

# Transcription enrichie dans un monde multilingue et multimodal /Rich Transcription in a Multilingual and Multimodal World/

---

Laurent BESACIER

Soutenance de DHDR /*HDR Defense*/

11 Janvier 2007.

# Outline of this talk

---

Part I : Domain, Problems, Methods

Part II : Beyond the transcription : speakers, sounds, etc...

Part III : Automatic speech recognition for under-resourced languages

Part IV : Ongoing work and future developments

---

# Part I : Domain, Problems, Methods

---

## **I.1 Domain**

- From signal to symbols
- The speech recognition domain
- Where we go

## **I.2 Problems**

- Limits and open issues
- My research topics

## **I.3 Methods**

- Course of action
- Theoretical tools
- Toolkits and the free software community

# Domain

---

- Signals → Symbols
- Automatic extraction of **symbolic information** from a **signal**
- Signals :
  - Speech recordings
  - But not only : audio track of a video, video
- Symbols :
  - Word transcription
  - But not only : labels (speakers, sounds), breaks (segmentation)
- Multilevel automatic signal labelling

# The speech recognition domain

---

- Signals :
  - **Speech**
- Symbols
  - **Word transcription**
  
- Core technology : Automatic Speech Recognition (ASR)
  
- ➔ Speech Transcription
  - Automatic Document Analysis
- ➔ Speech Interaction
  - Human / Machine Interfaces
  - Machine-augmented Communication

# Where we are...

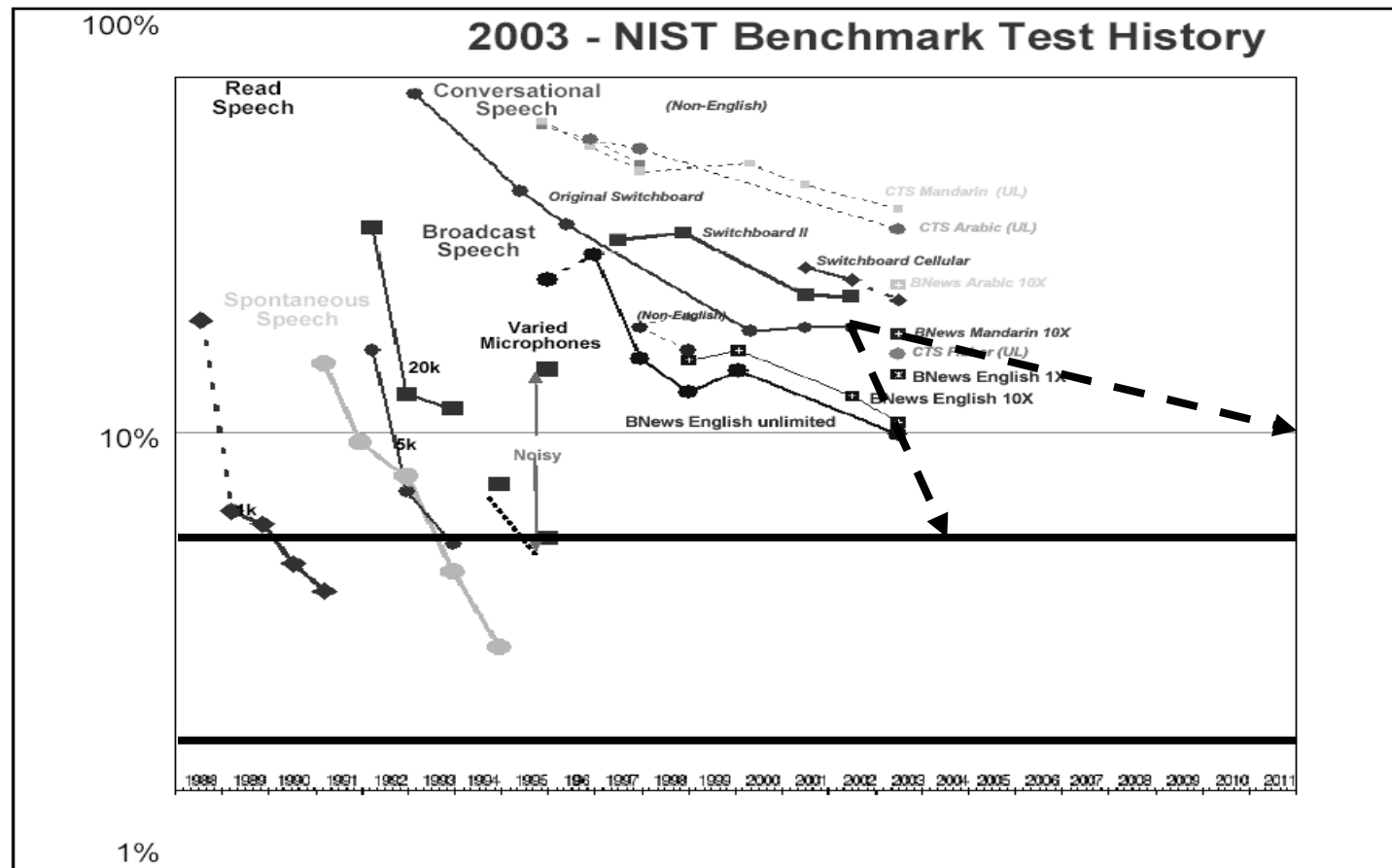
---

- Best systems achieve\*
  - ~10-12% WER for English on European Parliament Speeches or Broadcast News Data !
  - ~20% WER for English on broadcast or telephone conversations
- Large Improvements over the years  
See DARPA & NIST evaluations...

---

\*sources: TCSTAR & GALE projects

# Where we are...



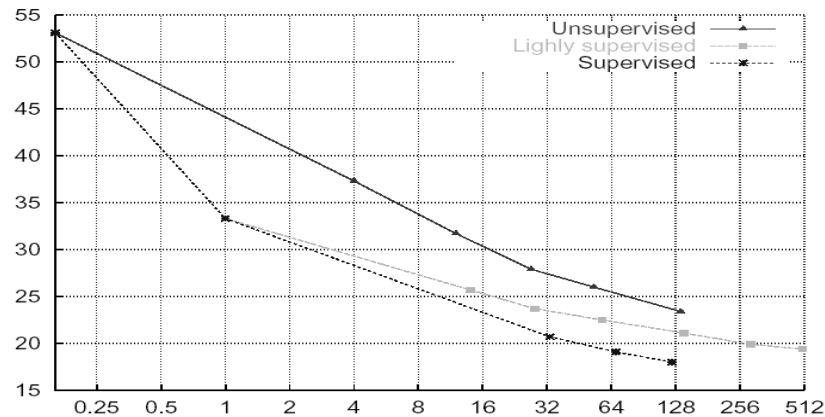
# Where we are...

---

- Improvements over the last 15 years mostly due to...
    - Better modeling : discriminant approaches (MMI,MPE), tying (mixtures, states)
    - Adaptation techniques (MAP,MLLR,VTLN)
    - Computational power : for multipass decoding and multiengine approaches (ROVER)
    - And last but not least...
-

# Where we are...

- **More Data !!**
- ***“There’s no data like more data”,***  
Robert L. Mercer



From LIMSI, Lamel (2002)

Training (hrs)	141	297	602	843
WER(%)	17.2	15.4	14.7	14.5

From RT03 (BBN)

# Where we go...

---

## □ Evolution of the domain

- 'Simple' Transcription → Rich Transcription
- Controlled Audio Stream → Continuous Audio Stream
- One sensor → Multiple sensors
- Monolingual → Multilingual
- Audio only → Multimodal

## □ Increasing difficulty of the tasks



# Limits and open issues

---

- Rich Transcription
  - Mark speaker turns, disfluencies, ...
- Continuous Audio Flaw
  - Need for sentence breaks, punctuation, ...
- Multiple sensors
- Multilingual
  - Portability to new languages, non native speaker
- Multimodal
  - Multiple data streams, asynchronism

# My research topics

---

- Related to the evolution of the domain
  - **Rich transcription**
  - **Multilingual ASR**
  
- New challenges
  - **Extraction of non linguistic informations from speech** → **Part II**
  - **Problem of under-resourced languages (no or few data available) for ASR** → **Part III**
  
- Innovative / alternative approaches still have to be proposed
  - **Moving beyond and below the word level in spoken language modeling** → **Part III & IV**

# Course of action (my 3 commandments)

---

- Keep a good balance
  - between exploratory work and operational systems
- Working with students
  - Graduate students supervision
- Compare to others
  - Participation to evaluation campaigns

# Working with students

---

- 6 PhDs (and 6 Masters)
  - Co-supervision
    - **C. Nguyen** (30%): *Automatic Speech Recognition in Vietnamese*. PhD INPG Grenoble, **defended** in June 2002.
    - **D. Vaufreydaz** (50%) : *Using the Web for Statistical Language Modeling for ASR*. PhD University J. Fourier, Grenoble, **defended** in January 2002.
    - **D. Istrate** (50%) : *Sound detection and recognition for medical telemonitoring*. PhD INPG Grenoble, **defended** in December 2003.
    - **V-B Le** (70%) : *ASR for under-resourced languages*. PhD University J. Fourier, Grenoble, **defended** in June 2006.
  - Full supervision
    - **D. Moraru** : *Speaker diarization for audio and audiovisual documents*. PhD INPG Grenoble, **defended** in December 2004.
    - **P. Mayorga** : *ASR in the context of VoIP networks : diagnosis and proposals*. PhD INPG Grenoble, **defended** in January 2005.

# Evaluation campaigns

<b>Task \ Year</b>	<b>2002</b>	<b>2003</b>	<b>2004</b>	<b>2005</b>	<b>2006</b>
<b>Speaker diarization</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 (ASR)</b>				ESTER BN 6/8	
<b>Information retrieval</b>	Feature detection Speech 7/13 Monologue 3/9	Feature detection Person X 4/4	Story segmentation 3/6		
<b>Speech-to-speech translation</b>					DARPA Transtac 1/6**

(\* = collaboration with LIA ; \*\* = during my stay at IBM ;  
BN = Broadcast News)

# Statistical modelling

---

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

Sequence of acoustic observations

- *Signal frames*
- *Filterbank coefficients*
- *Cepstral coefficients*
- *Time-frequency principal components*
- *...*

Sound object (or class) hypothesis

- *Sound type (speech / music / ...)*
- *speaker / language / channel*
- *phone / syllable / word*
- *Sound event (jingle)*
- *Past or future of a break (ex: speaker change)*
- *...*

→ Generic Approach

# Bayes

---

- $x$  : observation (signal)
- $c_i$  : class to be recognized

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

- Automatic Speech Recognition (ASR)

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

Acoustic model ↑  
Language model ↓

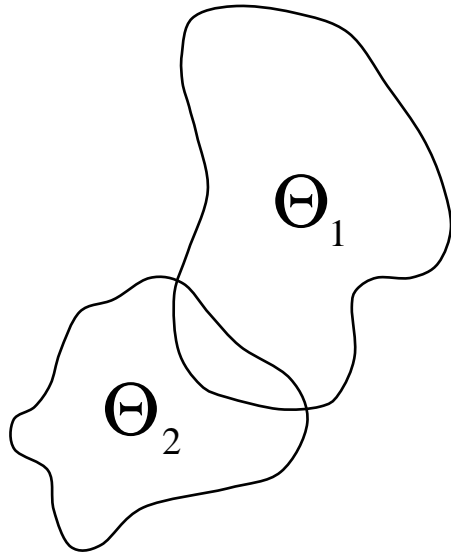
- Statistical Machine Translation (SMT)

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

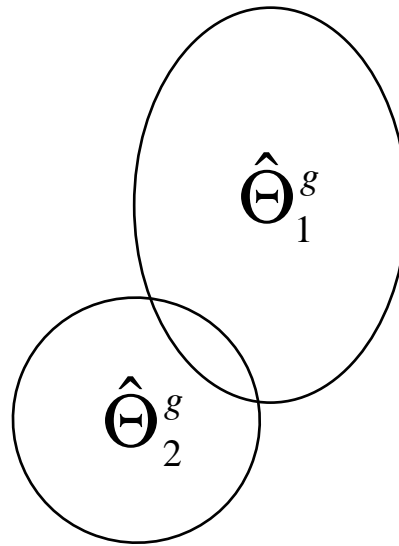
Translation model ↑  
Language model ↓

# Gaussians

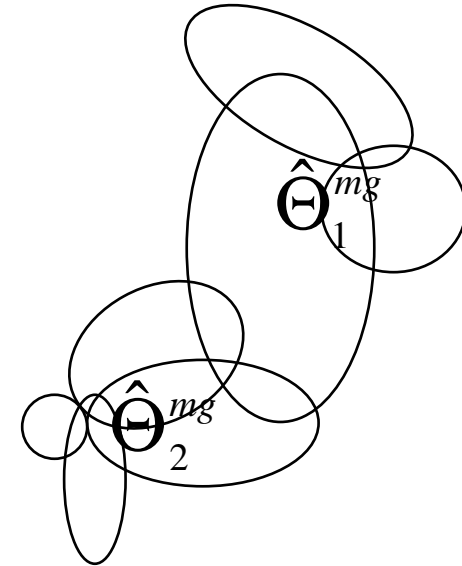
---



Real distribution



Gaussian model



Gaussian mixture model  
(GMM)

# Automata

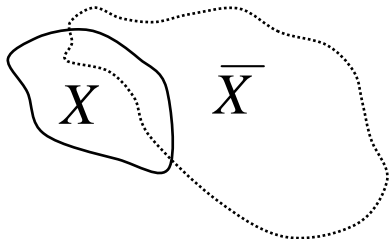
---

- For sequence processing
- Complex sequential patterns decomposed into piecewise stationary segments
- Each segment : deterministic or stochastic function
- Can describe grammar, lexicon, phone models...
- Example : Hidden Markov Models (HMMs)
  - 2 concurrent stochastic processes :
    - Sequence of HMM states (sequential structure of the data)
    - State output processes (local characteristics of the data)
    - Example : left-right HMM phone model with gaussian mixture output distributions

# Different problems

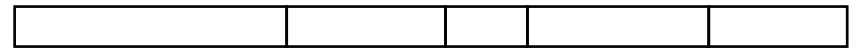
---

## Detection



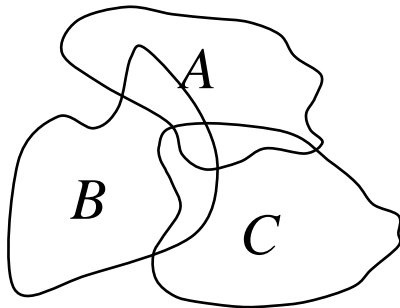
→ Binary decision tests

## Segmentation



→ Change point detection

## Clustering



→ Maximum A Posteriori

## Decoding



→ State sequence search

# Toolkits & Free Software Community

---

- From others
  - ASR
    - Janus, Sphinx
  - Language modeling
    - SRI-LM, FSM library (AT&T)
  - SMT
    - GIZA++, Pharaoh
  
- From CLIPS or from a project involving CLIPS
  - CLIPS-Text-Tk : extract, filter and select data for language modeling from web documents
  - EMACOP : speech corpora acquisition environment
  - ALIZE (free open tool for speaker recognition) : GMMs, ...

# Outline of this talk

---

Part I : Domain, Problems, Methods

**Part II : Beyond the transcription :  
speakers, sounds, etc...**

Part III : Automatic speech recognition for  
under-resourced languages

Part IV : Ongoing work and future  
developments

---

# Part II : Beyond the transcription : speakers, sounds, etc...

---

## **II.1 Rich transcription**

## **II.2 Speaker Information**

- Biometrics
- Speaker segmentation

## **II.3 Other informations**

- Sounds
- Jingles
- Question Marks

## **II.4 Exploiting Multimodality**

- Multimodal biometrics
- A/V signatures
- A/V segmentation

# Rich Transcription

```
<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

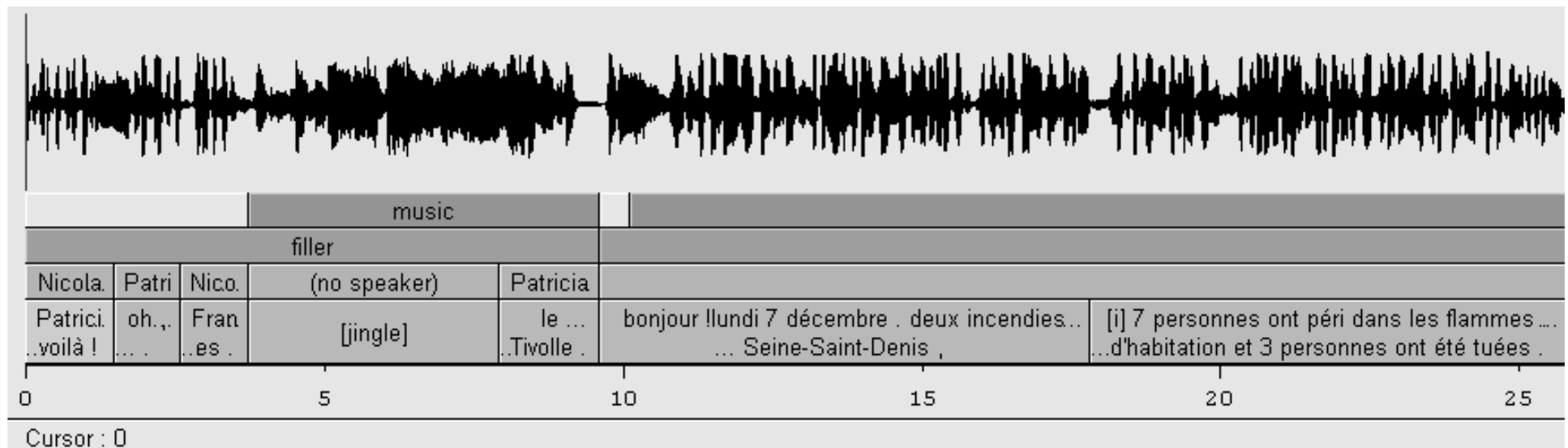
Sound event

Speaker info.

Speaker turn

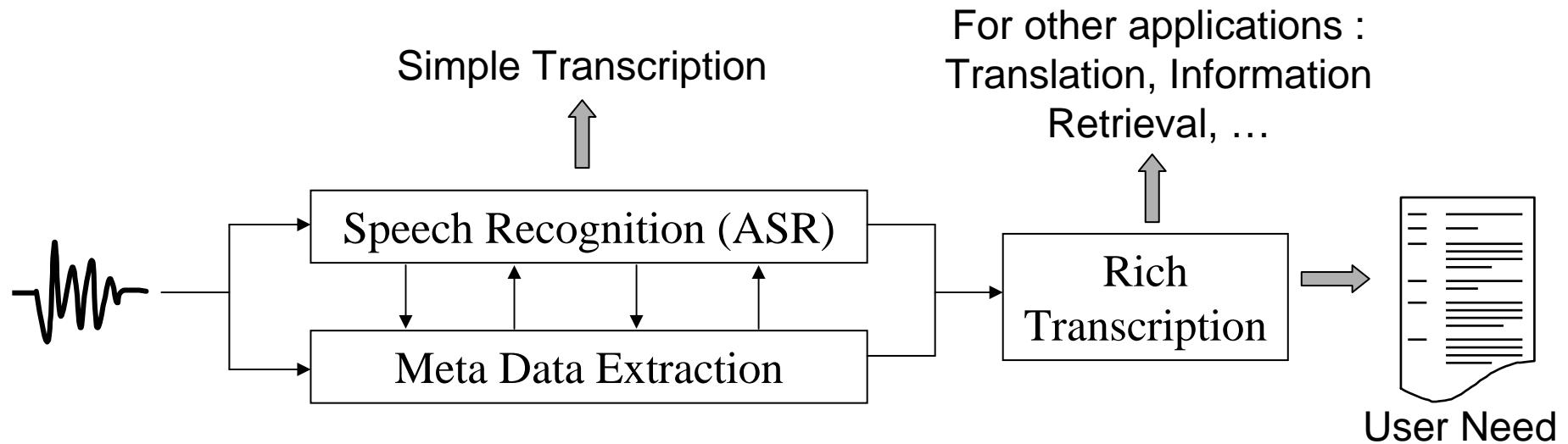
# Rich transcription

---



# Rich Transcription

---



Meta data :

- ❖ non linguistic information
- ❖ example : disfluencies, sentence breaks, **speaker turns**, **speaker labels**, **sounds**, **jingles**, laughs, **punctuation**, etc.

# Speaker Information

---

*Who spoke when ?*

❖ Speaker recognition (biometrics)

- give an identity to a speech segment (or a group of segments)

❖ Speaker segmentation (diarization) :

- segmenting an audio document into homogeneous parts which contain the voice of only one speaker
- grouping together all the segments that correspond to the same speaker

# Speaker Recognition (Biometrics)

---

- PhD on this topic defended in 1998
  - selection and localization of speaker specific information for automatic speaker recognition
- Different aspects of the domain investigated since 98
  - Robustness over telecommunication channels (wireless, VoIP)
  - Multimodal authentication systems
  - Real applications & Evaluation
- Internationally recognized expert on this topic
  - work published widely on international conferences and journals (speech communication journal, signal processing journal, applied signal proc. journal)

# Speaker Segmentation

---

- PhD Daniel Moraru (full supervision)
  - *Speaker diarization for audio and audiovisual documents.* December 2004.
- Contributions to some hard problems for the task
  - p1 : automatically estimate the number of speakers in a document
  - p2 : process an heterogeneous audio flow (speech, music, noise, etc.)
  - p3 : process spontaneous speech with speaker overlap
- Won the NIST/RT speaker segmentation task in 2002 and 2004 on meeting data (2d in 2003...)
- Application to multimedia document analysis and retrieval
- Work published in *Speech Communication Journal*

# Speaker Segmentation

---

Data	Perf (%err.) 2002	Perf (%err.) 2003	Perf (%err.) 2004
Telephone	16,58 %	-	-
Broadcast News	30,33 %	19,25 %	-
Meetings	50,20 %	-	22,6 %*
Problems solved	-	p1 + p2	p1+p3

\* Multiple microphones

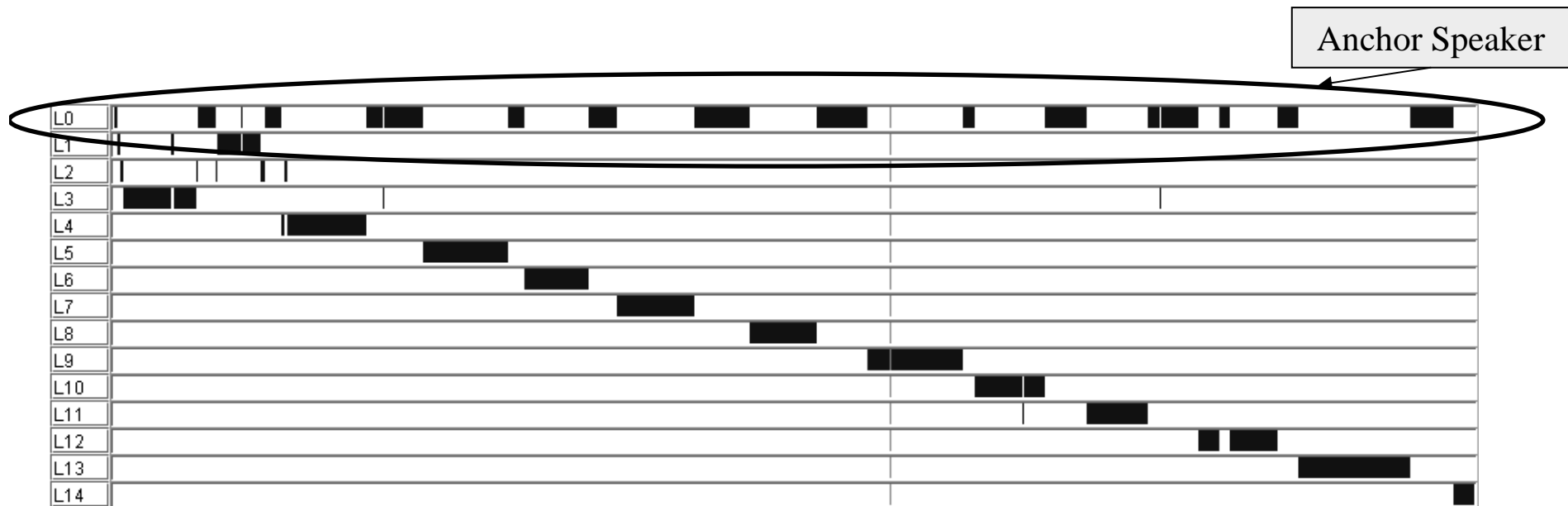
**p1 : automatically estimate the number of speakers in a document**

**p2 : process an heterogeneous audio flow (speech, music, noise, etc.)**

**p3 : process very spontaneous speech with speaker overlap**

# Application to multimedia document analysis

---



- ❖ From signal to high level information
- ❖ Structure of a Broadcast News Document

# Sound Recognition

---

- Sound detection and recognition for medical telemonitoring
  - *PhD Dan Istrate* (december 2003)
  - Migration of methods speech => sound
    - features from the speech and music analysis community, generative models (GMM, HMM)
  - Work published in *IEEE Transactions on Information Technology in Biomedicine* (2006)
  
- Key sounds detection for video analysis
  - Jingles
  - Audio signatures
  - Strong matching criteria

# Question Marks

---

- Extracting questions from a speech document
- From signal to high level information
  - Speech summarization
  - Adding punctuation to a transcription
- Features extracted from the intonation curve + Decision trees
- With / without automatic transcription
- Problem studied for tonal / non tonal languages (vietnamese / french)
- Not published yet

**Example of performance for French (234Q and 234^Q from meetings)**

<b>Approach</b>	<b>F measure</b>
<b>Prosody</b>	<b>73%</b>
<b>Transcription</b>	<b>65%</b>
<b>Combined</b>	<b>77%</b>

# Multimodality

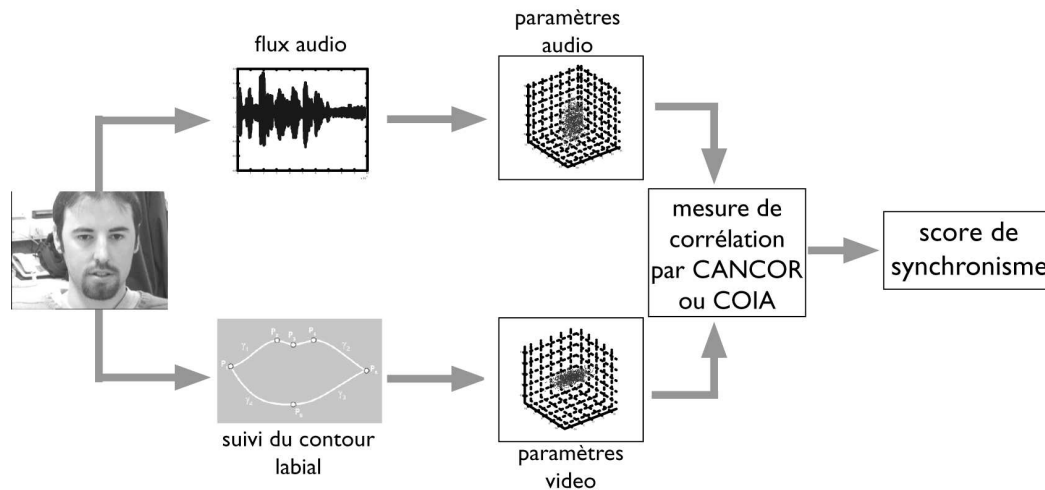
---

- Multimodal (MM) tasks and environments
  - MM biometrics
  - MM information retrieval
- 3 case studies
  - Synchronism score for multimodal biometrics
  - Audio-video signatures for video-clip detection
  - Audio-video segmentation of documents

# MM biometrics

---

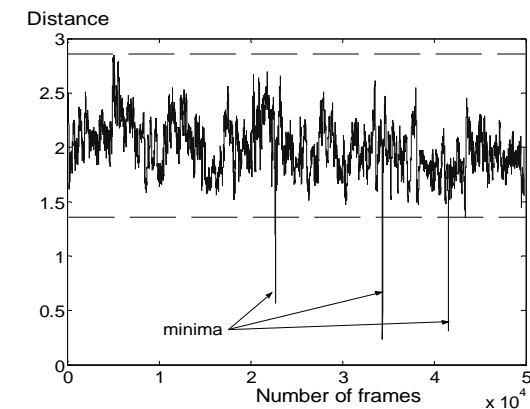
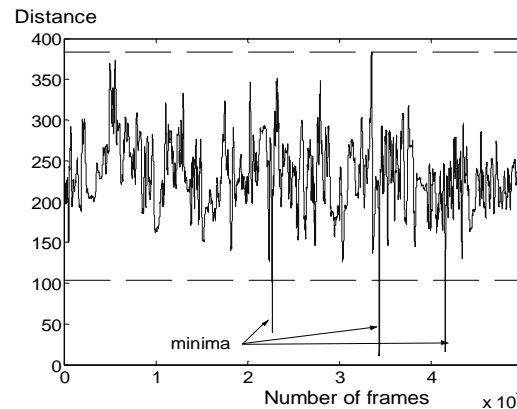
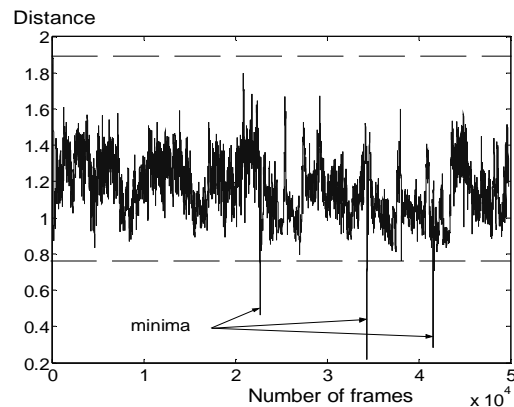
- Detect impostors and thwart replay attacks
  - Liveness test for biometric systems
- Analyzing the statistical dependency between audio and video
  - Synchronism score between lip and speech features
  - 12% EER on playback detection



# Audio-visual signatures

---

- Time step synchronized between video and audio signatures
- Normalize both audio and video distance curves
- Combine audio & video distance curves
- A+V Improves Precision / Recall on a Video clip detection task



**A**

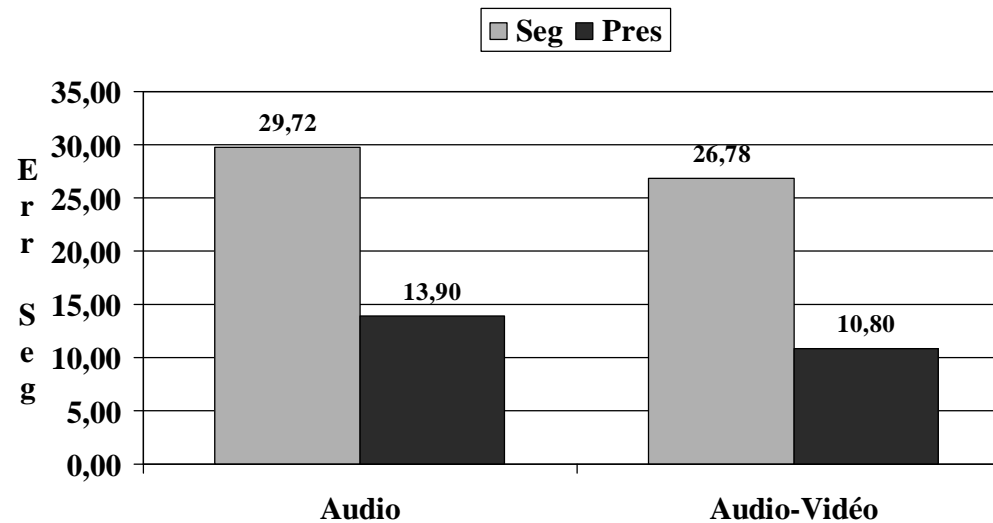
**V**

**A+V**<sub>β6</sub>

# Audio / Video Segmentation

---

- Use video information for “audio” tasks
  - Speaker diarization (Seg)
  - Anchor speaker tracking (Pres)
- Video information : automatic shot boundary detection



# Summary of part II

---

- Extraction of non linguistic informations for automatic document analysis
  - Speakers
  - Sounds
  - Jingles
  - Questions
  
- Case studies show the benefit of multimodality

# Outline of this talk

---

Part I : Domain, Problems, Methods

Part II : Beyond the transcription : speakers,  
sounds, etc...

**Part III : Automatic speech recognition  
for under-resourced languages**

Part IV : Ongoing work and future  
developments

---

# Part III : ASR for under-resourced languages

---

## **III.1 Collecting resources**

- Collecting text
- Collecting speech

## **III.2 Bootstrapping of acoustic models**

- Crosslingual Acoustic Modeling
- Application to vietnamese and khmer

## **III.3 Reduce models complexity**

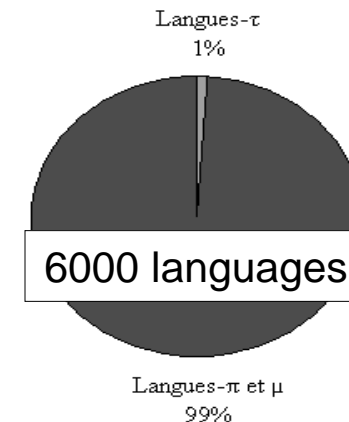
- Moving below the word level in spoken language modeling

# A Multilingual World

---

- In 2005, less than 1 % of the 6000 languages of the world have a high level of computerization, including a broad range of services going from text processing to machine translation...
  - Under-resourced languages
  - Low density languages

Cf. V.Berment thesis : «*Methods to computerize "little equipped" languages and groups of languages*»



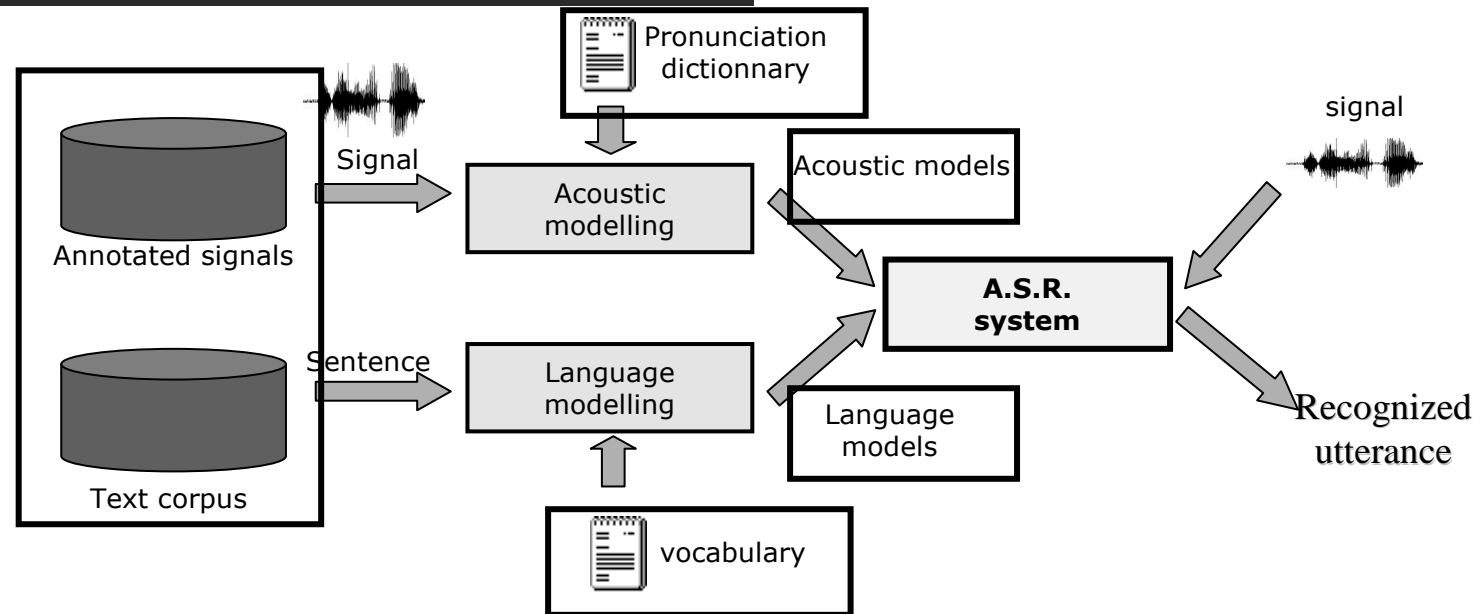
- Diversity of the written systems
- Mainly spoken (and not written) languages

# A Multilingual World

---

- Languages with low resources
  - Few data available
  - Need for innovative methods beyond simple retraining of acoustic and language models
    - Data collection methodology
    - Acoustic models bootstrapping
    - Reduce models complexity

# Resources needed for ASR : overview



- Text and Speech Corpora
- Pronunciation dictionary
- Acoustic model
- Language model

# Data collection

---

## □ Collection of text resources

- D. Vaufreydaz : *Using the **Web** for Statistical Language Modeling for ASR*. PhD University J. Fourier, Grenoble, **defended** in January 2002.
- Potential for under-resourced languages
  - Sometimes the only way to collect text data
  - But mainly news web sites
- Exemple : [www.voanews.com](http://www.voanews.com)

---

	<b>#sent</b>	<b>#words</b>	<b>#size</b>
<b>indonesian</b>	<b>116k</b>	<b>2.4M</b>	<b>17M</b>
<b>korean</b>	<b>405k</b>	<b>7M</b>	<b>67M</b>
<b>pashto</b>	<b>7k</b>	<b>0.2M</b>	<b>2M</b>
<b>kurdish</b>	<b>24k</b>	<b>0.6M</b>	<b>8M</b>
<b>hindi</b>	<b>73k</b>	<b>2M</b>	<b>28M</b>
<b>Persian(farsi)</b>	<b>212k</b>	<b>5.8M</b>	<b>54M</b>

# Data collection

---

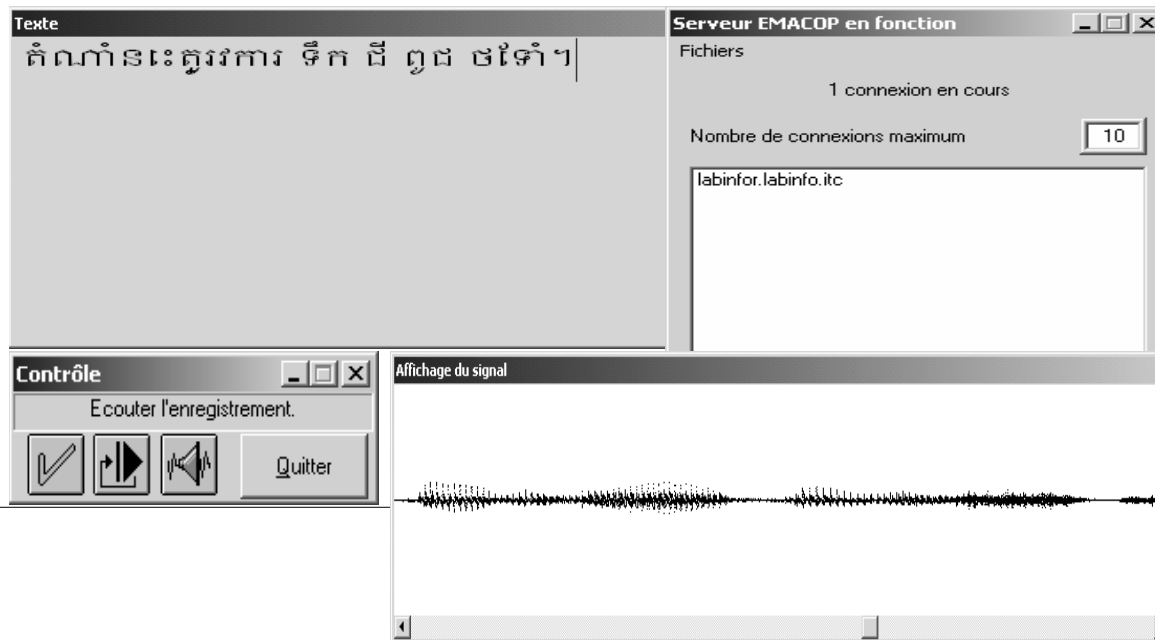
□Collection of text resources

□**Collection of speech data**

-Build local collaborations (MICA/Hanoi ; ITC/Phnom-Penh)

-Local recording with CLIPS tool EMACOP : Multimedia Environment for Acquiring and Managing Speech Corpora

-Local transcriptions of Radio or TV Broadcasts



# Bootstrapping of acoustic models

---

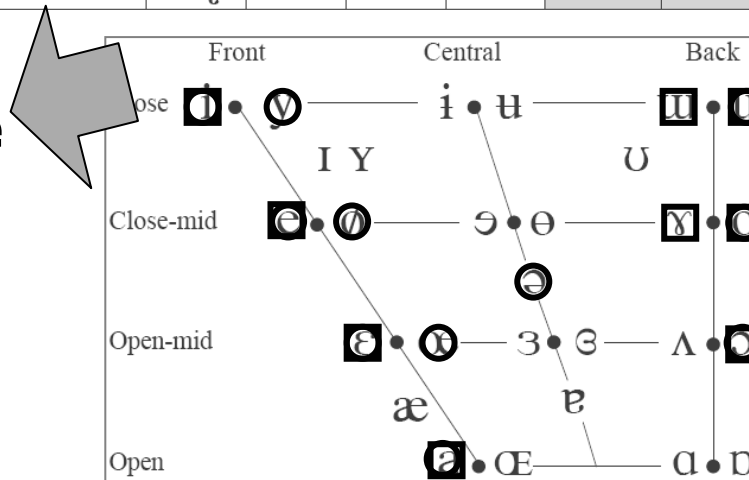
- Collection of text resources
- Collection of speech data
- Bootstrapping of acoustic models**
  - Crosslingual acoustic modeling

# Crosslingual Acoustic Modeling

	Bilabial	Labiodental	Dental	Alveolar	Post alveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	Ⓚ Ⓛ			Ⓣ Ⓝ		Ⓝ Ⓞ	Ⓚ Ⓜ	Ⓚ Ⓞ	q ɢ		ʔ
Nasal	Ⓝ	ɱ		Ⓝ		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			ʀ					ʀ		
Tap or Flap				ɾ		ɽ					
Fricative	ɸ β	ɸ ɸ	θ ð	ʃ ʒ	ʃ ʒ	ʃ ʒ	ç ʝ	x ɣ	ħ ʕ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				ɹ		ɻ	ʎ	ʟ			

○ Phoneme FR  
 □ Phoneme VN

- FR/VN ~63% coverage
  - If multiple source languages (ex: multilingual model from CMU : 7 languages)
  - => 87% coverage
- Benefit of multilingual coverage



# Crosslingual Acoustic Modeling

---

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

- Propose new phone (or polyphone) similarity measures to rapidly bootstrap acoustic models for new languages
- $\Phi_S$  and  $\Phi_T$  models in source and target language :
  - Monophones, polyphones, clustered polyphones
- $d$  : knowledge-based or data-driven distances
- V-B Le : *ASR for under-resourced languages*. PhD University J. Fourier, Grenoble, **defended** in June 2006.

# Application

---

- Collection of text resources
- Collection of speech data
- Acoustic models bootstrapping
- Application to vietnamese and khmer ASR**

ASR performance for vietnamese (%syllable accuracy)

**Dialog corpus**

Source	Distance	Adapt 1h	Adapt 2h
		WA	WA
French	Knowledge	60.4	63.6
	Data driven	61.6	63.8
Multilingual (CMU, 7 languages)	Knowledge	64.6	66.3
	Data driven	63.8	65.3

**Samemethodology applied  
to khmer : ASR system  
developed in few months  
WA=73.6% on read  
speech**

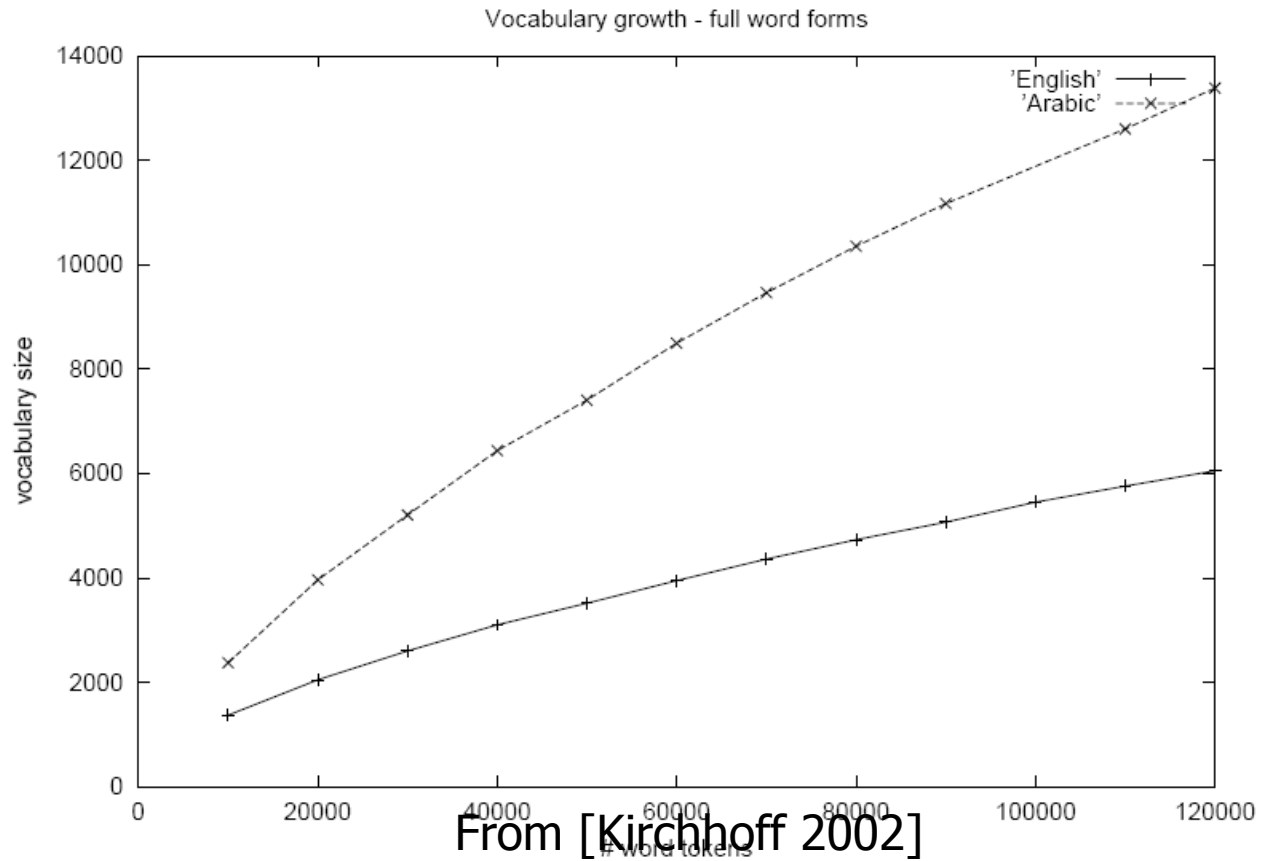
# Reduce models complexity

---

- Collection of text resources
- Collection of speech data
- Acoustic models bootstrapping
- Application to vietnamese and khmer ASR
- Reduce models complexity**
  - Moving below the word level in spoken language modeling**
  - Work done during my stay at IBM Watson (09/2005=>11/2006)
    - Dialectal arabic (Iraqi) : ASR and MT

# Exemple of standard arabic

---



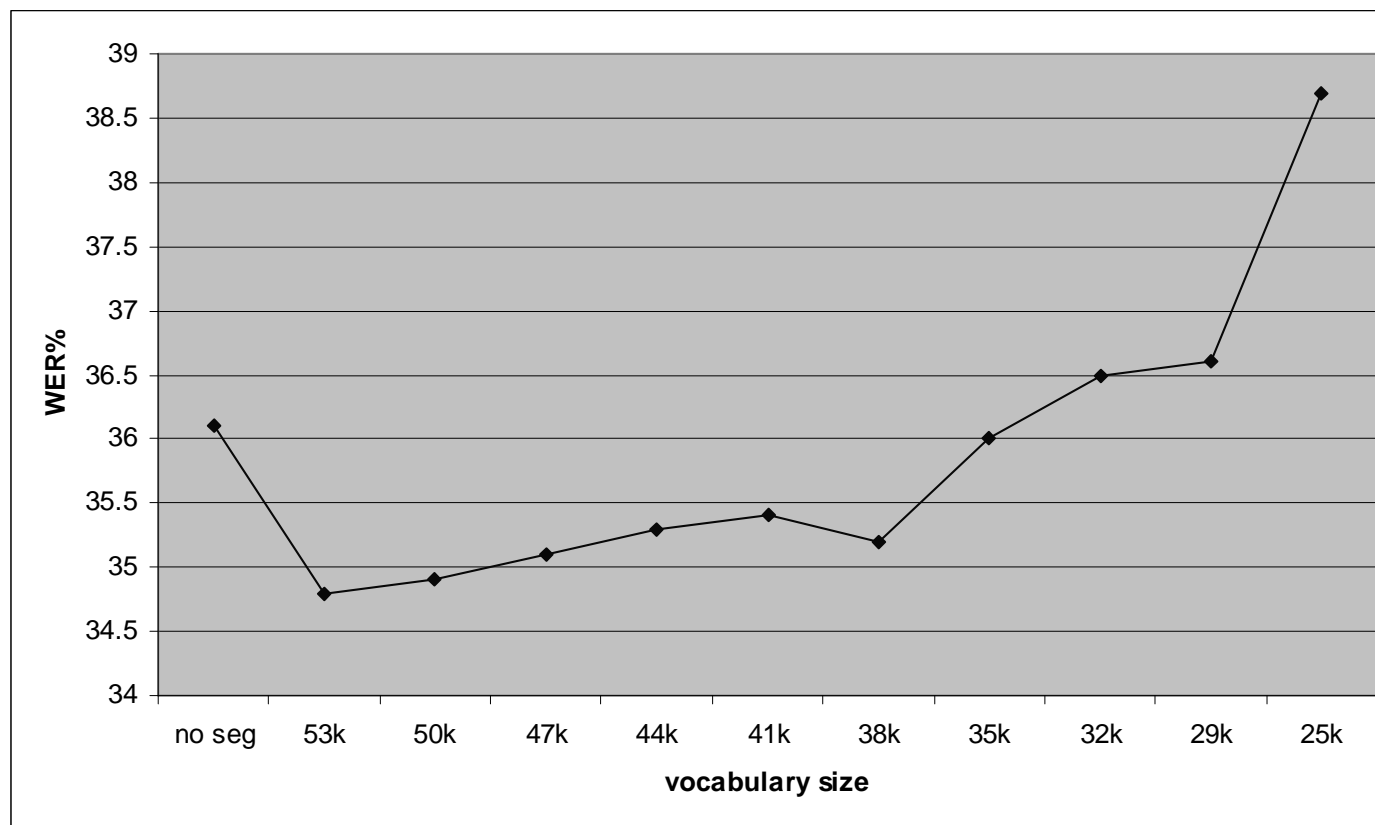
# Automatic morphological analysis for Iraqi ASR

---

- Few text data available (dialect)
- Segment words in pref-base-suff to reduce the vocabulary *إذا عندك قول لي حتى أ#روح أ#صيح اخوان+ي*
  - And consequently reduce the complexity of the LM
- Data driven approach using a segmented corpus
  - Train a model to predict the prefixes and suffixes symbols in an unsegmented stream
  - Train a morphological LM from the segmented data
  - Reparse LM data to keep top-N words and build new LM
    - Vary N => control the vocabulary size

# Results for Iraqi ASR

---

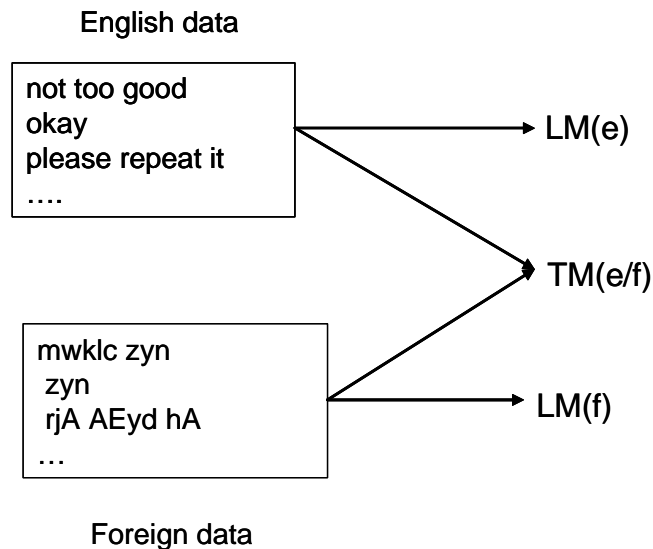




# Baseline *versus* Phone-based Approach

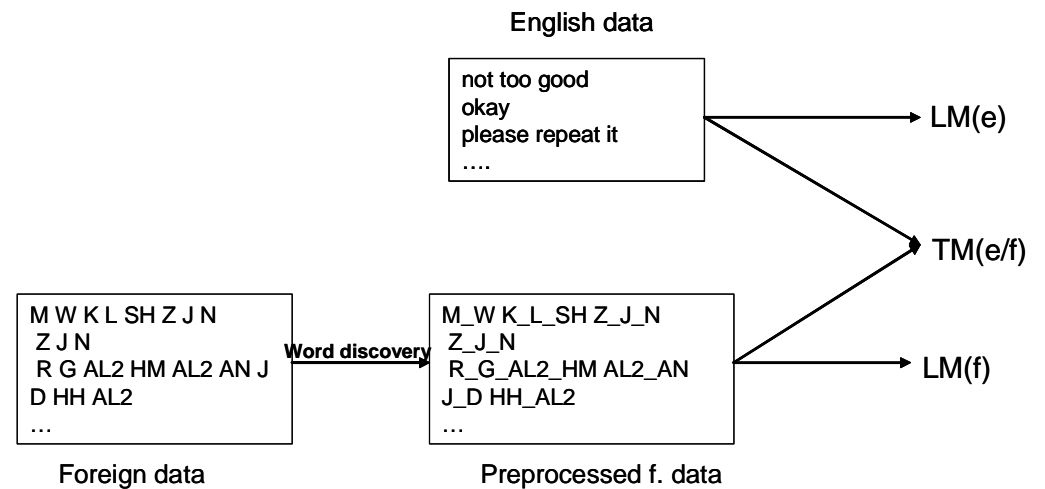
---

## □ Baseline



Foreign ASR vocab : 43k (cu2)

## □ Phone-based



Foreign ASR vocab : 36k (cu2)

# Experimental Results

---

- *Phone-based* approach equivalent in performance to the *baseline*
  - Human Evaluation (binary decisions : correct/incorrect)
  - Baseline 58% correct / Phone-based 54% correct
- Possible to overpass the written form *f* for S2S
  - Potential to reduce the vocabulary while keeping the same coverage
  - Details in « Towards speech translation of non written languages» **Laurent Besacier**, Bowen Zhou, Yuqing Gao. IEEE / ACL SLT 2006. Aruba, December 2006.

# Summary of Part III

---

## **ASR for under-resourced languages**

- Collection of text resources
- Collection of speech data
- Bootstrapping of acoustic models
- Application to vietnamese and khmer ASR
- Reduce models complexity
  - Moving below the word level in spoken language modeling

# Outline of this talk

---

Part I : Domain, Problems, Methods

Part II : Beyond the transcription : speakers, sounds, etc...

Part III : Automatic speech recognition for under-resourced languages

**Part IV : Ongoing work and future developments**

---

# Part IV : Ongoing work and future developments

---

## **IV.1 Multimodality**

- Ongoing projects
- Future

## **IV.2 Multilingual Aspects**

- Generalized cross-lingual acoustic modeling
- Multilevel spoken language modeling
- Future

# Multimodality : ongoing projects

---

- Multimodal gesture recognition for cued speech
  - Complement of lip-reading for hearing-impaired people
  - Project ANR TELMA (phone for hearing impaired people)
  - PhD of Nourredine Aboutabit (co-supervision)
- Multimodal Biometrics
  - Project ANR MISTRAL (start in 2007)
  - Thales, Calistel, LIA, IRIT, CLIPS, LIUM
  - Open source framework for biometric authentication
  - Migration methods speaker ID => other modalities (signatures, faces, ...)

# Multimodality : future

---

- Different feature extractors available for automatic document analysis
  - Transcription, speakers, sounds, jingles, questions
- Futur : multimodal scene analysis or multimedia document analysis
  - Consolidate / add new feature extractors
  - Introduce mathematical tools to combine these elementary descriptors
  - Experiment on tasks like multimedia information retrieval

# Multilinguism : generalized crosslingual acoustic modeling

---

- phone (polyphone) similarity measures
  - Already applied to acoustic modeling for under-resourced language (bootstrapping)
  - Potential for other tasks
    - Define distances between languages (between phone sets)
    - **ASR on non native speech**
    - Language identification

# Multilinguism : ASR on non native speech

---

- PhD Tien-Ping Tan
- Use speaker's mother tongue (L1) to improve the target language (L2) acoustic models
  - find non-native speaker's phone substitutions
  - use native speaker speech data for adapting acoustic models
    - "Acoustic model interpolation for non-native speech recognition", T.P. Tan, L. Besacier, Accepted to ICASSP 2007
  - Calculate pronunciation scores for rescoreing nbest lists (or include a pronunciation score into the decoding process)

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)

# Multilinguism : multilevel spoken language modeling

---

- Different units (levels) for spoken language modeling
  - words, syllables, characters, ...
- Joint use of these different levels
  - modeling, decoding
- Exemple :

Hôm nay, chúng tôi đến trường bằng xe hơi.

**syllables**

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

**characters**

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

**words**

# Multilinguism : future

---

- Multilinguism = core activity
  - New GETALP team in the laboratory
  - Multilingual ASR
  - Non native speakers
  - Formalization of the crosslingual acoustic modeling concept
  - Speech-to-speech translation
    - B. Zhou, L. Besacier, Y. Gao "On efficient coupling of ASR and SMT for speech translation", accepted to ICASSP 2007.
- Projects planned about under-resourced languages
  - French ANR or Europe
  - Translation assistants involving under-resourced languages

