Discovering speech reductions across speaking styles and languages

Martine Adda-Decker, Lori Lamel

CNRS-LPP & *Spoken Language Processing Group CNRS-LIMSI*

November 30, 2017, Paris

Outline

- Temporal speech reduction
- Automatic speech processing as tools for linguistic studies
- Speech corpora & methodology
- Results
- Discussion

Temporal speech reduction

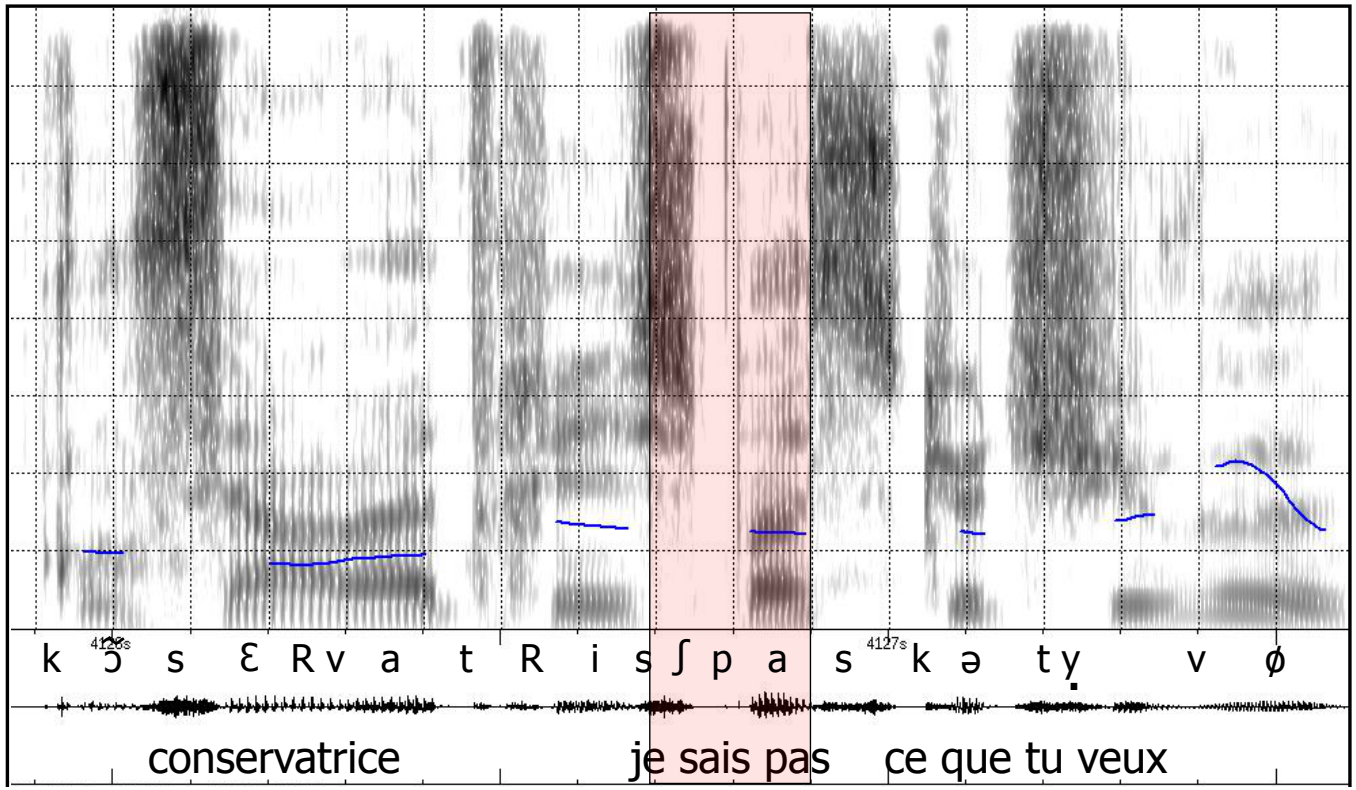
- Speech reduction: vowel reduction, consonant lenition, consonant cluster reduction, syllabic restructuring (e.g. Ernestus 2000, Duez 2003, Adda-Decker et al. 2005, Dilley & Pitt 2007, Van Son & Pols 2013)
- Traces of speech reduction in written language:
  - gonna be (going to be)
  - ça [sa] (cela [səla] 'that'), 'y a [ja] (il y a [ilija] 'there is')
  - ins [ɪns] (in das [ɪn das] 'in the')
- Temporal speech reduction (Adda-Decker & Snoeren 2010):
  includes any reduction process resulting in fewer segments in the produced speech

Speech reductions are known to frequently occur in spontaneous, interactive, less formal speech.
Speech reductions seem to first affect least informative speech portions (Jurafsky et al. 2001) and/or highly predictable from the context:

- function words
- discourse markers
- idioms
- dates...

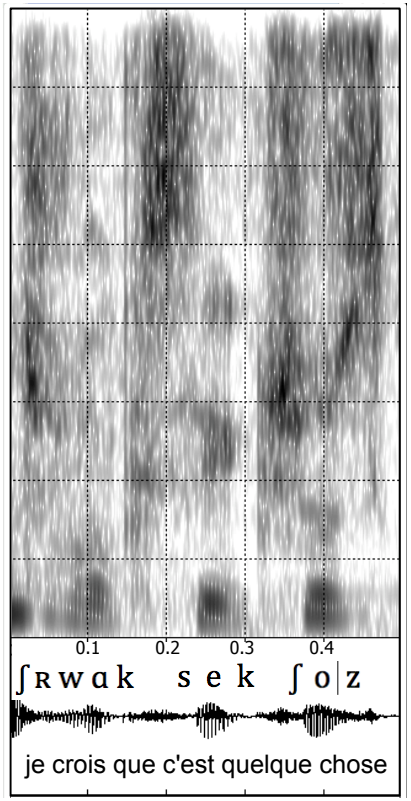
Question: may speech reduction occur
in more formal (prepared) speech?
in more informative speech portions?

Examples - French - multiword reduction



Note : the sound [ʃ] stands for 2 words 'je' (I) and 'sais' (know)

Examples - French - multiword reduction



'je crois que c'est quelque chose' (I believe it is something)
/ʒə kʁwa kə sɛ kɛlkə ʃoz/ [ʃʁwak sek ʃoz]

Temporal reduction phenomena raise issues:

- for automatic speech processing (both recognition, synthesis)
- for human processing in psycholinguistics
- for language learning/teaching...

Big data (large speech corpora) help to answer questions:

- Where do temporal reductions occur?
- Do they frequently occur?
- Are they conditioned by language, by speaking style?

Hypothesis

Temporal reduction mainly appears in unstressed speech segments and is conditioned by speech style

1. careful speaking

clearly uttered

2. casual speech

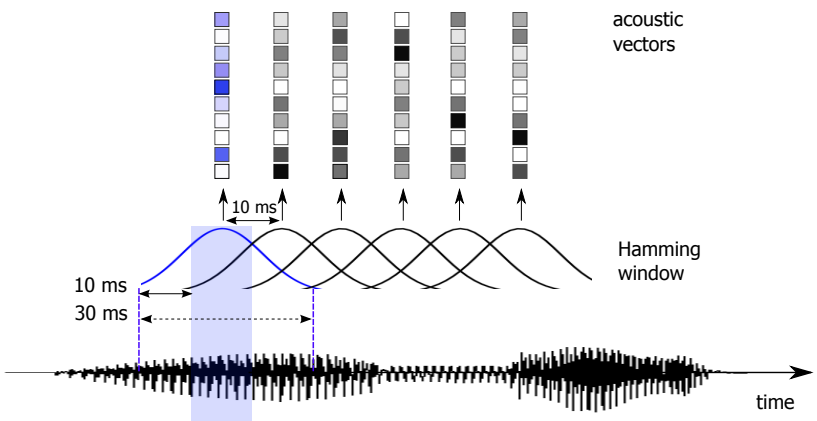
temporally reduced

black boxes : stressed segments
grey boxes : unstressed segments

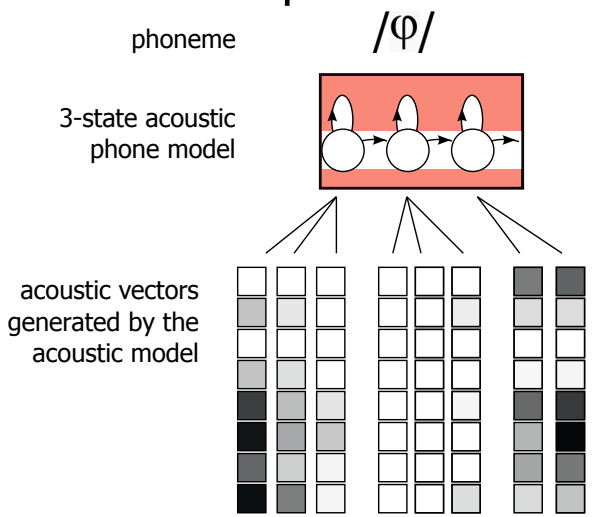
ASR as tool for linguistic studies

A bird-eye's view of acoustic modelling of speech in ASR (automatic speech recognition)

1. Parametrization of the signal



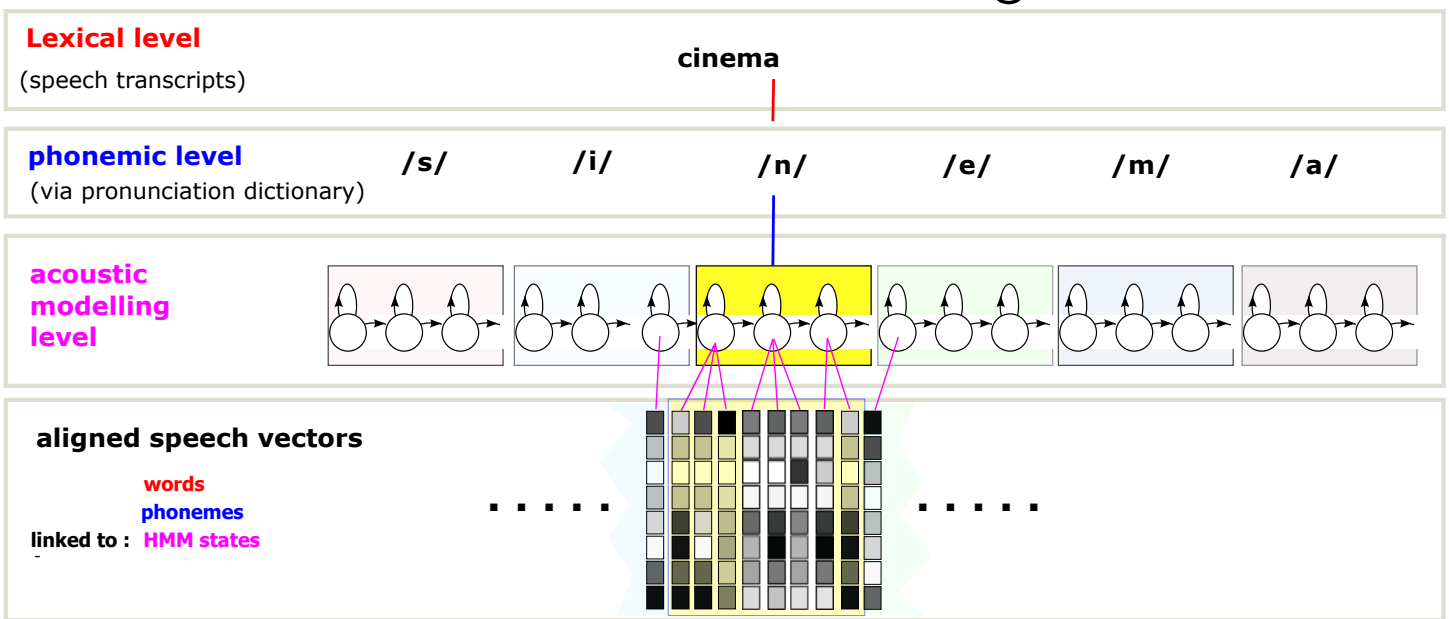
2. Acoustic phone modelling



ASR as tool for linguistic studies

A bird-eye's view of acoustic modelling of speech in ASR

3. Multi-level modelling



Pronunciations are given by a pronunciation lexicon
Variants may be introduced (e.g.: quatre [katʁə katʁ kat])

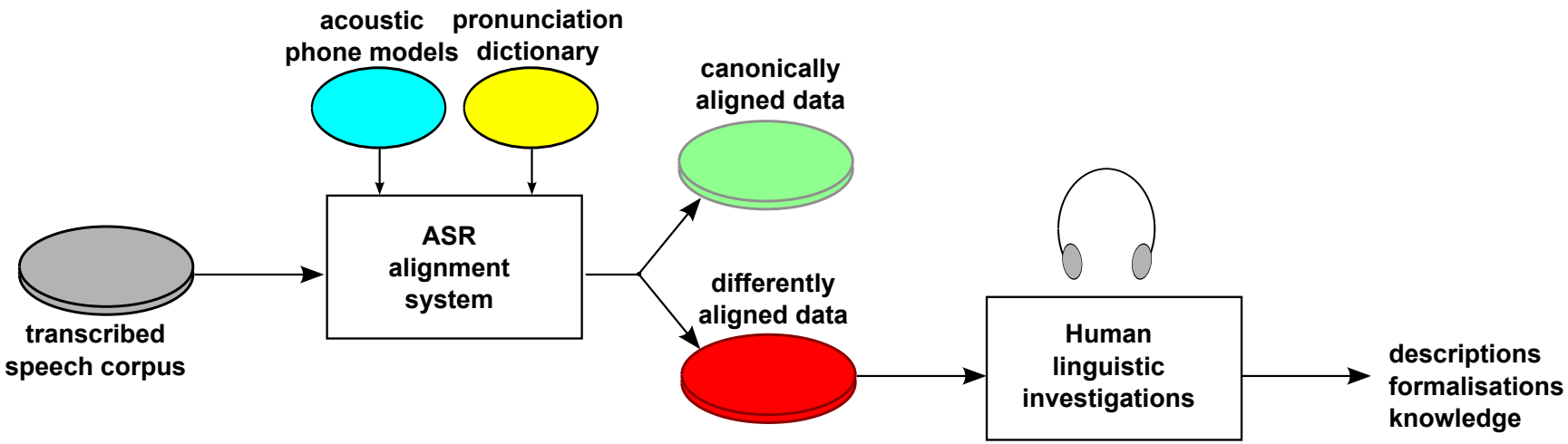
ASR as tool for linguistic studies

ASR: automatic speech recognition

→ converts speech to text

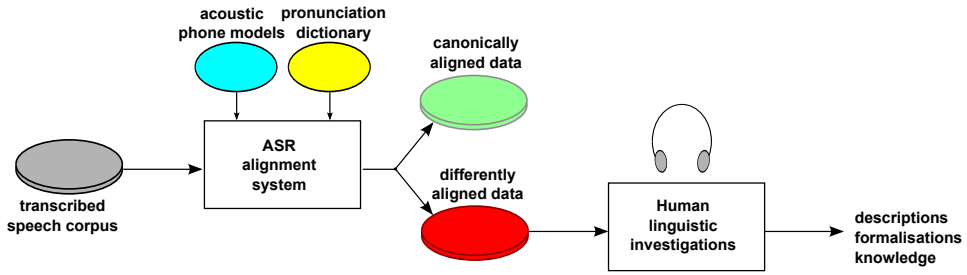
alignment mode:

→ transcriptions constrain the matching process between speech and text



Define the meaning of canonically / differently aligned data!

ASR as tool for linguistic studies

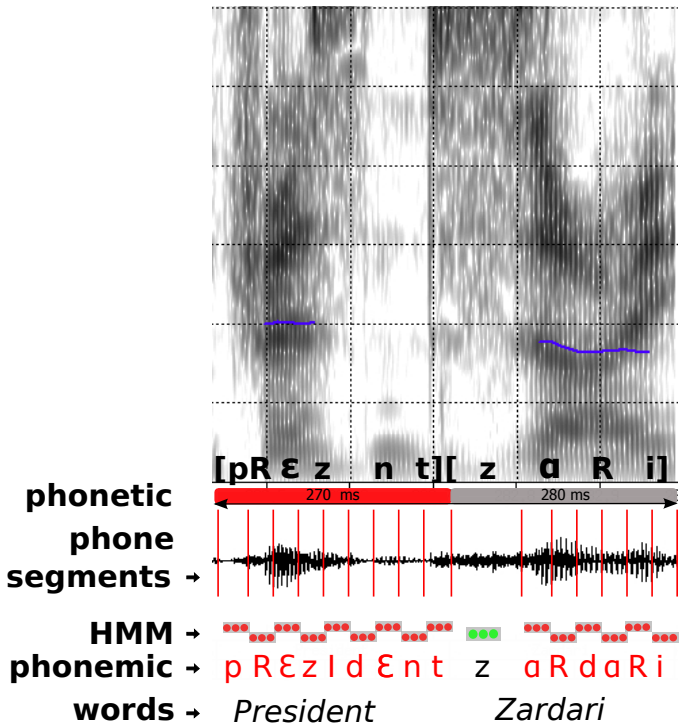


Define the meaning of canonically / differently aligned data!

- in case of (known/hypothesized) variants:
  - introduce relevant variants into pronunciation lexicon
  - canonically aligned = full form (reference) pronunciation
  - differently aligned = one of the variants
- for temporal reduction:
  - hypothesis: might occur anywhere (even though...)
  - introduce a segmental duration threshold on the aligned output
  - e.g. 30 or 40 ms segments

Illustration

Speech sample : 'President Zardari' (strong temporal reduction)
Forced alignement using reference pronunciations:
→ sequence of short segments (30 ms duration) and false/misplaced labels



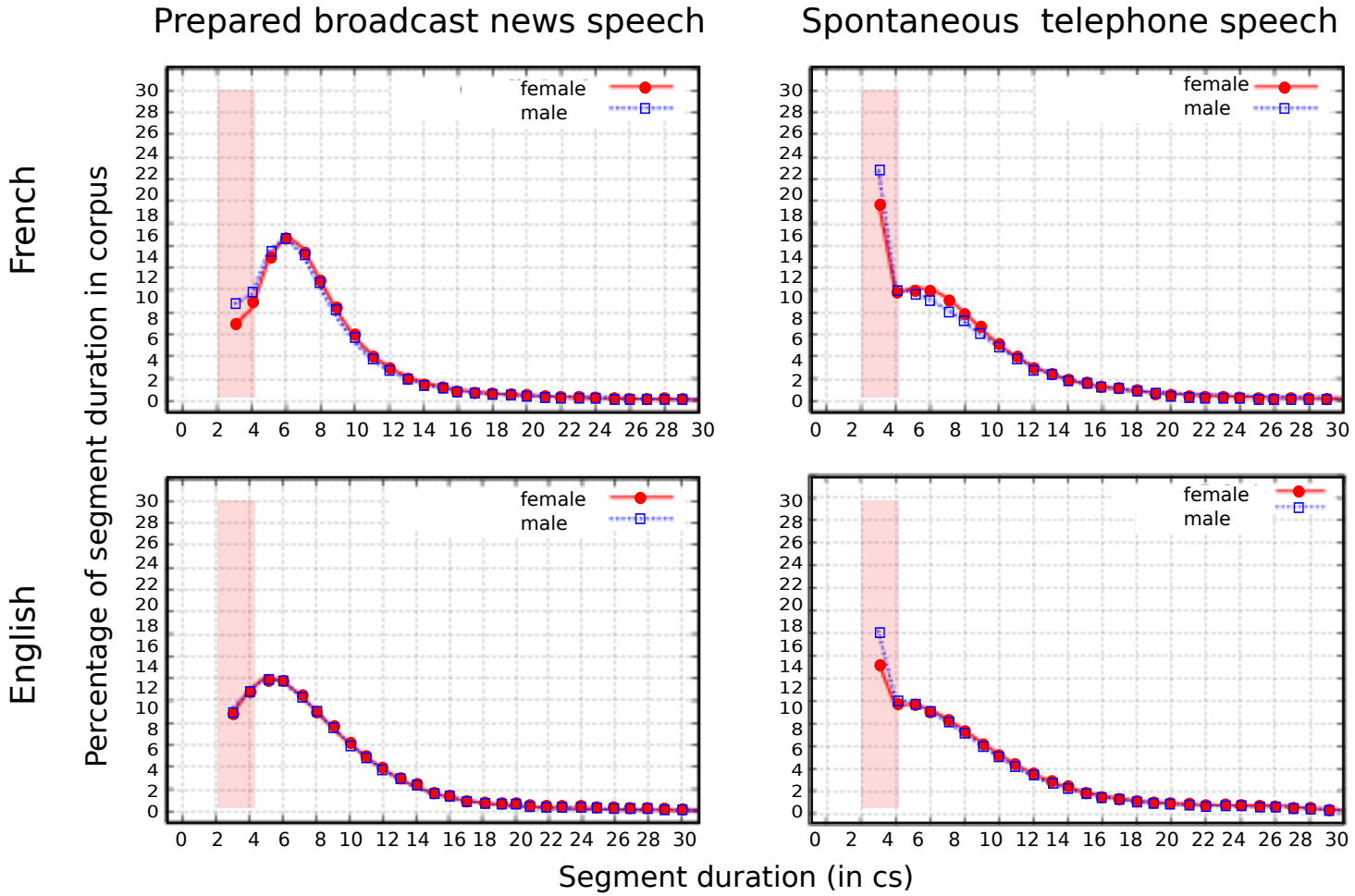
Phonetic: manual labelling
Phonemic: automatic labelling using the reference pron.

Corpora

Different speech styles:
BN careful, prepared (news) and conversational (conv) speech
Casual telephone (tel) and face-to-face speech

French			English		
	# word tokens	duration		# word tokens	duration
BN-news	3600 k	360 h	*BN-news*	7200 k	720 h
BN-conv	600 k	44 h	*BN-conv*	1500 k	124h
Casual-tel	1000 k	100 h	*Casual-Tel*	25000 k	2300 h
Casual-f2f	350 k	31 h			

Phone segment duration distributions



Pink box : short segments (3-4 cs) involving potential temporal reduction
20% min. dur. segments for BN, 25% En / 30% Fr for casual speech

Focus on segment duration variation

Case-study: English /t/
Position-dependent analysis in some typical frequent English words

word-initial position (stressed*)
word-medial position (stressed* / unstressed)
average phone durations are given in ms.

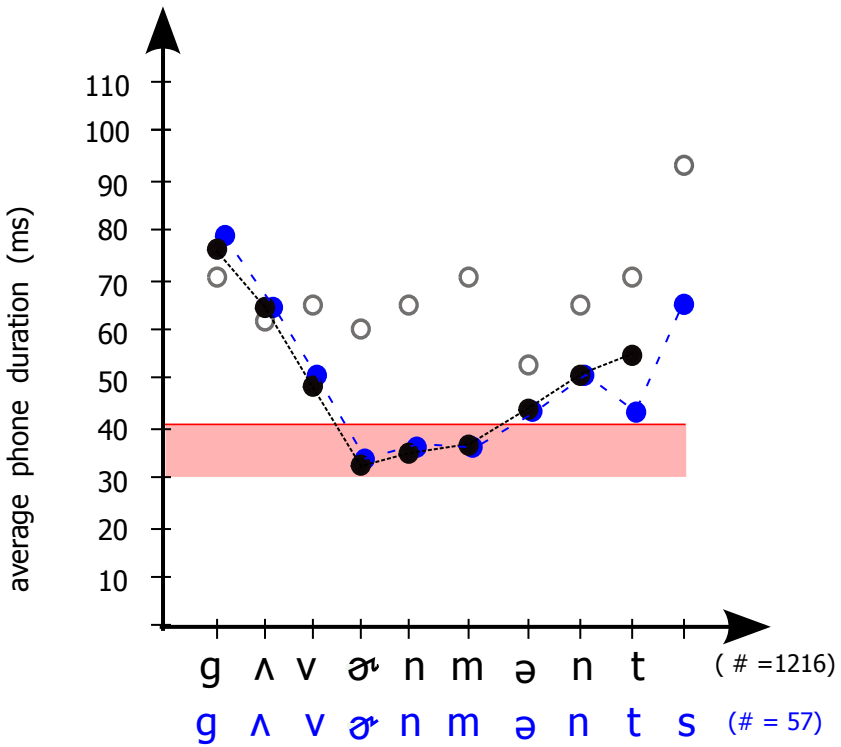
	/t/ position	Broadcast Conversations			SWB/Fisher Conversations		
		#token	avrg. dur.	**% min. dur.**	#token	avrg. dur.	**% min. dur.**
talking	w-init*	814	95	**5**	4898	80	**11**
trying	w-init*	684	95	**6**	4464	85	**11**
nineteen	w-med*	560	80	**7**	706	89	**8**
hotel	w-med*	105	118	**0**	178	126	**1**
little	w-med	1041	59	**41**	9379	37	**91**
getting	w-med	803	59	**52**	5692	39	**86**
exactly	w-med	387	54	**43**	6328	39	**85**
ninety	w-med	323	70	**20**	210	43	**84**

Stress prevents from temporal reduction both in w-init. and w-medial pos., in both speech styles
In unstressed position, temporal reduction is high, especially in casual speech

Word-internal segment duration variation

Case-study: English words 'government', 'governments'

word-initial word stress
enpty circles: average phone duration (all segments pooled per phone)
coloured circles: word-position dependent average phone duration

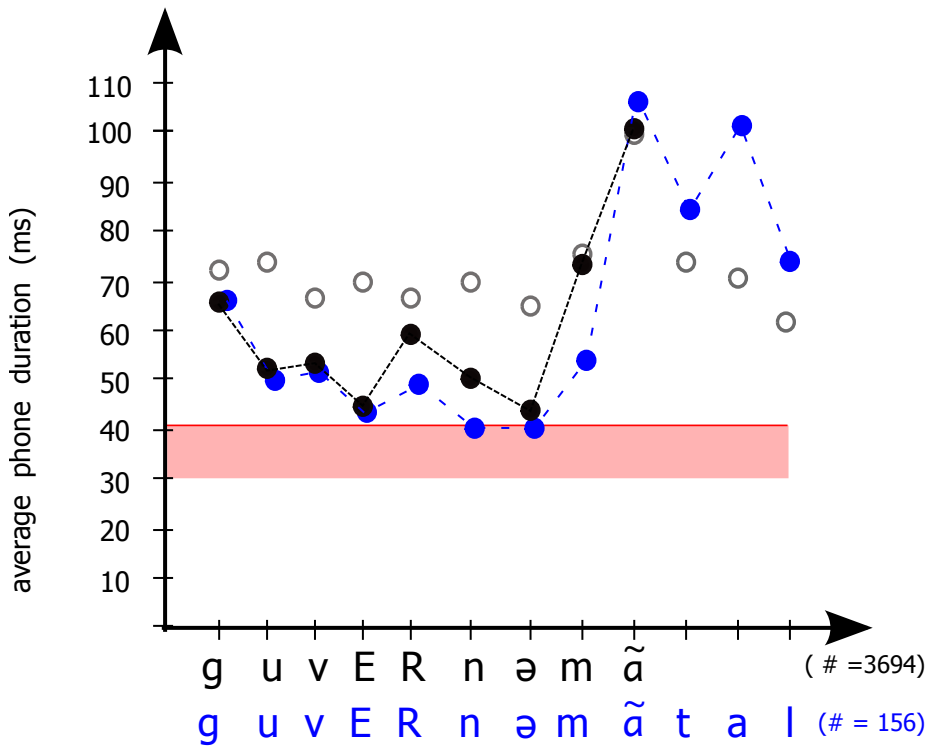


word-internal segments in minimum duration region (30-40 ms)

Word-internal segment duration variation

Case-study: French words 'gouvernement(s)', 'gouvernemental(e)'

stress on final syllable on these words
enpty circles: average phone duration (all segments pooled per phone)
coloured circles: word-position dependent average phone duration



word-internal segments close to the minimum duration region (30-40 ms)

Testing pronunciation variants : English

ASR multi-words with shortened pronunciation variants
English Switchboard data

Multi-word	#Total	*Full form* + *Variants*	#Align	%Align	Comments
did-not	2559	dɪd nɑt	103	4.0	full form
		+ dɪdn̩t	275	10.7	n(ɑ→ ə)
		+ **dɪdn̩**	1175	45.9	+ final-/t/ deletion
		+ dɪn	1006	39.3	+ coda /d/ deletion
we-have	3257	wihæv	1500	46.1	full form
		+ wiəv	205	6.3	onset /h/ del. + (æ→ ə)
		+ **wiv**	1552	47.7	+ V-deletion

Testing pronunciation variants : English

Multi-words with shortened pronunciation variants
English Switchboard data

Multi-word	#Total	*Full form* + *Variants*	#Align	%Align	Comments
going-to-be	750	gɔɪng tʊbi	73	9.7	full form
		+ **gɔnəbi**	432	57.6	complex: ɪng t → n
		+ gəbi	245	32.7	+ complex: ɔnə→ ə
wants-to	157	wɔntstu	15	9.6	full form
		+ wɔnstu	78	49.7	coda C-cluster simplification
		+ wɔntsə	7	4.5	onset /t/-deletion
		+ wɔnsə	57	36.3	both /t/-deletions

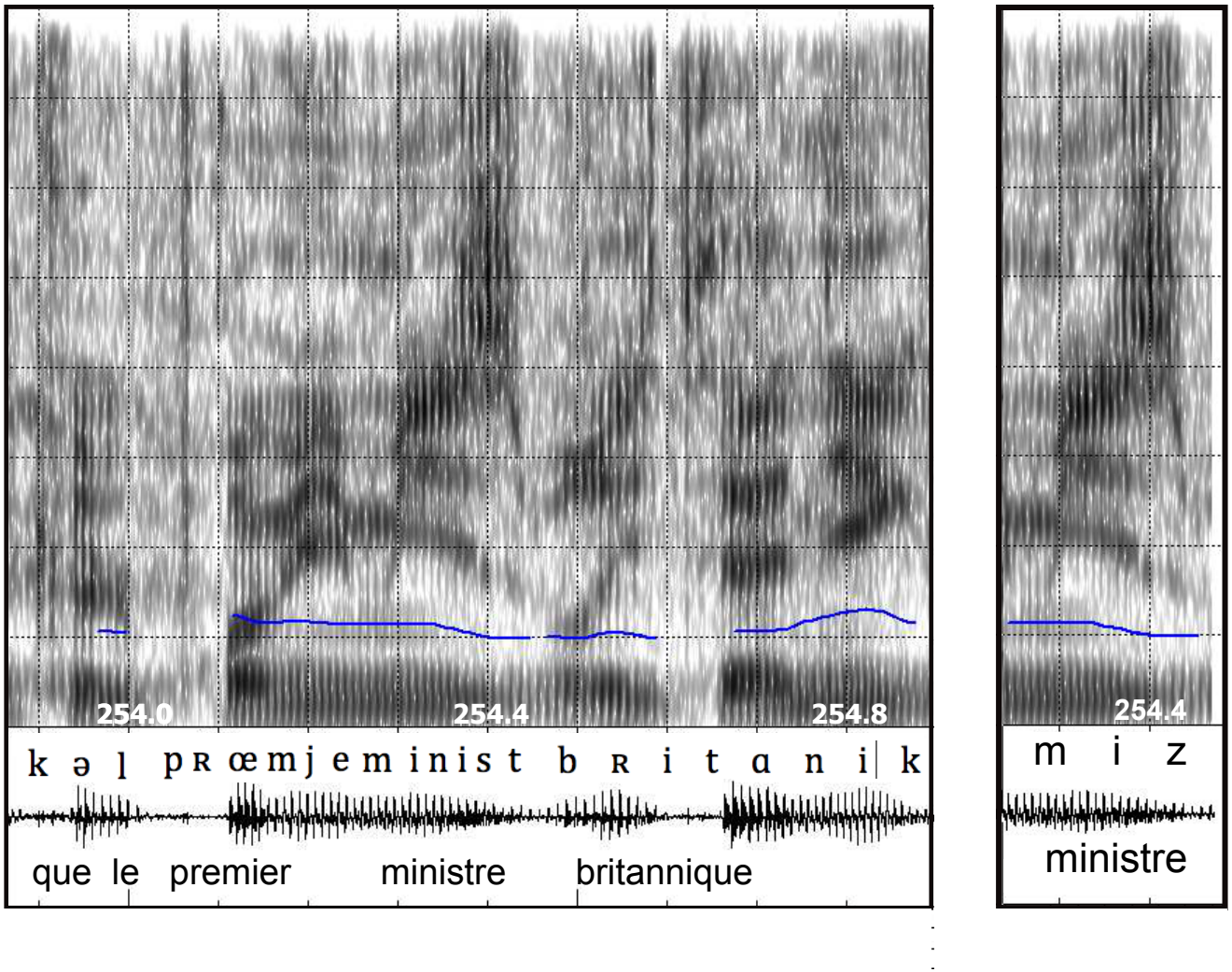
Testing pronunciation variants : French

Words with shortened pronunciation variants
French casual speech corpus (NCCFr)

Word	#Total	*Full form* + *Variants*	#Align	%Align	Comments
parce (que)	2590	pɑʀsə	4	0.2	full form
'because'		+ pɑʀs	45	1.7	no final schwa
		+ **pas**	1309	50.6	+ C-cluster simplification
		+ ps	1232	47.6	+ vowel deletion
quelques	56	kɛlkə	14	25	full form
'some'		+ **kɛkə**	28	50	+ /l/-deletion
		+ kɛ(k—g)	14	25	+ schwa deletion

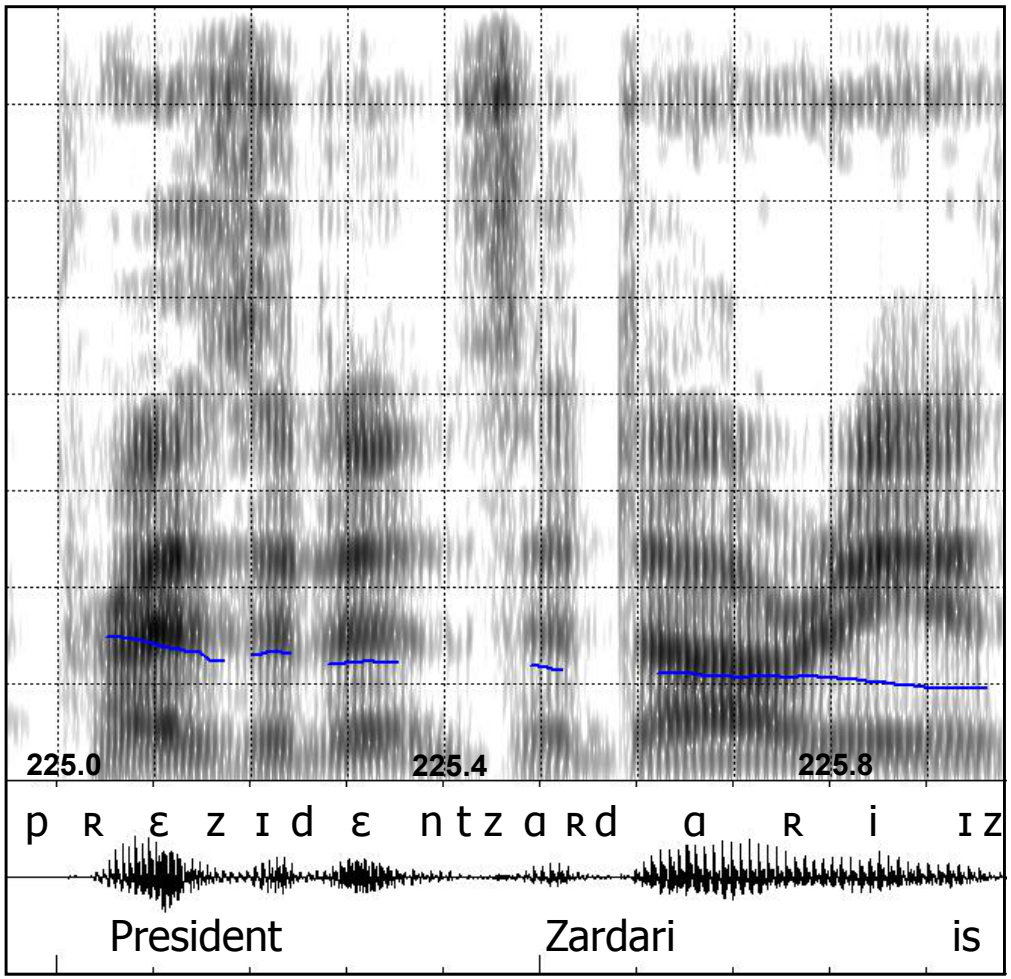
Example - French - reduction...

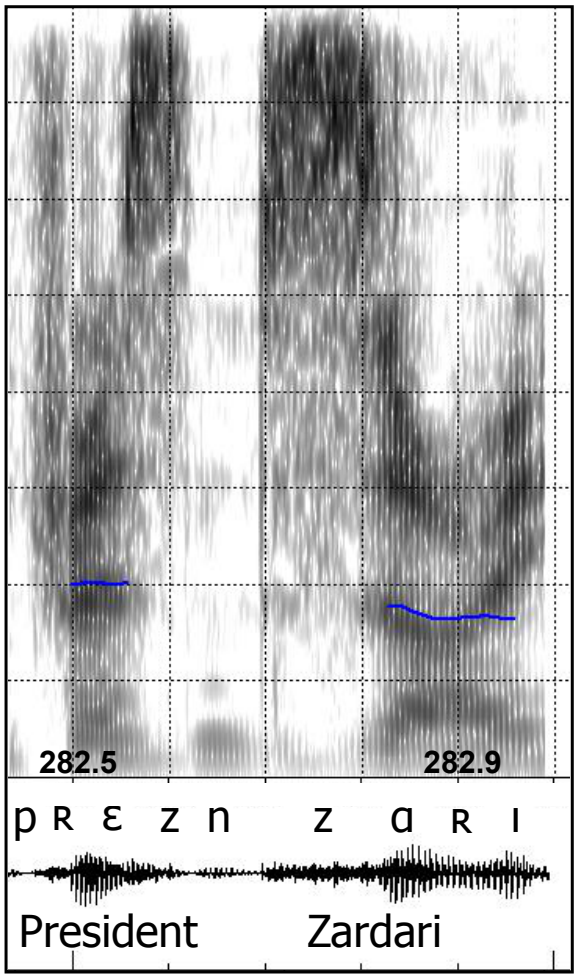
...in more formal, more informative speech portions



Example - English - reduction...

... in more formal, more informative speech portions





Conclusion - Perspectives

- Big Data: automatic speech processing generates ever increasing transcribed speech corpora
- Method: use of forced alignment to locate temporally reduced sequences in fluent speech $\rightarrow$ sequences of minimal duration segments reveal potentially reduced productions
- Temporal reduction observed in different speaking styles (broadcast conv., casual speech)
- Temporal reduction often involves unstressed stretches of speech
- Frequency and recency favor reduction

Research using Big Data in speech sciences

- Need to further foster interdisciplinarity between linguistics, phonetics, medecine, mathematics, signal processing, IT computer sciences...
- Foster team research using shared data and shared research questions
- Blend of expertise