Projet ESTER

Technolangues

Transcription automatique de grands corpus radiophoniques

Catégorisations phonétiques par
- système supervisé : MLP
- non supervisé : ART

Hervé Glotin
glotin@univ-tln.fr
SIS / Univ. Toulon-Var

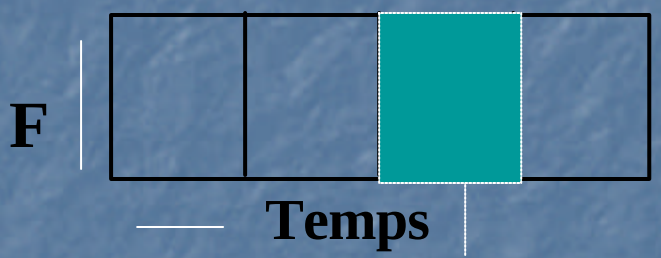
1/ Définition des traits acoustiques

160 coefficients pour 7 trames :

Filtrage	Décimation à 8KHz, filtrage [300 3500] Hz (tout le système est conçu pour la bande passante téléphonique)
Fenêtres	25ms, shift 10ms
Coeff	13 Coefficients log rasta + **indice de voisement**
Contexte	Contexte de 3 fenêtres gauches, et 3 droites.
Delta 1 et 2	Coefficients dynamiques : Dérivées premières et secondes des coefficients.
Compression	Moyennes en temps et fréquence pour les fenêtres les plus excentrées pour réduire le nombre de coefficients.

Le voisement comme indice de qualité de la parole

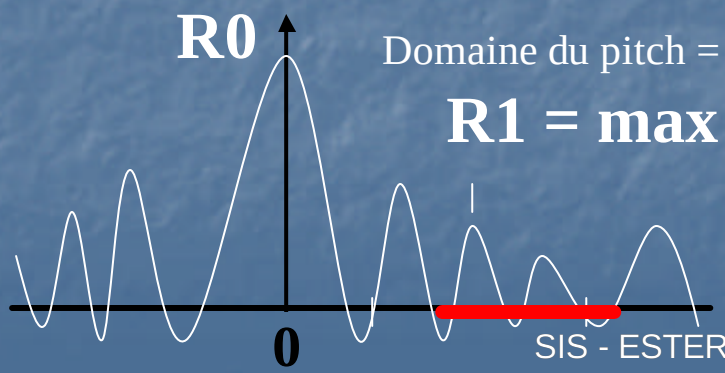
Analyses par trames de 25 ms à 100 ms



DEMODULATION

Indice de voisement :

$$R = \frac{R1}{R0}$$

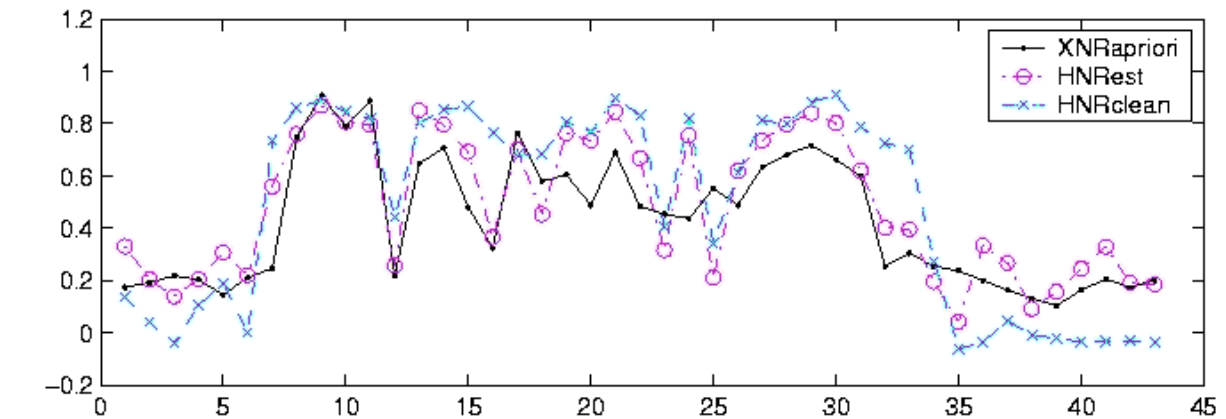


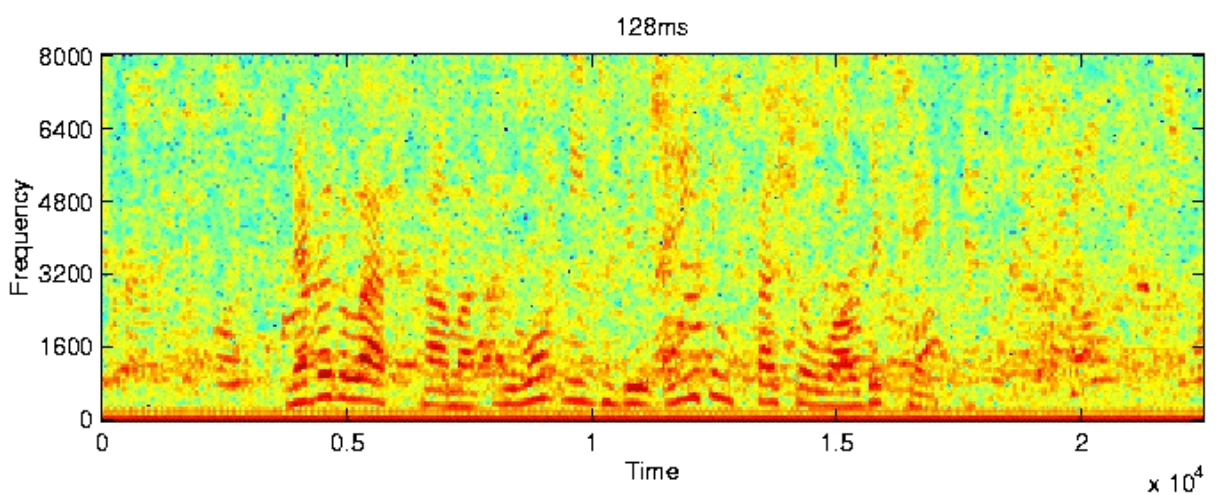
Domaine du pitch = [1/350 , 1/90]

R1 = max

[d'après Glotin PHD 2001]

Corrélation entre RSB (XNR) et R (HNR) (=0.84) avec du bruit de cafétéria 8.5 dB RSB



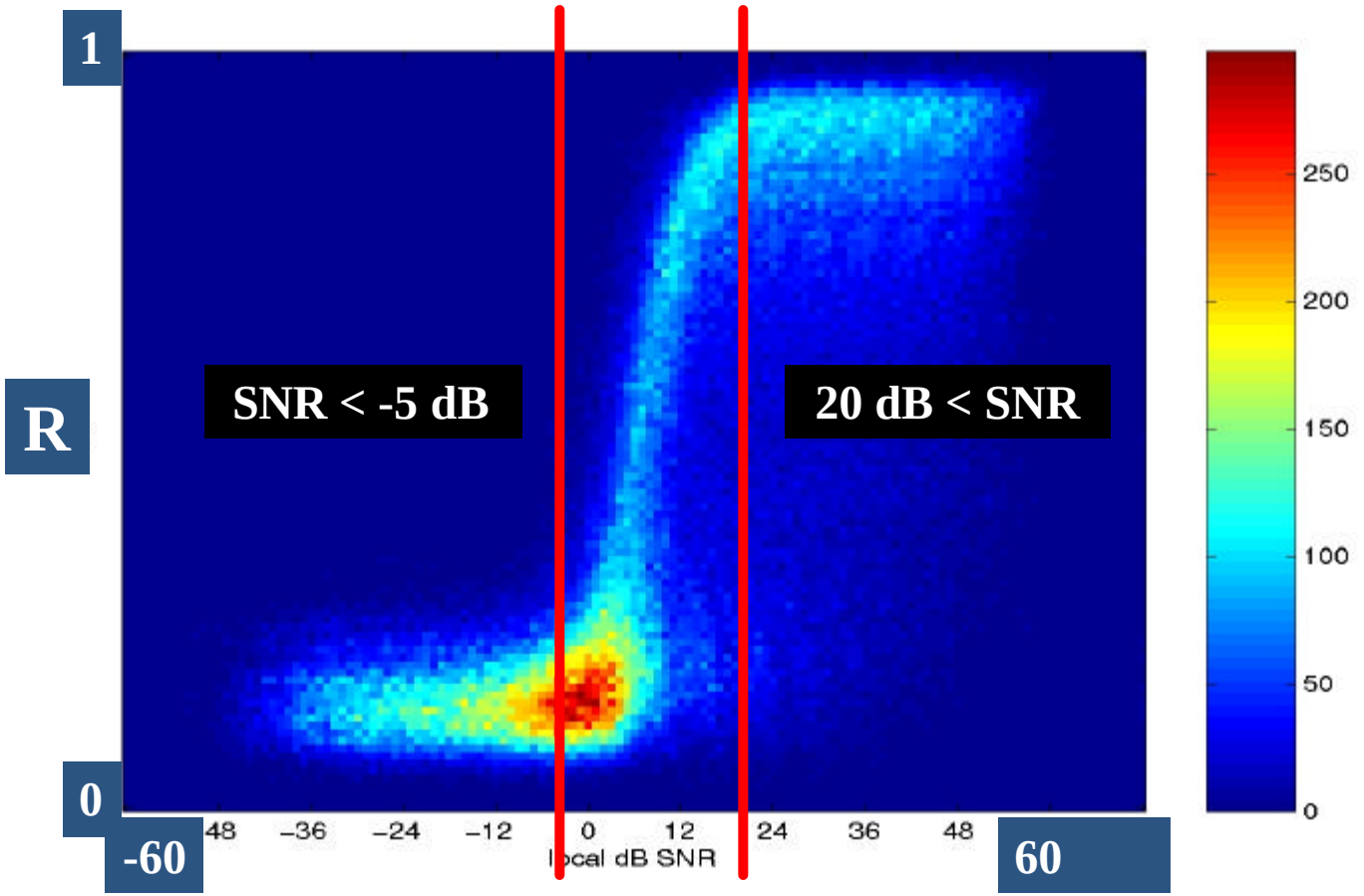


[d'après Glotin et al ICASSP 2001]

Corrélation entre R et RSB local

(sur 1000 paves = [0 600] Hz x 128ms)

[d'après PHD Glotin 2001]



2/ Définitions du Perceptron Multicouche (MLP)

- 41 sorties (catégories de la segmentation IRISA & ENST)
- 2000 neurones dans la couche cachée
- 400 000 paramètres : (41 + 160) * 2000
- Codes issus en partie de la toolbox IDIAP, modifiés au SIS.

Les 41 catégories phonétiques utilisées, code IRISA & ENST

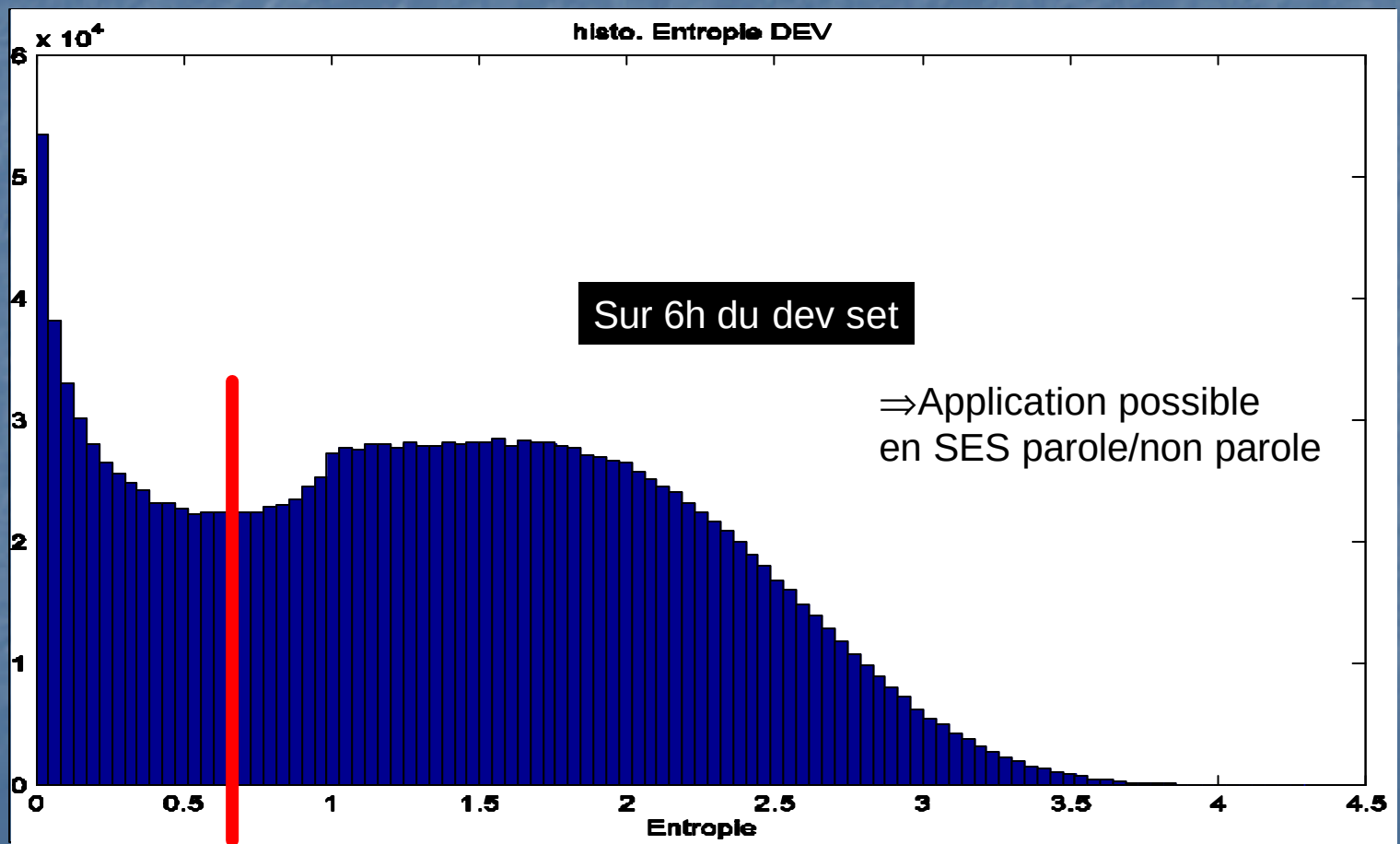
- # ;; short pause
- ## ;; long pause or silence
- 2 ;; deux=/d2/
- 9 ;; neuf=/n9f/
- </s> ;; sentence start silence
- <s> ;; sentence end silence
- @ ;; justement=/Zyst@ma~/
- E ;; seize=/sEz/
- H ;; juin=/ZHU~/
- J ;; oignon=/oJo~/
- N ;; camping=/ka~piN/
- O ;; comme=/kOm/
- R ;; rond=/Ro~/
- S ;; champ=/Sa~/
- U~ ;; vin=/vU~/, brin=/U~/
- Z ;; gens=/Za~/
- a ;; pâte=/pat/, patte=/pat/
- a~ ;; vent=/va~/
- b ;; bon=/bo~/
- cl ;; unvoiced closure
- d ;; dans=/da~/
- e ;; ses=/se/
- f ;; femme=/fam/
- g ;; gant=/ga~/
- i ;; si=/si/
- j ;; pierre=/pjER/
- k ;; quand=/ka~/
- l ;; long=/lo~/
- m ;; mont=/mo~/
- n ;; nom=/no~/
- o ;; gros=/gRo/
- o~ ;; bon=/bo~/
- p ;; pont=/po~/
- s ;; sans=/sa~/
- t ;; temps=/ta~/
- u ;; doux=/du/
- v ;; vent=/va~/
- vcl ;; voiced closure
- w ;; coin=/kwU~/
- y ;; du=/dy/
- z ;; zone=/zon/

Performances du MLP
Entraîné seulement sur 6h du train set
Bande passante [300 3500 Hz]

% Taux de reco Phonétique	Nom de la source (non prise en train)
0.40	19981207_0800_0900_inter_fm_dga
0.43	20000414_0930_1030_rfi_fm_dga
0.61	20000414_0930_1030_rfi_fm_dga **

48 % de reco phonétique en moyenne sur ces tests

Histogrammes de l'entropie des sorties du MLP



Conclusions sur MLP :

48 % de reco phonétique sur DEV

- Inconvénients :
- Dépendant de la qualité de la segmentation
- Bande passante actuellement réduite à celle du téléphone, la version suivante prendra tout le spectre.

- Avantages :
- Rapide :
  - 1/2 fois temps réel en training
  - 1/10 fois temps réel en forward

- Mesures sur distribution des sorties => SES Parole / non parole

3/ Réseau non supervisé
« Adaptive Resonance Theory» Carpenter & Grossberg 1991

Il existe plusieurs variantes mais toutes ont cette partie commune :

- une couche de cellule F1

- une couche de cellule F2 formées de N nœuds pour N catégories au maximum

- deux modules de contrôle :

Apprentissage : définit le nombre d'itérations dans le réseau

pour chaque exemple présenté.

Vigilance : = 0 => aucune catégorie n'est crée

= 1 => autant de catégories que d'exemples

Disposition des différentes couches d'un réseau ART-2

Chaque couche joue un rôle différent au sein du réseau :

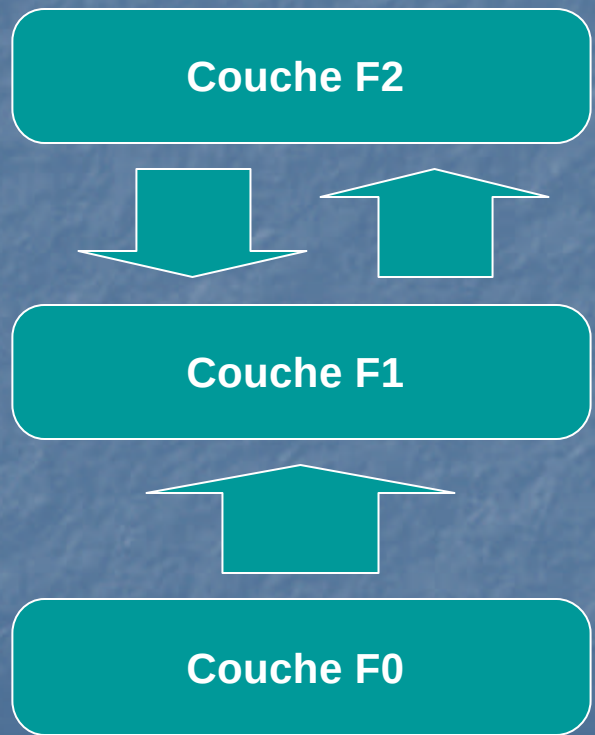
- ✓ **un réservoir de neurones destinés à mémoriser un nouveau modèle,** — **Couche F2**
- ✓ **une zone de traitement des données venant de la couche inférieure,** — **Couche F1**
- ✓ **une zone de filtrage et de réception des données.** — **Couche F0**

Connexion des différentes couches d'un réseau ART-2

Entre ces couches il y a deux espaces de connexion particuliers :

- ✓ **une connexion feed-forward : chaque neurone de la partie supérieure de F1 est connecté avec chaque neurone de la partie inférieure de F2 (et inversement),**

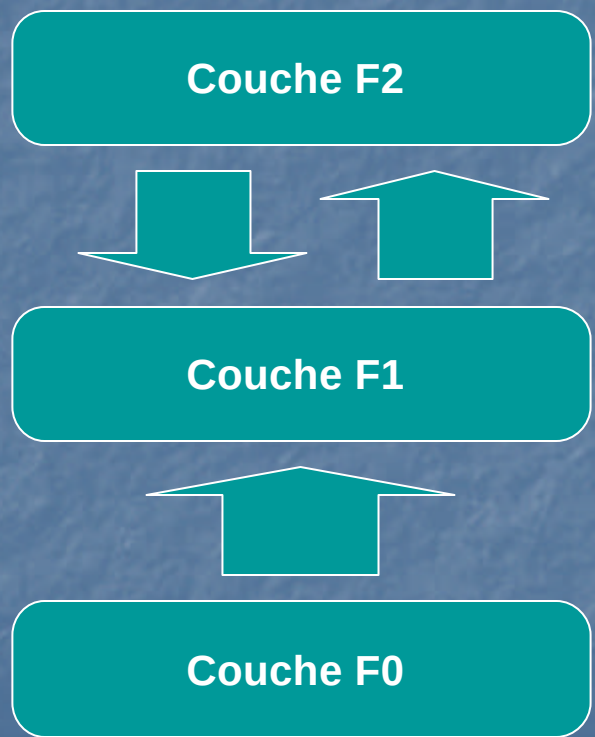
- ✓ **une connexion montante des neurones de F0 vers les neurones de la partie inférieure de F1.**



Principe de fonctionnement d'un réseau ART-2

Un réseau ART-2 fonctionne en 5 phases essentielles :

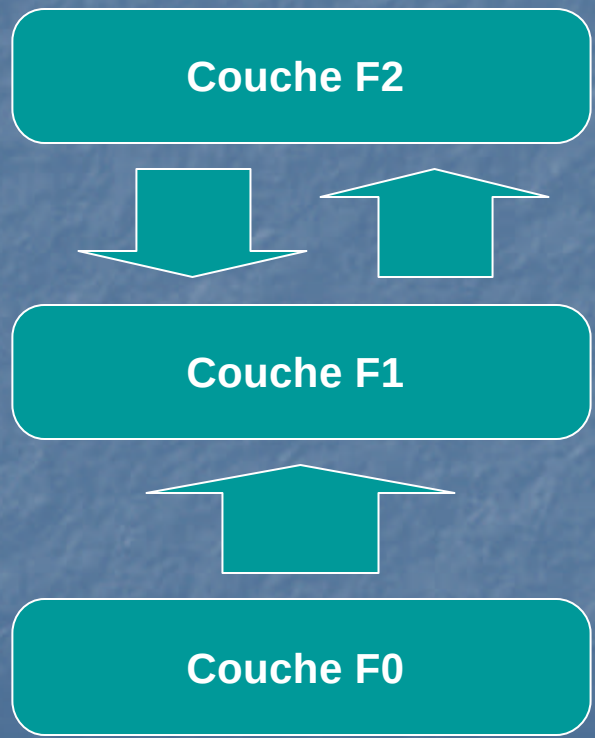
- ✓ **<u>présentation d'un vecteur :</u> la couche F0 reçoit la donnée à traiter et la transmet à la couche F1,**
- ✓ **<u>traitement du vecteur :</u> la couche F1 normalise la donnée et supprime le bruit,**
- ✓ **<u>compétition :</u> F1 organise une compétition par mis les neurones de F2. Le gagnant est celui qui entre en résonance avec la donnée traitée,**
- ✓ **<u>contrôle :</u> si le gagnant est trop éloigné de la donnée on recrute un nouveau neurone dans F2 pour la mémoriser,**
- ✓ **<u>mise à jour :</u> si le gagnant est assez proche de la donnée il sera modifié pour s'en rapprocher.**



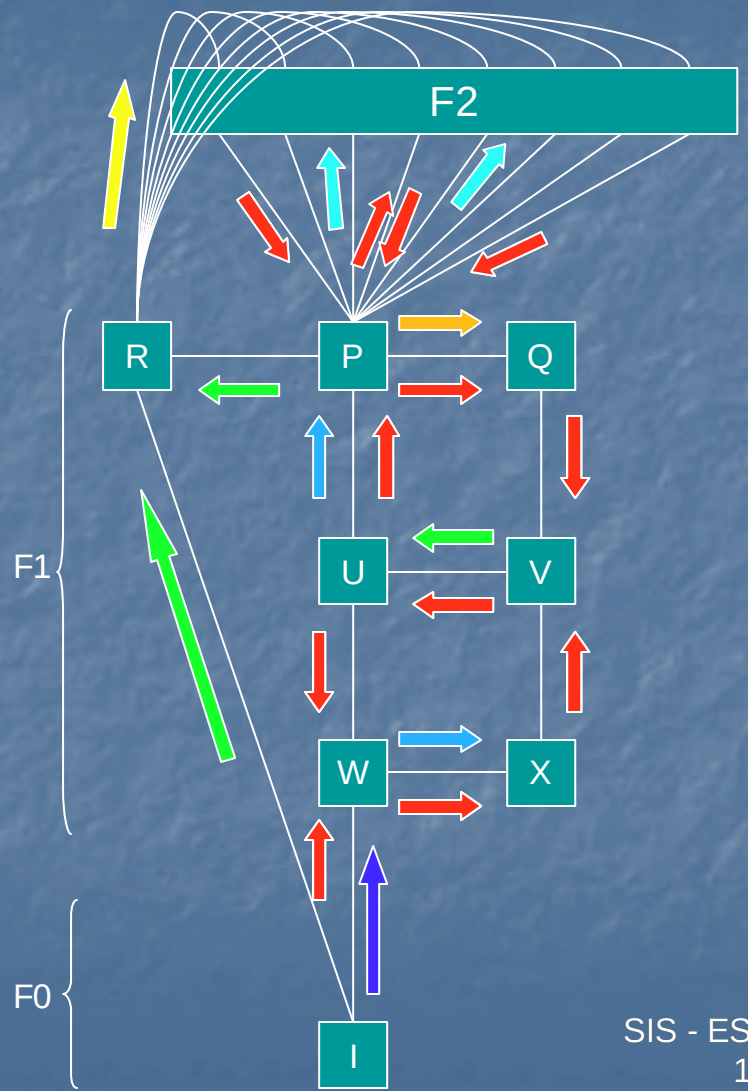
Choix du rôle des couches et connexions du réseau

Chaque couche pouvant hériter de certaines tâches d'une couche adjacente, il faut distribuer ces tâches. Pour la suite nous choisirons :

- ✓ **Couche F2 : réserve de neurones « challengers » et de neurones inhibiteurs.**
- ✓ **Connexion F2/F1 : mémorisation des connectivités fortes/faibles entre neurones.**
- ✓ **Couche F1 : 7 types de neurones pour filtrer le bruit, normaliser et organiser les compétitions.**
- ✓ **Couche F0 : transmet les vecteurs présentés au réseau.**

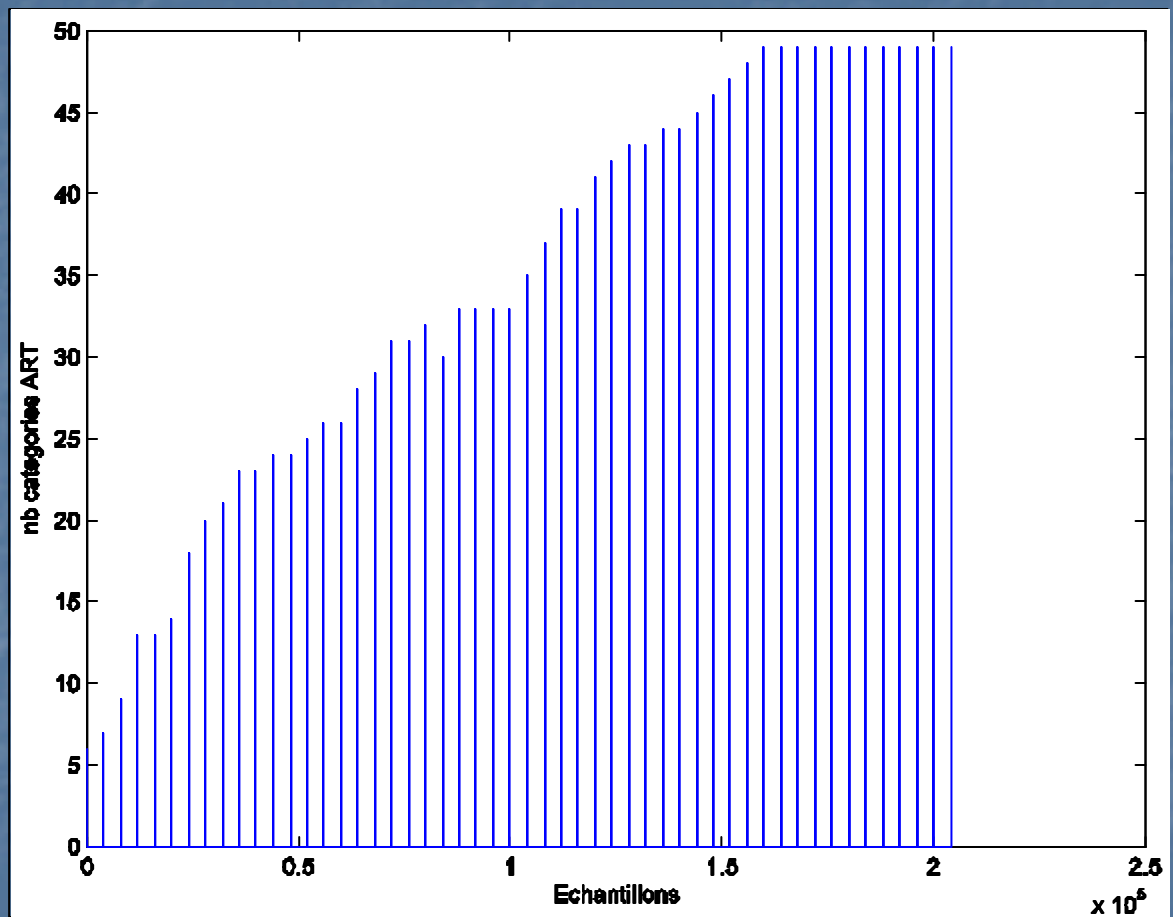


Traitement des données



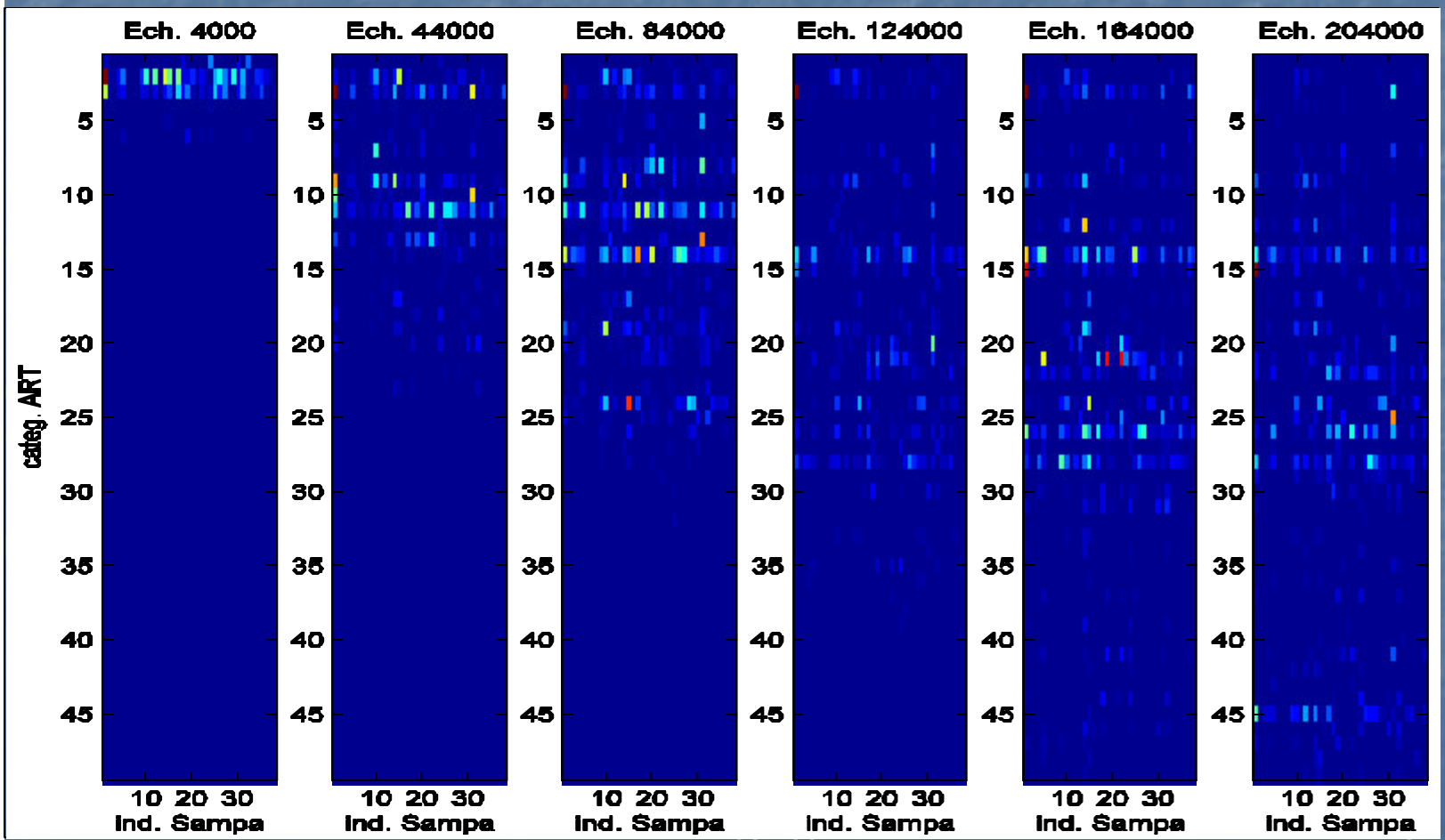
- On filtre une première fois le bruit.
- On répand la donnée dans le réseau tout en filtrant d'éventuels bruit internes.
- Synthèse des données réseau et entré. Initialisation des neurones de F2.
- Détermination d'un neurone résonant avec la donnée.
- Mécanisme de disqualification de la compétition d'un neurone résonant mais trop éloigné de la donnée.
- Mécanisme de stabilisation du réseau pour le préparer à une mise à jour.
- Mise à jour des poids ascendants et descendants ainsi que de la couche F1.

Evolution sur 1 h du train set du nombre de catégories

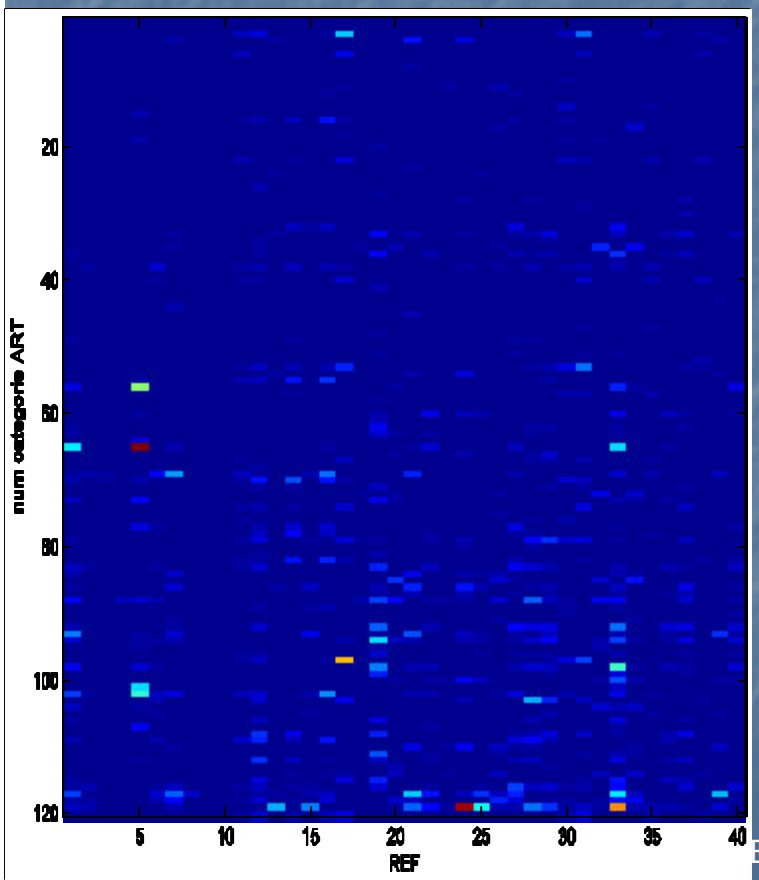


Evolution des matrices de confusions par époques d'une min

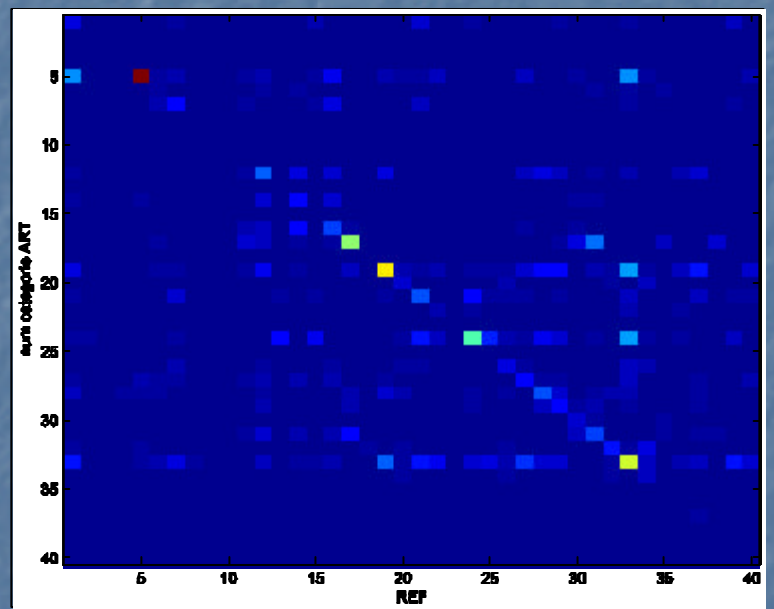
catégories ART (au max. 50) / phon. IRISA&ENST



Appareillement phonétique / Unification des catégories ART par maximisation de la diagonale de la matrice de confusion

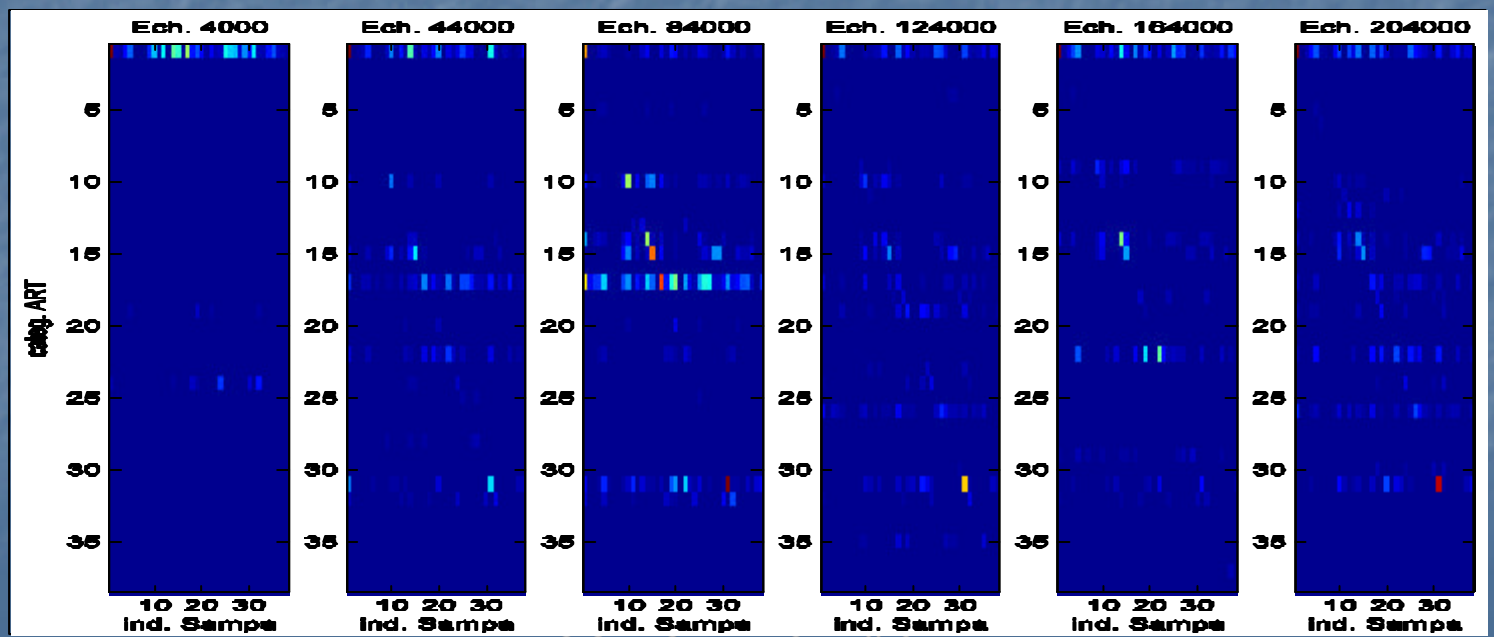


Après réduction :

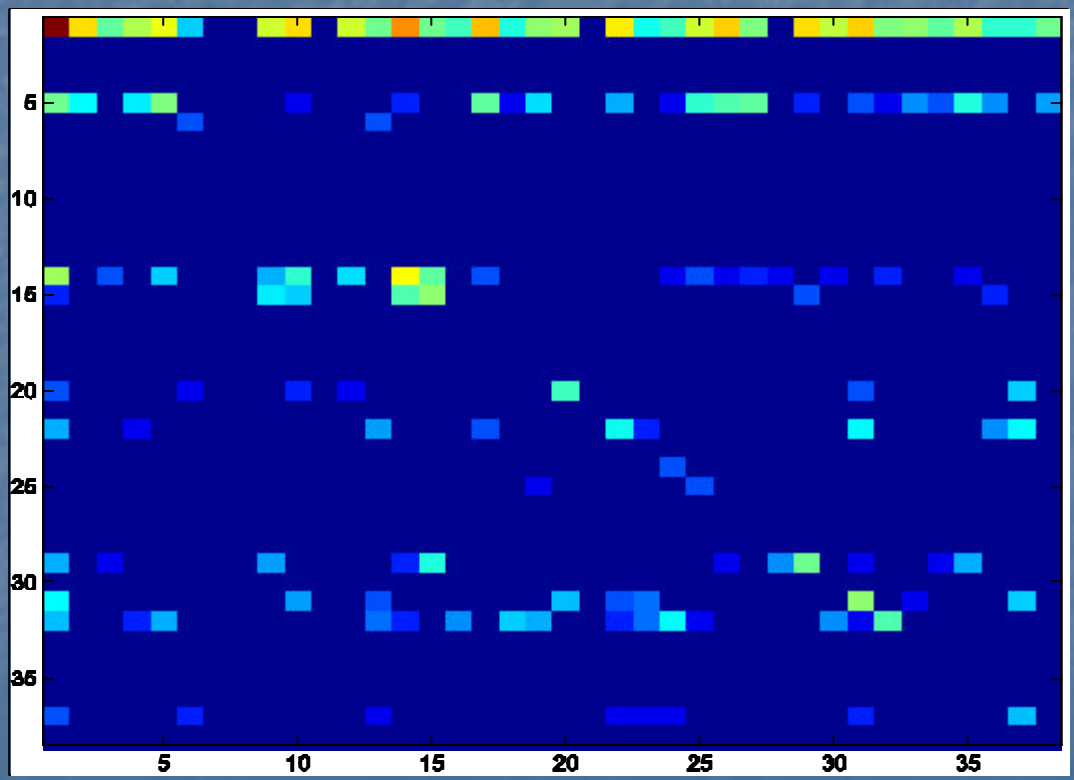


⇒ Après présentation de 1h de parole
⇒ 31 % de reco phonétique sur 1h de dev

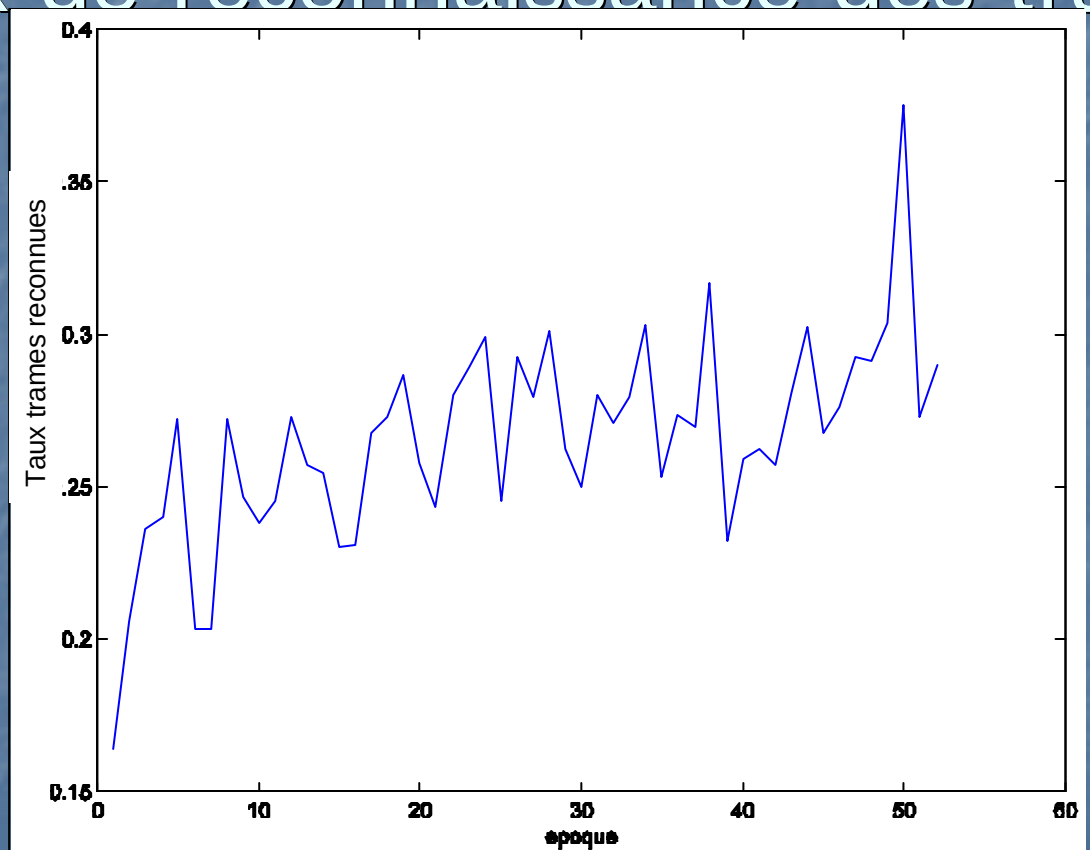
En phase de reconnaissance
la relation catégories / phonèmes est figée, mais pas en phase d'apprentissage... Evolution des matrices sur train set :



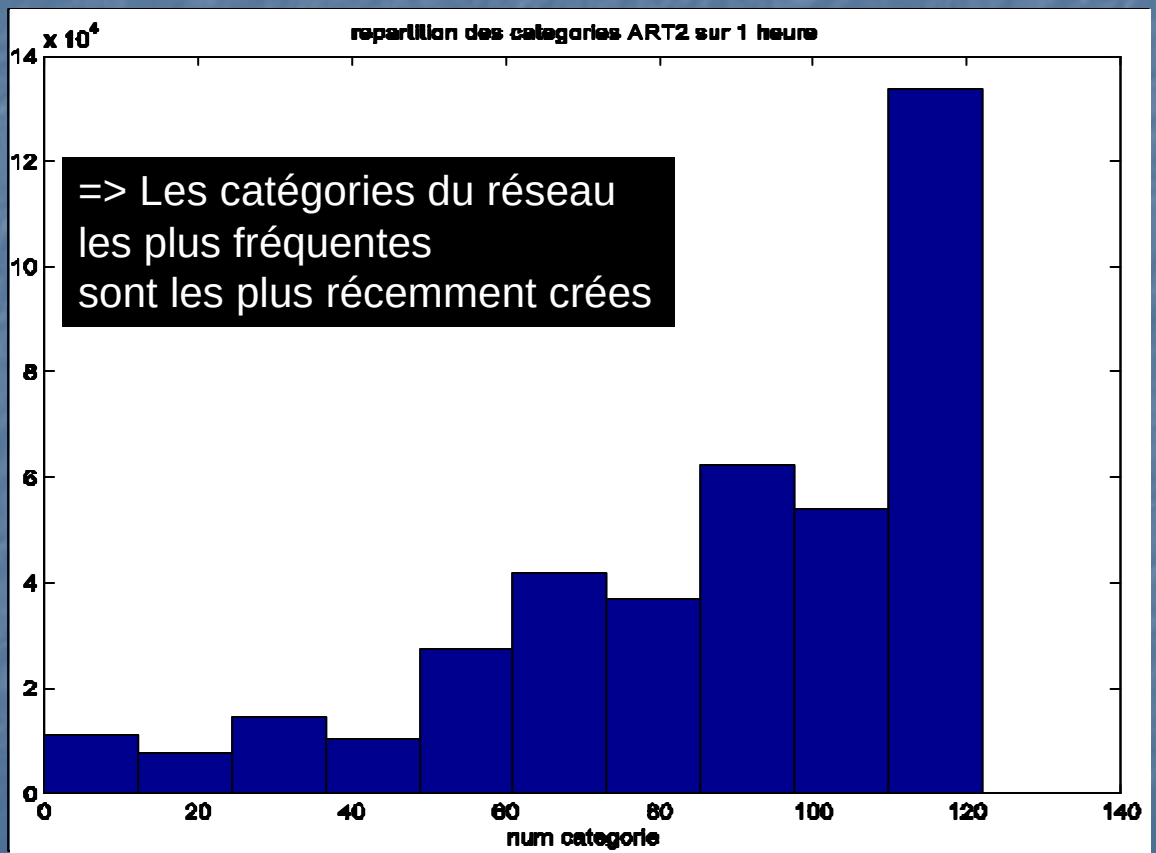
Matrice de confusion de l'époque 50 avec les catégories appareillées



Evolution sur 1h du train set du taux de reconnaissance des trames



Problème de stabilisation du réseau



Conclusions sur ART :

30 % de reco phonétiques,
sans superviseur
(aucun usage de segmentation)

- Inconvénients :
- Nécessité d'une étude de la stabilité des catégories en fonction des paramètres d'apprentissage et de vigilance.

- Avantages :
- Rapide : 1/10 fois temps réel
- Segmentation inutile, pourrait donc servir à segmenter...
- Variations sur les traits : pas de limite a priori car leur pertinence aux classes est gérée par ART.
- => Application à différentes échelles de temps (syllabique...), SES

- Possibilité après stabilisation d'induire la détection de nouveaux événements sonores correspondant à la naissance de nouvelles catégories.

Conclusion Générale

- Implémentation efficace des modèles
  - 1/10 * Temps réel forward MLP ou ART

- Différentes Qualités et différents défauts

entre MLP et ART...

⇒ La fusion pondérée des deux modèles pourrait être avantageuse ...