Are Audio or textual training data more important for ASR in less-represented languages?

Thomas Pellegrini, Lori Lamel

LIMSI-CNRS, UNIVERSITÉ PARIS SUD
Spoken Language Processing Group

May 5th 2008

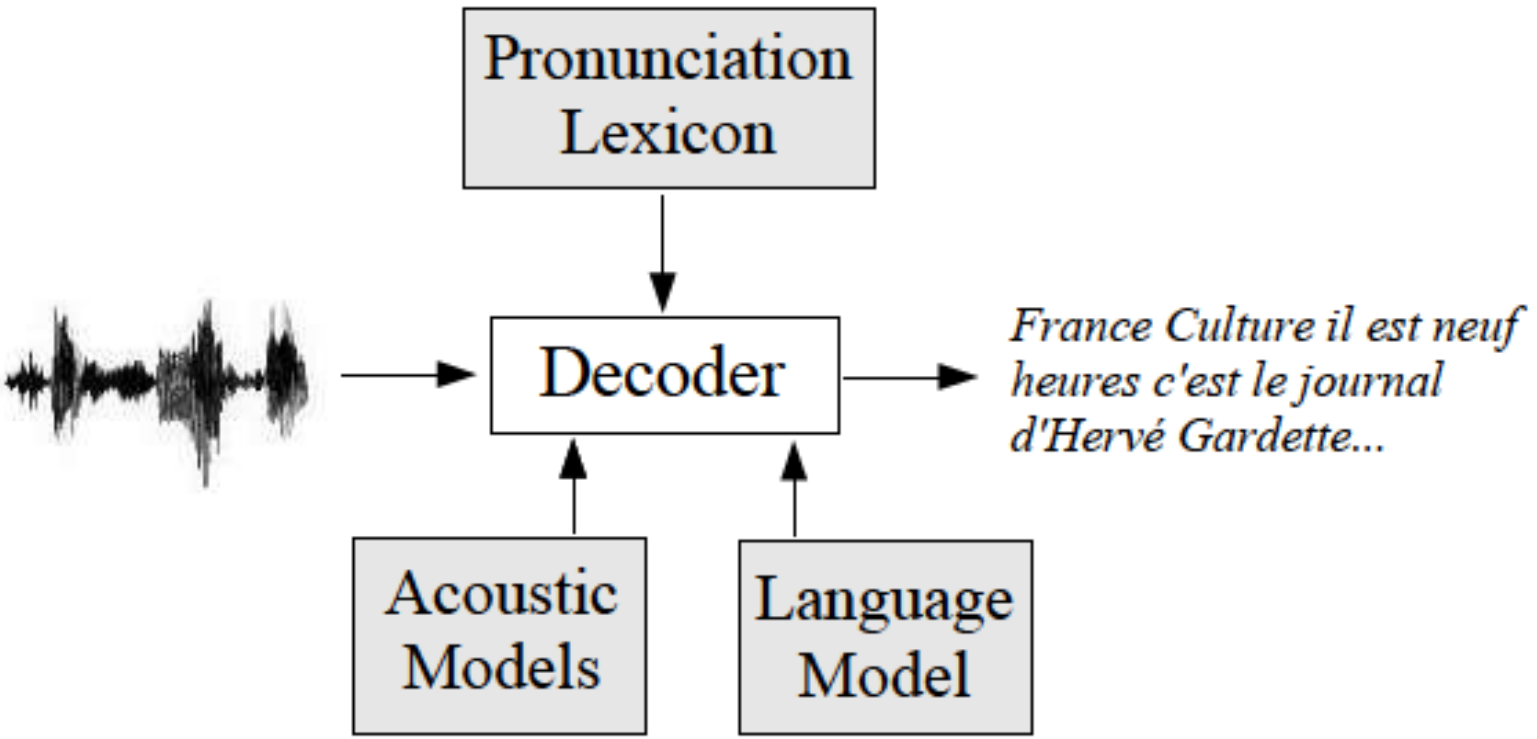
Introduction

- ASR: Automatic Speech Recognition
- More than 30 years of research
- Primary applications: dictation, spoken language dialogue, and transcription for information archival and retrieval systems
- State-of-the-art systems for about 10 languages

$\longrightarrow$ Focus: ASR for less-represented languages

Outline

Bird's eye view of Automatic Speech Recognition



State-of-the-art overview

- NIST Evaluations (1992-)

Task	*Word Error Rate (WER in %)*
Read speech (1992-1995)	7
Broadcast news (1996-)	[10-20]
Conversational speech (1997-)	[20-70]

For which languages?

English, Mandarin, Arabic, Spanish

Outline

Quantitative definition

Well-represented languages

Example : American English

Transcribed Audio: hundreds (thousand) hours

Texts: hundreds of million words (3000 M)

Less-represented languages

Example: Amharic

Transcribed Audio: a few hours (37h)

Texts: a few million words (5M)

Questions

- How much data is needed to train the models?
- What performance can be expected with a given amount of resources?
- Are Audio or textual training data more important for ASR in less-represented languages (languages with texts available)?

Impact on the WER

1. Speech material (acoustic modeling)
2. Texts (transcriptions)
3. Texts (newspapers, newswires available on the Web)

- Case study: Amharic, Broadcast news transciption

Outline

Corpus sizes

- Audio

Source	*Train*	*Devtest*
Deutsche welle	24h06	1h20
Radio Medhin	11h08	0h37
# Speakers	100	15
# Words	233k	14.1k

- Texts

Type	*# Words*	*Distinct Words*
Web	4.6M	200k
Transcriptions	247.1k	50k

Configurations

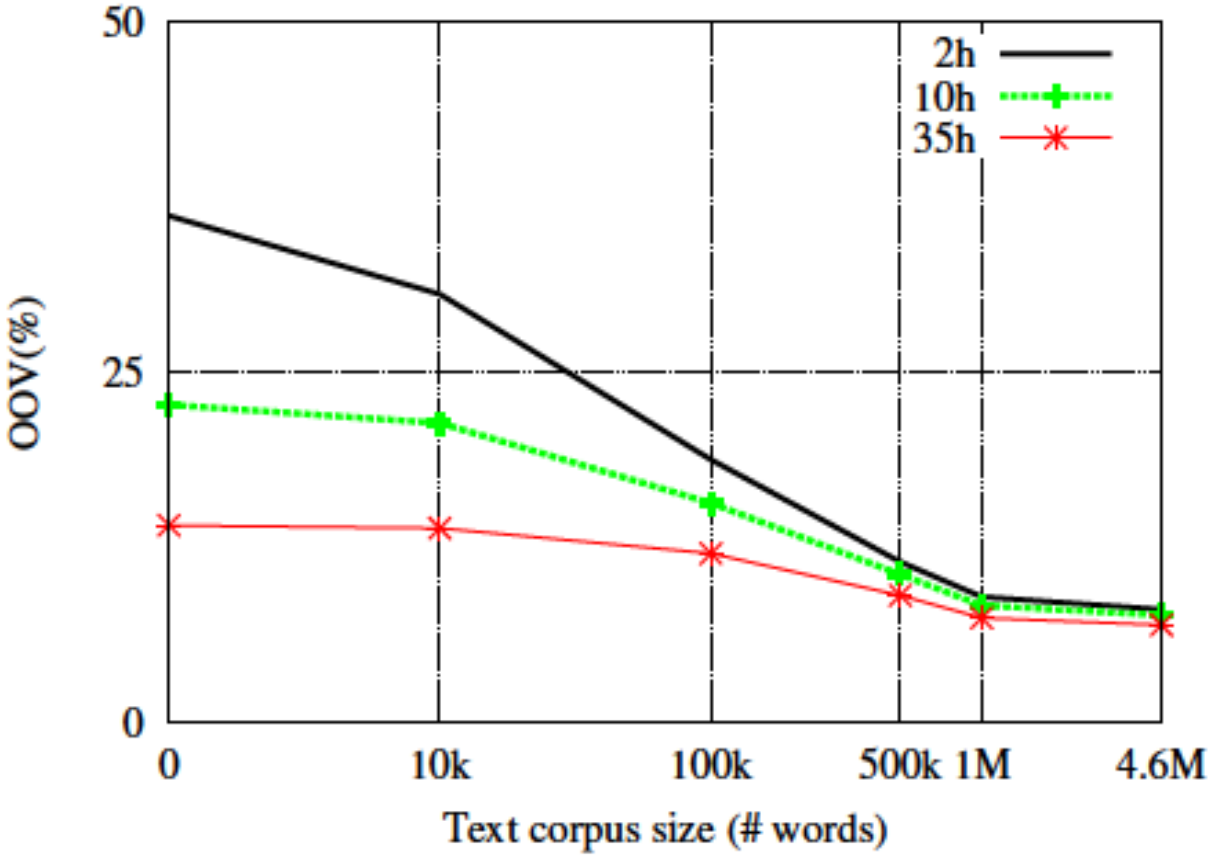
- Acoustic models: 2h, 5h, 10h, 35h
- Language models:
  - Transcriptions: 17k, 35k, 70k, 240k words +
  - Web texts: 10k, 100k, 500k, 1M, 4.6M words

Lexicons

- Lexicon sizes

Transcriptions		*Words (texts)*				
Hours	*Words/Types*	*10k*	*100k*	*500k*	*1M*	*4,6M*
2h	*17k / 7k*	11k	36k	96k	142k	114k
5h	*35k / 12k*	16k	39k	98k	144k	116k
10h	*70k / 21k*	24k	45k	103k	148k	119k
35h	*240k / 50k*	52k	69k	120k	163k	133k

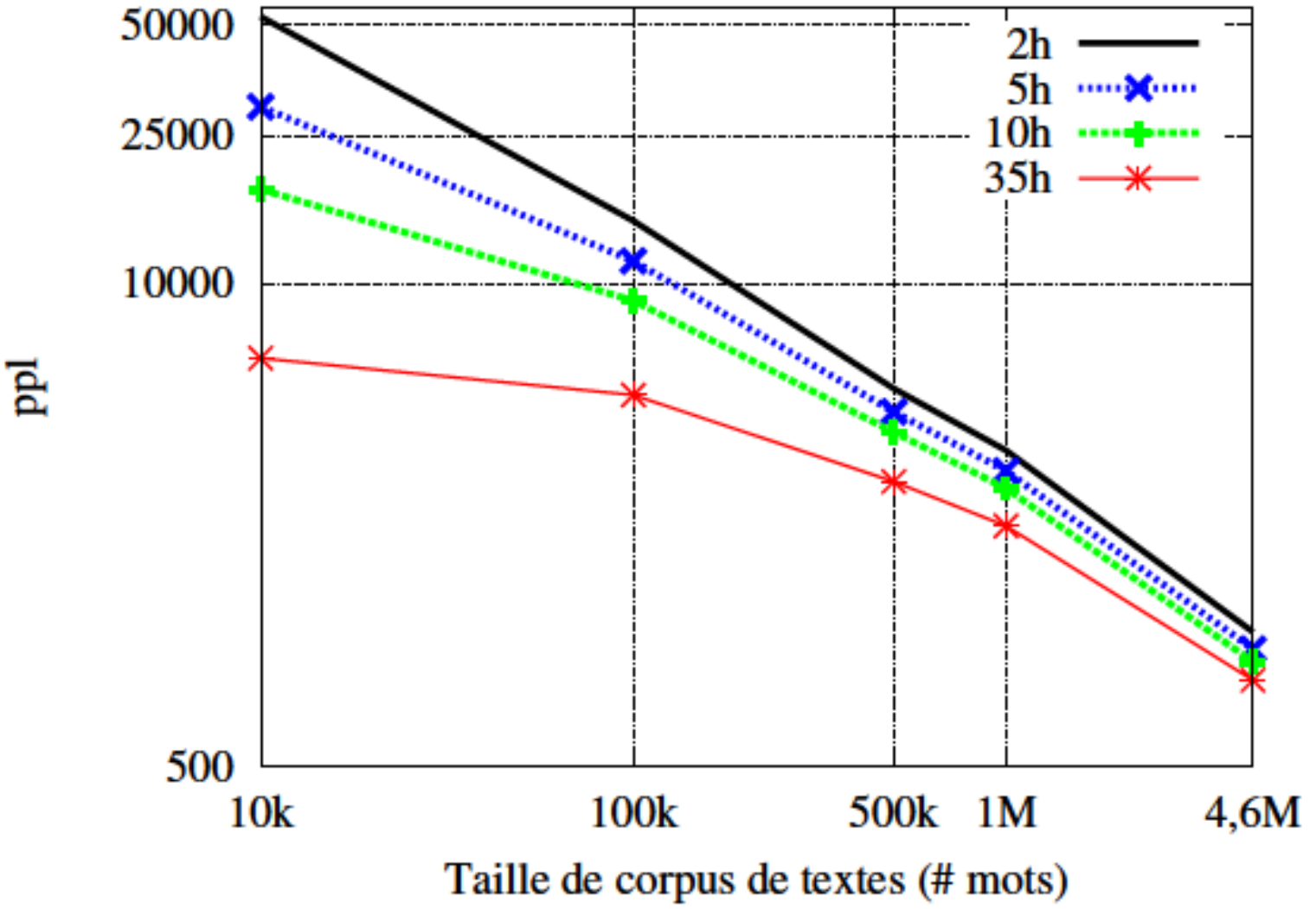
OOV rates



- Out-Of-Vocabulary (OOV) rates

Perplexities

- Normalized perplexities (500k word virtual lexicon)

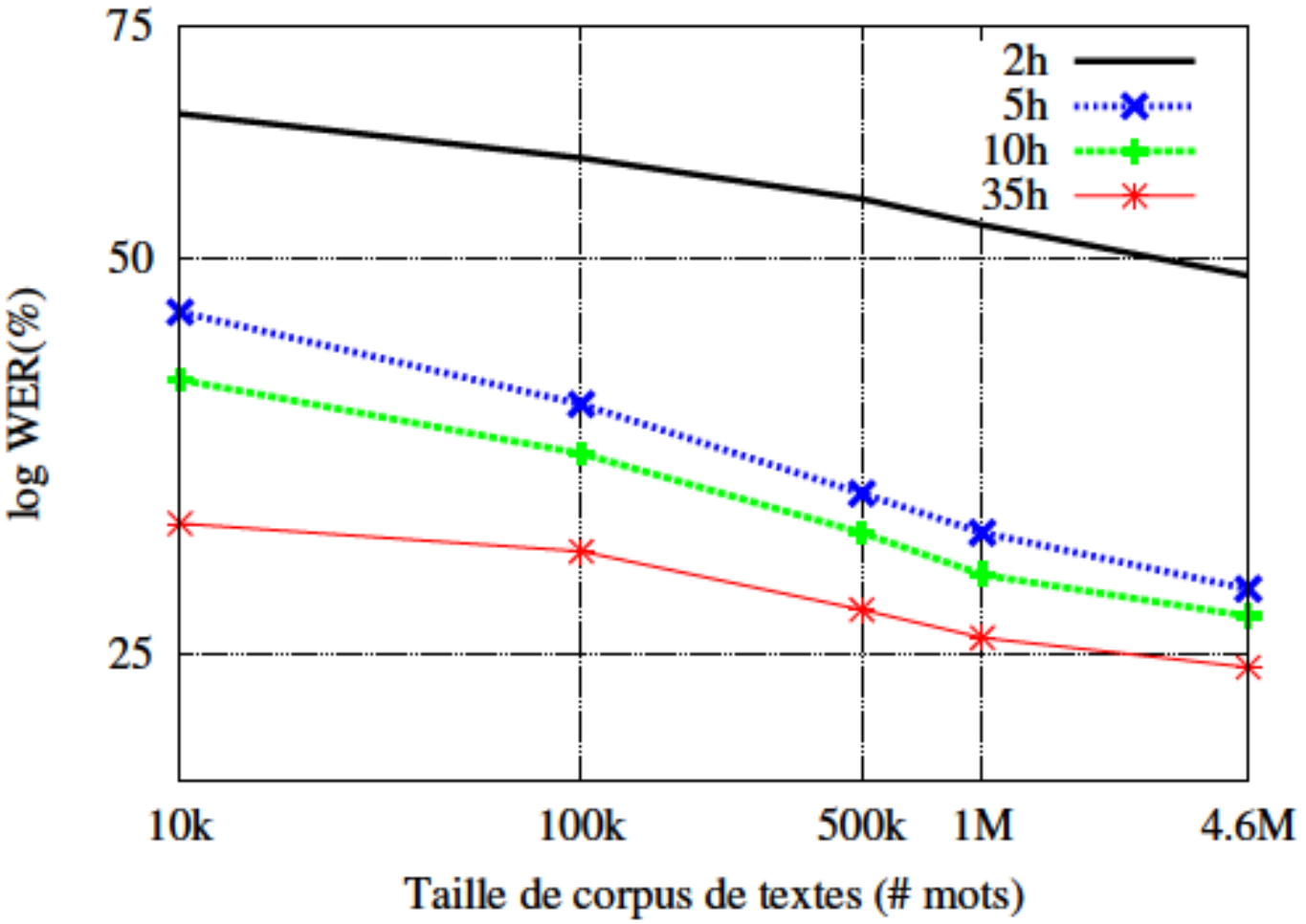


Acoustic models

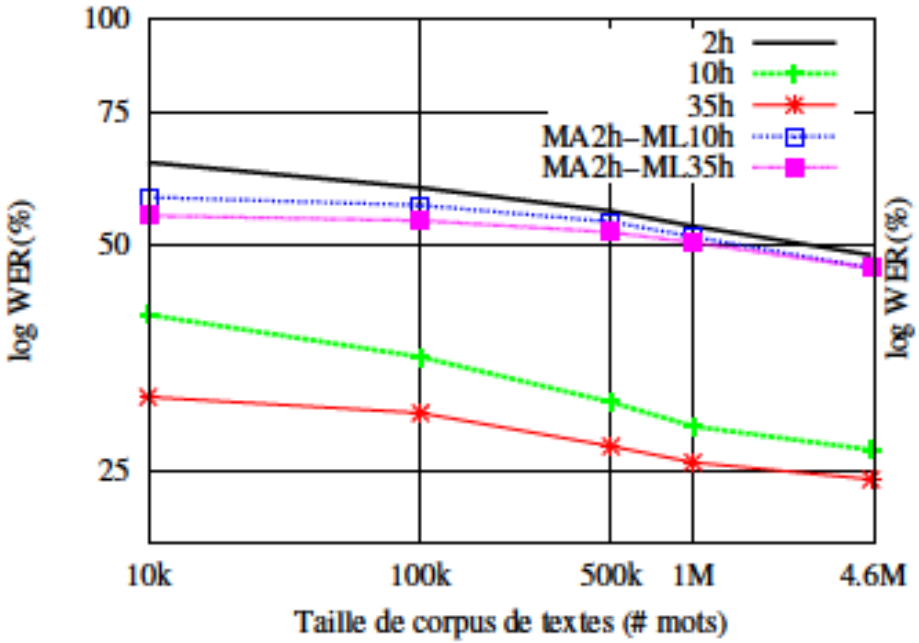
- Tied-state intra-word and cross-words HMMs, with 3 states per model and 32 Gaussians/state
- Specific acoustic models built for each audio training corpus subset

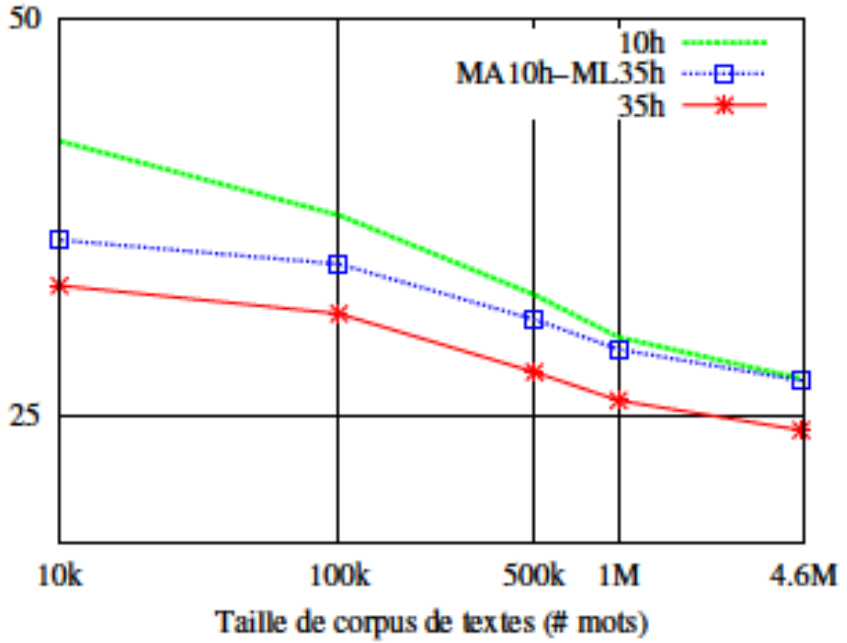
Transcriptions	*2h*	*5h*	*10h*	*35h*
# Contexts	3027	4557	6323	10726
# Tied states	1187	2286	3861	8554

Performances



Crossed experiments





- Influence of the transcription LM part

Conclusions

- 2h of audio training data not sufficient
- Good operating point: 10h audio, 1M word texts
- Collecting texts may be more important than transcribing audio at this operating point, particularly if combined with unsupervised acoustic model training

Perspectives

- Amharic: simple grapheme-to-phoneme conversion
- Similar experiments with other languages (e.g. syllable-based languages)
- Attention: these studies used broadcast news data, behavoir may be different for conversational speech or meetings
- Semi- and unsupervised training for acoustic models

Thank you for your attention!

Interpolation coefficients

Transcriptions	*# Words (texts)*				
	10k	*100k*	*500k*	*1M*	*4.6M*
2h	0.71	0.40	0.27	0.23	0.22
5h	0.84	0.53	0.38	0.33	0.28
10h	0.90	0.63	0.46	0.40	0.33
35h	0.95	0.81	0.61	0.52	0.43

- Interpolation coefficients

Unsupervised experiments

- Lamel et al, 2002