8

Discovering speech reductions across speaking styles and languages

Martine Adda-Decker[1,2], Lori Lamel[2]

(1) LPP-CNRS UMR 7018 Université Paris Sorbonne Nouvelle, Paris,

(2) LIMSI-CNRS, Université Paris Saclay, Orsay, France

8.1 Introduction

In this chapter we propose a dual investigation of speech reduction embracing both technological and linguistic aspects. This double-sided approach aims at combining our experience in automatic speech recognition (ASR) with efforts to relate the observed variation to linguistic structure and processes. The study of speech reduction has attracted increasing interest in linguistic and psycholinguistic research as witnessed by the Nijmegen 2008 workshop on this topic (Ernestus and Warner 2011). Speech reduction is also an acknowledged challenge in automatic speech processing. The last decades have witnessed major progress in ASR, largely due to the widespread use of statistical models combined with the availability of very large speech and language corpora. ASR systems require a large volume of spoken and written data to estimate models of spoken language, with so called language models for ordering speech sounds into meaningful word sequences, and acoustic models that represent the audio signal corresponding to the message. These models capture the average properties of phoneme realizations and include statistics about word and pronunciation frequencies. Speech recognition research has progressively addressed more challenging speech data, moving from well-prepared speech to spontaneous conversations. Studies of ASR transcription performance reveal important differences across speaking styles, attributable to lexical choices, wording and phrasing, as well as to the acoustic realization of a given word.

Since the eighties of the last millennium, the ASR research community has faced growing difficulties in processing progressively less controlled speech. A major bottleneck was the lack of written material truly reflecting spontaneous speech. Another observation was that part of the difficulty was related to pronunciations: acoustic models estimated using read speech data performed poorly on spontaneous speech: the same word, when read aloud and when occurring in natural discourse is not uttered quite the same. With respect to the former, important initiatives were launched to manually transcribe various sources of natural, more or less spontaneous speech, and hundreds of hours of transcribed speech have been made for improved ASR modeling. These data are also very interesting for large scale studies to get a better view of acoustic realization differences between speaking styles, and since such large corpora are only available for a few languages, various efforts are underway to apply the findings across languages. In the present chapter, we develop some observations highlighting commonalities and differences as a function of language and speaking style.

In the following, speech reduction is approached as a temporal reduction or duration shortening. Measured shortening may reflect a variety of processes, such as vowel reductions, consonant cluster reductions, assimilation and lenition processes, segmental and syllabic deletions and restructuring. Many speech scientists share the belief that much knowledge can be gained from studying characteristics of casual speech (Greenberg and Chang 2000; Greenberg et al. 2003; Nakamura, Furui, and Iwano 2006; Strik et al. 2006, Schuppler et al. 2014). For instance, Greenberg and colleagues (2000, 2002) investigated syllabic structures in casual speech from the SwitchBoard data. Nakamura, Furui and Iwano (2006) compared spectral properties of careful and casual speech on large Japanese corpora, thereby highlighting spectral reduction. Strik and colleagues (2006) studied reduction phenomena in Dutch, with a focus on the problem of disappearing sounds, especially in multiword expressions.

Speech reductions seem to first affect the least informative speech portions (Jurafsky et al. 2001), for example function words that are predictable from the context, idioms, morphological items (in particular endings), discourse markers etc. Speech reduction can be manifested in various ways, such as producing different (e.g. centralized) phonemes, fewer phonemes, or even fewer syllables (Ernestus 2000; Van Son and Pols 2003; Duez 2003; Adda-Decker et al. 2005).

As far as phonemic segmentation and labeling is concerned, it is far from obvious that an automatic speech recognizer will prefer the same options as a human expert. A human listener can not always tell for sure whether a phoneme is deleted since some of the missing phoneme's acoustic features may be present in adjacent phonemes, and may even be perceived. Moreover, it is well-known that human speech perception may sometimes be biased by higher-level language knowledge and understanding (see, e.g., Ganong 1980; Elman and McClelland 1988; Samuel and Pitt 2003). By contrast, an ASR system, for a given parameterization, will consistently make the same decisions over the entire corpus.

In this contribution, we investigate temporal reduction via a cross-lingual study in English and French. The basic idea is to identify speech regions that are prone to reduction using a forced speech alignment tool (Adda-Decker and Lamel 1999) based on the LIMSI speech recognition system (Gauvain et al. 1994). We extend the methods proposed in Adda-Decker and Snoeren (2011) to provide evidence of temporal speech reduction with the help of automatic speech alignment using full form pronunciation dictionaries and global descriptors, such as distributions of phone segment durations. Increasing proportions of short segments are often indicative of a higher degree of temporal reduction. The forced alignments are used to quantify speech reduction in large corpora of various speaking styles, ranging from broadcast news to telephone and face-to-face conversations. By studying large speech corpora, the extent of speech reduction can be quantified as a function of various factors, such as speaking style and language, broad phonemic category and syllable position, or social variables including gender, age or status.

The study of such speech regions may lead to deeper insight into the complexity of reduction phenomena specific to spontaneous speech, increase our understanding of the general mechanisms underlying pronunciation variation and last but not least, contribute to better acoustic speech models for ASR in the future. Before we turn to our corpus-based study, however, we first provide a brief introduction to speech reduction highlighted with a few examples, followed by a short overview of speech modeling of reduction in ASR systems. The remainder of the paper presents results and discusses the implications of the outcomes of the corpus-based study.

8.2 Temporal speech reduction

A considerable amount of research has been devoted to the study of speech reduction phenomena, including consonant lenition, consonant cluster simplification, vowel reduction and syllable restructuring (see, e.g., Van Son and Pols 2003; Duez 2003; Dilley and Pitt 2007; Ernestus 2000; Tseng 2005). Frequent "phonological words" reflecting temporal structure reduction can be found in written English (e.g., *isn't*, *it's*, *gonna* in informal writing) and in written German (*ins* instead of *in das*, 'in the'). In French, similar reduction phenomena occur (*ç*a instead of *cela*, 'it'). In this chapter we are interested in temporal reduction phenomena, in particular those that are not reflected in written language.

Reduced pronunciations are often observed on common word sequences which usually are easily predictable from the context. Some examples in French are: *il y a* [ilija] 'there is' which is most often uttered as *y a* [ja], and *je ne sais pas*, [ʒənəsɛpa] 'I don't know' which may have an acoustic realization close to [ʃɛpa] or even [ʃpa], where the *ne* in the negative form *ne ... pas* is completely omitted and /ʒə/ and /s/ are merged to form a single fricative segment that is [ʃ]-like. The /ɛ/-vowel may also become devoiced and merge with the preceding fricative segment.

Similar examples can be cited for English, where some reductions have even been widely adopted in written language. The word sequence *I do not know* is generally written as *I don't know*, and is often further reduced in speech to simply *I'd know* or *dunno*. From an ASR perspective, this problem has been addressed by adding reduced forms as such as *wanna, dunno, gonna* as lexical items or including “multiwords” (sequences of words that tend to frequently co-occur) in the pronunciation dictionary as a single entry (see Strik and Cucchiarini 1999; Strik, Binnenpoorte, and Cucchiarini 2005). Multiwords will have multiple pronunciations ranging from a concatenation of canonical forms to strong reductions. For English, *want to* can match a range of pronunciation variants from [wantu] to [wʌnə].

Figure 8.1 shows examples of typical reductions as observed in spontaneous speech. The left example in Figure 8.1 is taken from a casual conversational and the righthand one from a broadcast interview with politicians. These examples illustrate that the scope of sequential reductions often surpasses word boundaries, typically involving one or more short function words.

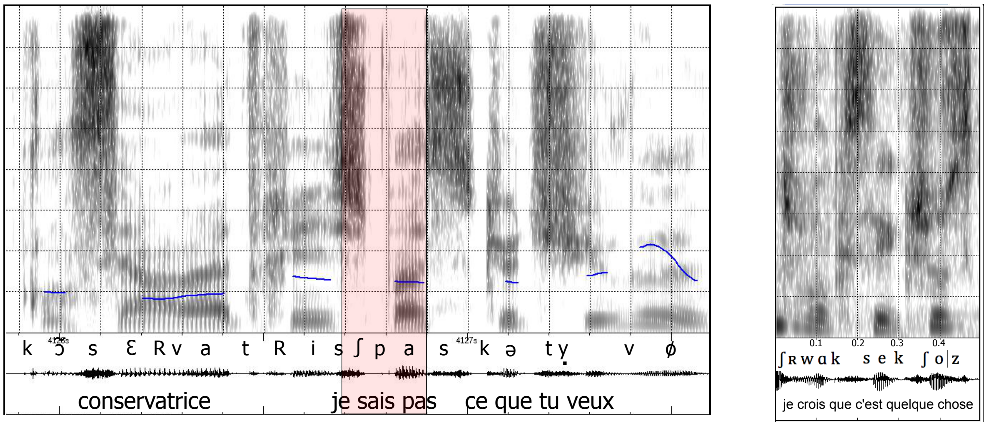

Figure 8.1 Speech signals of common reduction phenomena in French. Left: *je sais pas* ’I don’t know’ /ʒəsɛpa/ in context *conservatrice, je sais pas, ce que tu veux* ’conservative, dunno, what you want’, is approximately produced as [ʃpɑ]. Right *je crois que c’est quelque chose* ’I believe it is something’ /ʒəkʀwɑkəsɛkɛlkəʃoz/, is approximately produced as [ʃʀwɑksekʃoz].

Strong temporal reductions may also be observed on content words. This type of reduction often occurs with words which are highly predictable in a given context. For example, in news reports, polysyllabic words such as *(prime) minister, president* may be uttered very quickly with only parts of the underlying form recognizably uttered in the surface form, especially when they are followed by the person’s name. Figure 8.2 illustrates strong temporal reductions for content words in prepared speech taken from a French broadcast news recording, where the reduction is particularly strong in the excerpt shown in the right part.

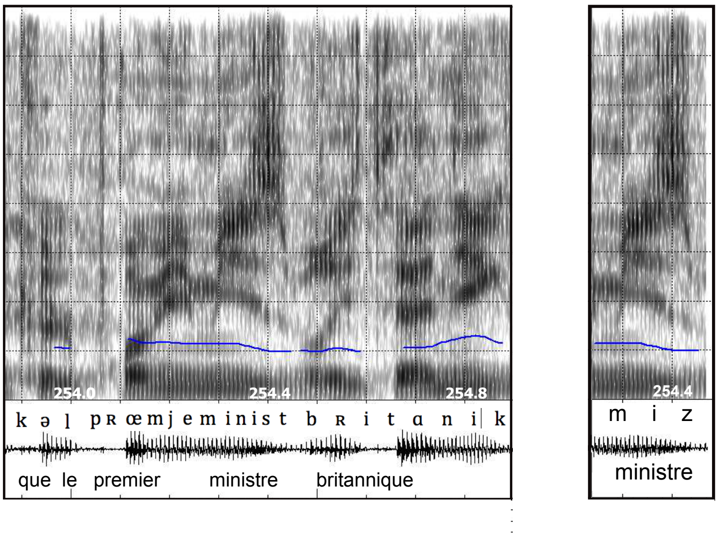

Figure 8.2 Speech signal illustrating French content word reduction in *ministre*. Left: word in context: *que le premier ministre britannique* ’that the British prime minister’ aligned as [mɪnɪst], the system’s shortest variant. Right focus on the word *ministre* /ministʀə/ approximately produced as [miz].

Figure 8.3 shows an English example with two different realizations of the word sequence *President Zardari*, occurring twice in the same conversation. While all of the phones of the word President are clearly articulated the first time and also clearly visible in the left spectrogram of Figure 8.3, the later production shown in the right panel is severely reduced, in particular the two word-final unstressed syllables of *president* (produced as [pʀɛzn̩]). It is noteworthy that also the president’s name is shortened with deletion of at least the two consonants /ʀ/ and /d/ and most probably also the preceding vowel /ɑ/ as the nucleus of an unstressed syllable.

In the following, the word “stress” is used to refer to accented parts in speech, be they due to lexical stress (as in English) or to phrase final accentuation (as typical for French which has no lexical stress) or due to the utterance’s focus structure. We thus use the word “stress”

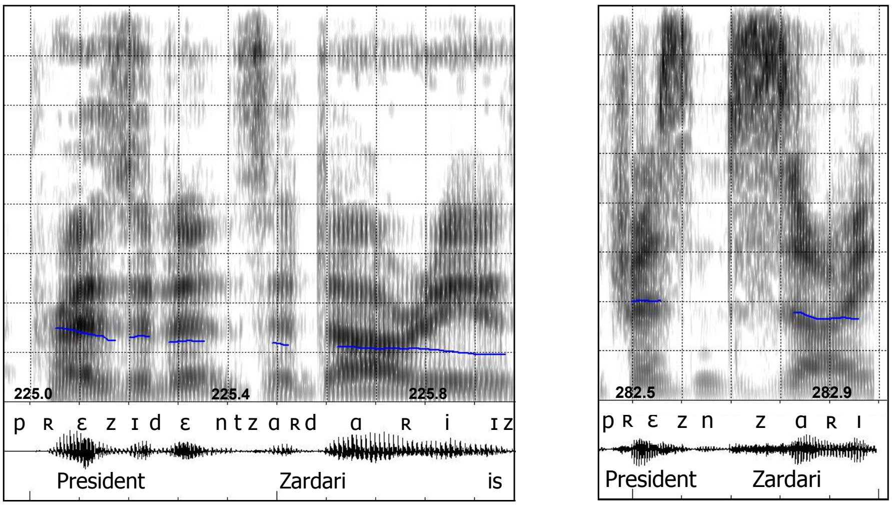

Figure 8.3 Speech signal illustrating content word reduction in *president* in English. Left: *President Zardari* clearly articulated. Right *President Zardari* strongly reduced approximately uttered as [pʀɛzn̩tzɑʀi].

with its general English definition corresponding to prominent regions in both languages. However, the interested reader should keep in mind that the most appropriate prosodic terminology for French would be “accent” (Jun and Fougeron 2002). Whenever necessary, and when speaking more specifically about French, we will use the term “accent” instead of “stress”. Our hypothesis is that although any stretch of speech may be shortened and altered, reduction is considered to affect most often the unstressed segments, whereas the prominent or stressed parts tend to remain more clearly articulated. Prominent or stressed regions may be considered as major anchor points attracting perceptual focus. These stressed regions enable or at least ease the restoration of the reduced, unstressed portions. Figure 8.4 gives a schematic view of our understanding of temporal speech reduction.

This view or “model” is compatible with temporal reductions as shown in spectrograms in Figure 8.1 and 8.2. In the first example, *je suis d’accord* ’I agree’, reductions mainly involve unstressed parts *je suis*. Corpus-based investigations can contribute to the validation of this hypothesis.

Before turning to our corpus-based study and the related methodology, we first give a

clearly uttered
temporally reduced

Figure 8.4 Schematic representation of spontaneous speech with prominent or stressed (black boxes) and unstressed (grey boxes) parts. Left: clearly uttered speech with both prominent and unstressed parts realized. Right: temporally reduced speech mainly affecting unstressed parts.

brief overview of speech modeling in automatic speech recognition and discuss the effects of speech reduction on ASR performance if it is not appropriately dealt with in a system's acoustic speech and pronunciation models.

8.3 Automatic speech processing as tools for linguistic studies

In this section, we present some basic processing and modeling steps in automatic speech recognition systems. In particular, we focus on temporal modeling aspects in order to demonstrate how temporally reduced speech may be detected by the system. In our opinion, it is essential to grasp these basic steps to understand the potential contributions and limits of forced alignment and correspondingly labeled data. We also briefly address the issues of segmentation accuracy and of segmentation labels as compared to those produced by human experts.

8.3.1 Speech modeling

A first processing step corresponds to the conversion of the acoustic signal to a sequence of acoustic parameter vectors used by the alignment system. Figure 8.5 (left) illustrates this conversion from the speech signal (bottom) to parameter vectors (top): a time window with a length of several pitch periods is required to compute meaningful spectral coefficients in voiced speech. The window size is of fixed length of typically 30 ms, and is shifted by a fixed step of usually 10 ms to produce a steady flow of acoustic parameter vectors. This window duration guarantees the inclusion of at least two pitch cycles in a deep male voice. The frame-based processing implies that automatically determined segment boundaries are no longer placed on a continuous time axis, but on a discrete grid with a regular (10 ms) spacing. Furthermore, fine details in the speech signal or the corresponding spectrograms

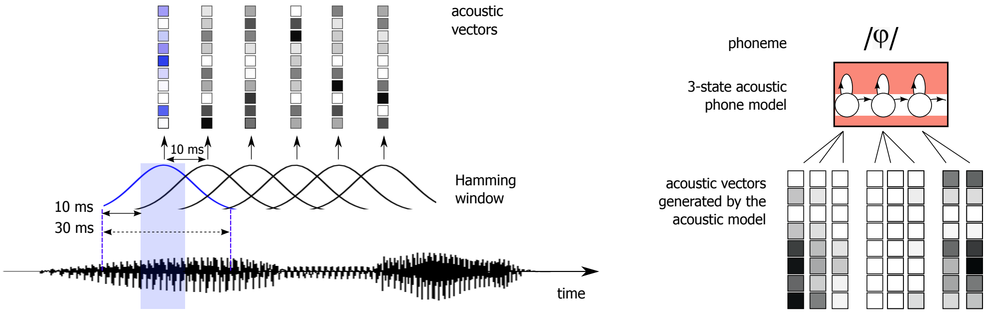
acoustic vectors
10 ms
Hamming window
10 ms
30 ms
time
phoneme
/φ/
3-state acoustic phone model
acoustic vectors generated by the acoustic model

Figure 8.5 Left: Speech parameterization: audio signal is converted to acoustic vectors with a 10 ms frame rate. Right: outline of acoustic phone modeling: 3-state HMM phone model linking the abstract phoneme level to an acoustic realization.

that may be essential cues for human experts are not available for boundary location. The segment boundaries of a given word are placed to globally optimize the location of the predicted segments (via the pronunciation dictionary) with respect to the observed signal. Although not used here, shorter steps of 5 ms have also been experimented with in the literature (Bartkova and Jouvet 2015) especially to address variant selection of temporally reduced speech variants.

Hidden Markov models (Rabiner and Juang 1986) are widely used to model the sequences of acoustic feature vectors, with acoustic units corresponding to phones as shown on the right side of Figure 8.5. Although Figure 8.5 shows only one single model, a given phoneme is typically modeled by a large set of context-dependent phone (allophone) models, as context strongly influences acoustic realizations. Figure 8.6 illustrates the speech modeling and alignment process. Each acoustic vector becomes part of a single phone segment (or a silence segment at word or phrase boundaries). Segment boundaries are typically located in transitional parts. Various studies have reported boundary location accuracy under 20 ms (Di Canio et al. 2012). Our experience with boundary location is in line with this result. The output of forced speech alignment is a sequence of contiguous phone segments with labels predicted by the pronunciation dictionary.

The quality of pronunciations included in the alignment system are of crucial importance in the production of automatically aligned speech data. However this leads to a host of

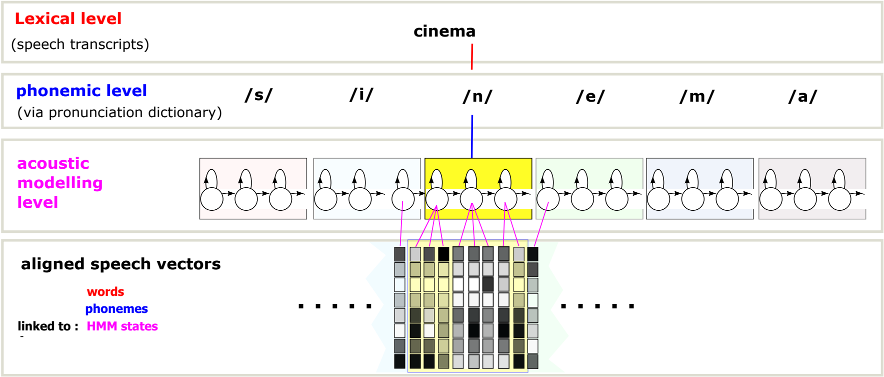

Figure 8.6 Multi-level speech modeling for automatic speech alignment: the lexical level with the written words links to a phonemic level with canonical pronunciations. Each phonemic symbol is associated with allophonic (context dependent) acoustic model to account for contextual variation. The central state of each 3-state HMM model corresponds to the center of the phone segment, the outer states to the phone's start and ending, with potential transitional frames to the neighboring segments.

questions about how to determine the pronunciation(s) that will be useful for further linguistic investigations? Should they reflect surface forms or underlying forms? phonetic or phonemic labels? A canonical pronunciation dictionary typically includes the full form pronunciations, which supposes that all possible segments are pronounced. For many languages, there is a strong correspondence between orthographic and spoken forms. Canonical pronunciations thus tend to produce phonemic labels of the underlying form. By introducing pronunciation variants in the dictionary, the alignment system can choose among different options to produce labels that are closest to the actually produced sounds. In this case, the automatic labeling may become closer to what is considered as a broad phonetic labeling as it tends to adjust to the observed production. However, even if pronunciation variants are added, the labels and segmentation options remain constrained to the actual options foreseen by the alignment system. How to ensure that major variants are included in the pronunciation dictionary? Generally speaking, the automatic system needs to learn the pronunciation variants, which consists of providing it with audio samples with corresponding transcripts, and cannot reliably make very fine level distinctions.

Table 8.1 Near-homophone errors with temporal reduction: the system's *Hypothesis* pronunciation is shorter than the pronunciation of the *Reference* transcription. The comment column suggests phonological processes underpinning the observed variation.

Reference	*Pronunciation*	*Hypothesis*	*Pronunciation*	*Comment*
ça avait	/saavɛ/	savaient	/savɛ/	V#V merger
semble que	/sɑ̃bləkə/	somme que	/sɔmkə/	word-final CC deletion
parce que	/paʁsəkə/	ce que	/səkə/	atone syll. deletion
près de Paris	/pʁɛdəpaʁi/	préparé	/pʁepaʁe/	clitic deletion

8.3.2 ASR errors and temporal structure

In this section, we briefly address ASR errors as these are one of the means of revealing mistakes and missing variants in the pronunciation dictionary. Previous studies for the French language reported word error rates of about 10% for careful (i.e., journalistic) speech and above 15% for casual telephone speech, using large corpora for system training (hundreds of hours of appropriate casual speech data) and complex system combinations (see Lefèvre, Gauvain, and Lamel 2005; Prasad et al. 2005). Approximatively 30-40% of the errors in automatic transcriptions of careful speech consist of homophone or near-homophone errors without temporal reduction. Several studies have focused on close homophone substitutions in terms of ASR errors and human perception (Vasilescu et al. 2009; Vasilescu et al. 2011).

Table 8.1 gives some examples of near-homophone errors with temporal reduction in prepared journalistic speech. The pronunciations of the hypothesized word sequences are shorter than those of the reference transcription. When analyzing casual, conversational speech, however, the proportion of errors due to temporal reduction increases significantly. Temporally reduced speech, corresponding to sequences of short words, such as discourse markers (*tu sais, tu vois* 'you know, you see') and markers of reported speech (*il m'a dit, je lui ai dit* 'he told me, I told him') is particularly frequent in this kind of data. Consequently, these sequences are often prone to recognition errors, unless specific shortened pronunciations are included in the pronunciation dictionary used for both acoustic model training and for decoding.

During forced alignment, full pronunciation models tend to be a poor match with temporally reduced speech. In such regions, the segmentation is characterized by several contiguous small segments of minimal duration, which can be automatically detected by looking for minimal duration phone segments in the forced alignments. These regions of short segments tend to reflect a mismatch of the system's speech model when a short surface form needs to be aligned with a longer underlying form. This situation may result in ASR transcription errors as can be illustrated by the following example: *quai de Seine* 'bank of the Seine' /kɛdəsɛn/ was uttered in two syllables (without the schwa vowel) and misrecognized as *quête saine* 'health quest' /kɛtsɛn/. The two sequences are almost homophonic, where the differences can be explained by a combination of French phonological processes such as schwa elision and regressive voice assimilation. French compound nouns are typically built as <noun>-*de* -<noun> sequences. In such constructs, the schwa of *de* 'of' is typically deleted before a consonant (here the /s/ of *Seine*) when preceded by an open syllable (here the /kɛ/ of *quai*). The /d/ may then become a devoiced [t] due to the following unvoiced /s/ (cf. Snoeren, Hallé, and Segui 2006).

In casual speech, it is common to find complex combinations of various reduction processes. Using large corpora provides the opportunity to elaborate a synthetic overview of the various reduction processes (cf. Schuppler et al. 2008). As a first step in this direction, we propose to quantify temporal reductions using forced alignment. This allows us to measure deviations from canonical temporal structures in terms of their phone segment duration distributions as well as in terms of minimum duration sequences. This approach is further explained in the methodology section (see 8.5).

8.4 Speech corpora

Several large speech corpora were used in these studies, containing different styles of data in French and English: broadcast news (BN), conversational telephone speech (CTS), and face-to-face conversations. The careful speech data set stems form French broadcast news and corresponds to 360 hours of various radio and TV shows that were used for the *Technolangue*-ESTER (Galliano et al. 2005) campaign distributed by ELDA (European Language Data Agency, http://www.elda.fr). Similar data for English are widely available, in particular the broadcast news data produced for the DARPA *Rich Transcription 2004 Broadcast*

News evaluation (Nguyen et al. 2004), distributed by LDC (Linguistic Data Consortium, http://www.ldc.upenn.edu/Catalog/). Some of the broadcast data were classified as broadcast conversations (BN-conv). This data is more spontaneous (less prepared) than BN-news, and are often interviews or debates.

The casual speech data set is comprised of about 120 hours of LIMSI internal French telephone conversations. These conversations are mostly between friends and/or family members, so the corpus therefore contains a highly casual speaking style. The casual speech data set for English comes from the Switchboard (Godfrey, Holliman, and McDaniel 1992) and Fisher data (distributed by LDC) including thousands of hours of speech. In these corpora, the telephone callers do not know each other and are supposed to speak about assigned topics. Therefore, the speech, although spontaneous, is less casual here than the speech in the French corpus. Each corpus includes hundreds of male and female speakers.

Furthermore, we studied a French corpus of face-to-face conversations between friends, the NCCFr – Nijmegen corpus of casual French – (Torreira, Adda-Decker, and Ernestus 2010), available at Nijmegen for research purposes.

Table 8.2 Corpora used in this study: BN careful, prepared (news) and conversational (conv) speech, and casual telephone (tel) and face-to-face (f2f) speech. French (left panel) and English (right panel).

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			

8.5 Methodology

Figure 8.7 illustrates how an ASR system can be used as an instrument for linguistic studies. The system can be used to align a word level transcription with the speech signal, given the pronunciations for each word. The provided pronunciations can allow the investigation of linguistic phenomena. Some previous investigations showed that major linguistic trends (e.g. vowel reduction, French liaison, voice assimilation, regional accent specificities) could be validated using automatically aligned speech data (Adda-Decker, Gendrot, and Nguyen 2008; Woehrling 2009). Previous work also compared segmenting various data types with full form canonical pronunciations as well as variants designed to detect vowel and consonant changes or deletions due to reductions (Adda-Decker and Lamel 1999). Building upon (Adda-Decker and Lamel 2005; Adda-Decker and Snoeren 2011), in this work the methodology is applied to highlight temporal reduction tendencies based on measures of phone durations (distributions, durations by phone classes, or phone sequences).

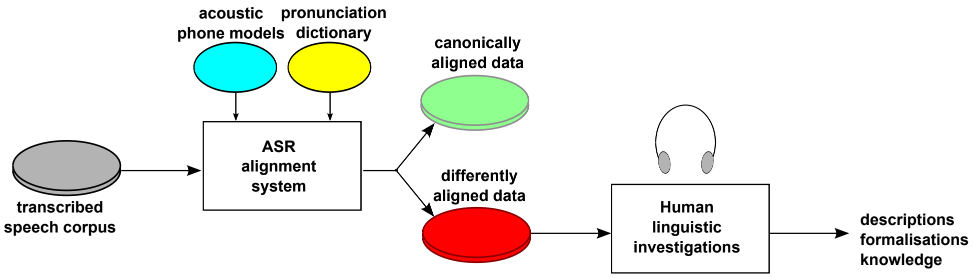

Figure 8.7 The automatic speech recognizer as an instrument to automatically select canonically and differently aligned subsets of speech deviating from expected representation. These subsets are of interest for more in-depth linguistic investigations.

The basic idea exploited is that when temporally reduced speech is aligned against full form pronunciations, there will generally be several contiguous phone segments of minimal duration (i.e. 30 ms here). An example of reduced speech together with an illustration of its automatic alignment using a full form pronunciation model is shown in Figure 8.8. Many of the aligned segments are of minimal duration. It is worth noting that in this case most segments are neither correctly located nor correctly labeled. However, a sequence of minimal duration segments highlights temporal reduction, which is the point of interest here, and is investigated by tracking such sequences of minimal duration in our corpora.

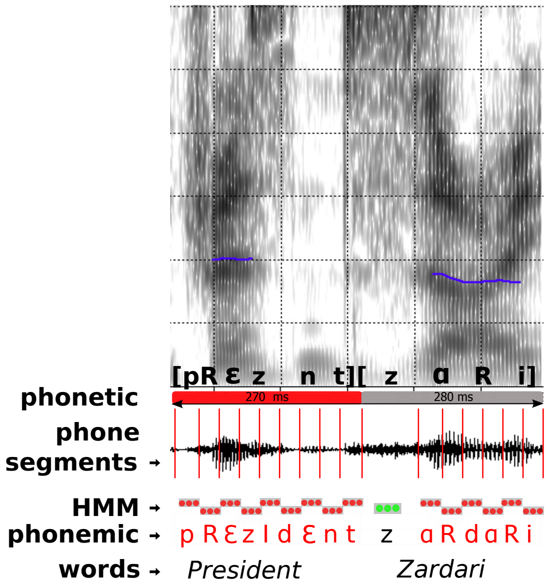

Figure 8.8 Minimal duration segment sequence in a temporally reduced English words President Zardari: /pʀɛzɪdɛnt zɑʀdɑʀi/, approximately produced as [pʀɛzntzɑʀi] /. As a result, the automatically aligned segments (except segment [z]) are of minimal duration (1 acoustic vector per state which results in 30 ms segments in our 3-state HMMs).

8.6 Results

A first investigation examines segmental durations produced by automatic speech alignments using full-form pronunciation dictionaries in order to localize temporal reduction in fluent speech. After comparing global segment duration distributions across speaking styles and languages, we will focus on duration variation fixing either the phone identity or the word identity.

The second investigation line aims at qualifying and quantifying the observed temporal reduction phenomena beyond the segment level. To this aim, we introduced shorter pronunciation variants into the dictionary. During forced alignment, the best matching variant was chosen. Our hypothesis is that this chosen variant, not only provides information about the presence of temporal reduction, but also uncovers possible clues about the reduction processes involved.

In the context of automatic speech processing, known temporal reduction phenomena may be accounted for in the pronunciation dictionary by adding pronunciation variants which are shorter than the canonical form (Lamel and Adda 1996; Lamel and Gauvain 2005; Karanasou and Lamel 2011) [1]. Assessing their usage during alignment can give an indication of the importance (frequency) of the phenomenon. However, our belief is that some of the temporal reduction phenomena still escape our inventory of explicit knowledge as they tend to be unnoticed by native speakers of the language. Sequences of phones that are of minimal duration point to such regions in the segmented data. Reduced sequences also tend to cause trouble to foreign language speakers who struggle to follow given the blatant mismatch between their learnt full form pronunciations and the reduced ones produced by native speakers. This also raises interesting cognitive processing questions which are beyond the scope of this chapter. What is perceived by listeners stems partly from the acoustic input and partly from the representations in their brain, reflecting among other things their past language experience and their current contextual situation.

[1] Note that the aligned pronunciation of `government` /gʌvɚmənt/ in Figure 8.10 is already reduced

8.6.1 Segmental duration

In the following, we provide a bird's-eye view of segmental duration variation, before detailing some illustrative examples at segmental and lexical levels.

Segment duration distributions

To provide a synthetic view of segment durations, Figure 8.9 shows the phone segment duration distributions of aligned data in French and English for both prepared broadcast news and spontaneous telephone speech. The speech alignments relied on full form pronunciations with only a small number of exceptions with shorter variants to account for well-known reduction phenomena (e.g. in English *hundred*: /hʌndrəd, /hʌnrəd/, /hʌnɚd/; in French *autre* ’other’: /otrə, /otr/, /ot/) . The alignments tend to find the best match between the proposed acoustic word models and the speakers productions. As highlighted in Figure 8.8, strongly reduced productions will result in sequences of minimal duration (30 ms) segments. High rates of short durations are thus indicative of temporal reduction.

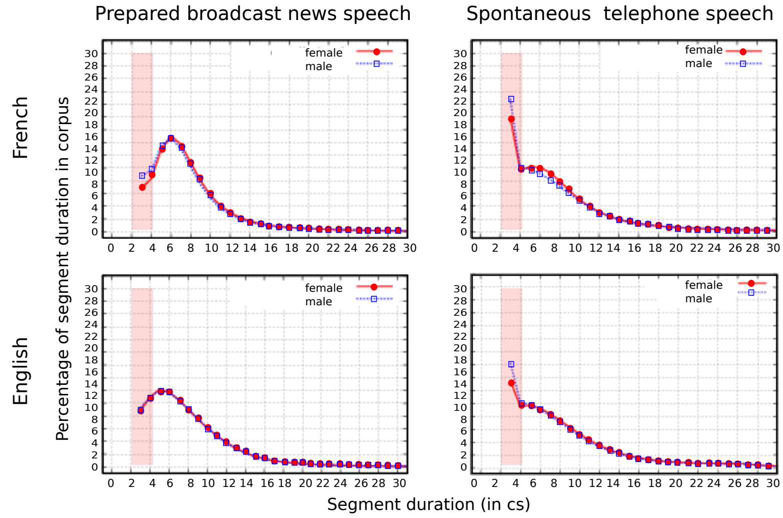

Figure 8.9 Phone segment duration distribution (all phone segments pooled). Comparison between broadcast news (left panels) and spontaneous telephone speech (right panels). The highlighted region (3-4 cs) corresponds to potentially reduced segments.

The top part of Figure 8.9 provides a histogram of proportions of segments in French as a function of segment duration, with corresponding histograms for English on the bottom. To save space on the abscissa, durations are given in centiseconds (cs) in the figures and not in milliseconds (ms) as in the text. The results are broken down into prepared (left) and conversational (right) speech styles. Concerning prepared speech, the largest number of segments (> 14% in French, 13% in English) have a duration of 60 ms in French and 50 ms in English. With respect to the spontaneous telephone speech, the French distribution has by far the most segments (>18%) in the shortest duration bin of 30 ms, with almost one-third of the segments have a duration of 30 or 40 ms. As for French, English spontaneous speech also has the highest proportion of segments (15%) in the minimal duration bin and 25% of the segments have a duration of up to 40 ms. The same trends are observed for male and female speakers. The high proportion of short duration segments (highlighted in pink in Figure 8.9) in spontaneous speech suggests that temporal reduction is an important issue to address in order to improve our knowledge of native pronunciations and related phonological processes in spontaneous speech and to increase the acoustic modeling accuracy in ASR. Even though the proportion of minimal duration segments is much lower in prepared speech, there still are 8% of all segments with 30 ms duration and 18% with 30 or 40 ms duration in both languages. If the minimal duration segments highlight temporal reduction, the distributions of Figure 8.9 reveal their strong presence in spontaneous speech. They also reveal that carefully prepared speech is also concerned, although to a lesser extent.

Position-dependent analysis of the English plosives /t/ and /k/

While the duration histograms indicate overall trends, hereafter we will examine the realizations of some English consonants in more detail. How do segment durations vary with their position in a word or a phrase? or as a function of lexical stress? Table 8.3 reports figures for some typical English words highlighting variation in duration of /t/ and /k/ consonants. Words are chosen so as to illustrate the influence on duration of word-initial and medial positions and of changing lexical stress. For each word type, the number of tokens in the corpus and the percentage of minimal (up to 40 ms) duration segments are shown. The chosen words occur rather frequently in both English corpora (BN and CTS). It can be seen that the duration of a phoneme's realization depends on its position in the word. For

Table 8.3 Position-dependent analysis of /t/ and /k/ in some typical English words in conversational broadcast data (left) and in the telephone Switchboard and Fisher conversations (right). Average phone durations are given in ms together with standard deviations. (*) indicates that /C/ is syllable-initial in a lexical stress position.

	/C/ position	Broadcast			SWB/Fisher Conversations		
	/t/	#tkn	avrg. dur. stdev	% min. dur.	#tkn	avrg. dur. stdev	% min. dur.
talking	w-init (*)	814	95 42	5	4898	80 34	11
trying	w-init (*)	684	95 43	6	4464	85 38	11
nineteen	w-mid (*)	560	80 23	7	706	89 21	8
hotel	w-mid (*)	105	118 29	0	178	126 32	1
ninety	w-mid	323	70 27	20	821	43 26	22
getting	w-mid	803	59 33	52	5692	39 21	86
little	w-mid	1041	59 31	41	9379	37 28	91
exactly	w-mid	387	54 29	43	6328	39 29	85
	/k/						
coming	w-init (*)	825	108 52	4	2301	96 35	2
conversation	w-init	110	92 28	3	610	88 27	6
doctor	w-mid	85	84 34	9	649	65 25	21
focus	w-mid	125	83 28	6	254	69 20	10
because	w-mid (*)	5342	80 35	16	32062	88 38	9
basically	w-mid	399	52 30	53	2499	64 28	30

example, in lexical stress-bearing syllables (marked (*) in Table 8.3; [t] in *talking, trying, nineteen, hotel*) or /k/ in *coming, because*), average durations of syllable-initial consonants are all higher than 80 ms, and rates of minimum duration segments remain low. In contrast, these rates tend to increase for consonants in syllable coda positions and in atone syllables in general. Similarly, average segment durations tend to decrease, more strongly for /t/ than for /k/. It can be observed that /t/ is more likely to undergo temporal reduction than /k/ in the shown examples. The highest minimum duration rates are observed for a [t] in a consonantal environment [k_l] (in *exactly*).

In general, it can be seen that segmental durations are lower for telephone conversation data than for broadcast data, exception made for /t/ in *hotel* and /k/ in the two last lines in Table 8.3. The latter need some additional comments: *because* and *basically* had proven to be often shortened in spontaneous English and thus had additional reduced pronunciations (/kɔz/ and /besɪkli/) in the CTS pronunciation dictionary. 40% of the *because* tokens were thus aligned with the 3-phone pronunciation (deletion of atone syllable *be-*) and all *basically* tokens preferred the shortest variant. The option of shorter pronunciations in CTS data during forced alignment thus resulted in somewhat longer segment durations as in BN data where these shorter pronunciations were not provided in the dictionary. This study suggests that it is interesting to consider also including such reduced variants in the BN pronunciation dictionary.

Word-internal duration variation

Another way of examining position dependence in the phonetic realization of segments is illustrated in

Figure 8.10, which depicts the temporal realization of sample polysyllabic words in English and French. Similar words were selected (*government, governments*, *gouvernement, gouvernementale*) in both English and French. The words were extracted from the BN corpora, pooling occurrences in all phrasal positions and were frequent enough so as to consider the measurements as speaker-independent. The hypothesis is that the average durations of unstressed segments are much shorter than those of stressed ones and are also shorter than the overall average segment duration. Some of them may even fall in the minimum duration zone (shown in red) for which adding shortened pronunciations to the pronunciation dictionary could be envisioned. It is interesting to see that the duration profiles are very similar for identical lemmas, even though the number of occurrences differs importantly. The longer words tend to have shorter segment durations in unstressed parts. It is also nice to observe that English words tend to have longer segments due to lexical stress in the word-initial part, and French longer segments in the word-final part which often co-occurs with phrase-final position. Some of the final lengthening may also be due to pre-pausal position, which is not explicitly denoted in our data.

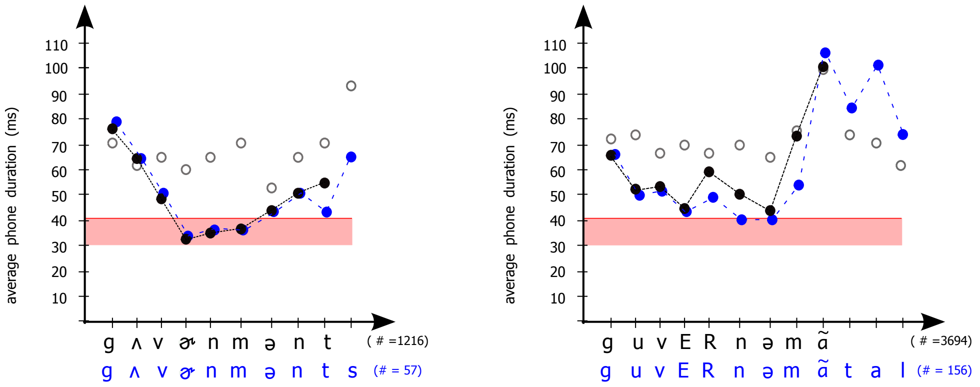

Figure 8.10 Average segment durations of polysyllabic content words as obtained by automatic alignment of BN data. Left: English *government* (in black), *governments* (in blue). Right: French: *gouvernement* (in black), *gouvernemental* (in blue). The abscissa shows the phone labels (and # of word occurrences) and the ordinate the average phone duration in ms. For comparison, the empty circles show the overall average phone durations.

8.6.2 Alignments using reduced variants

After having observed that many speech stretches are aligned with minimal durations, a sensible solution then consists of anticipating shorter pronunciation variants in the dictionary. In this part, we thus move away from segmental duration investigations to study the usage of these shorter variants during forced alignment.

Temporally reduced stretches of speech may correspond to relatively known phenomena (e.g. *dunno* in English). The use of "multi-words" which merge potentially reduced word sequences into one single pronunciation dictionary entry enable the introduction of shorter pronunciations accounting for cross-word phenomena. Multi-words were introduced in ASR (Stolcke et al. 2000, Strik et al. 2005) to tackle this spontaneous or fluent speech specific reduction problem. The rationale of “multi-words” is to limit the proposed reductions to these and only these word sequences preventing their broader usage in a general cross-word situation. Hence, they generally correspond to highly frequent word bigrams. Analysis of automatic speech recognition errors on spontaneous speech reveals that temporally reduced stretches of speech may also occur in less frequent word bigrams (see Table 8.1: *semble que* ’seems that’ recognized as *somme que* ’sum that’ due to word final CC cluster dropping in prosodic word internal position).

English variants

The effectiveness of shorter pronunciation variants in multi-words was studied via forced alignment in a subset of the English CTS Switchboard corpus (185 hours of speech). 250 multiwords were introduced for a vocabulary of about 25k word types. For this experiment, common reduced written forms (e.g. *didn’t, you’ve*) of the manual transcripts were matched and pooled with the corresponding full multi-word form *did-not, you-have*. Table 8.4 shows some typical examples, with the different pronunciations proposed in the dictionary: full form and shorter variants. To limit the number of variants to be displayed in the table, equal length variants which differ only in vowel quality (e.g. [tu], [tə]) were merged. The variants are ordered by decreasing length, with the most frequently aligned one shown in boldface. For each multi-word, the table indicates the frequency of each variant (i.e., the number of times it was used during alignment (#Align) and its corresponding percentage ($\frac{\#Align}{\#Total}$) in the total number of tokens of the multi-word (#Total). It is interesting to describe the different transformations when moving from the full form pronunciation to the more reduced ones. It can be observed that the widely studied vowel reduction process (change of vowel quality of peripheral vowels towards a more central schwa) often accompanies pronunciation shortening, unless the vowel completely disappears from the variant. Syllabic consonants reflect the merging of a schwa with following consonant (n,m,l,r). A metathesis can result in a /r/-vowel sequence (such as in the word *hundred*) being realized as /ɚ/. The consonants /t,d/ are easily deleted especially in homorganic consonant clusters and in coda positions. An inter-vocalic /h/ appears to be elidible even in onset position.

Table 8.4 Examples of ASR multi-words with shortened pronunciation variants to deal with temporal reductions in spontaneous English Switchboard data. For each multi-word type are given: the total number of tokens, the different pronunciation hypotheses (full form and variants) of the dictionary along with the number of tokens aligned with each one, and the corresponding ratio (#Align/#Total).

Multi-word	#Total	*Full form* + *Variants*	#Align	#Align / #Total	Comments
English spontaneous speech – Switchboard data					
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
going-to-be	750	gɔɪŋtʊbi	73	9.7	full form
		+ **gɔnəbi**	432	57.6	complex: ɪŋt → 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

French variants

For French, we have not yet investigated the use of multi-words for ASR. Unlike English, French standard writing does not tend to provide reduced written forms even though they may appear in oral productions (*je ne sais pas* 'I don't know' may be written as *je sais pas* in less formal writing, but never as *chais pas* even though this is a most common pronunciation in spontaneous French). For the NCCFr casual face-to-face speech, shortened pronunciations were introduced in the pronunciation dictionary to test whether they would be selected for temporally reduced words during speech alignment. Table 8.5 shows examples of spontaneous speech reductions in single French words taken from the automatic alignments of the NCCFr corpus.

Table 8.5 Examples of words with shortened pronunciation variants introduced to handle temporal reductions in spontaneous French NCCFr data. For each entry the total number of tokens, the different pronunciation hypotheses (full form and variants) of the dictionary are given. The number of tokens aligned with each variant, and the corresponding ratio (#Align/#Total) are specified. The most popular variant is shown in bold.

Word	#Total	*Full form* + *Variants*	#Align	$\frac{\#\text{Align}}{\#\text{Total}}$	Comments
French spontaneous speech – NCCFr data					
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
peut-être	636	pøtɛtʀə	18	2.8	full form
’maybe’		+ pøtɛtʀ	28	6.0	no final schwa
		+ p(ø|ə)tɛt	109	17.1	final cluster simplification
		+ **ptɛt**	481	75.6	+ unaccented vowel deletion
maintenant	352	mɛ̃tənɑ̃	8	2.3	full form
’now’		+ mɛ̃tnɑ̃	114	32.4	no internal schwa
		+ **mɛ̃nɑ̃**	230	65.3	+ /t/-deletion
quelques	56	kɛlkə	14	25	full form
’some’		+ **kɛkə**	28	50	+ /l/-deletion
		+ kɛ(k|g)	14	25	+ schwa deletion

Different phonological processes are seen to be active in reduced pronunciations in French. Among these, schwa-vowel deletion in final but also word-internal position is certainly the most pervasive one. As a result, the rhythmic pattern changes with a smaller number of more complex syllables. The French schwa is typically considered as optional: whether or not it is realized depends on the speaker, his/her regional origins, his/her speaking rate, the embedding context, the length of the prosodic word... Consonant clusters in syllable coda positions are often simplified. In particular liquids (/R/ and /l/) tend to disappear not only in word-final plosive-liquid clusters (*être → êt'* 'to be'; *montre → mont'*) 'show', but also in syllable coda position before another consonant (*parce que → pa'ce que* 'because'; *quelque → que'que* 'some'; *film → fi'm* 'movie'). In contrast, the schwa vowel, although often very short, is more systematically produced in English.

Table 8.5 also exemplifies /t/ deletions (in *main(te)nant* 'now'), however they tend to occur in homorganic consonant neighborhoods. Beyond schwa vowel deletion, we can observe that vowels in unaccented positions may disappear. For example, the frequent French word *peut-être* 'maybe' tends to be pronounced [ptɛt] in casual speech with a loss of the central rounded /ø/ vowel besides the simplification of the final consonant cluster. Another important process contributing to temporal reduction in spontaneous French (but not examined here) corresponds to vowel deletion (be it V1 or V2) in V#V contacts in cross-word situations. A typical example here is the *t'as* 'you've' production instead of *tu as* 'you have'. ASR error analysis often pointed out such cases be they located in highly frequent words such as *tu as* or in less frequent ones. V#V contacts are good candidates to undergo reduction with either vowel deletion or vowel merging. Temporal reduction may hence become more or less severe, depending on the cascade of phonological processes involved. We hope that the proposed descriptive work and methodology to spot temporal reductions in spontaneous speech will contribute to better disentangle the complexities of speech production and perception.

8.7 Discussion

In this chapter, we introduced the idea of using forced alignments to locate temporally reduced sequences in fluent speech. When used with full form pronunciation dictionaries, sequences of minimal duration segments reveal potentially reduced productions. Although the exact phone labels and time stamps of the aligned segments of the temporally shortened regions should not be taken as ground truth, the detected sequence is certainly pronounced differently than the predicted full form. The larger the number of contiguous minimum duration segments, the stronger the hypothesis of an actual temporal reduction.

Whereas reduction, and more specifically temporal reduction, is often considered to be specific to casual or at least spontaneous speech, our comparative investigations of both prepared and spontaneous speech in English and French reveal that temporal reduction exists in both speech styles, although to a lesser extent in the former as can be expected. We believe that similar mechanisms underly the production and processing of pronunciation variants in the different speaking styles. Temporal reduction involves unstressed stretches of speech more often than regions of focus. When examining the words most frequently included in minimal duration sequences, it is not surprising to find high frequency function words. Frequency might thus be one of the factors explaining such reduction. However, considering reduced sequences in low frequency words, major factors seem to be related to repetition (which is equivalent to a local boost in frequency) and to a prosodic grouping in an unstressed position. In all of the examined cases, reduced sequences are embedded in unstressed or non-emphasized portions of speech.

Prepared, formal, non interactive speech is generally uttered in a relatively steady tempo, whereas spontaneous speech undergoes substantial fluctuations in tempo. Interactive speech also includes more discourse markers which are particularly prone to temporal reduction. Many words and phrases may serve as discourse markers. For example, the high frequency of *I_don't_know* in English or *je_sais_pas* in French with their variously reduced surface forms in spontaneous speech is more often related to a discourse marker function, than to the expression of a lack of knowledge.

Temporally reduced speech is challenging for current ASR systems and for non native listeners, resulting in misrecognitions. A practical approach taken for ASR is the introduction of multiword expressions which allow shorter pronunciations to be associated with the word sequence. The English expression *sort of*, which is frequently observed in both BN and CTS speech corpora and tends to be produced very quickly, is a good candidate for a multi-word expression. Our investigations for French confirmed that the examples shown at the beginning of the chapter (*je crois que*, 'I believe that'; *je ne sais pas*, 'I don't know') are good candidates for multi-word modeling in ASR. They generally have a very low average phone duration and are discourse marker-like expressions comparable to multiwords in English.

These expressions and corresponding audio samples which can help improve the performance of ASR systems, could also be helpful for L2 training to better survive in a native speakers' environment.

By using forced alignment to quantify temporal reduction phenomena we have tried to demonstrate how ASR systems may serve as a tool to systematically investigate variations across different speaking styles and languages. We hope that the present results will shed some new light on the intrinsically complex nature of temporal processes in speech. In future work, we plan to refine the present approach and to further extend the analysis of the alignment results, with the aim of using this approach to discover new pronunciation variants attributable to temporal reduction. Studying linguistic phenomena from an ASR perspective using large corpora might also give us some clues about the encoding of information in speech. The speech signal is endowed with many fine phonetic details and features that the human listener is somehow able to rely on even in the face of ambiguity and noise. The perspectives available through an ASR approach are manifold. For researchers working in the domain of ASR, the ultimate goal is to uncover rules to improve pronunciation modeling. These rules can be applied to rarely observed or unobserved words, for which pronunciation variants cannot be estimated statistically. The framework developed should help to describe and quantify more or less well known linguistic phenomena on the phonemic and lexical levels, which is of relevance to linguists and cognitive scientists alike.

8.8 Acknowledgment

Parts of the research reported in this article have been funded by grants from the CNRS, the French Investissements d'Avenir - Labex EFL program (ANR-10-LABX-0083) and ANR Vera and Quaero. We would like to thank Jean-Luc Gauvain, and Gilles Adda for their help.

References

Adda-Decker, Martine & Natalie Snoeren (2011). Quantifying temporal speech reduction in French using forced speech alignment. *Journal of Phonetics*, vol. 39, pp. 261–270.

Adda-Decker, Martine, Cédric Gendrot & Noël Nguyen (2008). Contributions du traitement automatique de la parole à l'étude des voyelles orales du français. *Traitement Automatique des Langues*, vol. 49, no. 3, pp. 13–46.

Adda-Decker, Martine (2007). Problèmes posés par le schwa en reconnaissance et en alignement automatiques de la parole. In*Actes des 5es Journées d'Études Linguistiques de Nantes*, Nantes, pp. 211–216.

Adda-Decker, Martine, Philippe Boula de Mareüil, Gilles Adda & Lori Lamel (2005). Investigating syllabic structures and their variation in spontaneous French. *Speech Communication*, vol. 46, pp. 119–139.

Adda-Decker, Martine & Lori Lamel (2005). Do speech recognizers prefer female speakers? In *Proceedings of International Speech Communication Association (ISCA) Interspeech*, Lisbon, pp. 2205–2208.

Adda-Decker, Martine & Lori Lamel (1999). Pronunciation variants across system configuration, language and speaking style. *Speech Communication*, vol. 29, pp. 83–98.

Bartkova, Katarina & Denis Jouvet (2015). Impact of frame rate on automatic speech-text alignment for corpus-based phonetic studies. In *Proceedings of the 18th ICPhS*, Glasgow, paper no 667 (5 pages).

Campbell, Nick (1992). Segmental elasticity and timing in Japanese speech. In Tohkura, Vatikiotis-Bateson, and Sagisaka, Speech perception, production and Linguistic Structure. IOS Press, Amsterdam, Washington, Oxford, pp. 403–418.

Cole, Ronald, Beatrice T. Oshika, Mike Noel, Terri Lander & Mark Fanty (1994). Labeler Agreement in Phonetic Labeling of Continuous Speech. In *Proceedings of International Conference on Speech and Language Processing (ICSLP)*, Yokohama, vol. 2, pp. 2131–2134.

Cutler, Anne, Jacques Mehler, Dennis Norris & Juan Segui (1986). The Syllable's Differing Role in the Segmentation of French and English. *Journal of Memory and Language*, vol. 25, pp. 385–400.

Dauses, August (1973). Études sur l'e instable dans le français familier. Niemeyer Verlag. Tübingen.

Di Canio, Christian, Hosung Nam, Douglas H. Whalen, Timothy Bunnell, Jonathan D. Amith & Rey C. Garcia (2012). Assessing agreement level between forced alignment models with data from endangered language documentation corpora, *Proceedings of International Speech Communication Association (ISCA) Interspeech*, Portland.

Dilley, Laura C. & Mark Pitt (2007). A study of regressive place assimilation in spontaneous speech and its implications for spoken word recognition. *Journal of the Acoustical Society of America*, 122, pp. 2340–2353.

Duez, Danielle, (2003). Modelling Aspects of Reduction and Assimilation in Spontaneous French Speech. In *Proceedings of the IEEE-ISCA Workshop on Spontaneous Speech Processing and Recognition*, Tokyo.

Durand, Jacques, Bernard Laks & Chantal Lyche (2002). La phonologie du français contemporain usages, variétés et structure. In Claus D. Pusch and Wolfgang Raible (eds.) Romanistische Korpuslinguistik - Korpora und gesprochene Sprache / Romance Corpus Linguistics - Corpora and Spoken Language. Tübingen: Gunter Narr Verlag, pp. 93–106.

Durand, Jacques, Bernard Laks & Chantal D. Lyche (2005). Un corpus numérisé pour la phonologie du français. In G. Williams (ed.) La linguistique de corpus. Rennes: Presses Universitaires de Rennes. pp. 205–217.

Elman, Jeffrey L. and James L. McClelland (1988). Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory and Language*, vol. 27, pp. 143–165.

Ernestus, Mirjam (2000). Voice assimilation and segment reduction in casual Dutch, a corpus-based study of the phonology-phonetics interface. Utrecht: LOT.

Ernestus, Mirjam & Natasha Warner (Eds.). (2011). Speech reduction [Special Issue]. *Journal of Phonetics*, vol. 39.

Fougeron, Cécile, Jean-Philippe Goldman and Ulli H. Frauenfelder (2001). Liaison and schwa deletion in French: an effect of lexical frequency and competition. In *Proceedings of ESCA Eurospeech*, Aalborg, pp. 639–642.

Gahl, Susanne (2008). "Time" and "Thyme" Are Not Homophones: The Effect of Lemma Frequency on Word Durations in Spontaneous Speech. *Language*, vol. 84 no. 3, pp. 474–496.

Galliano, Sylvain, Edouard Geoffrois, Djamel Mostefa, Khalid Choukri, Jean-François Bonastre & Guillaume Gravier. (2005). The Ester phase II evaluation campaign for the rich transcription of French broadcast news. In *Proceedings of International Speech Communication Association (ISCA) Interspeech*, Lisbon, pp. 1149–1152.

Ganong, William F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, vol. 6, no. 1, pp. 110–125.

Gauvain, Jean-Luc, Lori Lamel, Gilles Adda & Martine Adda-Decker (1994). Speaker-independent continuous speech dictation. *Speech Communication*, vol. 15, pp. 21–37.

Gauvain, Jean-Luc., Gilles Adda, Martine Adda-Decker, Alexandre Allauzen, Véronique Gendner, Lori Lamel & Holger Schwenk (2005). Where are we in transcribing French broadcast news? In *Proceedings of International Speech Communication Association (ISCA) Interspeech*, Lisbon, pp. 1655–1658.

Gendrot, Cédric & Martine Adda-Decker (2005). Impact of duration on F1/F2 formant values of oral vowels: An automatic analysis of large broadcast news corpora in French and German. In *Proceedings of International Speech Communication Association (ISCA) Interspeech*, Lisbon, pp. 2453–2456.

Godfrey, John J., Edward Holliman & Jane McDaniel (1992). Switchboard: telephone speech corpus for research and development. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (IEEE-ICASSP)*, San Francisco, pp. 517–520.

Greenberg, Steven & Shuangyu Chang (2000). Linguistic dissection of Switchboard-corpus automatic speech recognition systems. In *Proceedings International Speech Communication Association ISCA-ITRW Workshop on ASR*, Paris, pp. 195–202.

Greenberg, Steven, Hannah Carvey, Leah Hitchcock & Shuangyu Chang (2003). Temporal properties of spontaneous speech – a syllable-centric perspective. *Journal of Phonetics*, vol. 31, pp. 465–485.

Hosom, Jean Paul (2009). Speaker-independent phoneme alignment using transition-dependent states. *Speech Communication*, vol. 51, no. 4, pp. 352–368.

Jun, Sun-Ah & Cécile Fougeron (2002). The realizations of the accentual phrase in French intonation. In *Probus (special issue on Intonation in the Romance Languages)*, J. Hualde, ed., vol. 14, pp. 147–172.

Jurafsky, Daniel, Alan Bell, Michelle Gregory & William D. Raymond, (2001). Probabilistic relations between words: Evidence from reduction in lexical production in *Frequency and the Emergence of Linguistic Structure*, Bybee and Hopper eds. John Benjamins, pp. 229–254.

Karanasou, Panagiota & Lori Lamel (2011). Pronunciation variants generation using SMT-inspired approaches. In *36th International Conference on Acoustics, Speech and Signal Processing, IEEE-ICASSP)*, Prague, pp. 4908–4911.

Lamel, Lori & Gilles Adda. (1996). On designing pronunciation lexicons for large vocabulary, continuous speech recognition. In *Proceedings of International Conference on Speech and Language Processing (ICSLP)*, Philadelphia, pp. 6–9.

Lamel, Lori & Jean-Luc Gauvain (2005). Alternate phone models for conversational speech. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (IEEE-ICASSP)*, vol. 1, Philadelphia, pp. 1005–1008.

Lefèvre, Fabrice, Jean-Luc Gauvain & Lori Lamel (2005). Genericity and portability for task-dependent speech recognition. *Computer Speech and Language*, vol. 19, pp. 345–363.

Levelt, Willem J.M. (1989). Speaking: From intention to articulation. Cambridge, MA: MIT Press.

Nakamura, Masanobu, Sadaoki Furui & Koji Iwano (2006). Acoustic and linguistic characterization of spontaneous speech. *International Speech Communication Association (ISCA) workshop on Speech Recognition and Intrinsic Variation*, Toulouse.

Long Nguyen, Sherif Abdou, Mohamed Afify, John Makhoul, Spyros Matsoukas, Richard Schwartz, Bing Xiang, Lori Lamel, Jean-Luc Gauvain, Gilles Adda, Holger Schwenk & Fabrice Lefèvre. (2004). The 2004 BBN/LIMSI 10XRT English broadcast news transcription system. In *Proceedings DARPA Rich Transcription Workshop (RT04)*, Palisades.

Prasad, Rohit, Spyros Matsoukas, Chia-Lin Kao, Jeff Ma., Dong-Xin Xu Thomas Colthurst, Owen Kimball, Richard Schwartz, Jean-Luc Gauvain, Lori Lamel, Holger Schwenk, Gilles Adda & Fabrice Lefèvre (2005). The 2004 BBN/LIMSI 20xRT English conversational telephone speech recognition system. In *Proceedings of International Speech Communication Association (ISCA) Interspeech*, Lisbon, pp. 1645–1648.

Rabiner, Lawrence R. & Biing-Hwang Juang (1986). An introduction to hidden Markov models. *IEEE Acoustics Speech and Signal Processing Magazine*, vol. ASSP-3, no. 1, pp. 4–16. January.

Samuel, Arthur G. & Mark A. Pitt (2003). Lexical activation (and other factors) can mediate compensation for coarticulation. *Journal of Memory and Language*, vol. 48, pp. 416–434.

Schuppler, Barbara, Mirjam Ernestus, Odette Scharenborg and Louis Boves (2008). Corpus of Dutch spontaneous dialogues for automatic phonetic analysis. In *Proceedings of International Speech Communication Association (ISCA) Interspeech*, Brisbane, pp. 1638–1641.

Schuppler, Barbara, Martin Hagmüller, Juan A. Morales-Cordovilla & Hannes Pessentheiner. (2014). GRASS: The Graz corpus of read and spontaneous speech. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, pp. 1465–1470.

Schuppler, Barbara, Martine Adda-Decker & Juan A. Morales-Cordovilla (2014). Pronunciation variation in read and conversational Austrian German. In In *Proceedings of International Speech Communication Association (ISCA) Interspeech*, Singapore, pp. 1453–1457.

Snoeren, Natalie, Pierre Hallé & Juan Segui (2006). A voice for the voiceless: Production and perception of assimilated stops in French. *Journal of Phonetics*, vol. 34, pp. 241–268.

Shriberg, Elizabeth (1994). Preliminaries to a theory of speech disfluencies. PhD thesis, University of California, Berkeley.

Stolcke, Andreas, Harry Bratt, John Butzberger, Horacio Franco, Venkata R. Rao Gadde, Madelaine Plauché, Colleen Rickey, Elizabeth Shriberg, Kemal Sönmez, Fuliang Weng, & Jing Zheng (2000). The SRI March 2000 hub-5 conversational speech transcription system. In *Proceedings NIST Speech Transcription Workshop*, College Park.

Strik, Helmer, Diana Binnenpoorte & Catia Cucchiarini (2005). Multiword expressions in spontaneous speech: Do we really speak like that? In *Proceedings of International Speech Communication Association (ISCA) Interspeech*, Lisbon, pp. 1161–1164.

Strik, Helmer, Anna Elffers, Dusan Bavcar & Catia Cucchiarini (2006). Half a word is enough for listeners, but problematic for ASR. In *Proceedings of International Speech Communication Association (ISCA) workshop on Speech Recognition and Intrinsic Variation*, Toulouse, pp. 101–106.

Strik, Helmer & Catia Cucchiarini (1999). Modelling pronunciation variation for ASR: A survey of the litterature. *Speech Communication*, vol. 29, pp. 225–246.

Van Son, Robert J.J.H. & Louis C.W. Pols (2003). An acoustic model of communicative efficiency in consonants and vowels taking into account context distinctiveness. In *Proceedings of the 15th ICPhS*, Barcelona, pp. 2141–2143.

Torreira, Francisco, Martine Adda-Decker & Mirjam Ernestus (2010). The Nijmegen corpus of casual French, *Speech Communication*, vol. 10, no. 3, pp. 201–212.

Torreira, Francisco & Mirjam Ernestus (2012). Weakening of intervocalic /s/ in the Nijmegen corpus of casual Spanish. *Phonetica,* vol. 69, pp. 124–148.

Tseng, Shu-Chuan (2005). Contracted syllables in Mandarin: Evidence from spontaneous conversation. *Language and Linguistics*, pp. 153–180.

Vasilescu, Ioana, Martine Adda-Decker, Lori Lamel & Pierre Hallé (2009). A perceptual investigation of speech transcription errors involving frequent near-homophones in French and American English. *Proceedings International Speech Communication Association (ISCA) Interspeech*, Brighton, pp. 144—147.

Vasilescu, Ioana, Dahlia Yahia, Natalie Snoeren, Lori Lamel & Martine Adda-Decker (2011). Cross-lingual study of ASR errors: on the role of the context in human perception of near homophones, *Proceedings of International Speech Communication Association (ISCA) Interspeech*, Florence, pp. 1949—1952.

Whalen, Douglas H. (1991). Infrequent words are longer in duration than frequent words. *Journal of the Acoustical Society of America*, vol. 90, no. 4, pp. 2311.