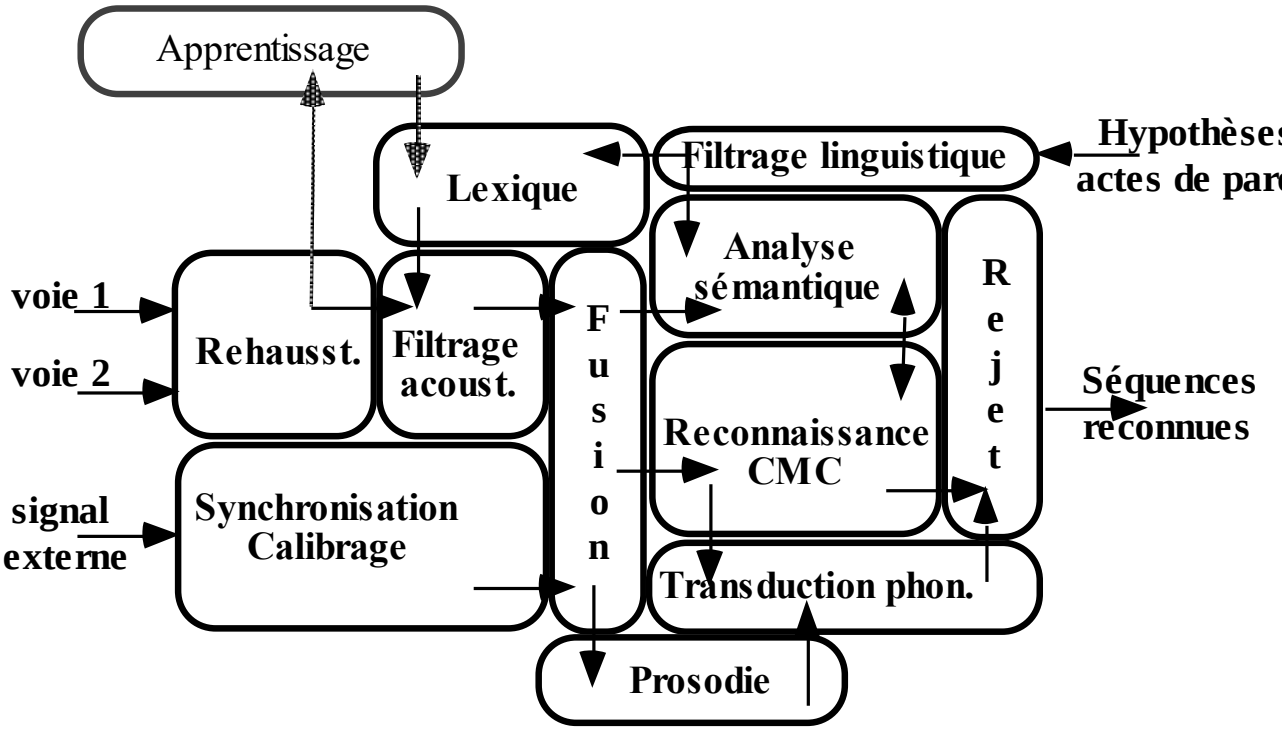
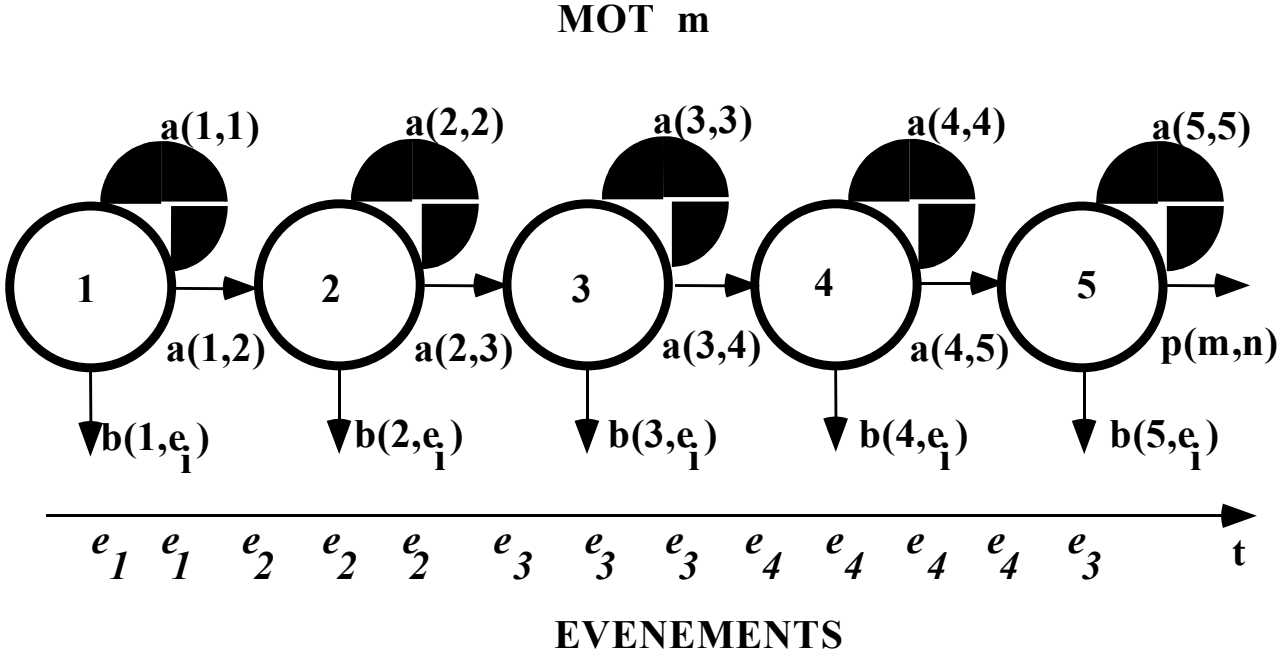
Architecture d'un système de compréhension de la parole

Jean Caelen



1. Reconnaissance par modèles de Markov (ou Réseau de neurones)
2. Filtrage acoustique (élimination des sons non vocaux)
3. Filtrage linguistique (élimination des séquences invalides)
4. Lexique (mots, modèles acoustiques)
5. Rehaussement (amélioration du rapport signal/bruit)
6. Rejet (élimination des solutions peu probables)
7. Analyse sémantique
8. Synchronisation, calibrage de sources
9. Fusion des informations issues de différents capteurs
10. Transduction phonétique (repérage de certains phonèmes)
11. Prosodie (contrôle des marques prosodiques)

1. La modélisation CMC



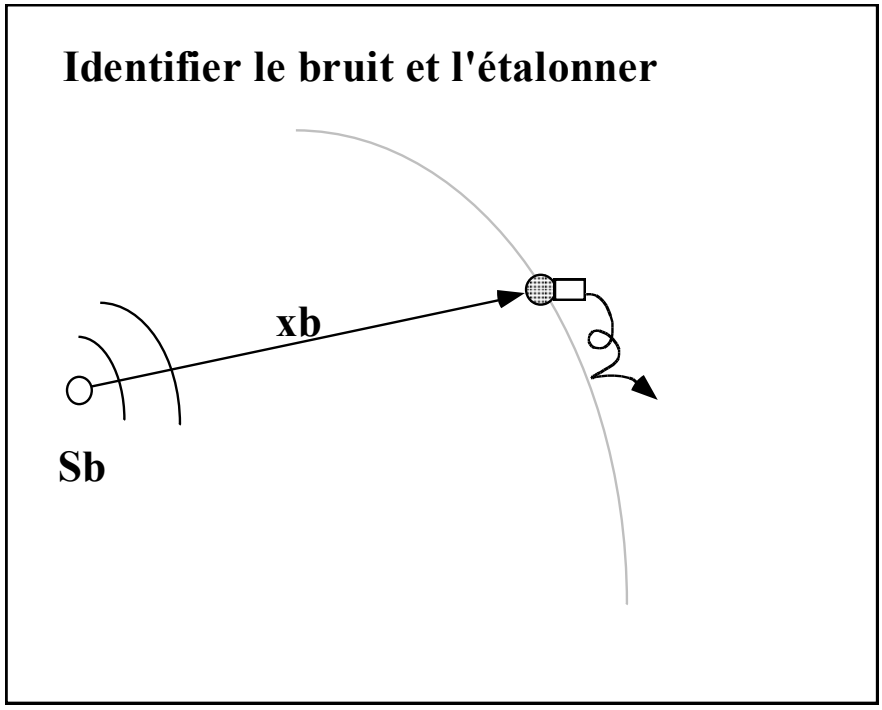
Entrée : signal s(t)

1. Paramètres des modèles : matrices a(.,.), b(.,.)
2. Paramètres du langage : matrices p(.,.), durée(.)
3. Conditions d'apprentissage
   Au moins 10 répétitions de chaque mot, 500 locuteurs
   => Vocabulaire de 1000 mots = 50 000 000 exemplaires

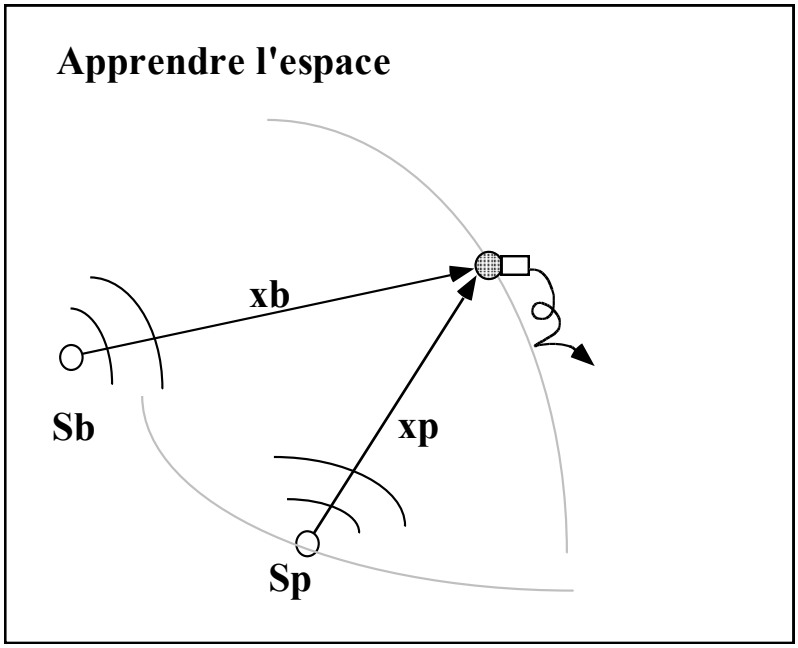
4. Apprentissage Baum-Welch
5. Reconnaissance Viterbi

Sorties : r meilleures séquences

2. Filtrage acoustique



Reconnaissance de bruits de l’environnement



Parcours selon les lignes du champ acoustique

3. Filtrage linguistique

Ce module reçoit les données d'un système de dialogue

M : voulez-vous dessiner un cercle ?
=> réponse attendue
confirmation .[complétive]
infirmation . [[justification]+[rectification]]
incidence
U : oui, rouge
=> vérification
linguistique + attentes

PREDICTION :

1. Focalisation du vocabulaire et de la syntaxe sur les actes attendus

VERIFICATION :

2. Elimination des séquences invalides sous contrôle de la syntaxe et de la sémantique

3. Elimination des interprétations non conformes aux attentes

4. Lexique

Mot	**Phonétique**	**Etiquette Morpho-syntaxique**	**μ-Sém**	**Traits**
Dessine	/desin/ /desinë/	V pers=1,3 Présent / Impératif	•Obj •Agt •[Loc]	/Evénement/ /Action/ /Graphique/

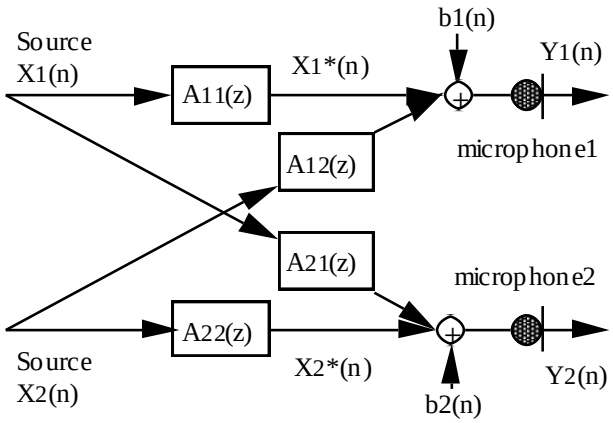
Composantes

- phonétique, description phonétique
- morpho-syntaxique, genre, nombre, personne, temps, etc.
- sémantique, actants + traits

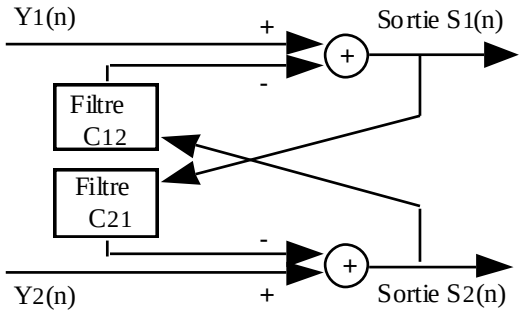
Cette base de données donne une représentation de tous les mots du langage considéré

5. Rehaussement de la parole par séparation de sources

1. Modèle de l'environnement sonore



2. Séparation des sources par un réseau monocouche



Convergence du réseau sur critère de minimisation des cumulants croisés d'ordre 4. Adaptation continue. 12 à 24 dB de gain.

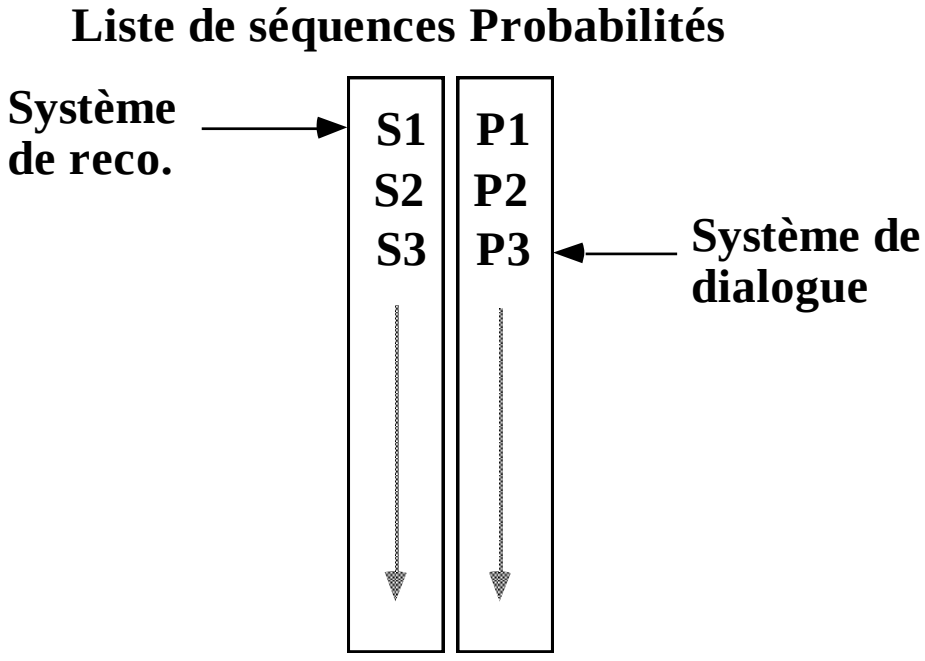
6. Rejet

1. Rejet amont

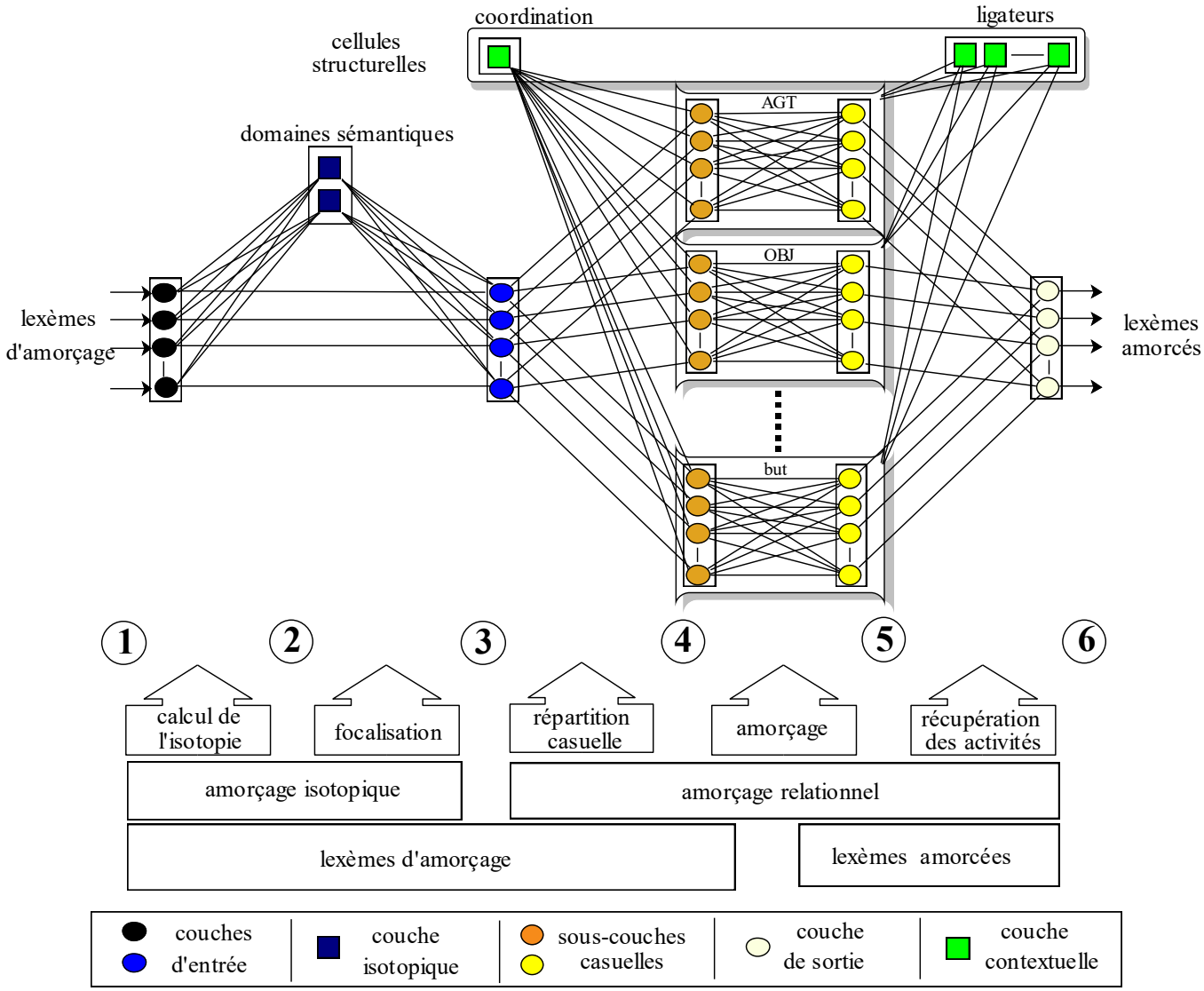
• sur r meilleures probabilités
• sur critères phonétiques
• sur critères prosodiques (positions des débuts et fins de mots, débuts et fins de syntagmes, etc.)

2. Rejet aval (par le système de dialogue)

=> Cela permet au système d'améliorer ses performances (on redemande l'apprentissage de mots qui sont systématiquement rejetés)



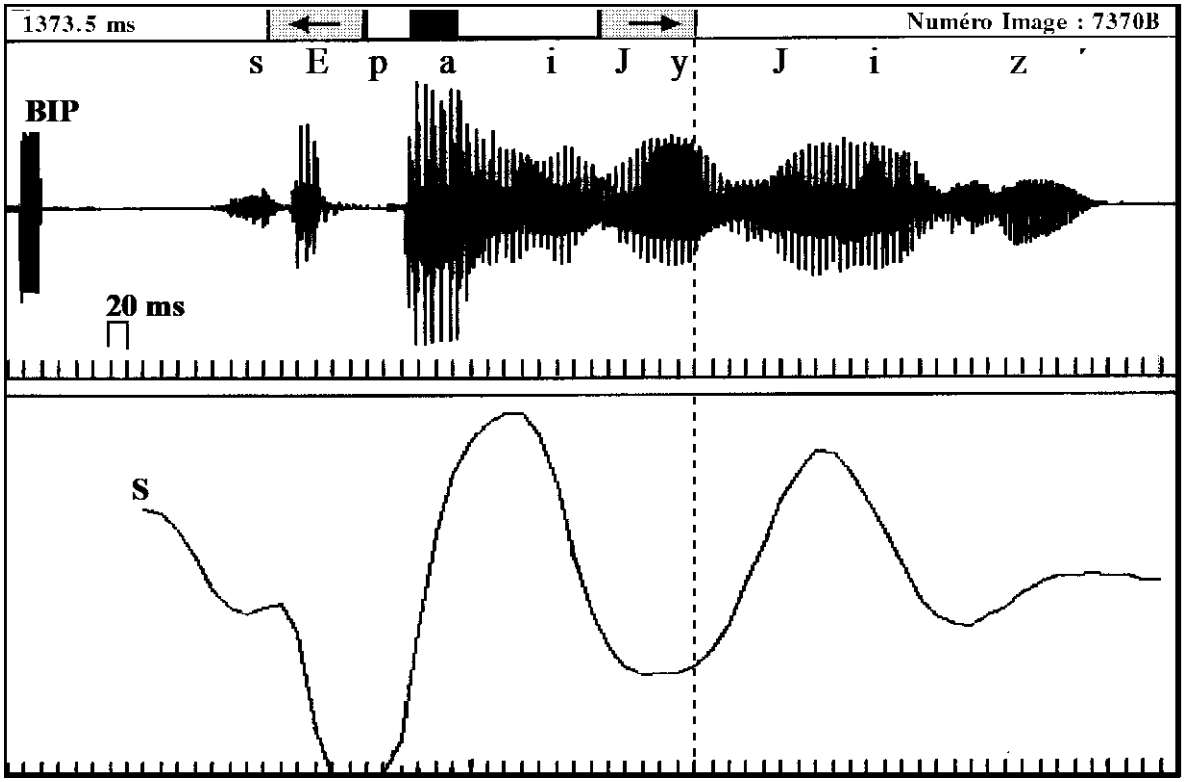
7. Analyse sémantique



Réseau de neurones multicouches pour l'amorçage sémantique et l'analyse micro-sémantique

8. Synchronisation de sources

Ce module permet de synchroniser des sources multimodales avec la source vocale (par exemple ici avec l'image de la bouche du sujet parlant)



Signal et aire aux lèvres S

1. Synchronisation

L'aire aux lèvres S est maximum dans la transition /a-i/ et non au cours de la réalisation du /a/

=> Variation contextuelle des paramètres visuels

=> Décalage par rapport à l'onde sonore

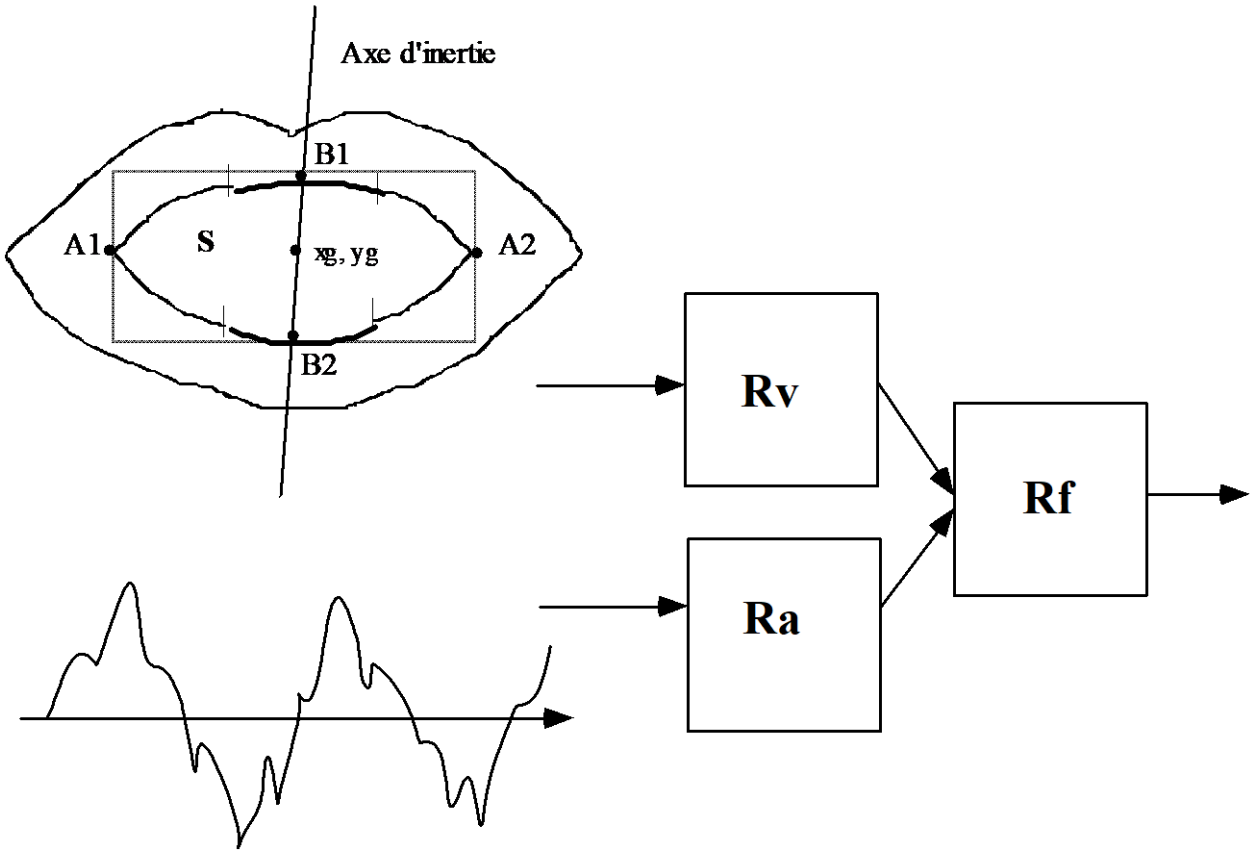
2. Calibrage

L'aire aux lèvres dépend

- de l'angle de prise de vue
- de la distance de la caméra

=> Calibrage des données relatives audio-vidéo

9. Fusion multi-capteurs

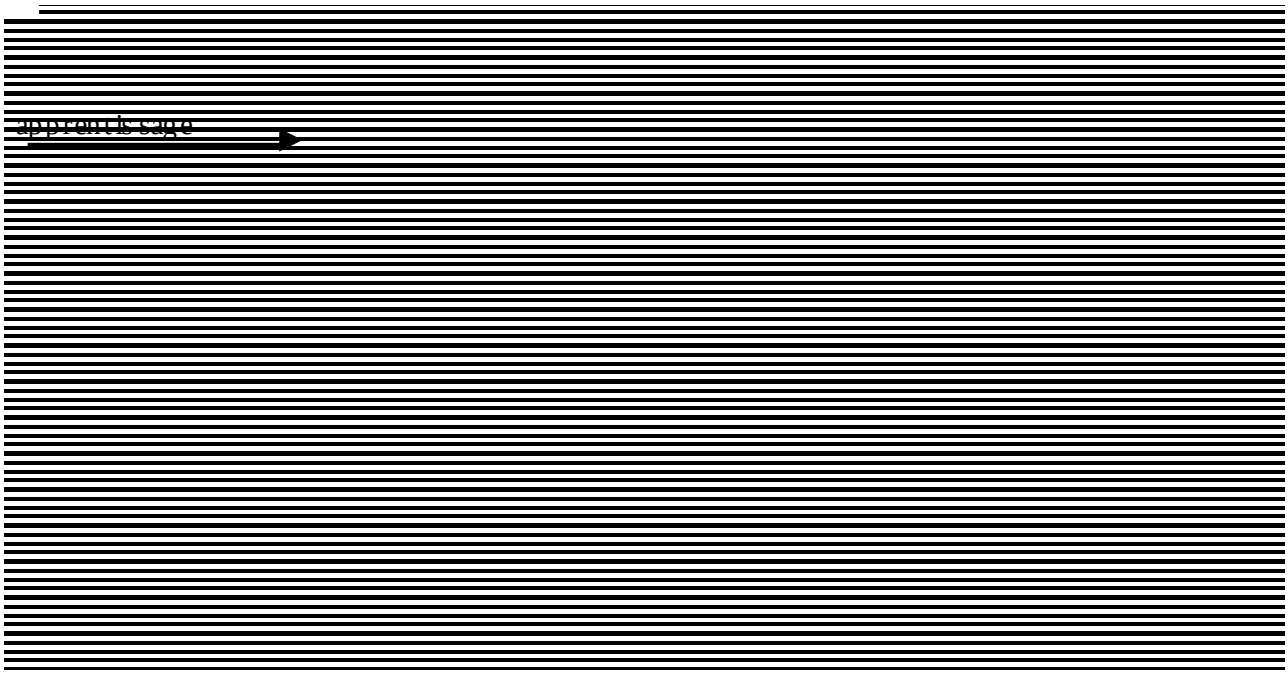


1. Rv : réseau de “mise en forme” des données visuelles
2. Ra : réseau de “mise en forme” des données audio
3. Rf : réseau de fusion

Résultat : gain de performance si l’une des deux sources est fortement bruitée

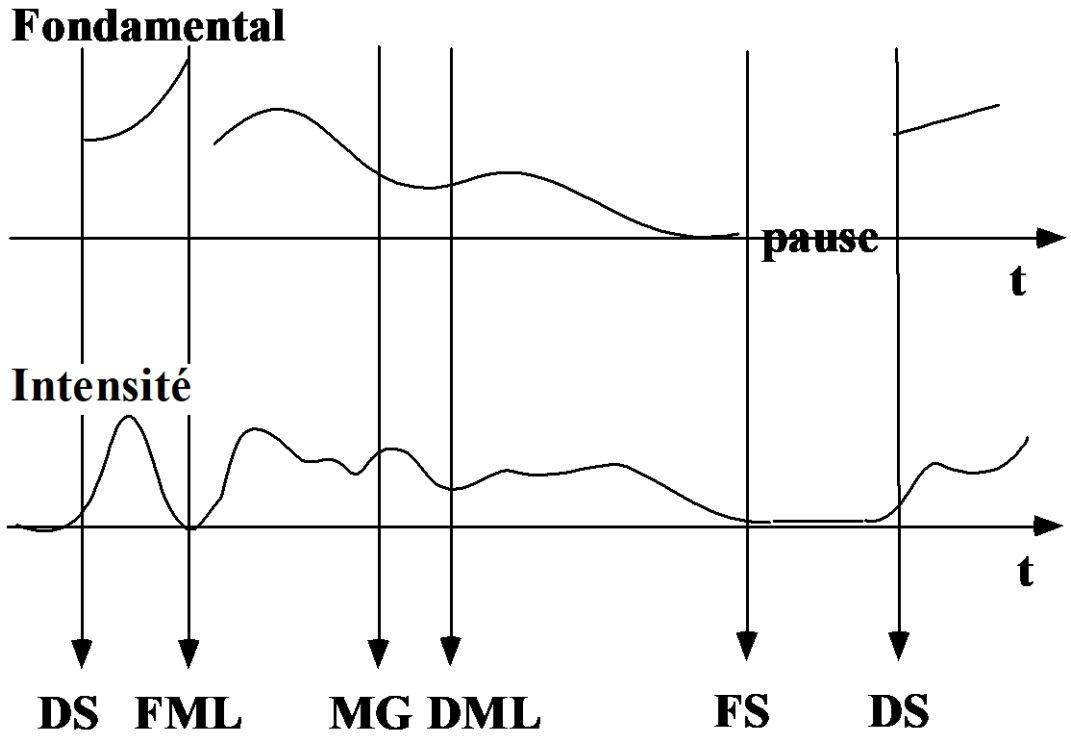
D’autres architectures ont donné de moins bons résultats.

10. Transduction phonétique et alignement



1. Transduction texte / chaînes phonétiques
2. Génération de modèles acoustiques
3. Recherche des modèles par DTW sur le signal
4. Alignement
5. Décision

11. Marques prosodiques



1. Repérages des débuts et fins de mots
2. Repérages des pauses

=> Démarcation de frontières de mots
=> Démarcation de frontières de syntagmes

3. Apprentissage automatique de règles
4. Appuyé sur une expertise