Reconnaissance automatique de la parole pour les langues peu dotées

Laurent BESACIER, Viet-Bac LE

LIG/GETALP (Grenoble, France)

PLAN

1 Collecte de données

- Collecte de données textuelles
- Collecte de parole

2 Amorçage des modèles acoustiques

- Modélisation acoustique translingue
- Application au vietnamien et au khmer

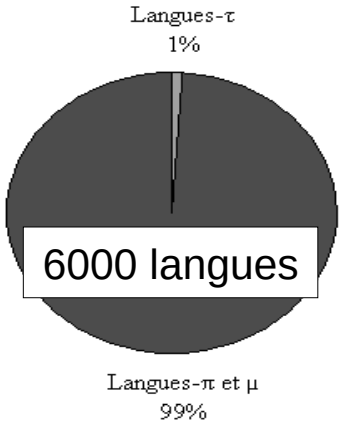
3 Réduction de la complexité des modèles

- 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

Exemple du khmer

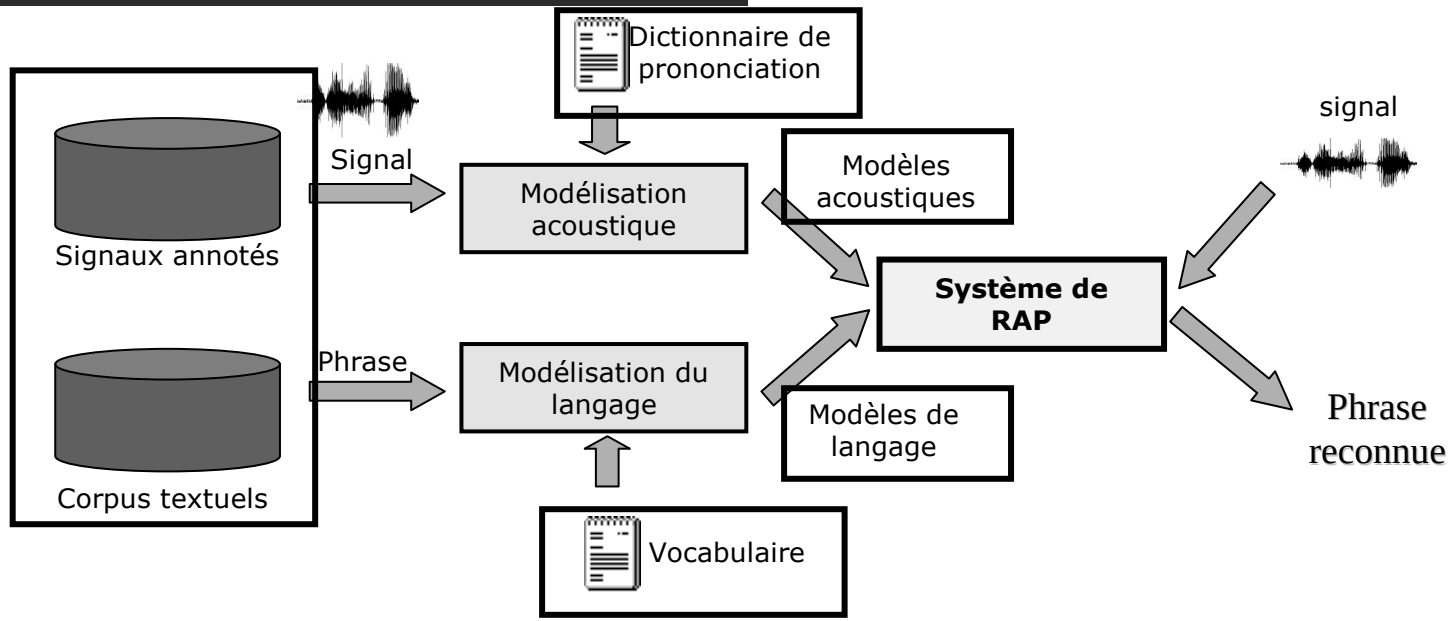
de [Berment, 2004]

	Services / ressources	Importance (/10)	Mark (/20)	Weighted mark (Importance x Mark)
Text processing				
	Basic input	10	16	160
	Visualization / printing	10	14	140
	Search and Replace	8	12	48
	Text selection	6	12	72
	Lexicographical sort	5	0	0
	Spelling Correction	2	0	0
Speech processing				
	Text-to-speech	5	0	0
	Automatic Speech Recognition	**5**	**0**	**0**
Translation				
	Tools for Automatic translation	8	4	32
OCR				
	Optical Character Recognition	9	0	0
Ressources				
	Bilingual dictionnary	10	4	40
	Usability dictionnary	10	0	0
Total				540 / 1760
Mean				31 / 100

Autre exemple : Vietnamien

	Services / ressources	Importance (/10)	Mark (/20)	Weighted mark (Importance x Mark)
Text processing				
	Basic input	10	16	160
	Visualization / printing	10	16	160
	Search and Replace	8	17	136
	Text selection	6	17	102
	Lexicographical sort	5	6	30
	Spelling Correction	2	6	12
Speech processing				
	Text-to-speech	5	0	0
	Automatic Speech Recognition	**5**	**0**	**0**
Translation				
	Tools for automatic translation	8	6	48
OCR				
	Optical Character Recognition	9	12	108
Ressources				
	Bilingual dictionnary	10	13	130
	Usability dictionnary	10	0	0
Total				886 / 1760
Mean				50 / 100

Ressources nécessaires pour la RAP



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

Collecte de données

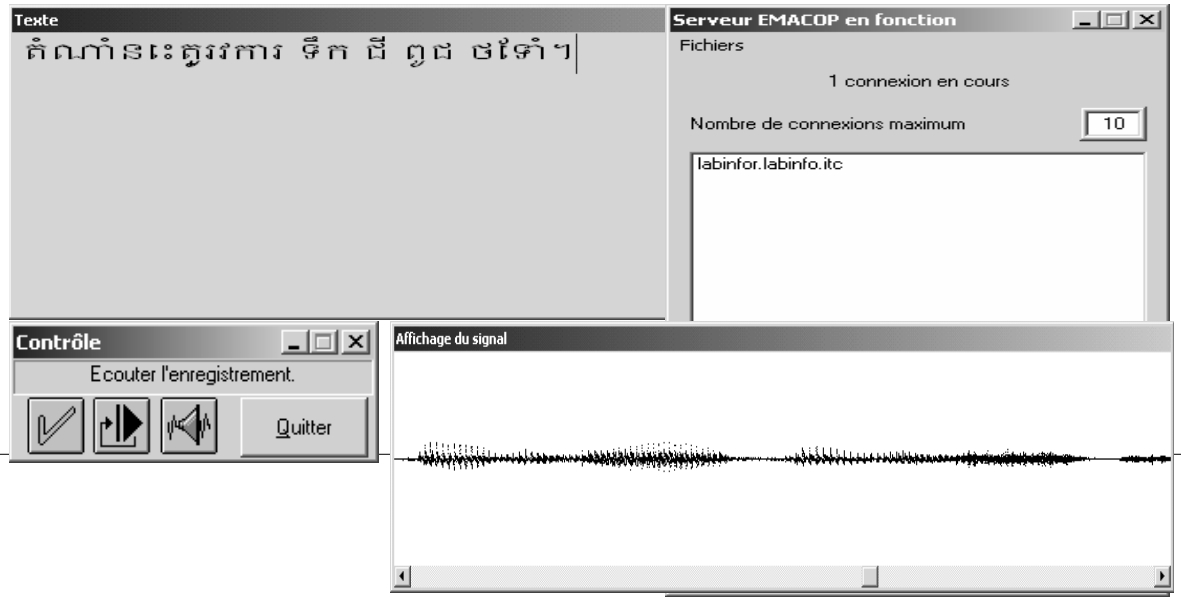
- **Collecte de données textuelles**
  - D. Vaufreydaz : *Modélisation statistique du langage à partir d'Internet pour la reconnaissance automatique de la parole continue*. Doctorat de l'Université J. Fourier, **thèse soutenue** en janvier 2002 .
  - Potentiel pour les langues peu dotées
    - Web parfois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

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

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

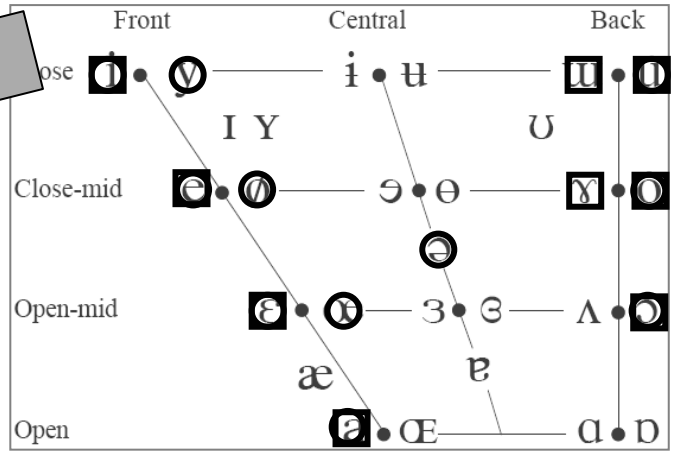
- Collecte de données textuelles
- Collecte de parole
- **Amorçage des modèles acoustiques (bootstrap)**
  - Modélisation acoustique translingue

Modélisation acoustique translingue

	Bilabial	Labiodental	Dental	Alveolar	Post alveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k ɡ	q G		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	N		
Trill	B			r					R		
Tap or Flap				ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

○ 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

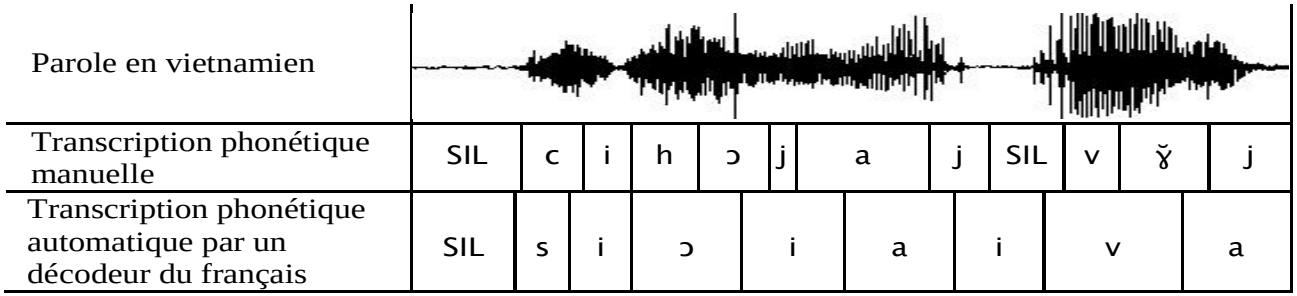
Modélisation acoustique translingue

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

- Proposition de nouvelles mesures de similarité entre phonèmes (ou polyphonèmes) pour l'amorce (*bootstrap*) rapide des modèles acoustiques dans une nouvelle langue
- Φs et Φt : 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.

Distance entre deux phonèmes : méthode automatique *(data-driven method)*

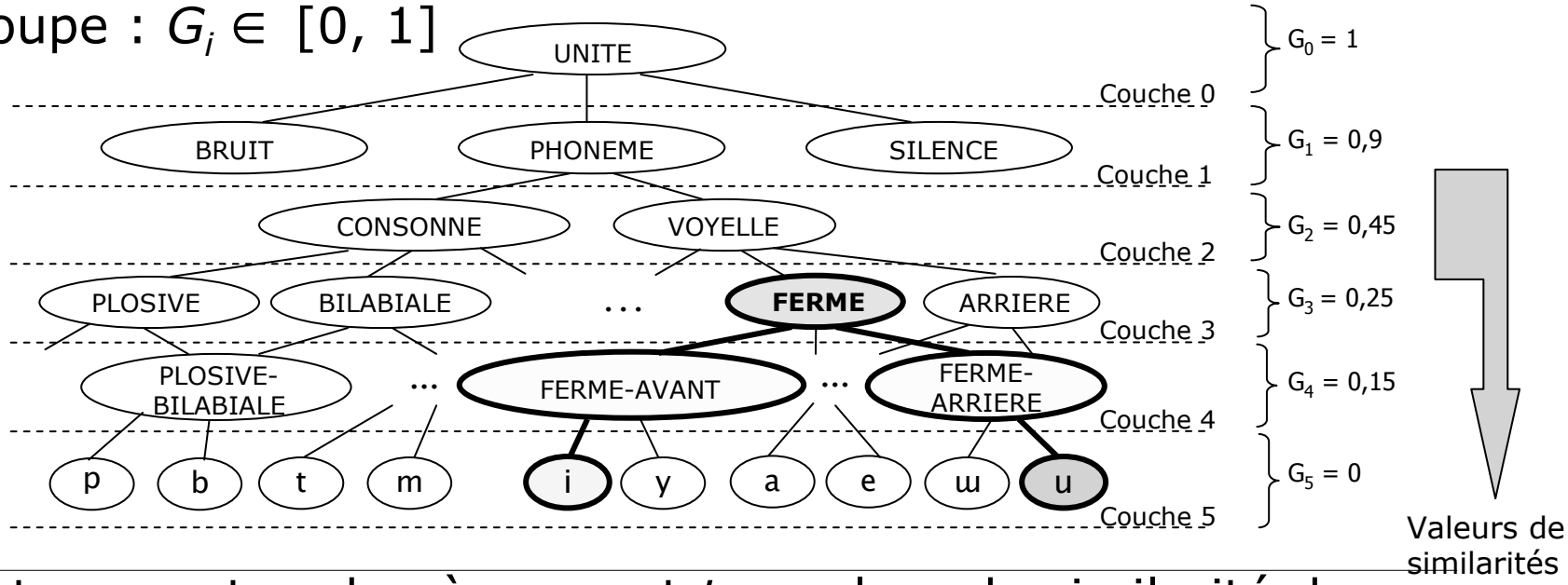
- Besoin d'un corpus vocal étiqueté en phonèmes en quantité limitée
- Utiliser un décodeur acoustico-phonétique en langue source
- Calculer une matrice de confusion (phonèmes source/cible)
- Développer un outil d'estimation automatique de distance entre deux phonèmes
  - PESCORING (Time-based source/target phoneme scoring)



Soit $A(M, N)$: matrice de confusion de phonèmes,
avec $0 \leq A_{i,j} \leq 1$ $d(s_i, t_j) = A_{i,j}$ où $A_{i,j} \in [0, 1]$ et $i=1..M$, $j=1..N$

Distance entre deux phonèmes : méthode à base de connaissances phonémiques (API)

- Construire un graphe hiérarchique où chaque nœud est attaché à un groupe de phonèmes
- Chaque groupe de phonème est assigné à une valeur de similarité qui représente la similarité des éléments dans ce groupe : $G_i \in [0, 1]$



- Distance entre phonèmes *s* et *t* = valeur de similarité du nœud père le plus proche de *s* et *t*

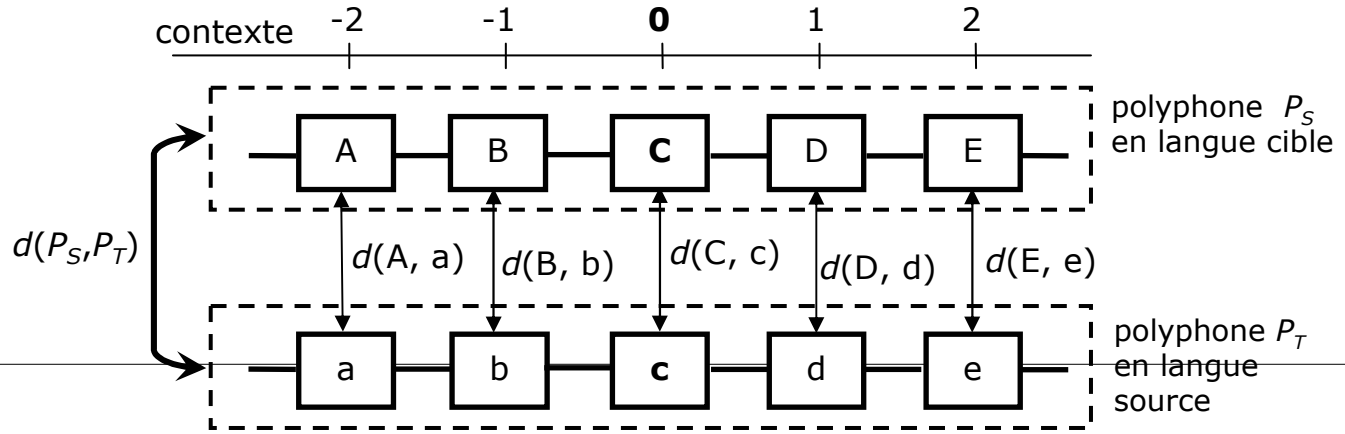
Distance entre deux polyphones

- *Distance entre le polyphone PS en langue source et le polyphone PT en langue cible est calculée à partir des distances entre les phonèmes source/cible dans les contextes correspondant*

$$d(P_S, P_T) = \propto_0 .d(s_0, t_0) + \propto_1 .[d(s_{-1}, t_{-1}) + d(s_1, t_1)] + \ldots + \propto_L .[d(s_{-L}, t_{-L}) + d(s_L, t_L)]$$

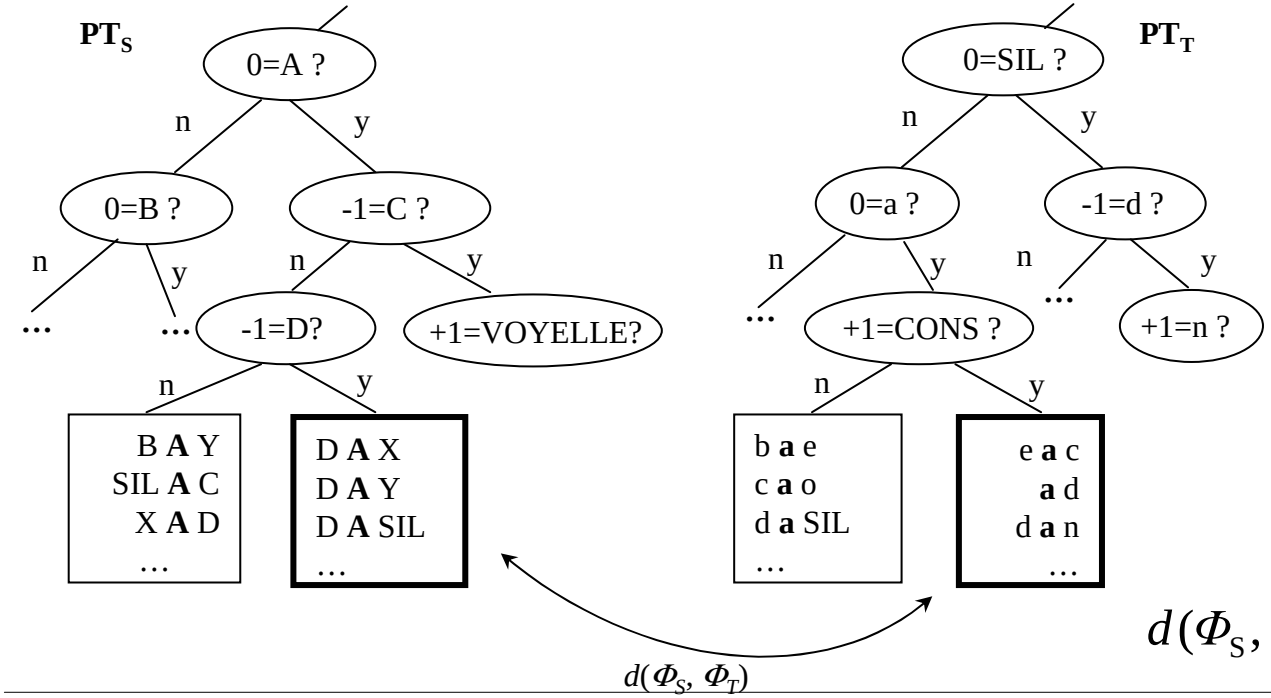
avec : - $\propto_0, \propto_1, \ldots, \propto_L$: coefficients de distance suivant la position dans le contexte

- $d(s_k, t_k)$: la distance entre deux phonèmes pour $k = -L, \ldots L$



Distance entre deux groupes de polyphones

- Distance entre deux groupes de polyphones est égale à la moyenne des distances entre tous les couples de polyphones composant les deux ensembles Φ_T et Φ_S



$$d(\Phi_S, \Phi_T) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} d(P_{Si}, P_{Tj})}{m.n}$$

Modélisation acoustique à base de graphèmes (1/2)

- Génération d'un dictionnaire de prononciation
  - Fractionner un mot en une suite de graphèmes (caractères)
  - Romaniser des caractères : convertir les caractères en une forme plus simple (ASCII)

Caractère khmer	Romanisation
ក	Ka
ខ	KHa
ឥ	QI
ៀ	YA

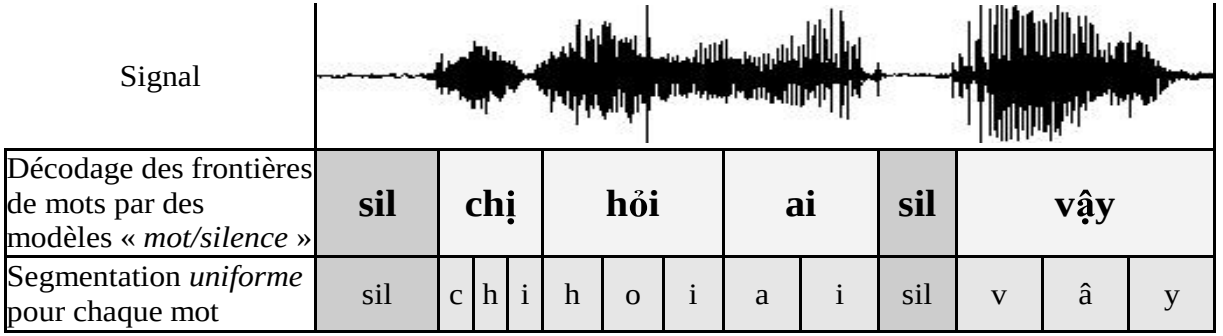
Romanisation

Mot khmer	Prononciation
កកិត	Ka Ka I Ta
កកិល	Ka Ka I Lo
តោងទាម	Ta OO NGo To AA Mo
តោតតួង	Ta OO Ta Ta UU Ngo

Dictionnaire de prononciation

Modélisation acoustique à base de graphèmes (2/2)

- Initialisation de modèles acoustiques graphémiques
  - Démarrage aléatoire, démarrage uniforme, …
  - Segmentation uniforme pour tous les graphèmes [Killer 2003]
  - **Méthode proposée :** initialisation des modèles acoustiques en utilisant une détection de frontières des syllabes



- Modélisation acoustique graphémique dépendante du contexte :
  - Méthode de « singleton » : chaque question linguistique consiste en un seul graphème
  - Méthode à base de relation *graphème-phonème*

Application

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

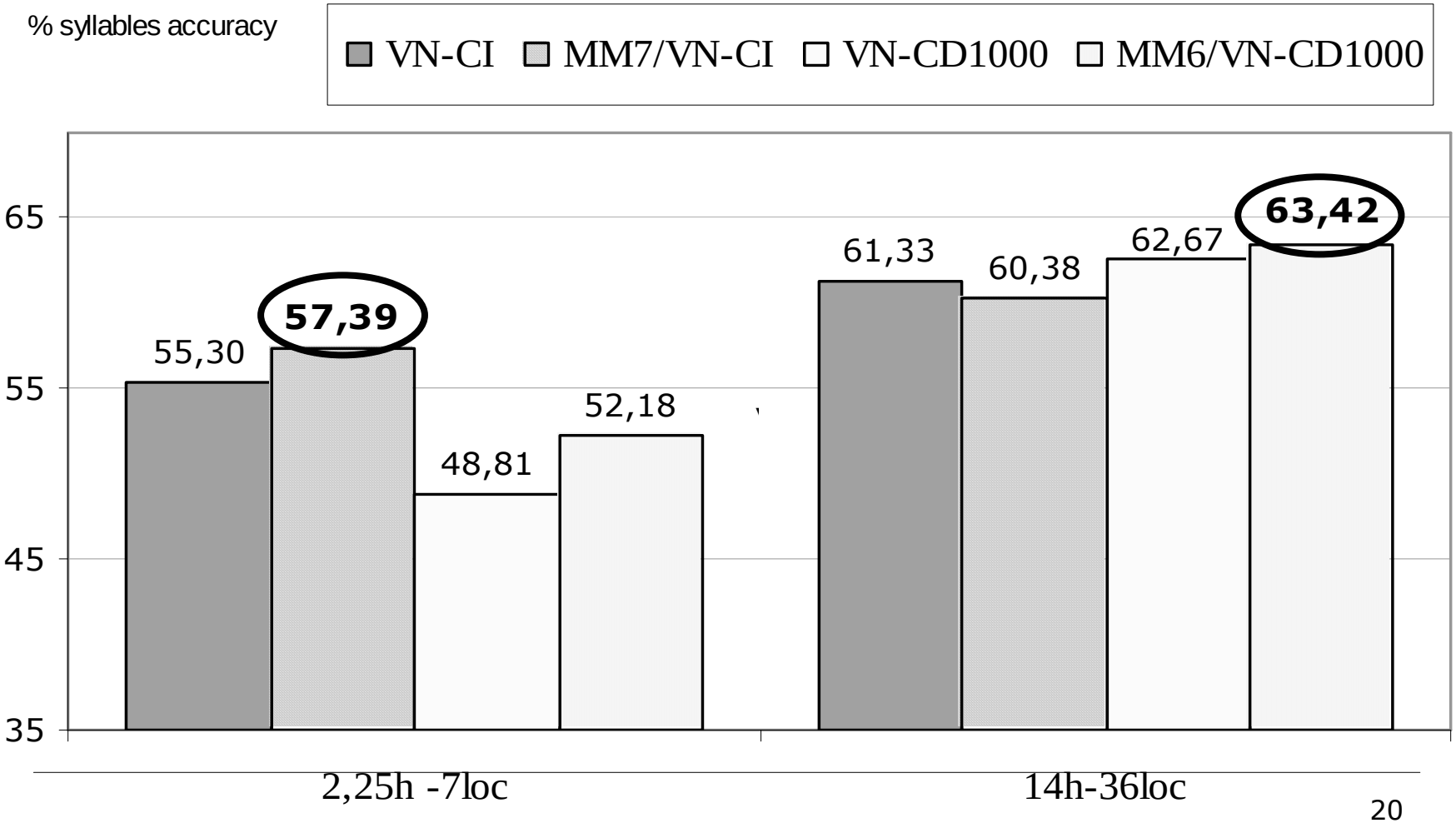
Performance de RAP pour le vietnamien (% syllabes correctes)
Corpus de dialogue

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

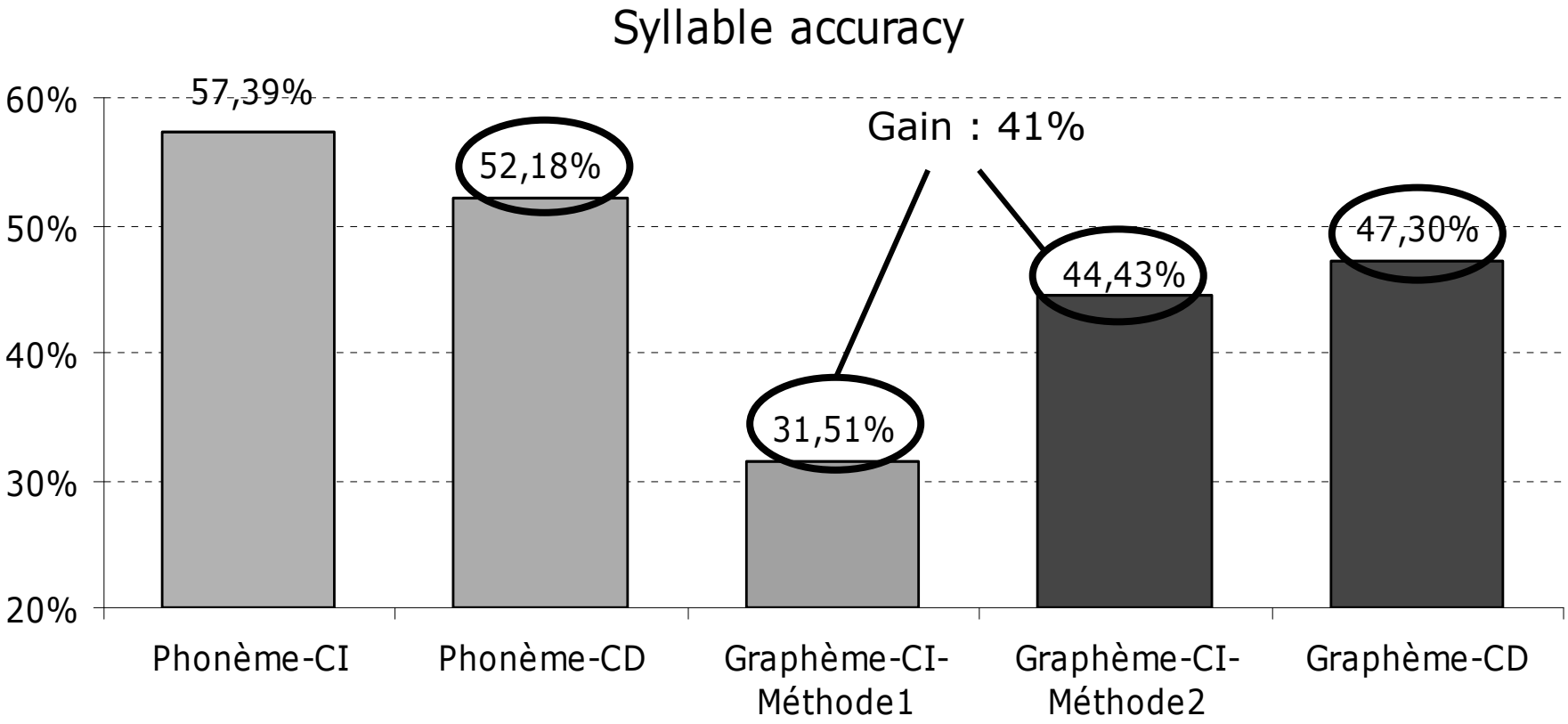
Même méthodologie appliquée au khmer : système de RAP développé en quelques mois : WA=73.6% sur des phrases lues

chi,cro,fr,ge,
jap,esp,turc

Résultats détaillés (Vietnamien)



Résultats détaillés (Vietnamien)



- Apprentissage : 2,5 heures – 7 locuteurs

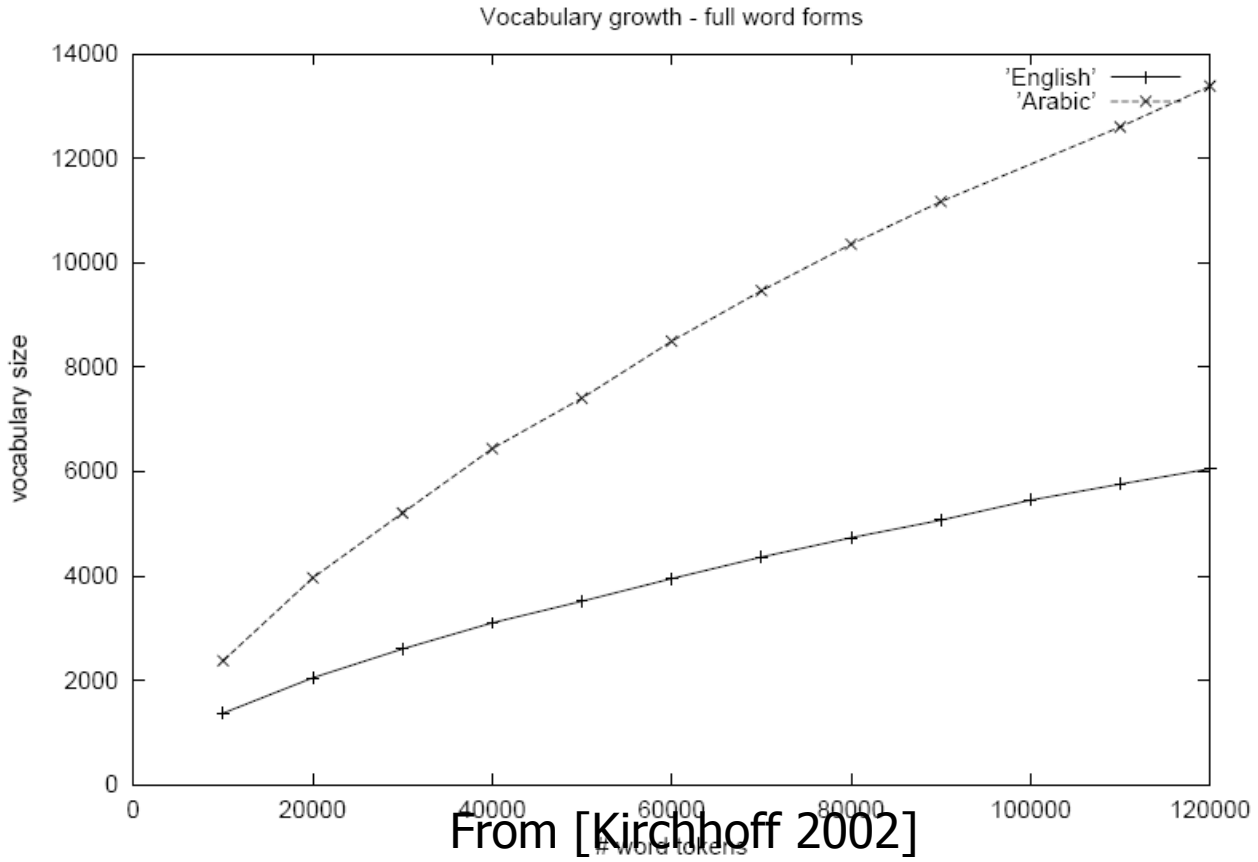
Méthode 1 : Segmentation uniforme pour tous les graphèmes [Killer 2003]

Méthode 2 : Initialisation des MA en utilisant une détection de frontières des syllabes.

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

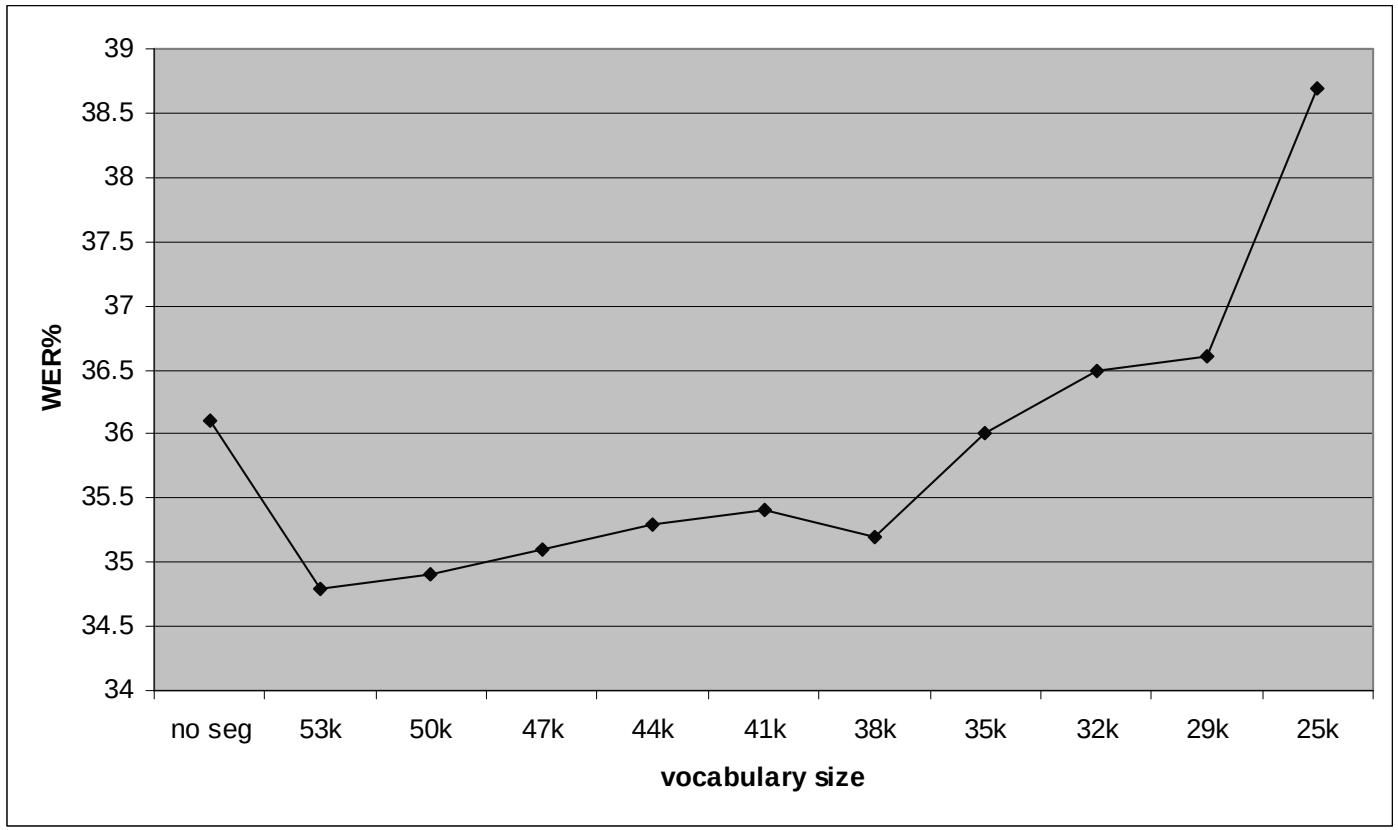


From [Kirchhoff 2002]

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 حتّى 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

$$\hat{e} = \arg\max_{e} P(e/x) = \arg\max_{e} \sum_{f} P(e, f/x)$$

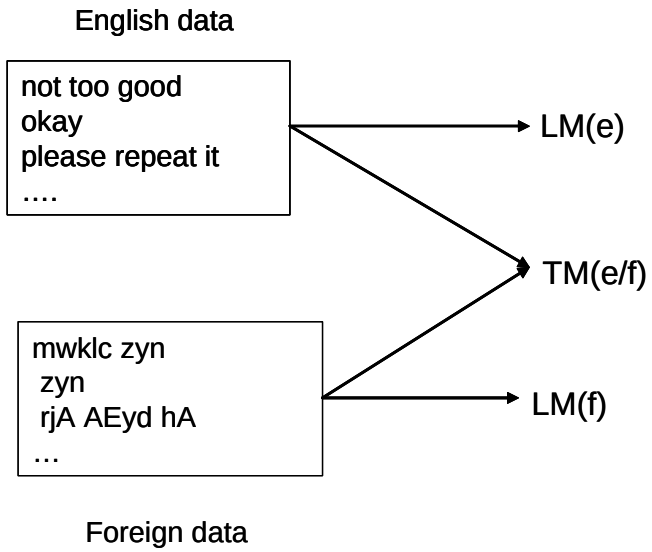
$$\approx \arg\max_{e} \sum_{f} P(e/f) P(f/x)$$

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

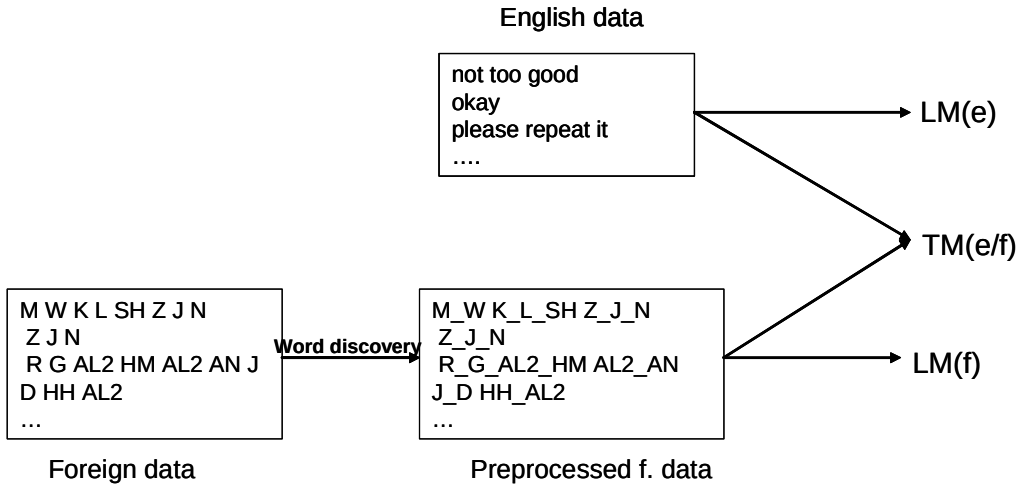
Mots *versus* Phonèmes

☐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

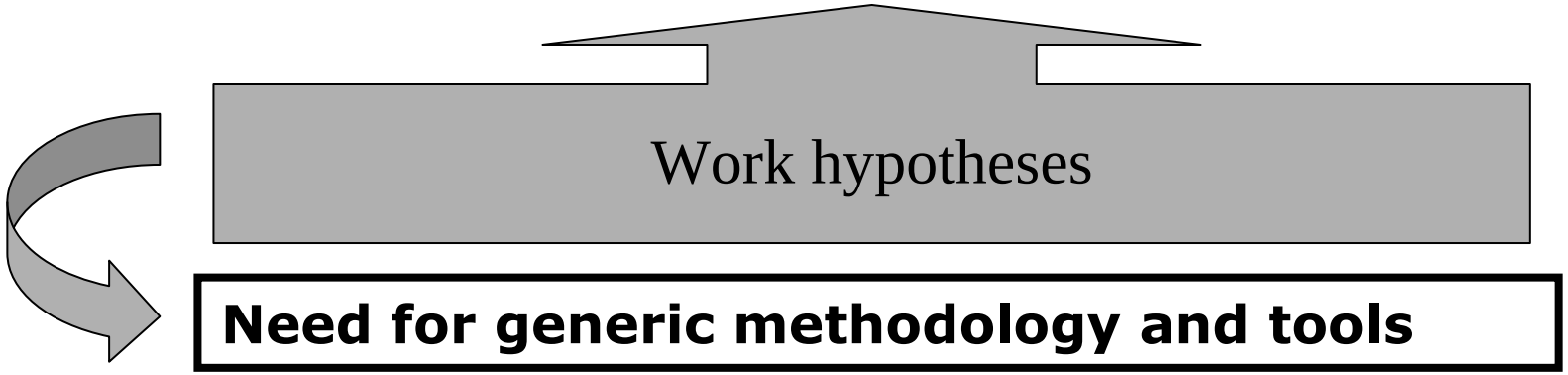
- **Contributions à la reconnaissance automatique de la parole pour les langues peu dotées**
  - Collecte de données textuelles
  - Collecte de parole
  - Amorçage des modèles acoustiques (bootstrap)
  - Application au vietnamien et au khmer
  - Réduction de la complexité des modèles
    - Langues peu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

Introduction (1)

- 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**
- A large number of languages in the world do not have an acknowledged written form and only 5 10% of all languages use one of about 25 writing systems
- => Obvious problems for ASR and MT

Introduction (2)

- No or few corpora (text or signal)
- No or few available informations on the language (linguistic or phonetic descriptions)
- No or few standards (character coding, IPA, …)
- …



Automatic morphological analysis for Iraqi ASR

- Different from a knowledge based approach were prefixes and suffixes of the language are known a priori
- Iraqi different from MSA : more prefixes and suffixes are informally introduced => Buckwalter morphological analysis did not help
- Data driven approach using a segmented corpus
  - Use segmented data provided by LDC (100K words)
  - Ex: Endy_bw hAy_bw Al#qTE+p_bw I#ryd_bw I#nTy_bw
  - Train a model to predict the prefixes and suffixes symbols in an unsegmented stream
  - Implementation using FSM toolkit (char.based algorithm)

FSM-based Segmentation

- Completely data-driven
  - No list of prefixes and suffixes specified.
- Train a 5-gram character LM
  - 100K words of manually segmented data provided by LDC.
  - Represent LM as an FSM.
- Create all possible segmentations of a word
  - Word "abcd" $\Rightarrow$ {abcd, a#bcd, ab#cd, abc#d}.
  - Represent alternatives as an FSM.
- Start from un-segmented LM data
  - Segment each sentence using the model
    - Compose both FSMs and search for best path.
  - Reparse LM data to keep top-N words and build new LM
- 95.4% word correct on a hand-labeled test set.
  - Words might have multiple segmentation.
- Decoded output has segmented words
  - Blind gluing. Mark prefixes and suffixes with Al# and

ASR Architecture

- Acoustic model
  - Nine frames of 24-dimensional cepstra reduced to 40 dimensions using LDA+MLLT.
  - 33 grapheme models (3 states per phone)
    - Vowelization didn't work well for dialect.
  - States are clustered using decision trees.
  - GMM built for each leaf.
    - Trained using ML and refined using MPE.
  - Rank models on top of the GMMs.
- Language model
  - Trigram LM built using deleted interpolation.
- The decoder uses a stack-based search algorithm.

ASR: Data and Models

- Acoustic model
  - About 200 hours of training data.
  - About 2K leaves and 60K Gaussians.
- Lexicons
  - About 90K unique words in corpus.
  - Baseline LM : Un-segmented&cutoff=1:
  - Morph. LM : after automatic segmentation, vary N (more frequent words that are kept unsegmented).
- LM data of size 1.5M words
  - Trigram built using deleted interpolation with 10% held-out data.
- Iraqi Test set has 15K words and 1.5 hours of speech.

Results for FSM Approach
Iraqi ASR