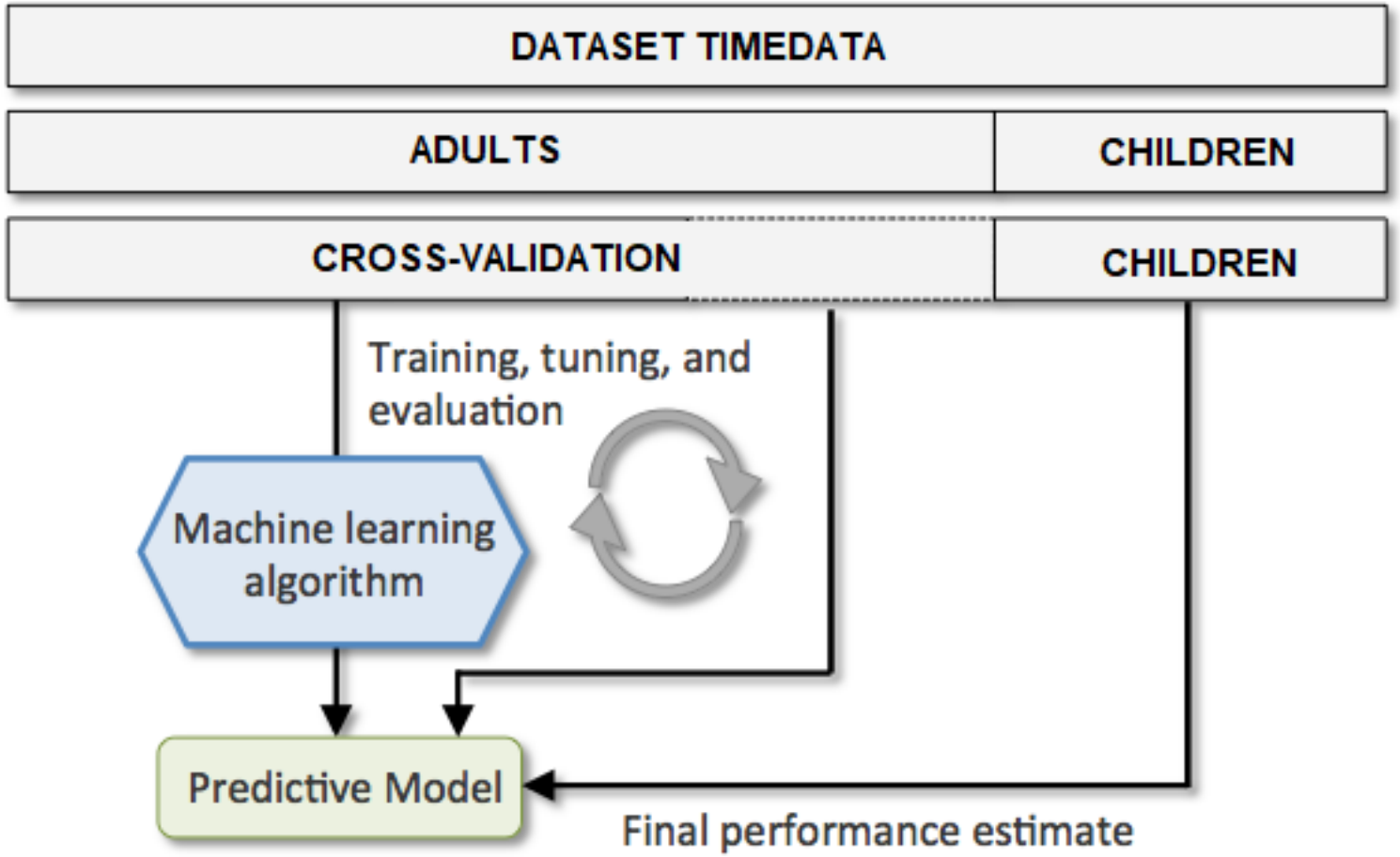
/perezy/phoneme

Protocole de test

Que doit faire mon protocole de test ?

- Déterminer si l'utilisation de phonèmes uniquement produits par des adultes pour l'entrainement d'un modèle suffit pour correctement classifier des phonèmes produits par des d'enfants
- Déterminer combien le fait d'ajouter des données d'entrainement est bénéfique à la résolution de la problématique
- Comparer les différents combinaisons d'algorithme d'extraction et de classification pour déterminer laquelle est la plus performante et la plus adaptée pour résoudre la problématique

Protocole de test



/perezy/phoneme

Ensemble d’entrainement, de validation et de test

- Séparer timedata en 2 sets de données : adultes et enfants
- Adultes : Ensemble d’entrainement et de validation
  - Entrainement ➔ Pouvoir prédire
  - Validation ➔ Ajustement des paramètres
- Enfants : Ensemble de test
  - ➔ Estimation non biaisée

Validation croisée : K-Fold (ensemble adultes) & Optimisation du classificateur

- Recherche exhaustive pour trouver les paramètres qui donne le meilleur score sur l'ensemble de validation

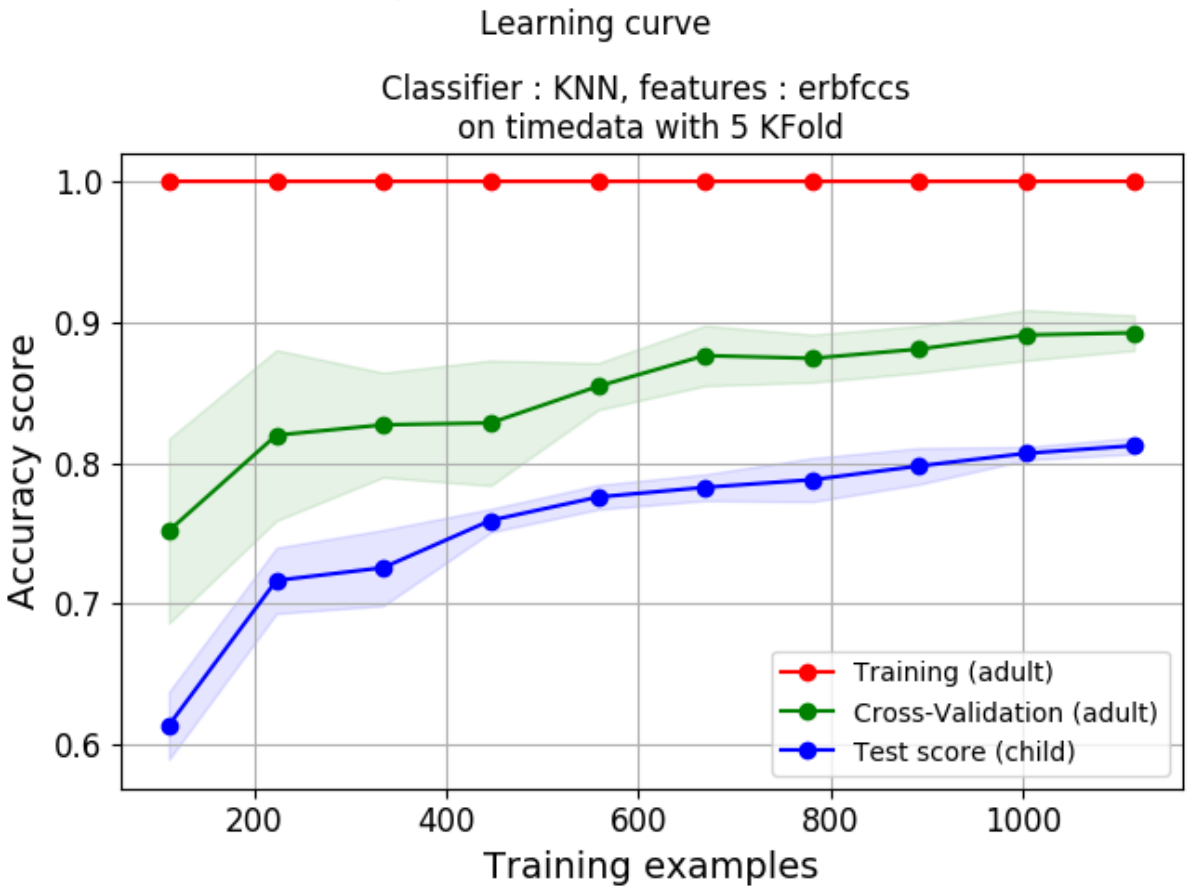
5-fold CV

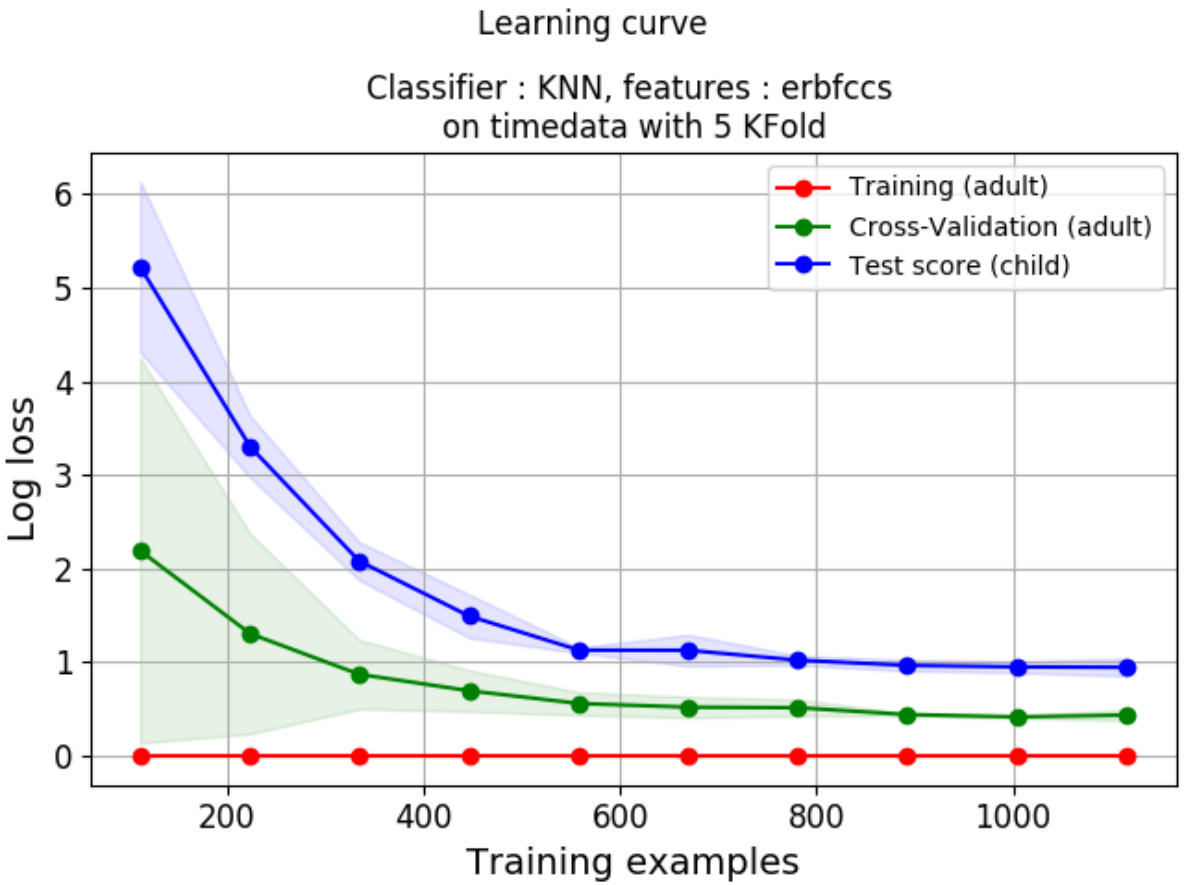
ADULTS

Estimation 1	Validation	Train	Train	Train	Train
Estimation 2	Train	Validation	Train	Train	Train
Estimation 3	Train	Train	Validation	Train	Train
Estimation 4	Train	Train	Train	Validation	Train
Estimation 5	Train	Train	Train	Train	Validation

```
svm_parameters = [{'kernel':['linear'],
                   'C':[1,10,100],
                   'gamma':np.logspace(-9,3,10)}]
grid = GridSearchCV(SVC(probability=True),
                    param_grid=svm_parameters,
                    cv=nfold,
                    scoring='accuracy_score')
```

Courbes d'apprentissage





/perezy/phoneme

Protocole de test

Évaluer la qualité de prédictions d'un modèle

- Métriques de classification
  - Précision
  - Incertitude
  - Temps

Procédures de validation

- Validation croisée
- Sets d'entrainement, de validation et de test
- Courbe d'apprentissage

Optimisation des classificateurs

- GridSearchCV

Résultats

Temps d’extraction

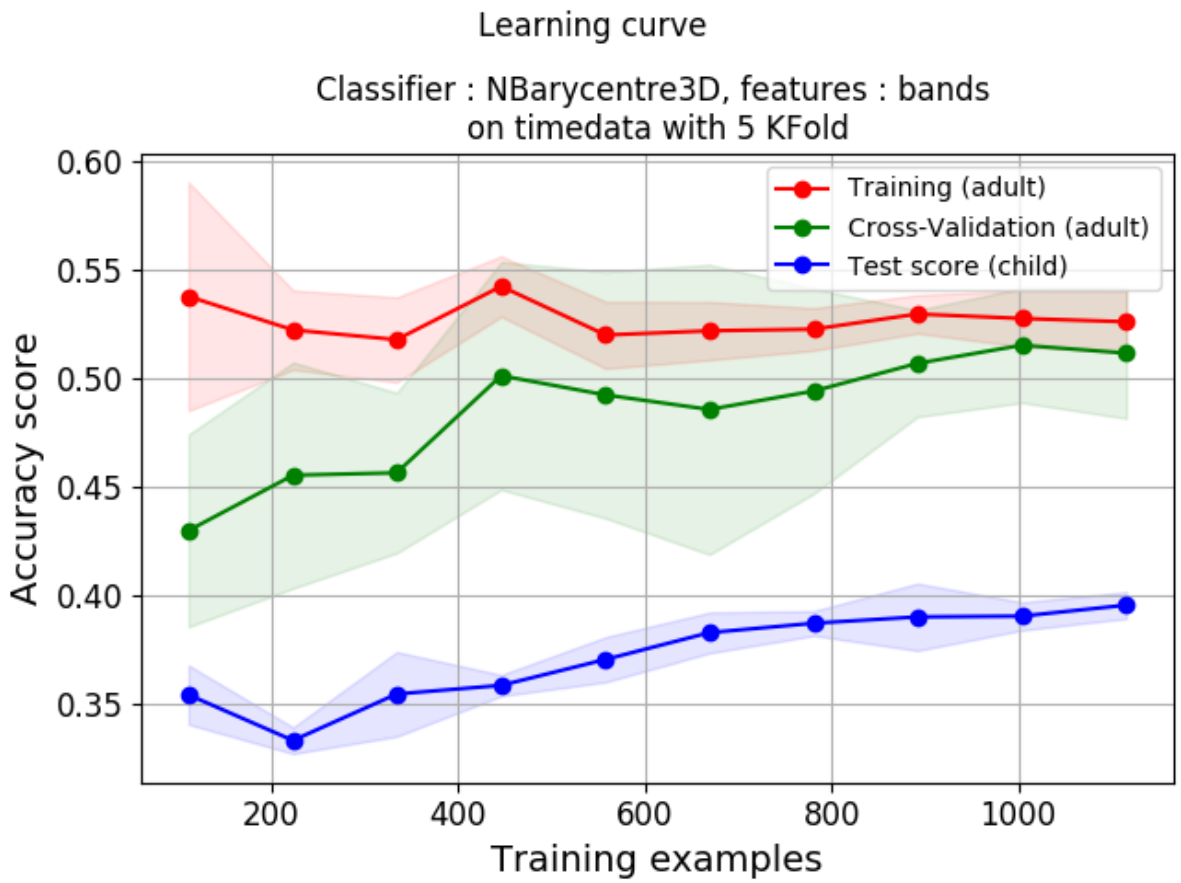
```
********** (timeit.timeit) **********
Temps d'extraction des caractéristiques acoustiques
Bandes de fréquences      0.0016907600599961369
MFCC                      0.006335410500000762
MWFCC                     0.007366854649999685
ERB-FCC                   0.006944680970000263
LFCC                      0.006945115169996825
BFCC                      0.006984381679999387
```

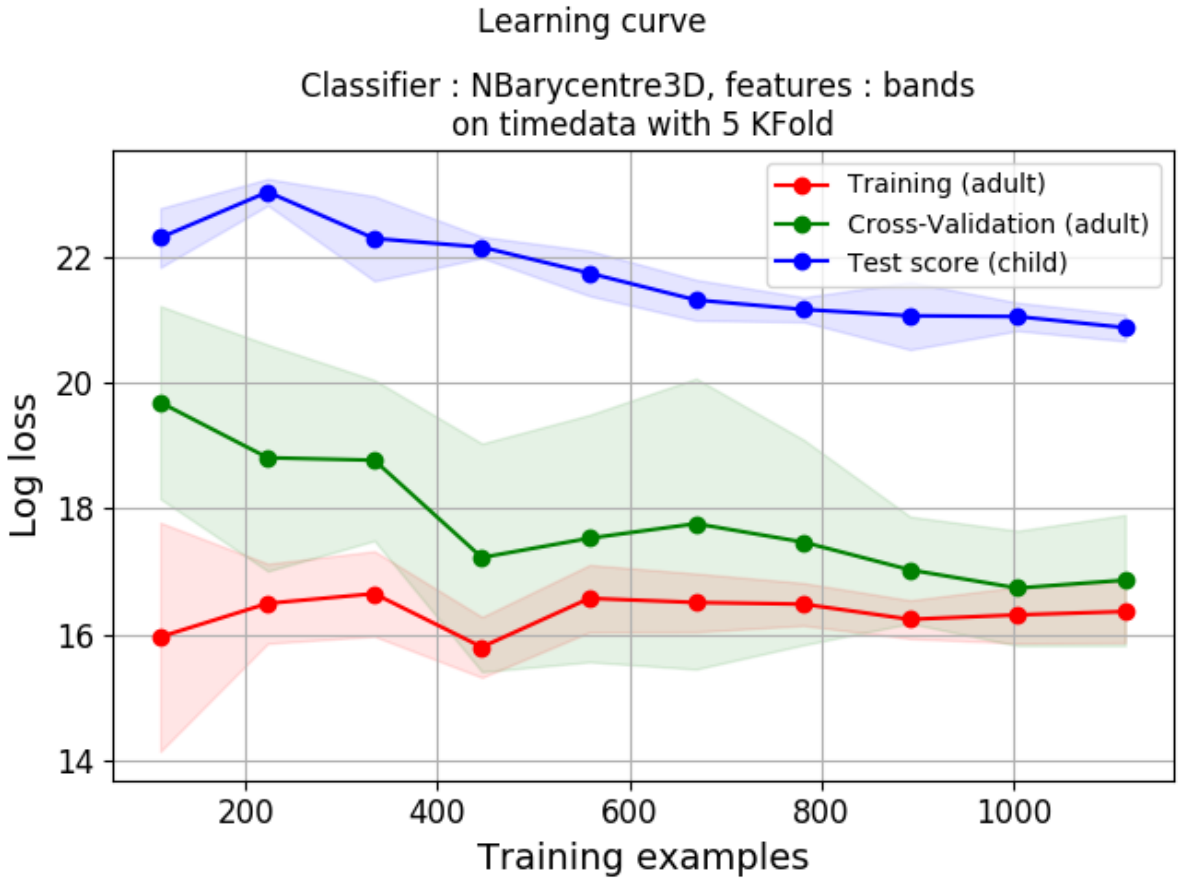
/perezy/phoneme

Résultats

- 5 caractéristiques acoustiques
  - Bandes de fréquences, Banc de filtres, MFCC et ses 3 variantes
- 5 algorithmes de classification
  - Plus proche barycentre, Plus proches voisins, SVM et 2 méthodes ensemblistes

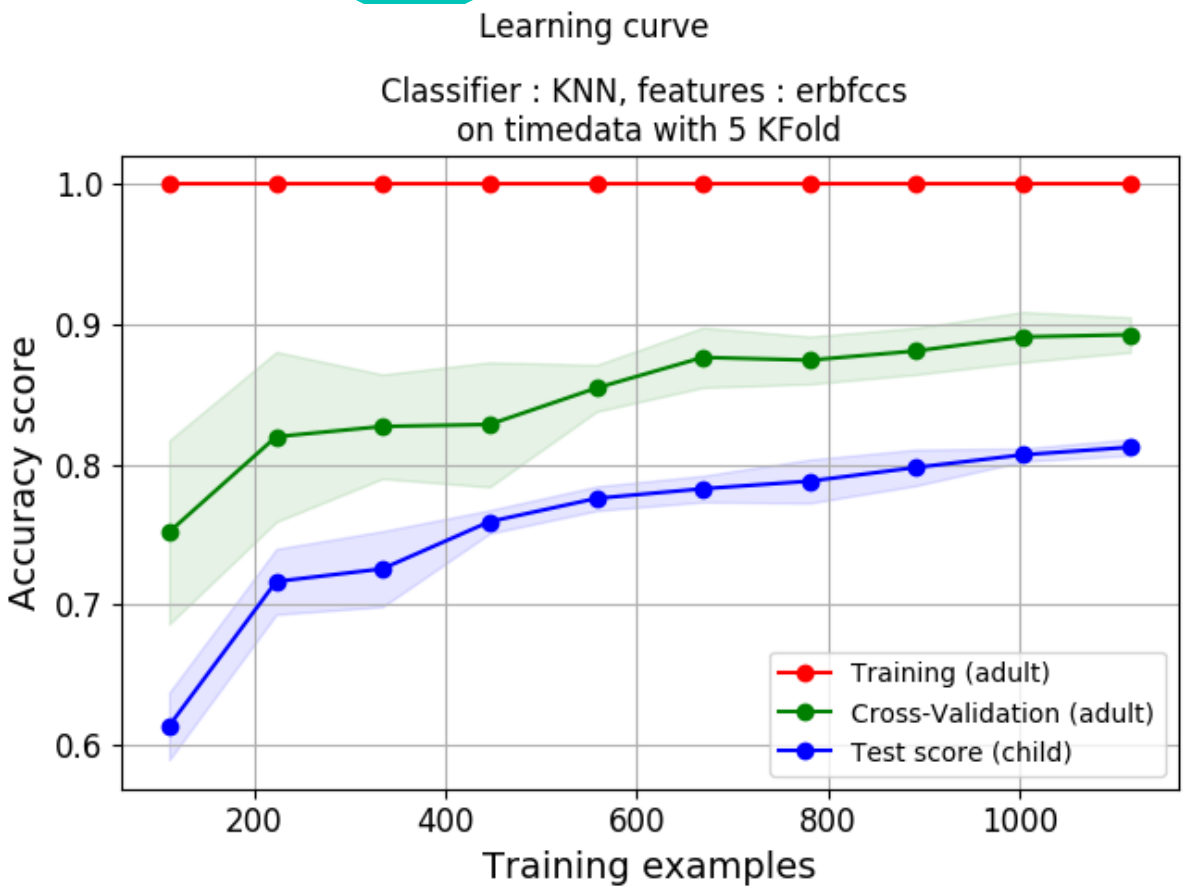
Algorithmes de FunSpeech
Plus proche barycentre + bandes de fréquences

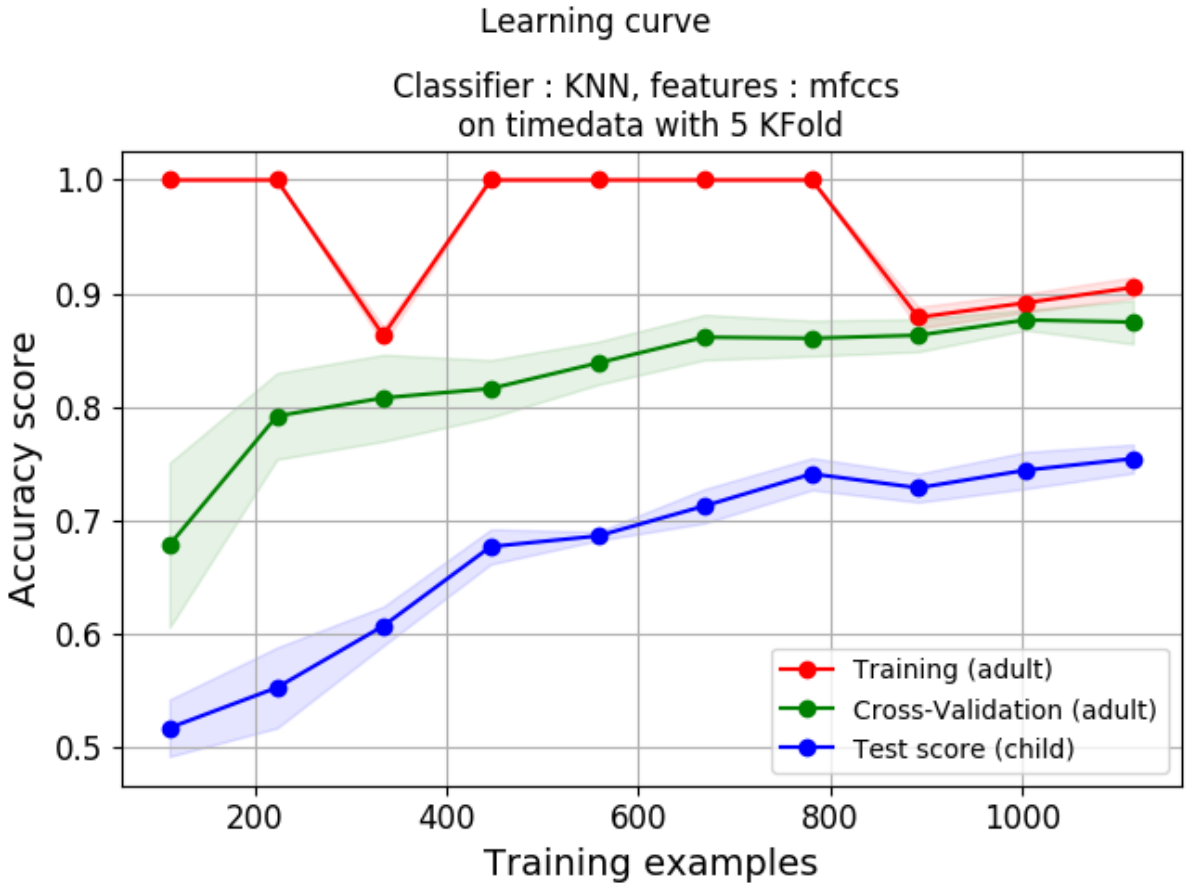




/perezy/phoneme

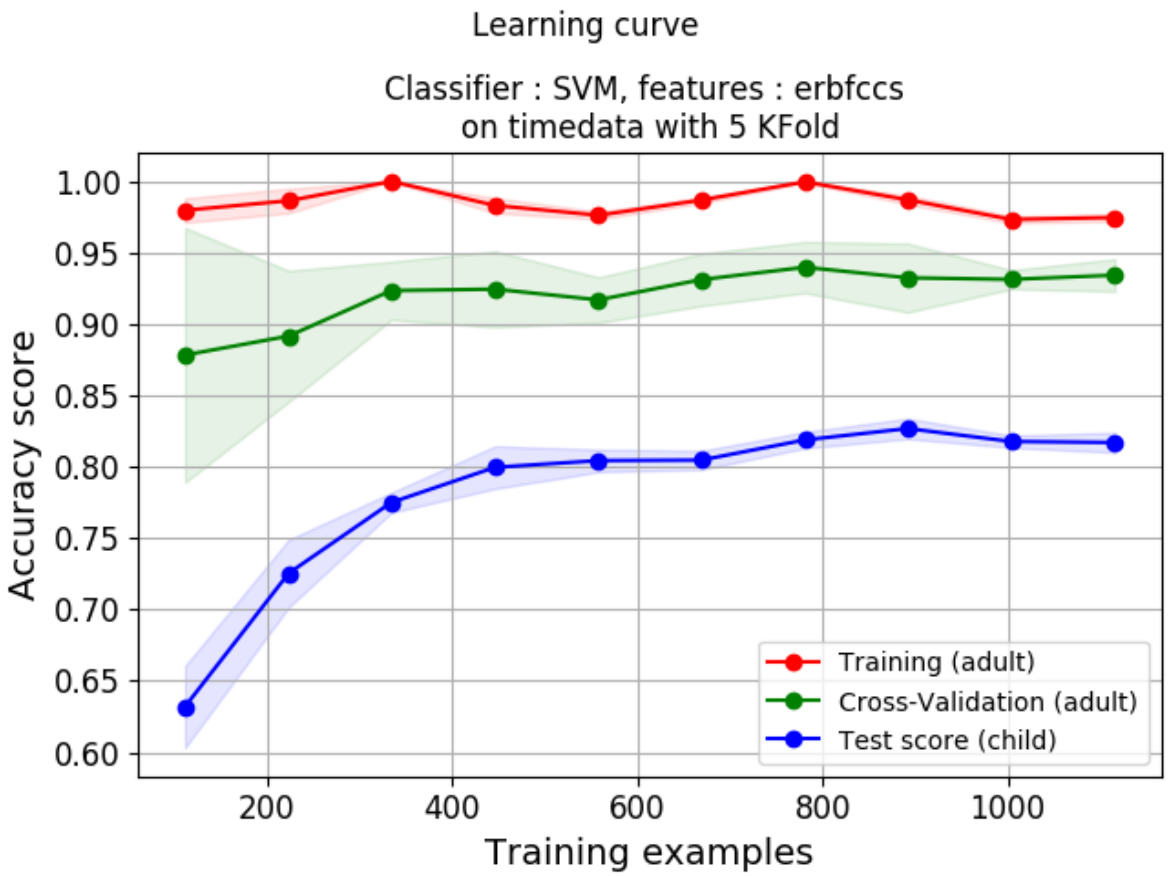
Précision : KNN + ERBFCC et KNN +MFCC

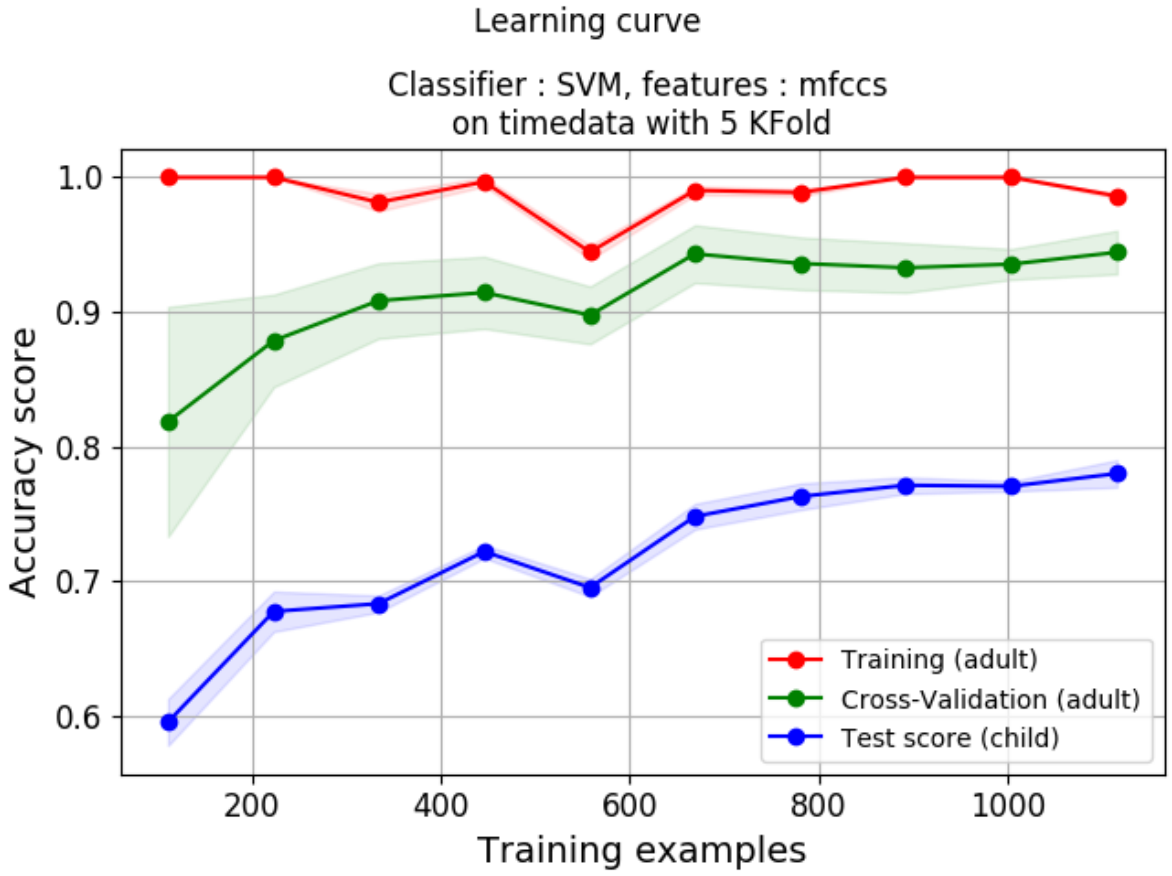




/perezy/phoneme

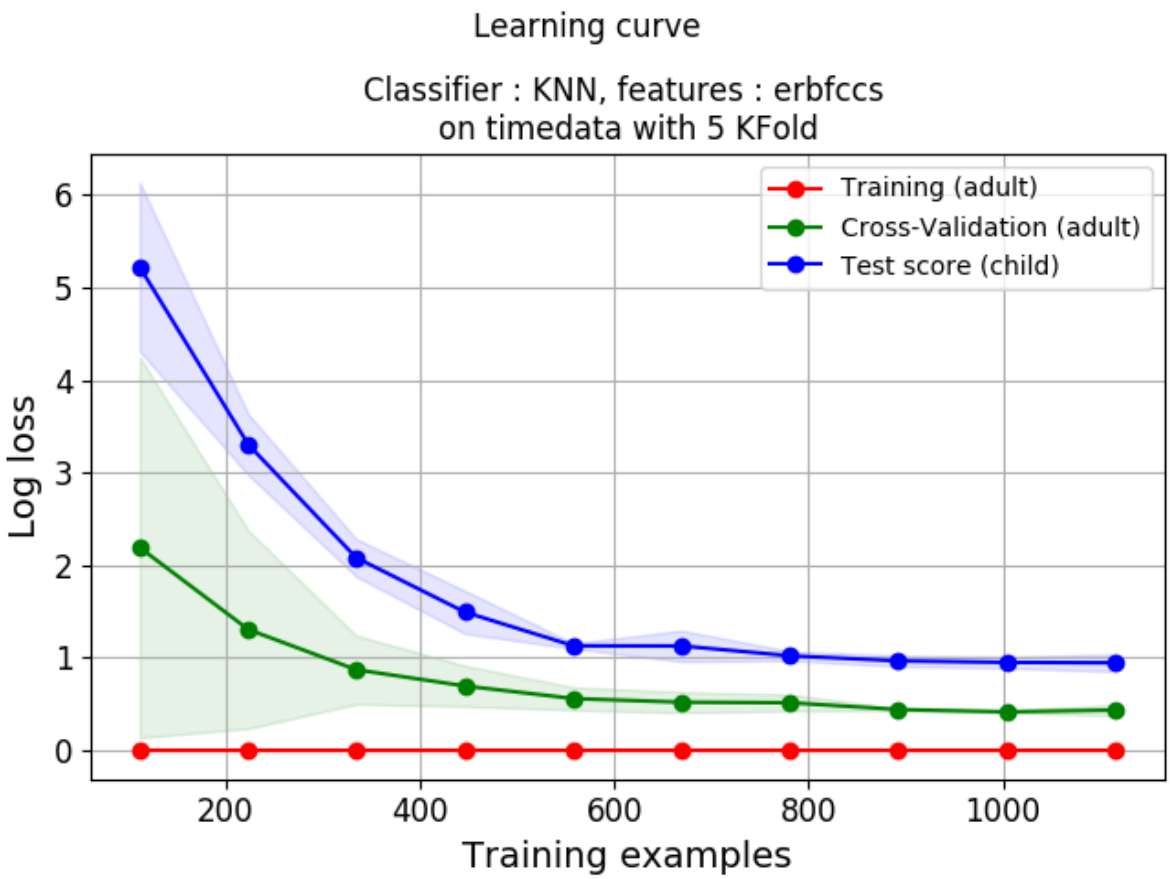
Précision: SVM + ERBFCC vs SVM + MFCC

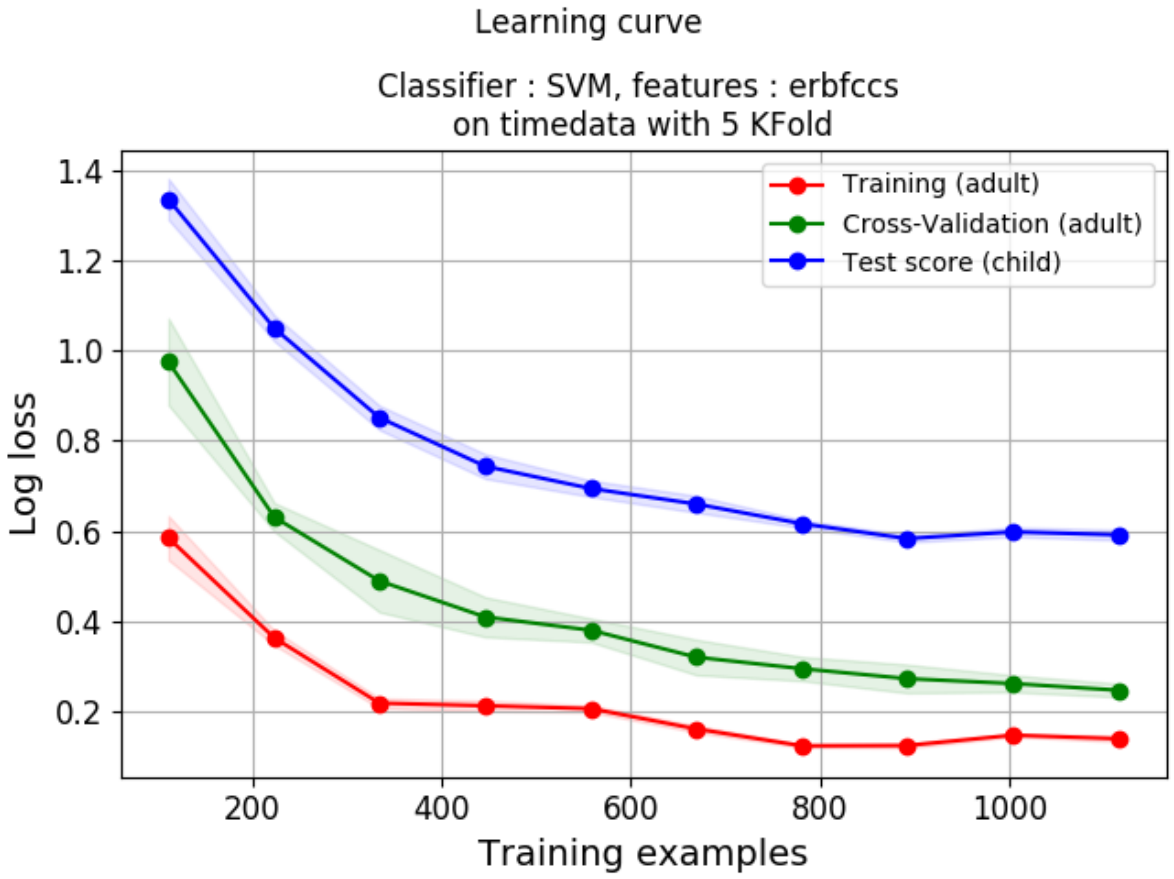




/perezy/phoneme

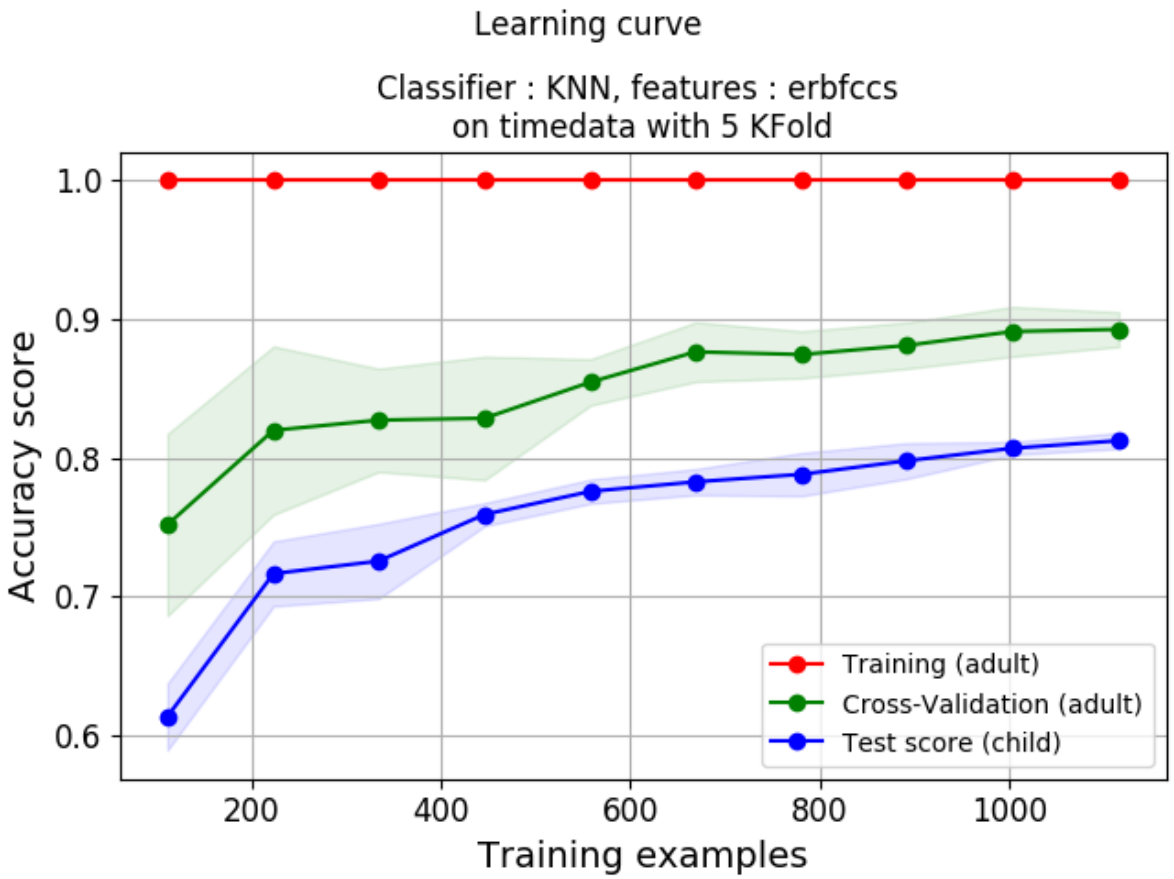
Incertitude: KNN + ERBFCC vs SVM + ERBFCC

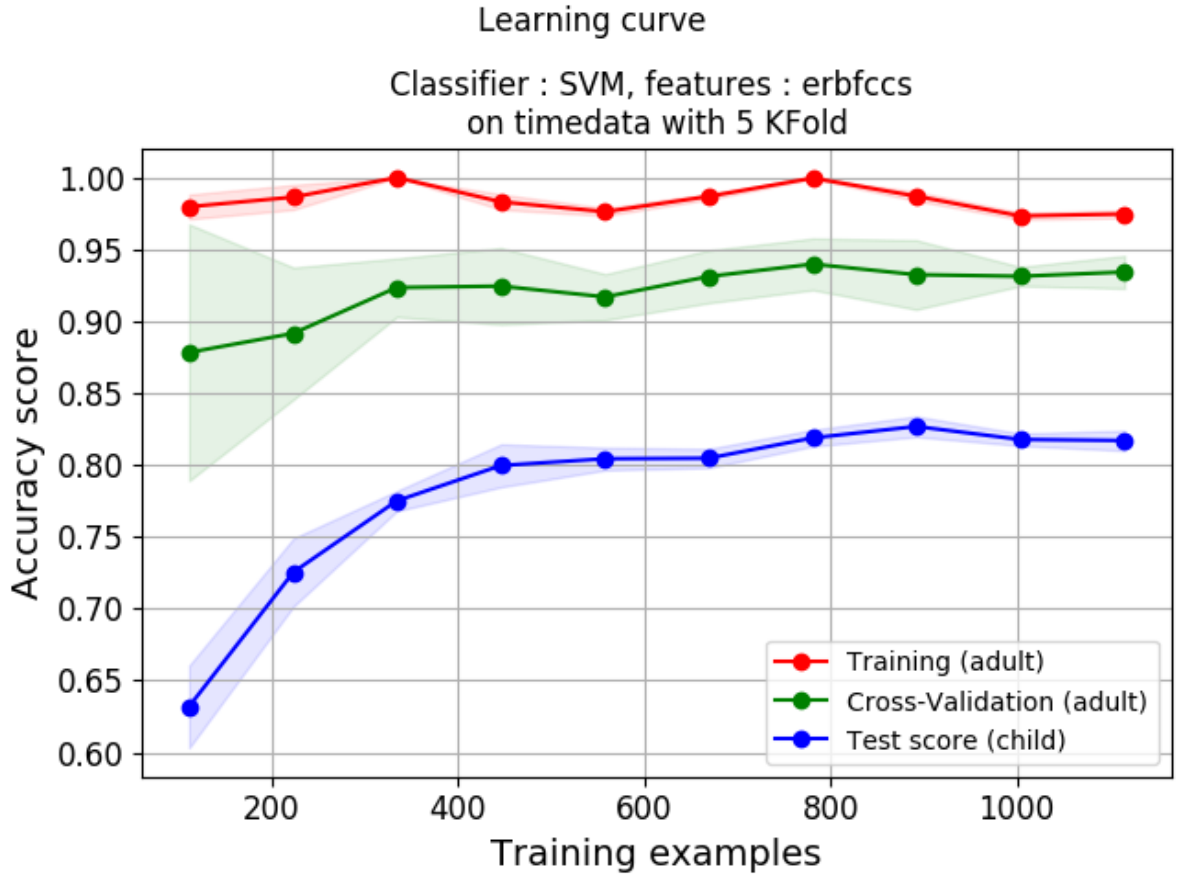




/perezy/phoneme

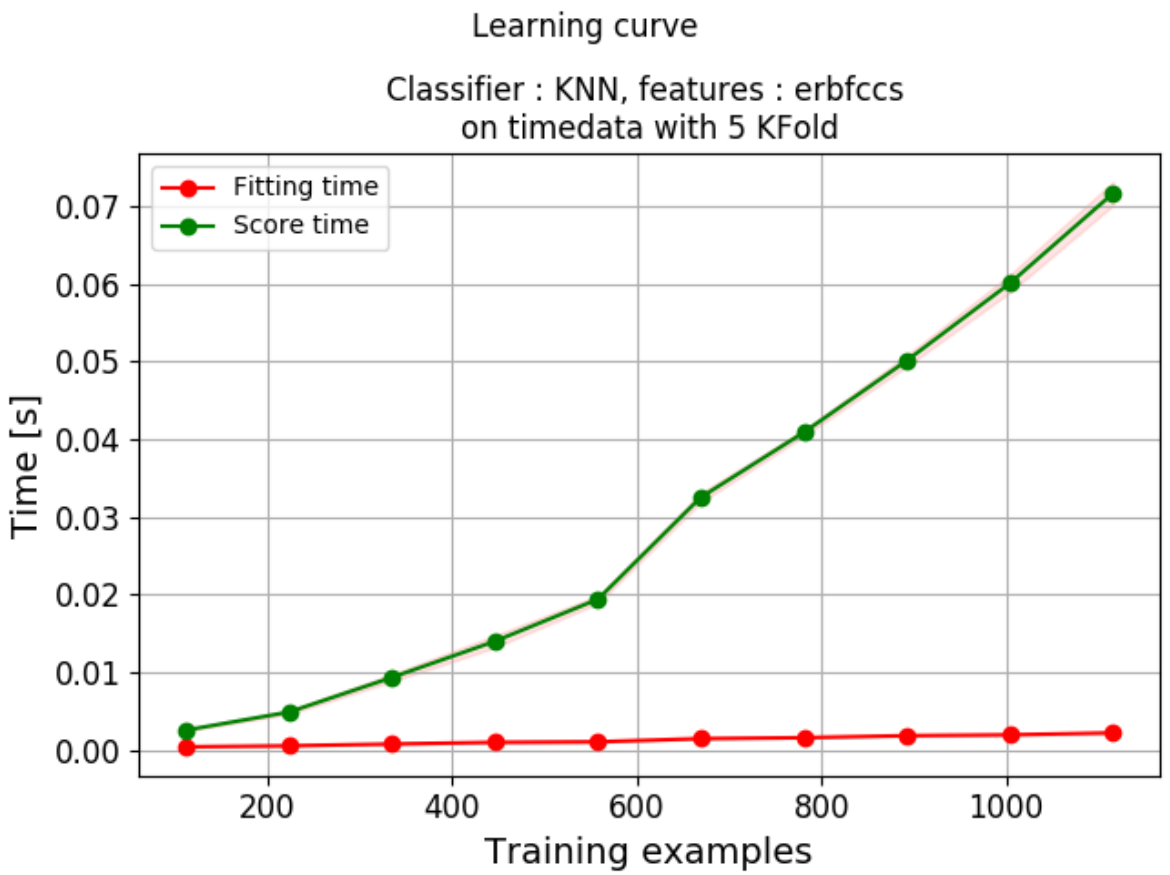
Précision : KNN ERBFCC vs SVM ERBFCC

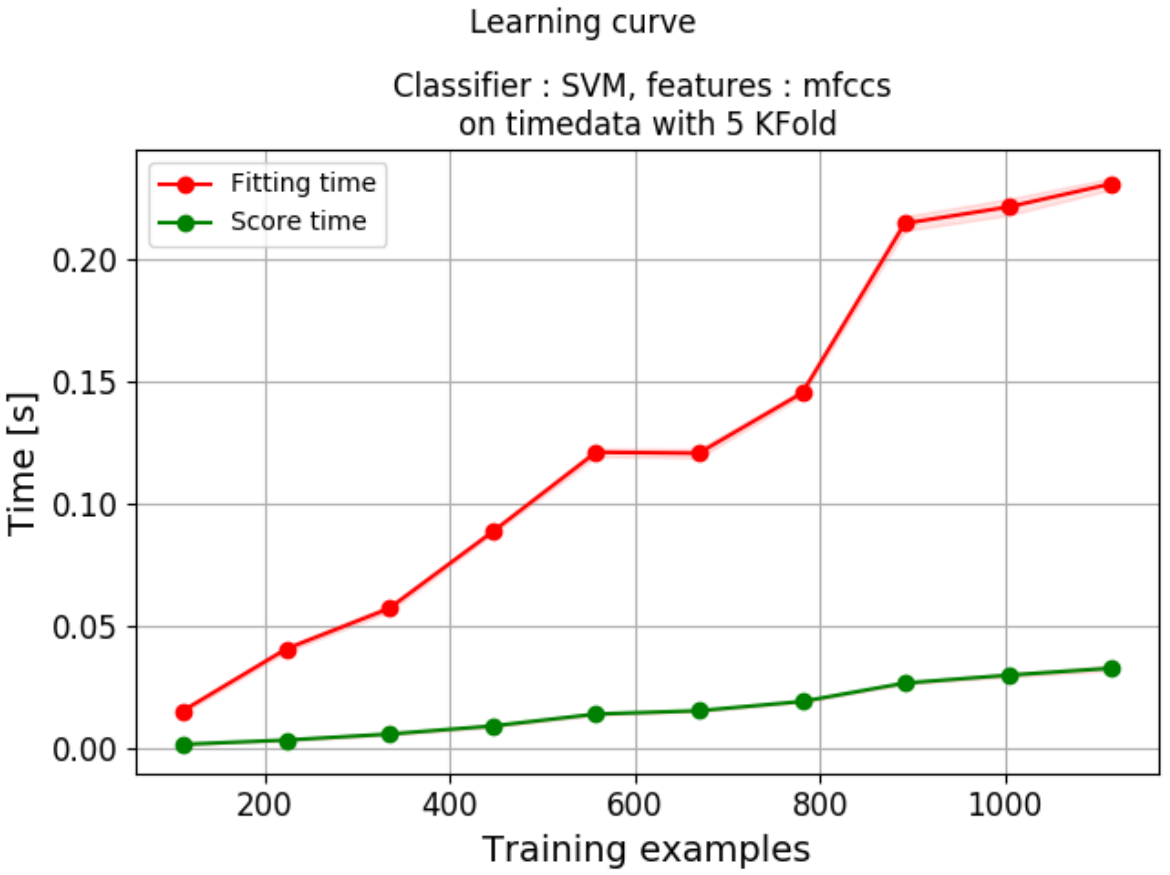




/perezy/phoneme

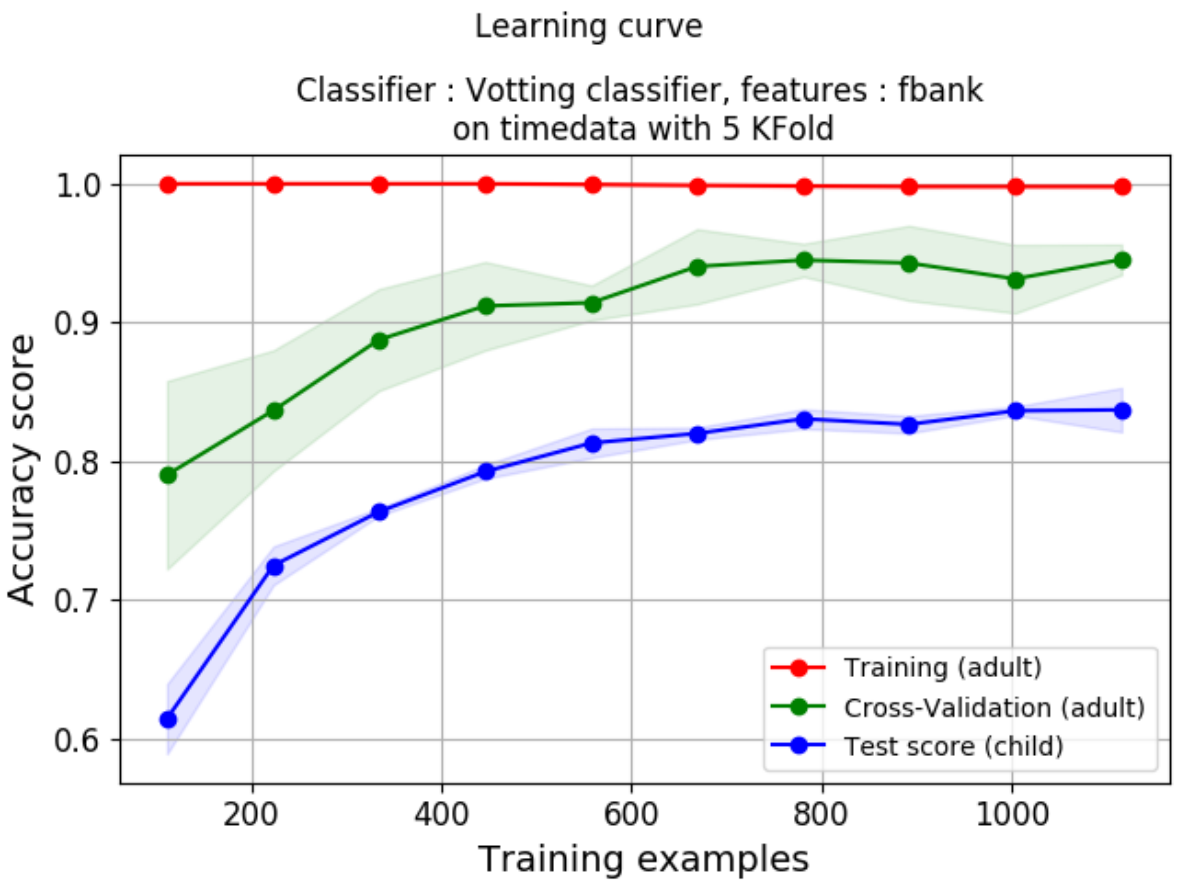
Time : KNN + ERBFCC vs SVM ERBFCC

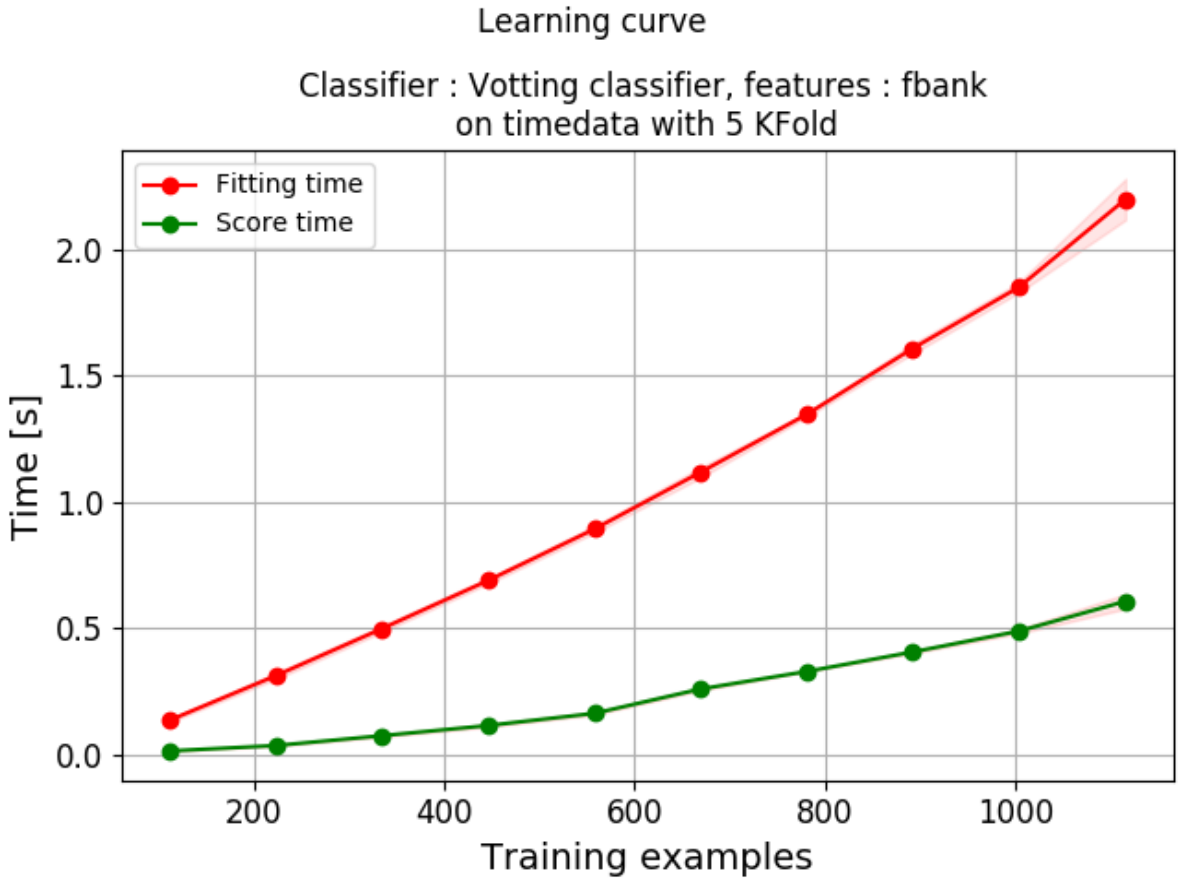




/perezy/phoneme

Combinaison de Linear SVM + RBF SVM + KNN + Bag. RBF SVM + Bag. KNN



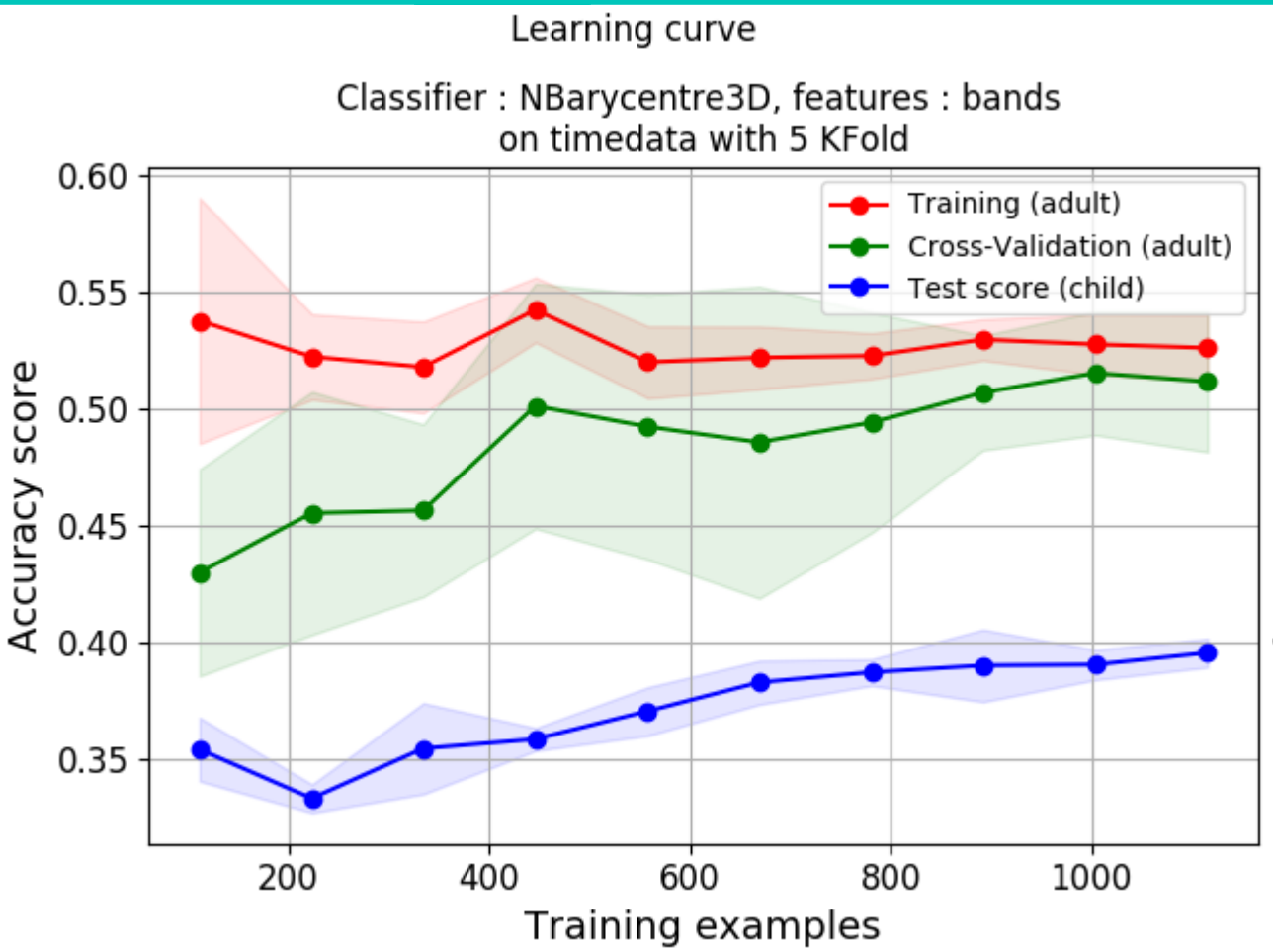


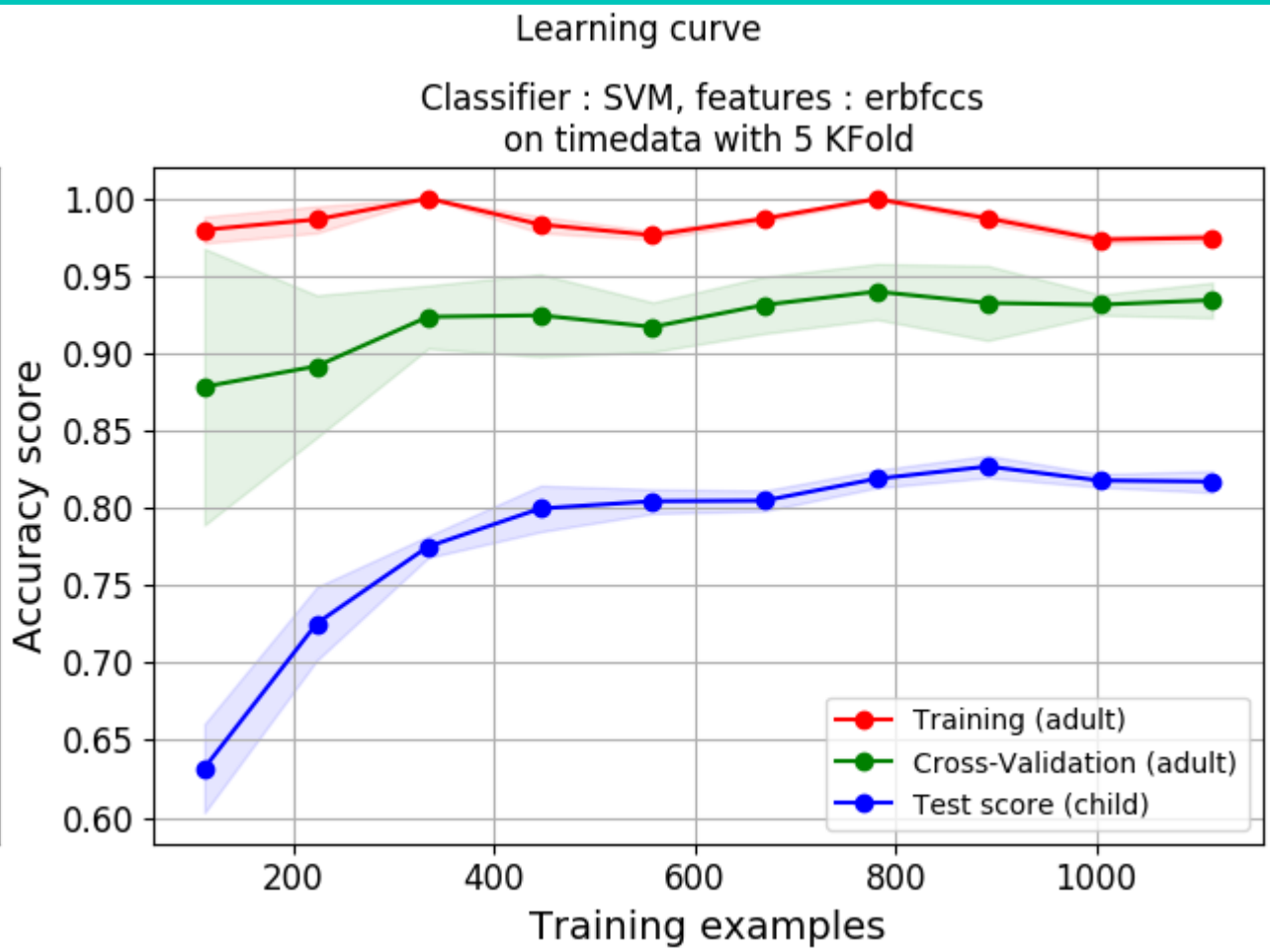
/perezy/phoneme

Résultats - Conclusion

- Algorithme de classification : SVM
- Caractéristiques acoustiques : ERB-FCC
- Des phonèmes d’adultes sont suffisants pour classifier des phonèmes d’enfants
- Limiter la taille de l’ensemble d’entrainement
- Stagne 82 % de reconnaissance

De 40 % à 82 % de reconnaissance





/perezy/phoneme