

# Transcription enrichie dans un monde multilingue et multimodal

---

Laurent BESACIER  
Soutenance de DHDR  
11 Janvier 2007.

# Vue d'ensemble

---

Partie I : Domaine, Problèmes, Méthodes

Partie II : Au-delà de la transcription :  
locuteurs, sons, etc.

Partie III : Reconnaissance automatique de la  
parole pour les langues peu dotées

Partie IV : Travaux en cours et axes de  
développement scientifique

---

# Partie I : Domaine, Problèmes, Méthodes

---

## **I.1 Domaine**

- Du signal au symbole
- La reconnaissance automatique de la parole
- Evolution du domaine

## **I.2 Problèmes**

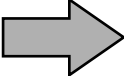
- Limites et problèmes ouverts
- Mes thèmes de recherche

## **I.3 Méthodes**

- Ligne de conduite
- Outils théoriques
- Outils logiciels

# Domaine

---

- Signal  Symboles
- Extraire automatiquement une information **symbolique** à partir d'un **signal**
- Signaux :
  - Enregistrements de parole pure
  - Mais pas uniquement : bande son d'une vidéo, vidéo
- Symboles :
  - Transcription orthographique
  - Mais pas uniquement : labels (locuteurs, sons), marques de ruptures (segmentation)
- Etiquetage automatique multi-niveaux de signaux

# Domaine

---

- Signaux :
  - **Parole**
- Symboles :
  - **Transcription orthographique**
  
- Technologie centrale : Reconnaissance Automatique de la Parole (RAP)
  
- ➔ Transcription de parole
  - Analyse automatique de documents
- ➔ Interaction Parlée
  - Interfaces homme/machine
  - Interfaces homme/homme médiatisées

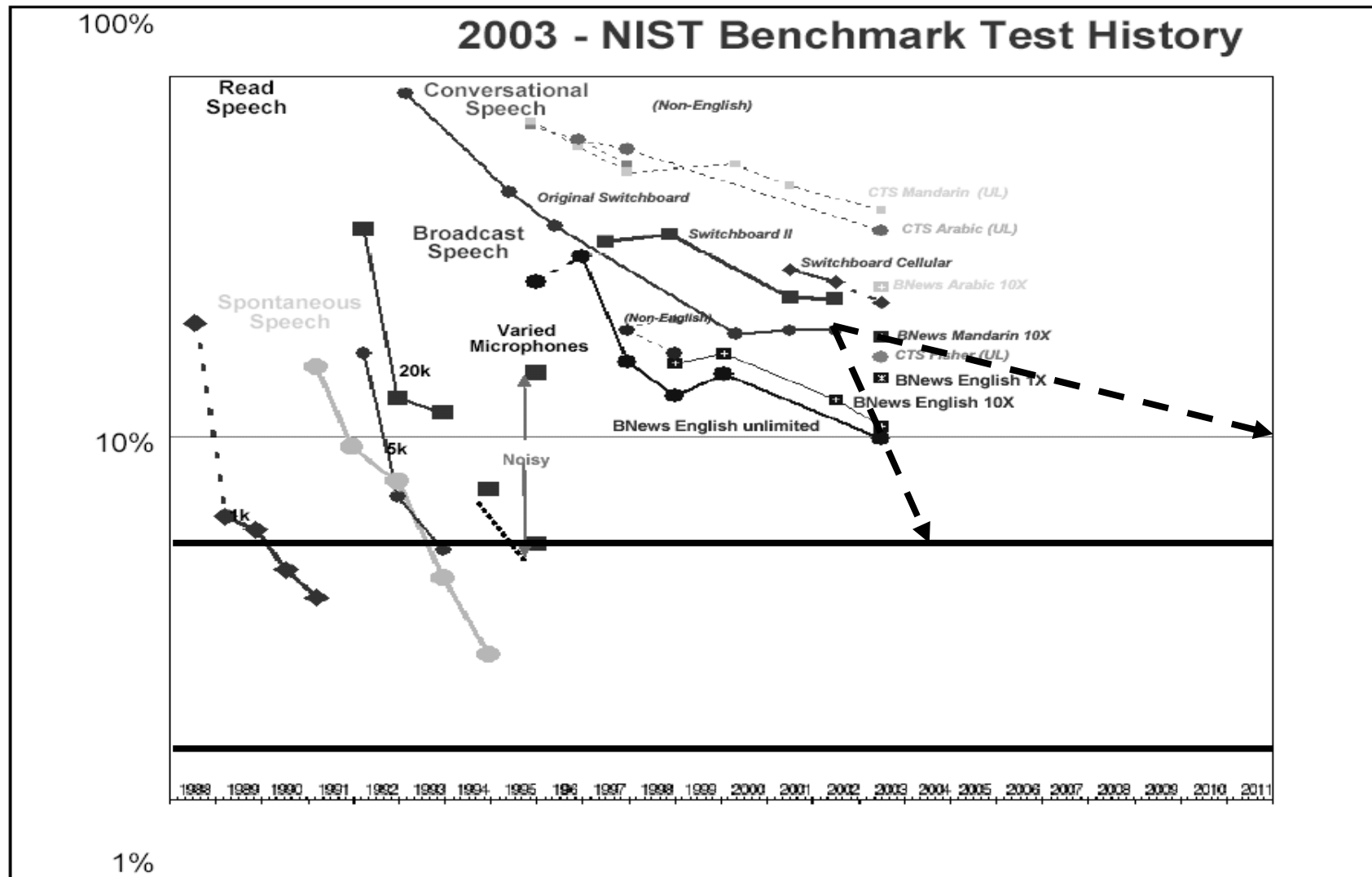
# La reconnaissance automatique de la parole

---

- Les meilleurs systèmes obtiennent\*
  - ~10-12% de taux d'erreur de mots pour l'anglais sur des documents de journaux télévisés ou des enregistrements du parlement européen
  - ~20% de taux d'erreur de mots pour l'anglais sur des conversations téléphoniques
- Progrès réguliers  
Voir évaluations DARPA & NIST ...

\*sources: projets TCSTAR & GALE

# La reconnaissance automatique de la parole



# La reconnaissance automatique de la parole

---

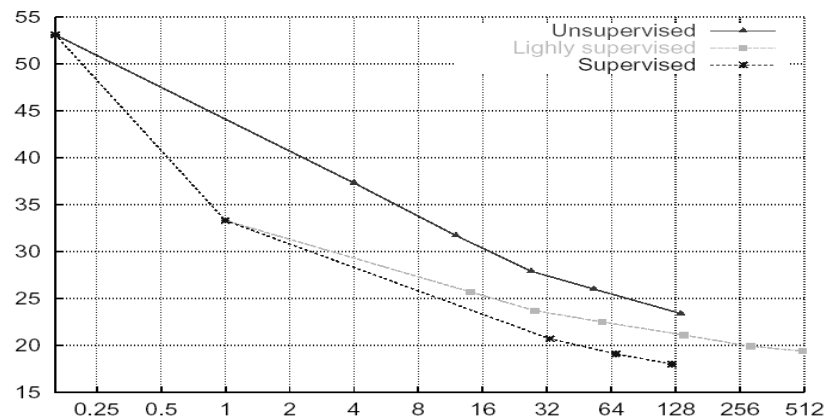
- Progrès des 15 dernières années essentiellement dus à...
  - Un affinement des modèles : approches discriminantes (MMI,MPE), partage (tying) de gaussiennes ou d'états
  - Des techniques d'adaptation (MAP,MLLR,VTLN)
  - Une puissance de calcul croissante : pour le décodage multi-passe ou les approches multi-reconnaisseurs (ROVER)

---

■ ... surtout...

# La reconnaissance automatique de la parole

- **Plus de données !**
- ***“There’s no data like more data”***, Robert L. Mercer



LIMSI, Lamel (2002)

<b>Training (hrs)</b>	<b>141</b>	<b>297</b>	<b>602</b>	<b>843</b>
<b>WER(%)</b>	<b>17.2</b>	<b>15.4</b>	<b>14.7</b>	<b>14.5</b>

RT03 (BBN)

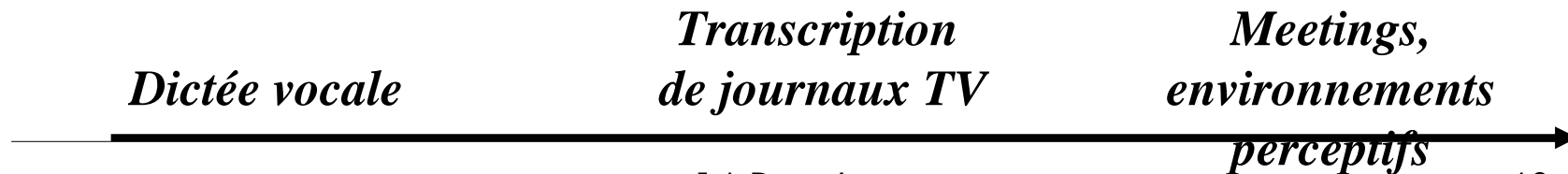
# Evolution du domaine

---

## □ Tendances

- Transcription 'Simple' → Transcription Riche
- Flux audio contrôlé → Flux audio continu
- Un capteur → Capteurs multiples
- Monolingue → Multilingue
- Audio seul → Multimodal

## □ Difficulté croissante des tâches



# Limites et problèmes ouverts

---

- Transcription enrichie
  - Marquer les tours de parole, les disfluences
- Flux audio continu
  - Besoin de marques de début / fin de phrases, ponctuation, pas uniquement de la parole
- Capteurs multiples
- Multilinguisme
  - Portage vers une nouvelle langue, locuteurs non natifs
- Multimodalité
  - Flux multiples et souvent asynchrones

# Mes thèmes de recherche

---

- En lien avec l'évolution du domaine
  - **Transcription enrichie**
  - **Reconnaissance de la parole multilingue**
  
- Défis
  - **Extraction d'informations non linguistiques à partir de la parole** → **Partie II**
  - **Problème des langues peu dotées (peu de données disponibles) pour la reconnaissance automatique de la parole** → **Partie III**
  
- Besoin d'approches alternatives
  - **Utilisation d'unités sous-lexicales pour la modélisation statistique du langage parlé** → **Parties III & IV**

# Ligne de conduite (mes 3 principes)

---

- Garder un bon équilibre
  - Entre aspects exploratoires et opérationnels
- Travailler avec les doctorants
  - Encadrement de plusieurs thèses
- Se comparer aux autres
  - Participation régulière à des campagnes d'évaluation

# Encadrement de thèses

---

□ 6 doctorats (et 6 Masters-R)

□ Encadrement partiel

-C. Nguyen (taux d'encadrement : 30%) : *Reconnaissance automatique de la parole en langue vietnamienne*. Doctorat de l'INPG, **thèse soutenue** en Juin 2002.

-D. Vaufreydaz (taux d'encadrement : 50%) : *Modélisation statistique du langage à partir d'Internet pour la reconnaissance automatique de la parole continue*. Doctorat de l'Université J. Fourier, **thèse soutenue** en Janvier 2002.

-D. Istrate (taux d'encadrement : 50%) : *Détection et Reconnaissance des Sons pour la Surveillance Médicale*. Doctorat de l'INPG, **thèse soutenue** en Décembre 2003.

-V-B Le (taux d'encadrement : 70%) : *Reconnaissance automatique de la parole pour des langues peu dotées*. Doctorat de l'Université J. Fourier, **thèse soutenue** le 1er Juin 2006.

□ Encadrement total

-D. Moraru : *Segmentation en locuteurs de documents audios et audiovisuels : application à la recherche d'information multimédia*. Doctorat de l'INPG, **thèse soutenue** en Décembre 2004.

-P. Mayorga : *Reconnaissance vocale dans un contexte de voix sur IP : diagnostic et propositions*. Doctorat de l'INPG, **thèse soutenue** en Janvier 2005.

# Campagnes d'évaluation

Tâches \ Année	2002	2003	2004	2005	2006
<b>Segmentation en locuteurs</b>	NIST meeting 1/4 NIST BN 2/4 NIST Tel 3/4	Rich Transcription (RT) BN 2/8*	Rich Transcription (RT) meeting 1/3*	ESTER BN 4/5 Rich Transcription (RT) meeting 2/3*	
<b>Transcription</b>				ESTER BN 6/8	
<b>Recherche d'informations</b>	Extraction de plans Parole 7/13 Monologue 3/9	Extraction de plans Personne X 4/4	Segmentation en histoires 3/6		
<b>Traduction de parole</b>					DARPA/Transtac 1/6**

(\*=collaboration avec le LIA ; \*\*=pendant mon séjour à IBM ;  
BN=Broadcast News)

# Modélisation probabiliste

---

$$\hat{P}(Y|X)$$

Séquence d'observations acoustiques  
(vecteurs multi-dimensionnels)

- *tranches (trames) de signal*
- *coefficients de bancs de filtres*
- *coefficients cepstraux*
- *composantes principales temps-fréquence*
- *...*

hypothèse de classe  
ou d'objet sonore

- *type de son (parole / musique / ...)*
- *locuteur / langue / canal*
- *phonème / syllabe / mot*
- *événement sonore (jingle)*
- *passé / futur d'une rupture*
- *...*

→ Approche générique

# Bayes

---

- $x$  : observation (signal)
- $c_i$  : classe à reconnaître

$$c^* = \arg \max_i p(c_i / x) = \arg \max_i \frac{p(x / c_i) \cdot P(c_i)}{p(x)} = \arg \max_i p(x / c_i) \cdot P(c_i)$$

- Reconnaissance automatique de la parole

$$w^* = \arg \max_i \frac{p(x / w_i) \cdot P(w_i)}{p(x)} = \arg \max_i p(x / w_i) \cdot P(w_i)$$

Modèle acoustique ↑  
↓  
Modèle de langage

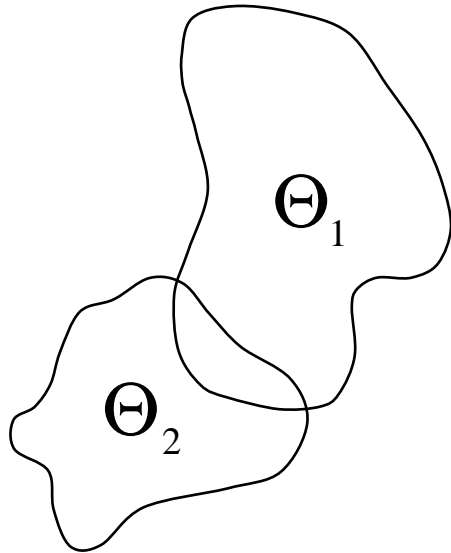
- Traduction automatique statistique

$$e^* = \arg \max_i \frac{p(f / e_i) \cdot P(e_i)}{p(f)} = \arg \max_i p(f / e_i) \cdot P(e_i)$$

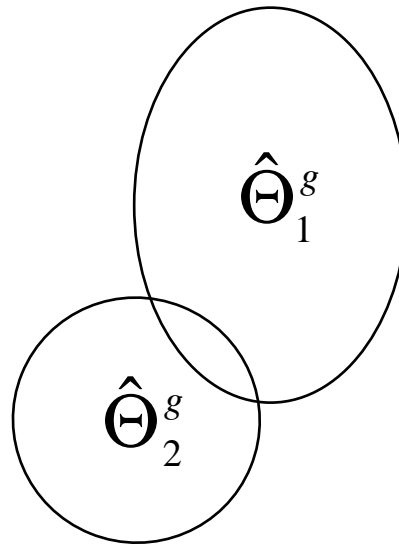
Modèle de traduction ↑  
↓  
Modèle de langage

# Gaussiennes

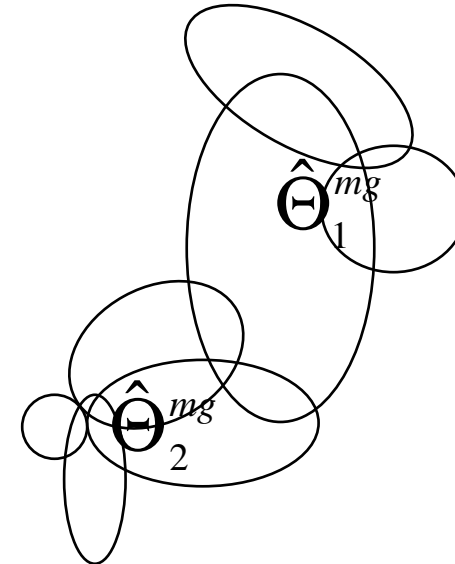
---



Distribution réelle



Modèle gaussien



Modèle multigaussien  
(GMM)

# Automates

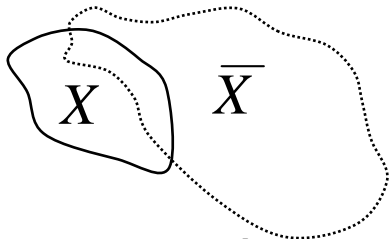
---

- Traitement / Modélisation de processus séquentiels
- Structures séquentielles complexes décomposées en segments élémentaires stationnaires
- Chaque segment : fonction déterministe ou stochastique
- Permet de décrire des modèles de langues, de lexiques, de phonèmes...
- Exemple : Modèles de Markov Cachés (HMMs)
  - 2 processus stochastiques :
    - Séquence d'états du HMM (structure séquentielle des données)
    - Probabilité d'émission de l'état (caractéristiques locales des données)
    - Exemple : modèle HMM gauche-droit de phonème avec distributions multigaussiennes

# Types de problèmes traités

---

## Détection



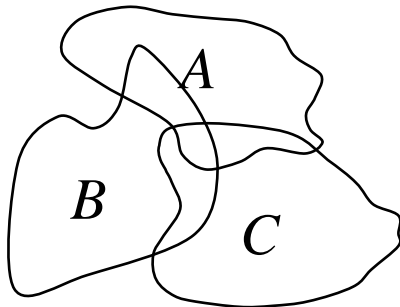
→ Tests d'hypothèses binaires

## Segmentation



→ Détection de ruptures

## Classification



→ Maximum A Posteriori

## Décodage



→ Recherche de séquences d'états

\*Transparent emprunté à F. Bimbot

# Boîtes à outils et communauté du logiciel libre

---

- Utilisées
  - RAP
    - Janus, Sphinx
  - Modélisation du langage
    - SRI-LM, FSM library (AT&T)
  - Traduction automatique statistique
    - GIZA++, Pharaoh
  
- Développées au CLIPS ou lors d'un projet impliquant le CLIPS
  - CLIPS-Text-Tk : extraction, filtrage et sélection de données pour la modélisation du langage à partir du Web
  - EMACOP : environnement d'acquisition et de gestion de corpus de parole
  - ALIZE (boîte à outils libre pour la reconnaissance automatique du locuteur) : GMMs, ...

# Vue d'ensemble

---

Partie I : Domaine, Problèmes, Méthodes

**Partie II : Au-delà de la transcription :  
locuteurs, sons, etc.**

Partie III : Reconnaissance automatique de la  
parole pour les langues peu dotées

Partie IV : Travaux en cours et axes de  
développement scientifique

---

# **Partie II : Au-delà de la transcription : locuteurs, sons, etc.**

---

## **II.1 Transcription enrichie**

### **II.2 L'information "locuteur"**

- Biométrie vocale
- Segmentation en locuteurs

### **II.3 Autres informations**

- Sons
- Jingles
- Questions

### **II.4 Exploiter la multimodalité**

- Biométrie multimodale
- Signatures audiovisuelles
- Segmentation audiovisuelle de documents

# Transcription enrichie

```
<Speaker id="sp1" name="Nicolas Stoufflet" check="yes" type="male"
  dialect="native" accent="" scope="global"/>
<Section type="filler" startTime="0" endTime="9.632">
<Turn startTime="0" endTime="1.5" speaker="sp1" >
<Sync time="0"/>
Patricia Martin , que voici , que
<Event desc="top" extent="instantaneous"/>
voilà !
</Turn>
<Turn speaker="sp53" startTime="1.5" endTime="2.624">
<Sync time="1.5"/>
oh , bonjour
<Event desc="top" extent="instantaneous"/>
Nicolas Stoufflet .
</Turn>
<Turn speaker="sp1" startTime="2.624" endTime="3.765">
<Sync time="2.624"/>
France-Inter
<Event desc="top" extent="instantaneous"/>
, 7 heures .
</Turn>
```

Transcription

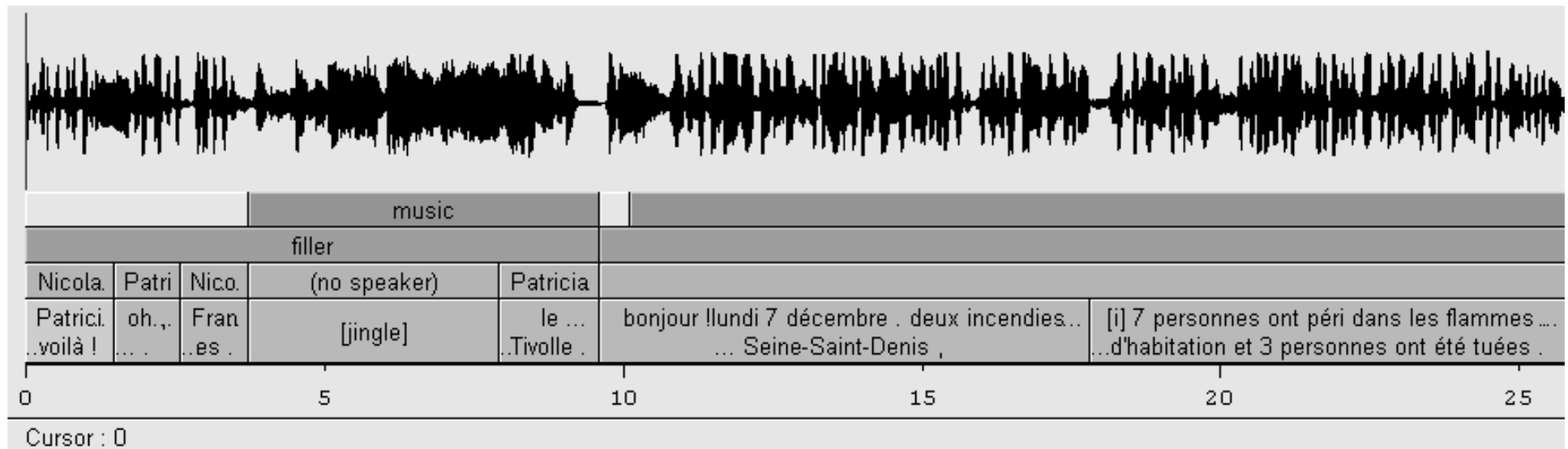
Son

Locuteur

Tour de parole

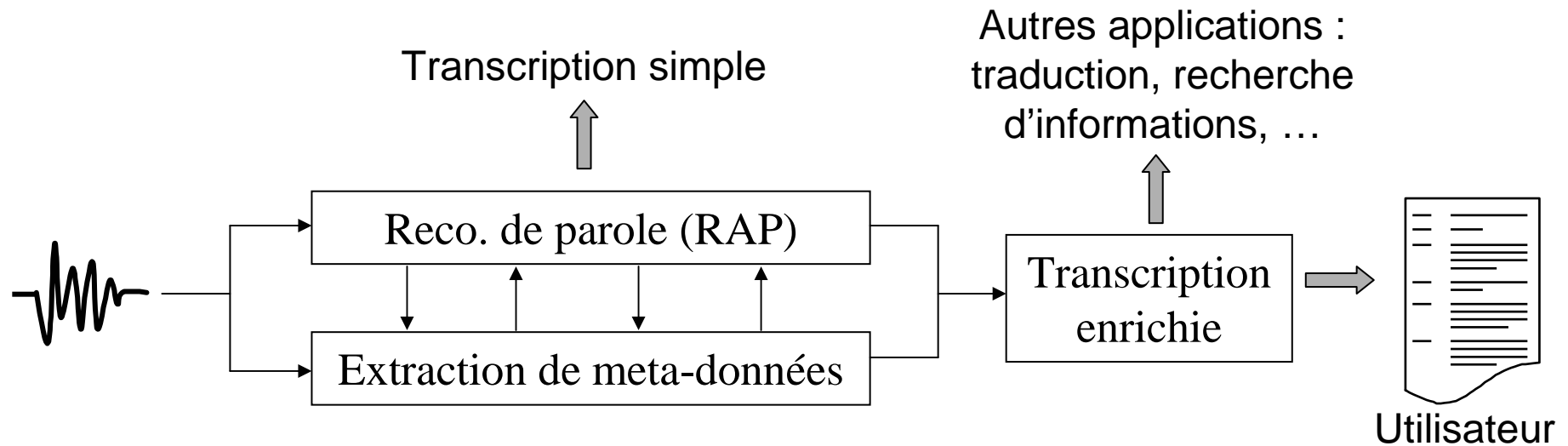
# Transcription enrichie

---



# Transcription enrichie

---



Meta données :

- ❖ Informations non linguistiques
- ❖ exemple : disfluences, marques de début / fin de phrases, **tours de parole**, **locuteurs**, **sons**, **jingles**, rires, **ponctuation**, etc.

# L'information « locuteur »

---

*Qui parle quand ?*

- ❖ Reconnaissance du locuteur (biométrie vocale)
  - Affecter une identité à un segment (ou un groupe de segments) de parole
- ❖ Segmentation en locuteurs
  - Segmenter un document audio en zones homogènes contenant chacune la voix d'un seul locuteur
  - Grouper entre eux les segments correspondant à un même locuteur

# Reconnaissance du locuteur (biométrie vocale)

---

- Doctorat sur ce thème soutenu en 1998
  - sélection et localisation d'informations pertinentes pour la reconnaissance automatique du locuteur
  - Multiples aspects du domaine abordés depuis 98
    - Robustesse à travers les réseaux de télécommunication (mobiles ou de voix sur IP)
    - Systèmes d'authentification multimodaux
    - Applications réelles & évaluation
- Travaux largement publiés dans les conférences internationales et les journaux (speech communication journal, signal processing journal, applied signal proc. journal)

# Segmentation en locuteurs

---

- Doctorat Daniel Moraru
  - *Segmentation en locuteurs de documents audios et audiovisuels : application à la recherche d'information multimédia.* décembre 2004.
- Apport de solutions à quelques problèmes liés à la tâche
  - p1 : estimer automatiquement le nombre de locuteurs dans un document
  - p2 : traiter un flux hétérogène (parole, musique, bruit, etc.)
  - p3 : recouvrement entre locuteurs (cas à plusieurs microphones)
- Vainqueur des évaluations RT-NIST 2002 et 2004 (pour les données de type *réunion*) de systèmes de segmentation en locuteurs (2ème en 2003...)
- Application à l'analyse automatique de documents multimédia
- Travaux publiés dans la revue *Speech Communication* (2006)

# Segmentation en locuteurs

---

Données	Perf (%err.) 2002	Perf (%err.) 2003	Perf (%err.) 2004
Téléphone	16,58 %	-	-
Journaux TV	30,33 %	19,25 %	-
Réunions	50,20 %	-	22,6 %*
Problèmes abordés	-	p1 + p2	p1+p3

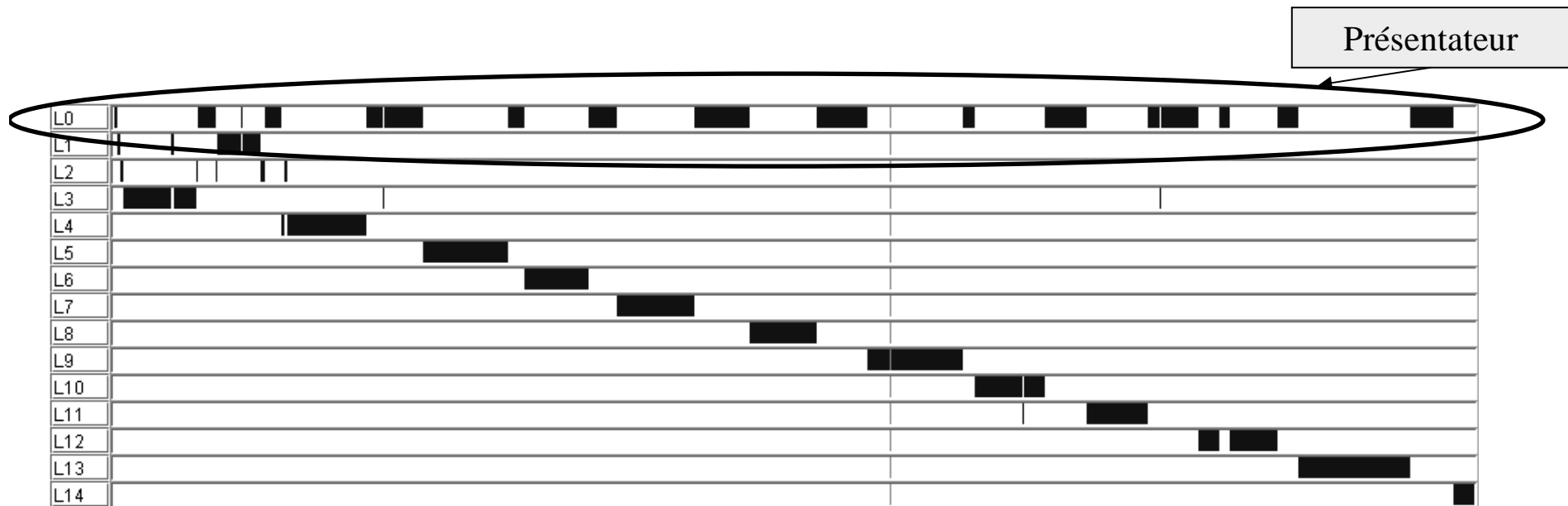
\* Plusieurs microphones

**p1 : estimer automatiquement le nombre de locuteurs dans un document**

**p2 : traiter un flux hétérogène (parole, musique, bruit, etc.)**

**p3 : recouvrement entre locuteurs (cas à plusieurs microphones)**

# Application à l'analyse automatique de documents multimédia



- ❖ Du signal vers une information de haut niveau
- ❖ Structure d'un enregistrement de journal télévisé

# Reconnaissance de sons

---

- Détection et reconnaissance des sons pour la surveillance médicale
  - Doctorat Dan Istrate (décembre 2003)
  - Transfert de méthodes parole => sons
    - paramètres issus des communautés "parole" et "musique", modèles génératifs (GMM, HMM)
  - Travaux publiés dans la revue *IEEE Transactions on Information Technology in Biomedicine* (2006)
- Détection de sons clés pour l'analyse de vidéos
  - Détecter des invariants de production (jingles)
  - Signatures audio
  - Critères forts d'appariement (pas de modèles)

# Questions

---

- Travail en cours de Vu-Minh Quang
- Trouver automatiquement les questions dans un enregistrement de parole
- Du signal vers une information de haut niveau
  - Résumé (*speech summarization*)
  - Ajout de ponctuation à une transcription
- Paramètres extraits de la courbe d'intonation + arbres de décision
- Avec / sans transcription auto.
- Problème étudié pour des langues tonales et non tonales (vietnamien / français)

## Exemple de performance pour le français (234Q et 234<sup>^</sup>Q issues de réunions)

Modèles	F mesure
Prosodique	<b>74%</b>
Lexical	<b>58%</b>
Combiné	<b>77%</b>

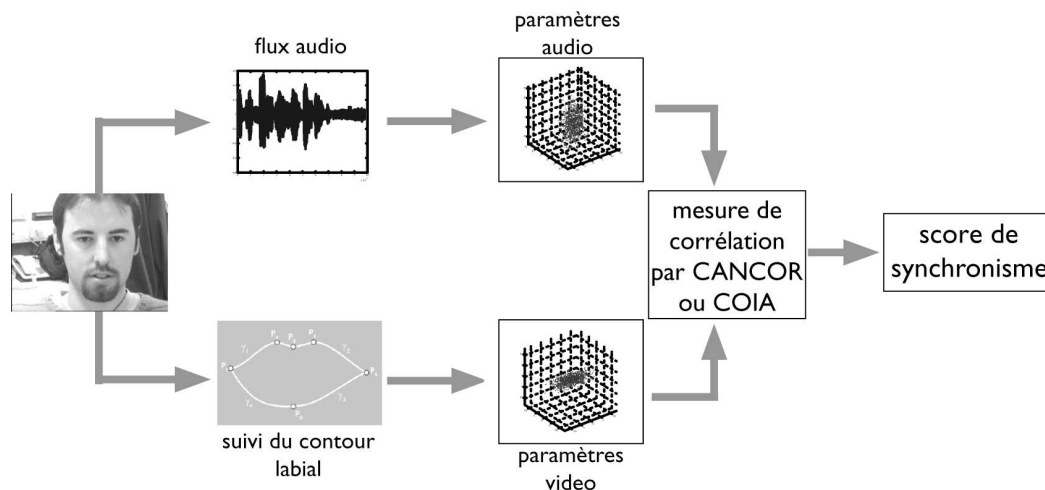
# Exploiter la multimodalité

---

- Tâches et environnements multimodaux
  - Biométrie multimodale
  - Recherche d'information multimédia
- Etude de trois cas
  - Définition d'un score de synchronisme pour la biométrie labiale multimodale
  - Signatures audiovisuelles pour la détection de séquences vidéo
  - Segmentation audiovisuelle de documents

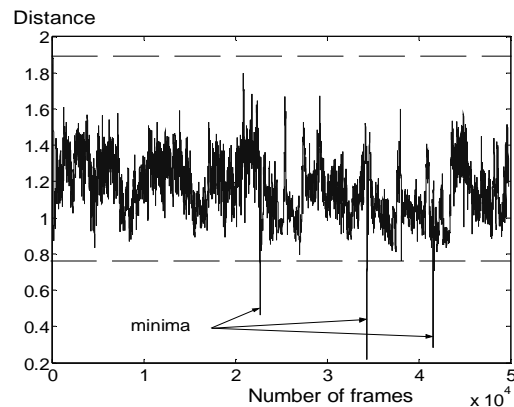
# Biométrie multimodale

- Détection de playback (imposteurs)
  - Test de présence humaine (*liveness test*) pour les systèmes biométriques
- Analyse de la dépendance statistique entre les flux audio et vidéo
  - Score de synchronisme entre les paramètres labiaux et de parole
  - 12% EER en détection de playback (ensemble de test construit à partir de la base M2VTS)

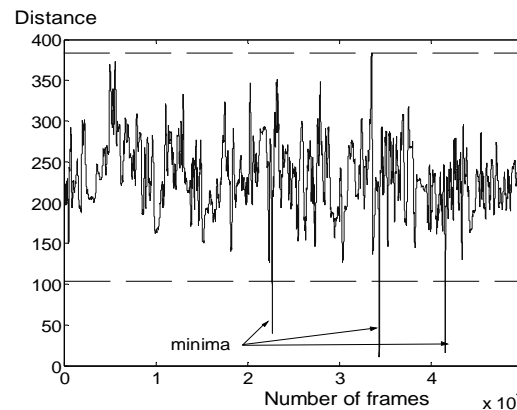


# Signatures audiovisuelles

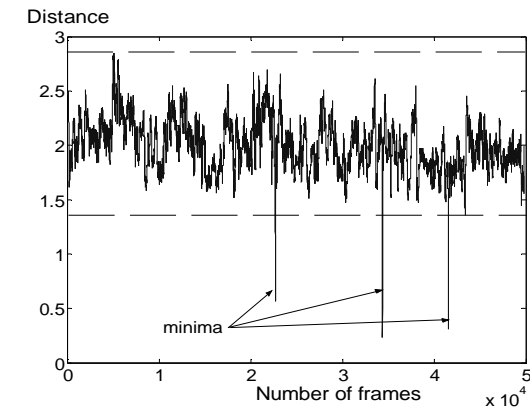
- Collaboration avec le LIS
- Extraction de signatures audio et visuelles
- Synchronisation, normalisation et combinaison des courbes de distance A et V entre une requête et un document
- L'utilisation conjointe des signatures audio et vidéo (A+V) améliore précision et rappel pour une tâche de détection de séquences vidéo



**A**



**V**

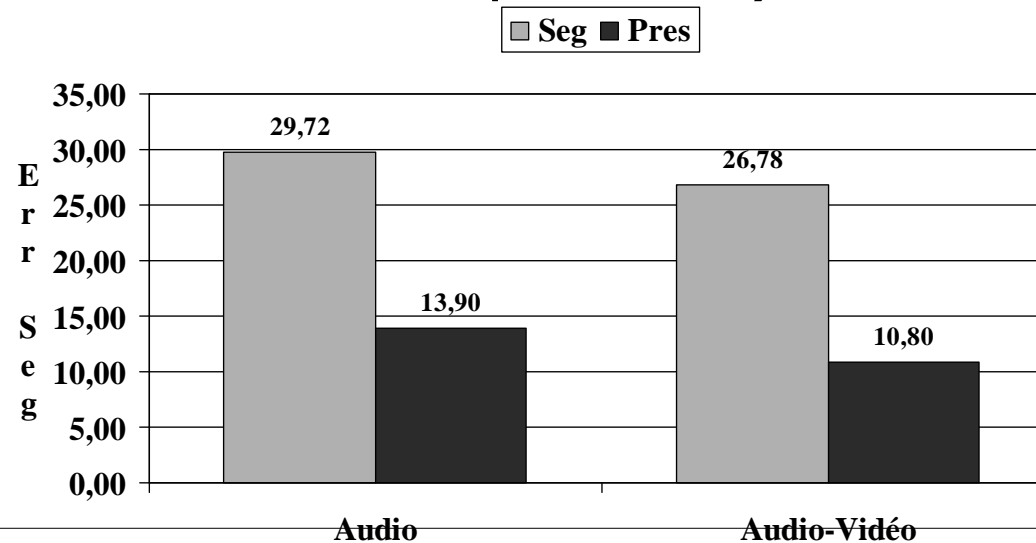


**A+V**<sub>β6</sub>

# Segmentation audiovisuelle

---

- Utiliser une information issue de la vidéo pour des tâches "audio"
  - Segmentation et regroupement en locuteurs (Seg)
  - Suivi du présentateur (Pres)
- Information vidéo : marques de frontières de plans (obtenues automatiquement)



# Bilan de la seconde partie

---

- Extraction d'informations non linguistiques pour l'annotation multi-niveaux de documents audio et audiovisuels
  - Locuteurs
  - Sons
  - Jingles
  - Questions
  
- Quelques études de cas montrent le bénéfice que l'on peut retirer en exploitant la dimension multimodale des signaux

# Vue d'ensemble

---

Partie I : Domaine, Problèmes, Méthodes

Partie II : Au-delà de la transcription :  
locuteurs, sons, etc.

**Partie III : Reconnaissance automatique  
de la parole pour les langues peu  
dotées**

Partie IV : Travaux en cours et axes de  
développement scientifique

---

# **Partie III : Reconnaissance automatique de la parole pour les langues peu dotées**

---

## **III.1 Collecte de données**

- Collecte de données textuelles
- Collecte de parole

## **III.2 Amorçage des modèles acoustiques**

- Modélisation acoustique translingue
- Application au vietnamien et au khmer

## **III.3 Réduction de la complexité des modèles**

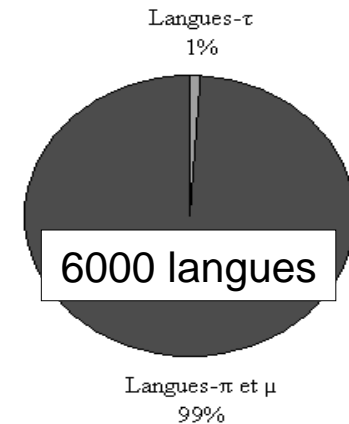
- Utilisation d'unités sous-lexicales pour la modélisation statistique du langage parlé
- Application à une langue peu écrite

# Un monde multilingue

---

- En 2005, moins de 1 % des 6000 langues du monde atteignent un haut niveau d'informatisation (services allant du traitement de texte à la traduction automatique)
  - Langues peu dotées (*under-resourced languages, low density languages*)

Cf. Thèse V.Berment : «Méthodes pour informatiser des langues et des groupes de langues peu dotées»



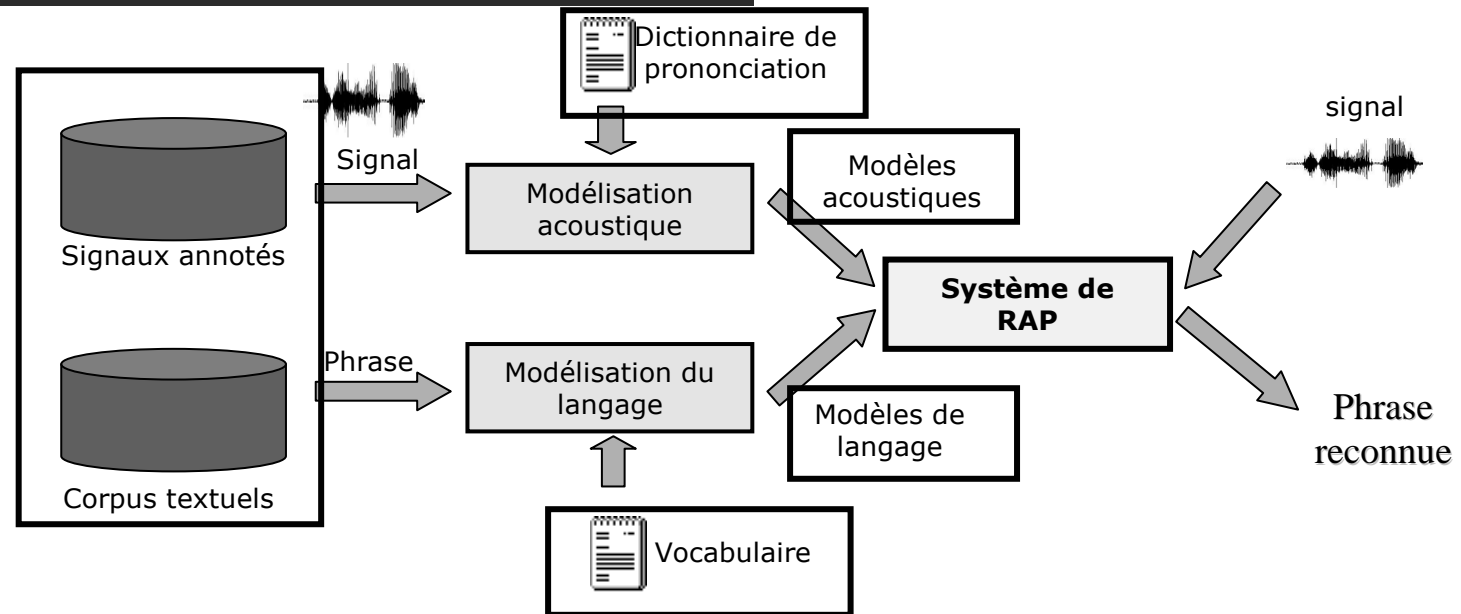
- Grande diversité des systèmes d'écriture
- Langues à forte tradition orale (langues peu écrites)

# Un monde multilingue

---

- Langues peu dotées
  - Peu de données disponibles
  - Besoin de méthodes innovantes qui vont au delà du simple ré-apprentissage des modèles acoustiques et de langage
    - Méthodologie de collecte
    - Amorçage (bootstrap) des modèles acoustiques
    - Réduction de la complexité des modèles

# Ressources nécessaires pour la RAP



- Corpus textuels et de parole
- Dictionnaire de prononciation
- Modèles acoustiques
- Modèles de langage

# Collecte de données

---

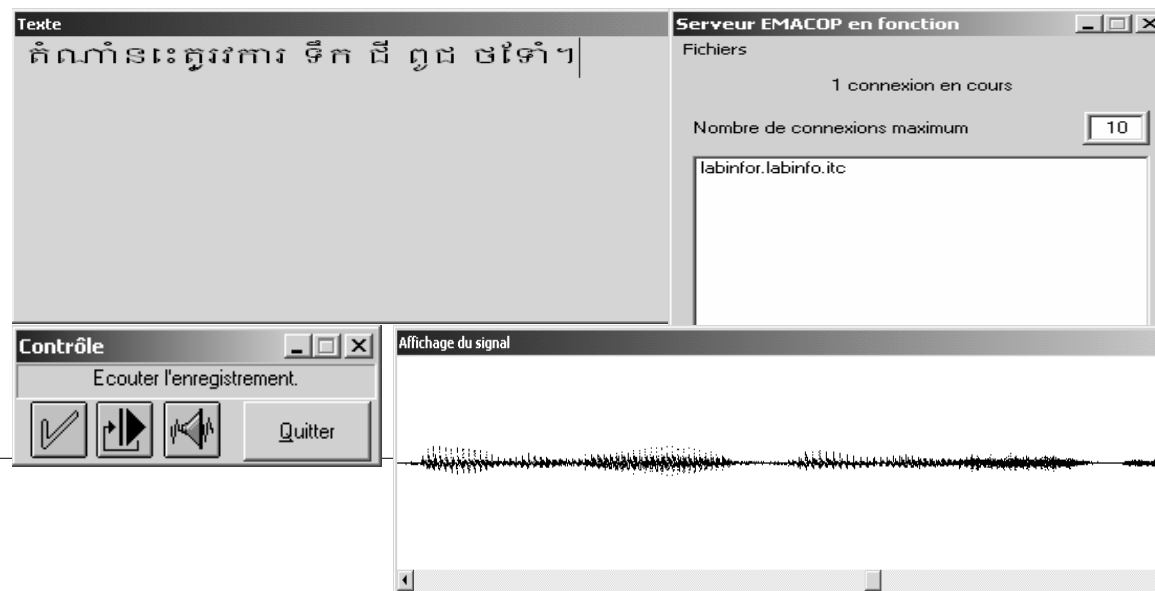
## □ Collecte de données textuelles

- D. Vaufreydaz : *Modélisation statistique du langage à partir d'Internet pour la reconnaissance automatique de la parole continue*. Doctorat de l'Université J. Fourier, **thèse soutenue** en janvier 2002 .
- Potentiel pour les langues peu dotées
  - Web parfois unique moyen de collecter des données texte
  - Mais principalement sites d'informations
- Exemple : [www.voanews.com](http://www.voanews.com)

	<b>#phr</b>	<b>#mots</b>	<b>#octets</b>
<b>indonésien</b>	<b>116k</b>	<b>2.4M</b>	<b>17M</b>
<b>coréen</b>	<b>405k</b>	<b>7M</b>	<b>67M</b>
<b>pachto</b>	<b>7k</b>	<b>0.2M</b>	<b>2M</b>
<b>kurde</b>	<b>24k</b>	<b>0.6M</b>	<b>8M</b>
<b>hindi</b>	<b>73k</b>	<b>2M</b>	<b>28M</b>
<b>farsi</b>	<b>212k</b>	<b>5.8M</b>	<b>54M</b>

# Collecte de données

- Collecte de données textuelles
- **Collecte de parole**
  - Collaborations locales (MICA/Hanoi ; ITC/Phnom-Penh)
  - Enregistrement sur place avec EMACOP (*Multimedia Environment for Acquiring and Managing Speech Corpora*)
  - Transcriptions locales d'enregistrements radio ou TV



# Amorçage des modèles acoustiques

---

- Collecte de données textuelles
- Collecte de parole
- **Amorçage des modèles acoustiques (bootstrap)**
  - Modélisation acoustique translingue

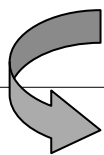
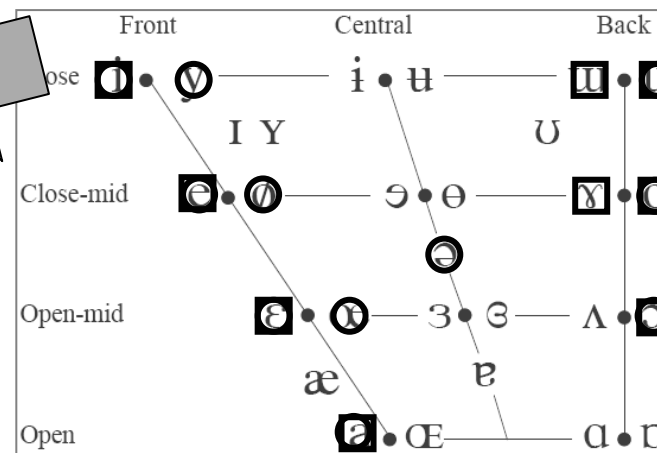
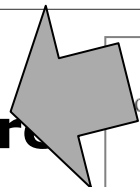
# Modélisation acoustique translingue

	Bilabial	Labiodental	Dental	Alveolar	Post alveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	◻ ◻			◻ ◻		◻ ◻	◻ ◻	◻ ◻	q G		ʔ
Nasal	◻	ŋ		◻		ŋ	◻	◻	N		
Trill	B			r					R		
Tap or Flap				ɾ		ɽ					
Fricative	φ β ◻ ◻		θ ð ◻ ◻	◻ ◻ ◻ ◻	◻ ◻	◻ ◻	ç ʝ	x ◻ ◻	◻ ◻ ◻	ħ ʕ	◻ ◻
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	ɟ	ɰ			
Lateral approximant				◻		ɭ	ʎ	L			

◻  
Phonème FR

◻  
Phonème VN

- FR/VN ~63% couverture
- Si plusieurs langues source (ex: modèle multilingue de 7 langues)  
=> 87% couverture



**Bénéfice d'une  
couverture  
multilingue**

# Modélisation acoustique translingue

---

$$\forall \Phi_S, d(\Phi_{S^*}, \Phi_T) = \min [d(\Phi_S, \Phi_T)]$$

- Proposition de nouvelles mesures de similarité entre phonèmes (ou polyphonèmes) pour l'amorce (*bootstrap*) rapide des modèles acoustiques dans une nouvelle langue
- $\Phi_S$  et  $\Phi_T$  : modèles en langue source et cible
  - Monophones, polyphones, groupes de polyphones
- $d$  : distances fondées sur les connaissances ou fondées sur les données
- V-B Le : *Reconnaissance automatique de la parole pour des langues peu dotées*. Doctorat de l'Université J. Fourier, école doctorale EDMI Grenoble, **thèse soutenue** le 1er Juin 2006.

# Application

---

- Collecte de données textuelles
- Collecte de parole
- Amorçage des modèles acoustiques (bootstrap)
- Application au vietnamien et au khmer**

Performance de RAP pour le vietnamien (% syllabes correctes)  
Corpus de dialogue

Système source	Distance	Adapt 1h	Adapt 2h
		WA	WA
Français	Connaissance	60.4	63.6
	Données	61.6	63.8
Multilingue (CMU, 7 langues)	Connaissance	64.6	66.3
	Données	63.8	65.3

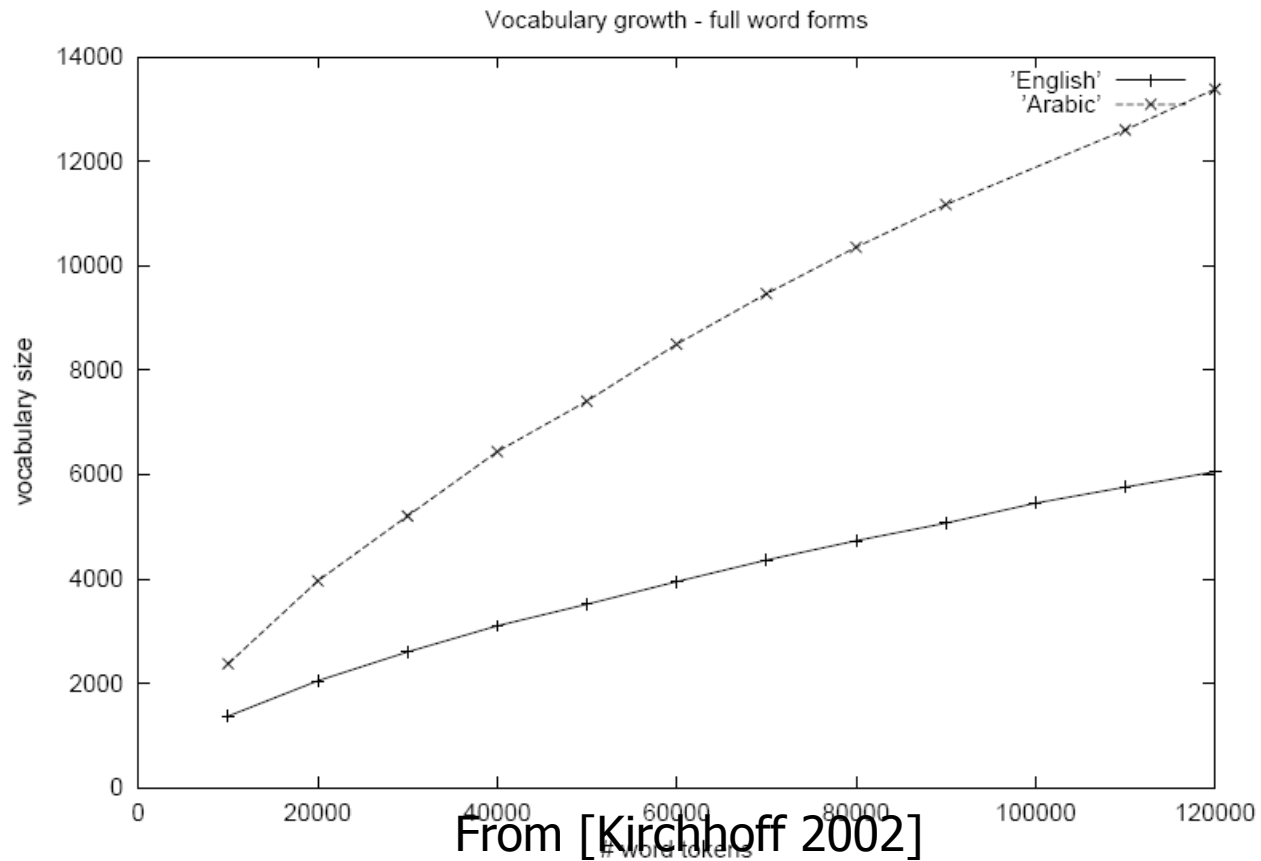
**Même méthodologie  
appliquée au khmer :  
système de RAP développé  
en quelques mois :  
WA=73.6% sur des  
phrases lues**

# Réduction de la complexité des modèles

---

- Collecte de données textuelles
- Collecte de parole
- Amorçage des modèles acoustiques (bootstrap)
- Application au vietnamien et au khmer
- **Réduction de la complexité des modèles**
  - **Utilisation d'unités sous-lexicales pour la modélisation statistique du langage parlé**
  - Séjour à IBM Watson (09/2005=>11/2006)
    - Arabe dialectal (Irakien) : reconnaissance et traduction
    - Langue peu écrite

# Exemple de l'arabe standard



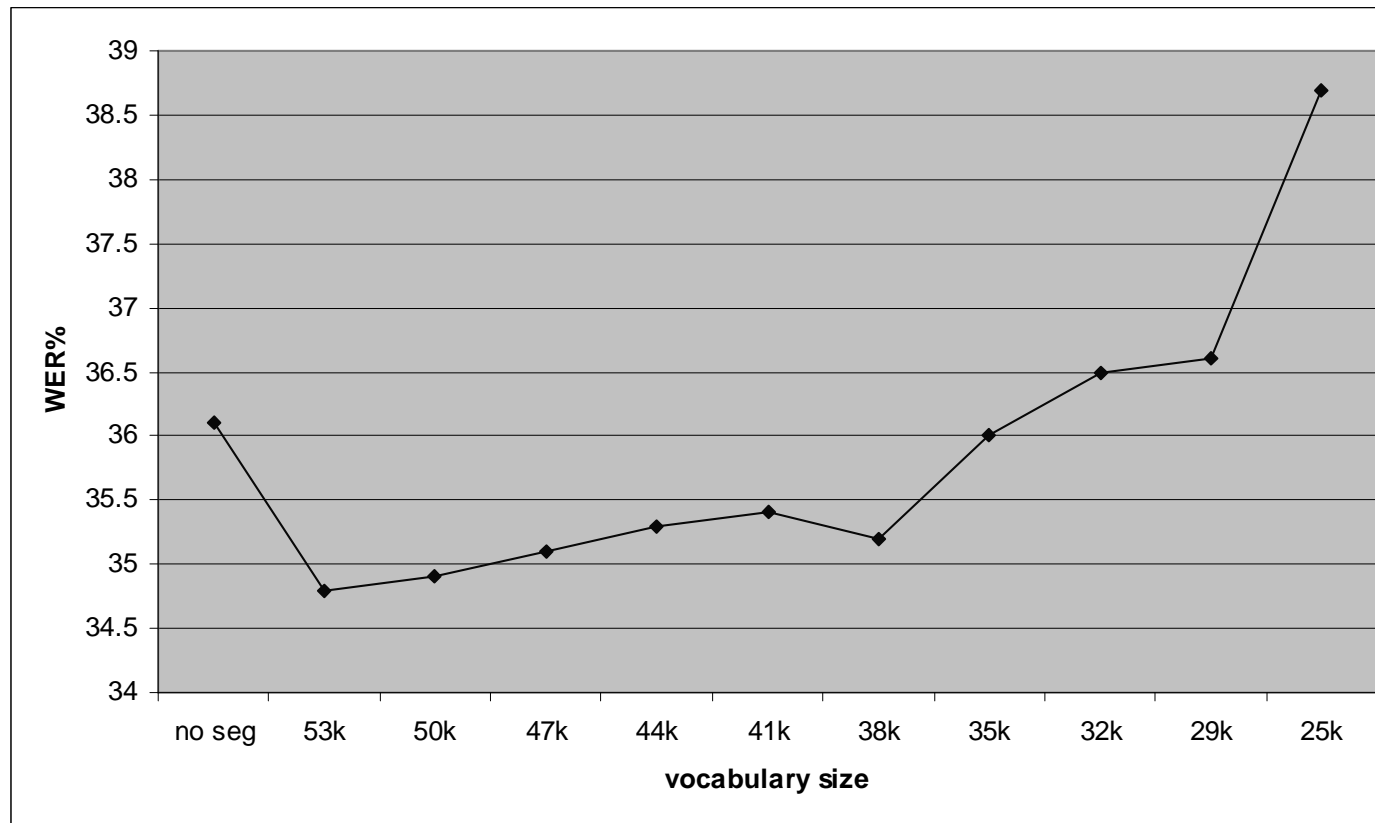
# Analyse morphologique pour l'arabe dialectal

---

- Peu de données textuelles disponibles
- Segmenter les mots en préfixe-base-suffixe pour réduire le vocabulaire de l'application *إذا عندك قول لي حتى أ#روح أ#صيح اخوان+ي*
  - Et donc réduire la complexité des modèles
- Approche fondée sur les données
  - Apprentissage d'un modèle qui prédit les marques de préfixes et de suffixes à partir d'une chaîne non segmentée
  - Ne pas segmenter les N mots les plus fréquents du corpus d'apprentissage
    - Problème de couverture des modèles de langage n-grammes
    - N => contrôle la taille du vocabulaire

# Performances de reconnaissance automatique de parole en irakien

---



# Vers une traduction automatique des langues peu écrites

---

- Idée : pour une tâche telle que la traduction de parole, la forme écrite  $f$  de la langue source pourrait être considérée comme secondaire

$$\hat{e} = \arg \max_e P(e / x) = \arg \max_e \sum_f P(e, f / x)$$

$$\approx \arg \max_e \sum_f P(e / f) P(f / x)$$

SMT

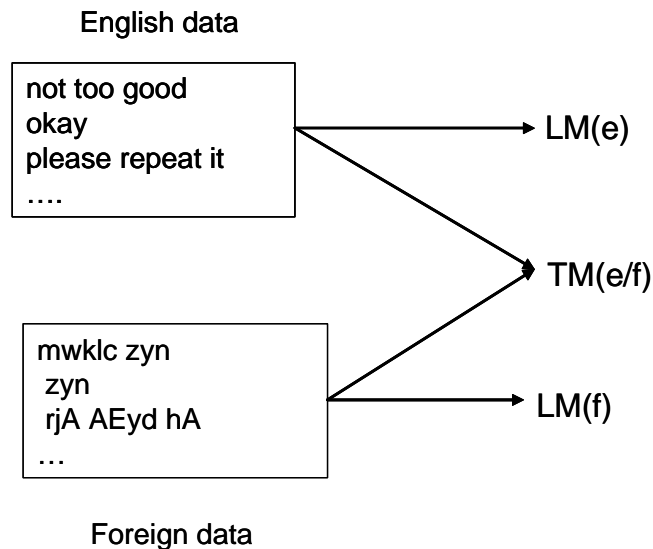
ASR

- Est-il possible de construire un système de traduction de parole à partir d'un corpus parallèle composé d'enregistrements d'une langue peu écrite et de leur traduction en anglais ?

- Hypothèse : enregistrements transcrits en symboles phonétiques

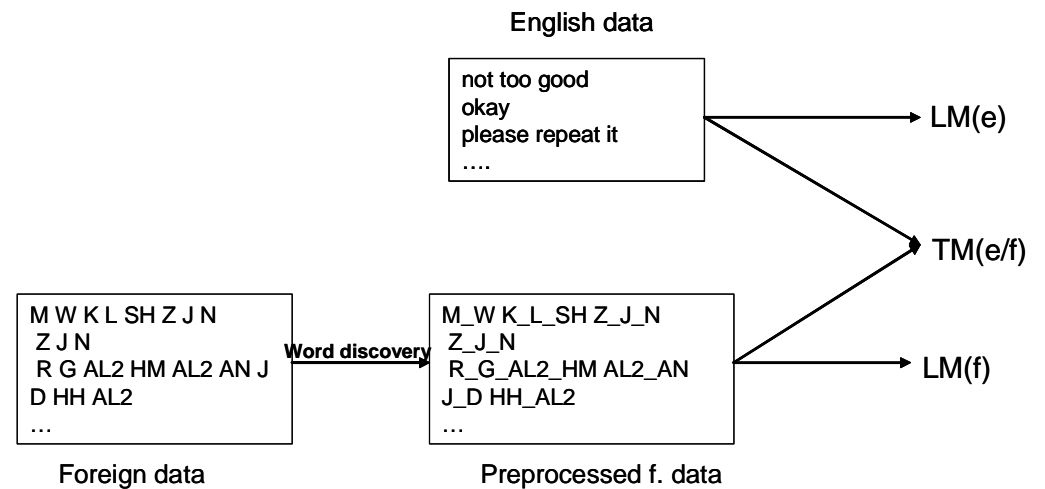
# Mots *versus* Phonèmes

## □ Mots



Taille vocabulaire : 43k

## □ Phonèmes



Taille vocabulaire : 36k

# Résultats expérimentaux

---

- La méthode utilisant les unités phonétiques (*phonèmes*) est pratiquement équivalente en performance à la méthode classique (*mots*)
- 54% phrases jugées correctement traduites (*phonèmes*) contre 58% (*mots*)
  - Potentiel pour les langues peu écrites
  - Potentiel pour réduire le vocabulaire de l'application (sans perte de couverture)
  - Détails publiés dans « Towards speech translation of non written languages» **Laurent Besacier**, Bowen Zhou, Yuqing Gao. IEEE / ACL SLT 2006. Aruba, December 2006.

# Bilan de la troisième partie

---

- **Contributions à la reconnaissance automatique de la parole pour les langues peu dotées**
  - Collecte de données textuelles
  - Collecte de parole
  - Amorçage des modèles acoustiques (bootstrap)
  - Application au vietnamien et au khmer
  - Réduction de la complexité des modèles
    - Langues peu écrites
    - Utilisation d'unités sous-lexicales pour la modélisation statistique du langage parlé

# Vue d'ensemble

---

Partie I : Domaine, Problèmes, Méthodes

Partie II : Au-delà de la transcription :  
locuteurs, sons, etc.

Partie III : Reconnaissance automatique de la  
parole pour les langues peu dotées

**Partie IV : Travaux en cours et axes de  
développement scientifique**

---

# **Partie IV : Travaux en cours et axes de développement scientifique**

---

## **IV.1 Multimodalité**

- Projets en cours
- Considérations pour le futur

## **IV.2 Multilinguisme**

- Extension de la modélisation acoustique translingue
- Modélisation multiniveau du langage parlé
- Considérations pour le futur

# Projets en cours

---

- Reconnaissance de gestes multimodaux pour le langage parlé complété
  - Complément de la lecture labiale pour les malentendants
  - Projet ANR TELMA (téléphonie pour les malentendants)
  - Doctorat de Nourredine Aboutabit à l'ICP (participation à l'encadrement sur les aspects "reconnaissance")
- Biométrie multimodale
  - Projet ANR MISTRAL, démarrage en 2007
  - Environnement "open source" pour l'authentification biométrique
  - Thales, Calistel, LIA, IRIT, CLIPS, LIUM
  - Transfert de méthodes utilisées en parole vers d'autres modalités (signatures, visages, ...)

# Considérations pour le futur

---

- **Constat : ensemble d'outils élémentaires pour l'annotation multiniveaux de signaux audio et audiovisuels**
  - Transcription, locuteurs, sons, jingles, questions
  
- **Futur : utilisation conjointe de descripteurs multi-niveaux pour l'analyse de scènes multimodales et l'analyse automatique de documents multimédia**
  - Consolider / compléter les descripteurs élémentaires existant
  - Introduire des outils théoriques permettant l'utilisation conjointe de ces descripteurs
    - Outils permettant de modéliser des processus temporels mettant en jeu des dépendances complexes entre paramètres
  - Expérimenter sur des tâches applicatives telles que la recherche d'information multimédia (collaboration avec l'équipe MRIM du LIG)

# Extension de la modélisation acoustique translingue

---

- Mesures de similarité entre phonèmes (ou unités plus complexes)
  - Déjà appliquée à la modélisation acoustique des langues peu dotées (amorçage ou *bootstrapping*)
  - Potentiel pour aborder d'autres problèmes
    - Définition de distances interlingues (entre systèmes phonétiques)
    - **Reconnaissance automatique de la parole non native**
    - Identification des langues

# Reconnaissance automatique de la parole non native

- Doctorat Tien-Ping Tan
- Utiliser la langue maternelle du locuteur (L1) pour améliorer les modèles acoustiques en langue cible (L2)
  - Trouver des confusions phonétiques pour un locuteur non natif
  - Utiliser des données en langue maternelle du locuteur pour adapter les modèles acoustiques de phonèmes en langue cible
  - "Acoustic model interpolation for non-native speech recognition", T.P. Tan, L. Besacier, Accepted to ICASSP 2007
  - Définir un *score de prononciation* et l'inclure dans le processus de décodage de parole

French Phoneme	Phoneme substitution (Vietnamese)
ø	ɣ (vn), ø
œ	ɣ (vn), œ
ə	ɣ (vn), ə, ø
ã	ã, ɔ (vn)
g	R
ɛ̃	ẽ, ẽ(vn)
ʃ	ʃ, ʒ (vn), s
œ̃	œ̃, a
ʒ	ʒ, z, z (vn)

# Modélisation multiniveau du langage

---

- Différents niveaux de description pour la modélisation statistique du langage
  - Mots, syllabes, caractères
- Utilisation conjointe de ces descriptions
  - Modélisation et décodage
- Exemple : 3 flux

Hôm nay, chúng tôi đến trường bằng xe hơi.

**syllabes**

H ô m n a y c h ú n g t ô i đ ể n t r u ờ n g b ằ n g x e h ơ i

**caractères**

Hôm\_nay chúng\_tôi đến trường bằng xe\_hơi.

**mots**

# Considérations pour le futur

---

- Multilinguisme = activité centrale
  - Nouvelle équipe GETALP du LIG
  - RAP multilingue (généricité)
  - Locuteurs non natifs
  - Formalisation de la modélisation acoustique translingue
  - Traduction de parole
    - B. Zhou, L. Besacier, Y. Gao "On efficient coupling of ASR and SMT for speech translation", accepted to ICASSP 2007.
  
- Projets en préparation autour des langues peu dotées
  - ANR Blanc
  - Europe (COST ou ICT)
  - Assistants de traduction pour des langues peu dotées

# Merci de votre attention

---