Quantifying Temporal Speech Reduction in French using Forced Speech Alignment

Martine Adda-Decker and Natalie D. Snoeren

LIMSI-CNRS
B.P. 133, 91403 Orsay, France
{madda,nsnoeren}@limsi.fr

Abstract

The processing of speech reduction remains one of the major challenges to automatic speech recognition (ASR) systems, as speech reduction often results in a considerable number of automatic transcription errors. Conversely, automatic transcription errors may indicate interesting reduction phenomena. The present article focuses on temporal speech reduction in spoken French. In a series of explorations, we examined large speech corpora with respect to production variation in different speech styles, and in particular, to shorter pronunciations and disappearing sounds. An ASR tool was used, forced speech alignment, that allows one to locate and to quantify speech regions prone to temporal reductions. Our study made use of various styles of large speech corpora, including broadcast news, as well as telephone and face-to-face conversations, thereby including both casual and careful speech. The results highlight the increasing impact of temporal speech reduction with less formal, more spontaneous and interactive speaking styles. In a broader sense, our study provides a demonstration of how ASR systems can be employed to consistently explore variations in speech in virtually unlimited speech corpora.

Key words: Automatic speech recognition, forced alignment, speech reduction in French, pronunciation variants, speech style, segment duration distribution.

1. Introduction

An important bottleneck to a large-scale expansion of automatic speech recognition (ASR) devices is the efficient processing of spontaneous or ca-

sual speech. Within the ASR community "spontaneous" generally refers to all kinds of phenomena that make the speech signal deviate from a carefully articulated sequence of words and sounds. Problematic topics in casual speech include word fragments and various sorts of disfluencies (Shriberg, 1994), speech sounds that overlap with other sounds, such as laughter or crying, produced by the same speakers, phonological variants, underarticulated speech and, last but certainly not least, speech reductions resulting in missing sounds. Many speech scientists share the belief that much knowledge can be gained from studying characteristics of casual speech (Greenberg & Chang, 2000; Greenberg, Carvey, Hitchcock, & Chang, 2003; Nakamura, Furui, & Iwano 2006; Strik, Elffers, Bavcar, & Cucchiarini, 2006). For instance, Greenberg and colleagues (2000, 2002) have investigated syllabic structures in casual speech from the SwitchBoard data, Nakamura et al. (2006) compared spectral properties of careful and casual speech on large Japanese corpora, thereby highlighting spectral reduction. Strik and colleagues (2006) have 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, Bell, Gregory & Raymong, 2001), i.e., function words that are predictable from the context, idioms, morphological items (in particular endings), dates, discourse markers etc. Speech reduction produces either different (centralised) phonemes, fewer phonemes, or even fewer syllables (Ernestus, 2000; Van Son & Pols, 2003; Duez, 2003; Adda-Decker, Boula de Mareuïl, Adda, & Lamel, 2005).

As far as phonemic segmentation and labelling is concerned, it is far from being 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 & McClelland, 1988; Samuel & Pitt, 2003). A given ASR system, on the other hand, will consistently make the same decisions over the entire corpus, and can be parameterized to best fit the investigator's needs.

Many studies have addressed the issue of phone boundary reliability, as well as agreement between several manual and/or automatic annotations. In early work, Cole, Oshika, Noel, Lander and Fanty (1994) observed that

boundary location achieved 80% inter-annotator agreement regarding manually labelled phone boundaries with a 10 ms tolerance. More recently, over 90% agreement with a 20 ms tolerance has been reported between several automatic alignments and manual labelling with a 20 ms tolerance (cf. Hosom, 2009). The overall reliability of automatic labelling through forced alignment can be considered close to the one achieved by human experts. Following these results, ASR systems can be used for a large panel of empirical studies in that they allow one to consistently investigate variations in large speech corpora in terms of known influential parameters, such as speaking style, gender, dialectal accents, and emotions. In this paper, however, we propose a method to provide evidence of temporal speech reduction with the help of global descriptors, such as phone segment duration distributions. Increasing proportions of short segments are considered as indicative of a higher density of temporal reduction. Corresponding speech regions need to be further studied in order to gain deeper insight in the complexity of spontaneous speech specific reduction phenomena, to increase our understanding of the general mechanisms underlying pronunciation variation and last but not least, to contribute to better acoustic speech models for ASR in the future.

In the current paper, we are concerned with uncovering some of these processes, in particular temporal speech reduction in French. The goal of the present research is to identify speech regions that are prone to reduction using the forced speech alignment tool (Adda-Decker & Lamel, 1999) based on the LIMSI speech recognition system (Gauvain, Lamel, Adda, & Adda-Decker, 1994). It is demonstrated that forced speech alignment can be employed to quantify speech reduction in large spoken corpora. We carried out the investigations with a special focus on various speech styles, ranging from broadcast news to telephone and face-to-face conversations. By using large speech corpora, we aimed to find out whether the extent to which speech reduction is observed varies as a function of different speech styles and languages (French and English). Moreover, we looked into the question of whether vowels and consonants are more or less prone to temporal reduction. The final section summarizes the main results and discusses the implications of the outcomes of the corpus-based study. Before we turn to our corpus-based study, however, we will first provide a short overview of speech reduction and the type of ASR errors it may give rise to in the French language, as the proposed investigations are originally motivated by ASR error analyses.

2. Speech reduction and ASR errors

2.1. Speech reduction

A considerable amount of research has been devoted to the study of speech reduction phenomena, including consonant lenition, consonant cluster simplifications, vowel reduction and syllable restructuring (see, e.g., Van Son & Pols, 2003; Duez, 2003; Dilley & Pitt, 2007; Ernestus, 2000; Tseng, 2005). Temporal structure reduction is frequently observed in spoken English (e.g., *isn't it* or *it's*) and spoken German (*ins* instead of *in das*, 'in the'). In French, similar reduction phenomena occur. However, in the remainder of this paper we are solely concerned with less explicit temporal reduction phenomena. Such reduced pronunciations are generally not reflected in normative written sources: *il y a* [ilija] ('there is') is most often uttered as *y a* [ja], and *je ne sais pas*, [ʒənəsɛpa] ('I don't know') may have acoustic realisations close to [ʃɛpa] or even [ʃpa], where the *ne* in the negative form *ne ... pas* is being omitted and /ʒə/ and /s/ are merged to form a mere fricative segment with a [ʃ]-like sound. Moreover, the /ɛ/-vowel may become devoiced and merged with the preceding fricative segment. As these examples illustrate, the scope of sequential reductions in French often surpasses word boundaries. Typically one or more short function words are involved. One common means of addressing this problem from an ASR perspective in speech processing is by adding "multiwords" in the pronunciation dictionary for observed sequences of words that tend to co-occur more frequently than chance (see Strik & Cucchiarini, 1999; Strik, Binnenpoorte, & Cucchiarini, 2005). For English, *want to* can thus accept a pronunciation variant like [wʌnə] and for French *je ne sais pas* can even receive a [ʃpa] reduced pronunciation variant. Strong temporal reductions may also appear within groups of content words, such as dates and compound nouns, as is illustrated by an example in English (see Figure 1). The temporally reduced portion *student athletes* is further detailed in Figure 2: a manual phonetic transcription of *student* is given together with a canonical full form obtained by forced alignment. We will come back to this example in the Method section. Before turning to our corpus-based study and the related methodology, we will first discuss some of the most frequent transcription errors that are observed for French.

2.2. Typical transcription errors in French

Previous studies have reported about 10% word error rates for French careful (i.e., journalistic) speech and above 15% for casual telephone speech,

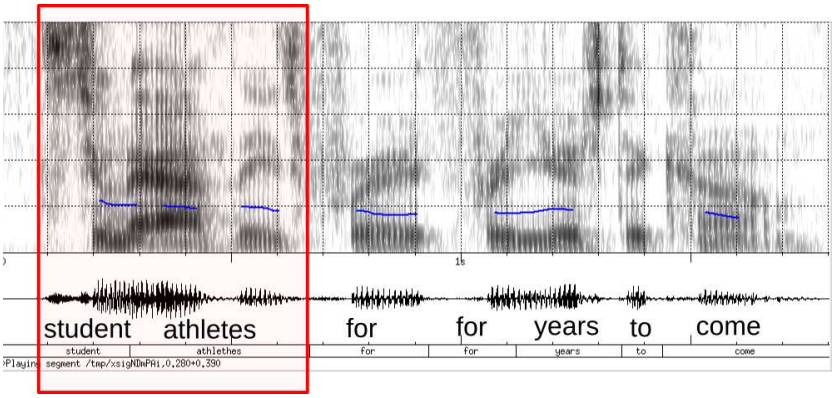

Figure 1: Speech signal of a reduction phenomenon in an English compound: *student athletes* /stjudənt æθlits/, is approximately being produced as [stjunæθlits] within the carrier sentence *it's gonna benefit student athletes (for) for years to come*.

Reference	*Pronunciation*	*Hypothesis*	*Pronunciation*
a	/a/	à	/a/
sait	/sɛ/	c'est	/sɛ/
cesse	/sɛs/	seize	/sɛz/
rentre	/ʁɑ̃tʁə/	rendre	/ʁɑ̃dʁə/

Table 1: (Near-)homophone substitutions in French that do not contain temporal reduction. Both the reference words and their transcription output hypotheses are indicated with their corresponding pronunciations.

with hundreds of hours of appropriate casual speech data and complex system combinations (see Lefèvre, Gauvain & Lamel 2005; Prasad et al, 2005). Among automatic transcription errors in careful speech, approximatively 30-40% of errors consist of homophone or near-homophone errors without temporal reduction. In Table 1, some examples are given of typical confusions that involve frequent and less frequent (near)-homophones. Reference words are substituted by the more frequently occurring homophone hypotheses. For example, *sait* ('knows') may thus be replaced by *c'est* ('that is'), the latter having a higher prior probability in the ASR language model. Table 2 shows examples of near-homophone transcription errors due to typical temporal reduction phenomena in French, including cross-word vowel merging, word-final consonant cluster simplification and short function word deletion.

The above-mentioned examples correspond to journalistic, i.e. carefully

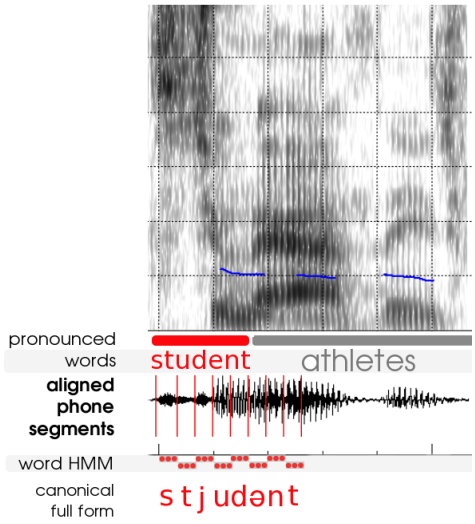

Figure 2: Zoom on the mismatched automatic alignment of the temporally reduced speech portion from the previous Figure. The noun *student* approximately being produced as [stjʊn] corresponds to the red bar portion. This is much shorter than the minimal duration of the automatically aligned word HMM based on the canonical full form /stjudənt/.

Reference	*Pronunciation*	*Hypothesis*	*Pronunciation*
ça avait	/saavɛ/	savaient	/savɛ/
semble que	/sɑ̃bləkə/	somme que	/sɔmkə/
parce que	/paʁsəkə/	ce que	/səkə/
près de Paris	/pʁɛdəpaʁi/	préparé	/pʁepaʁe/

Table 2: Near-homophone errors with temporal reduction: the system's *Hypothesis* pronunciation is shorter than the *Reference* pronunciation.

prepared speech. When analysing casual, conversational speech, however, the proportion of errors due to temporal reduction increases significantly. The proportion of short word sequences, in particular 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 said to him')) is particularly high. Consequently, these sequences are often prone to recognition errors, unless specific acoustic models have been trained on these specific sequences.

Finally, in French the schwa is known to contribute to temporal structure variation, giving rise to a wealth of automatic transcription errors. In Table 3, some illustrations of errors are listed that are due to insertions and deletions of the French schwa (see Verney Pleasants, 1956; Dauses, 1973; Fougeron,

Goldman & Frauenfelder, 2001; Adda-Decker, 2007). The listed examples are chosen from the French ESTER-2005 ASR system evaluation campaign and nicely illustrate the impact of relatively simple structure variations on recognition performance. About 5% of the errors can be related to appearing or disappearing schwas, leading to sequences that are simply incompatible with what was actually predicted by the acoustic word models.

Reference word string	**Decoded word string**	**Comment**
Absence of schwa in dict. & acoustic model, but **produced** by speaker		
Marc Blondel	**marque** Blondel	consonant cluster
en fait	en fait **de**	final release
le week-end pascal	le week-end pascal **le**	phrase boundary
Presence of schwa in dict. & acoustic model, but **not produced** by speaker		
tout le temps	tout _ temps	idiom
temps de leur installation	temps _ leur installation	noun phrase
quai de Seine	**quête saine**	compound + assimilation
c'était le même marasme	c'**est elle** même marasme	homophone
confiance appréciable le tandem	confiance appréciable _ tandem	homophone

Table 3: Transcription errors arising from mismatches in temporal structures between the observations and the model due to word-final schwa. In the upper panel, the schwa was missing in the model, but observed in the speech signal. The lower panel illustrates the reverse situation.

The examples in the top panel show contexts where a schwa is being produced in the speech signal without any evidence in the written forms. The epenthetic schwa is due to contextual effects such as the creation of a complex consonant cluster at word boundaries (here /ʁkbl/ of the proper name *Marc Blondel*), or phrase-final releases. In this particular case, the French language permits near-homophone insertions of short and frequent function words such as *de* ('of') and *le* ('the').

The examples in the lower panel correspond to the more frequent situation of a mismatch due to shorter productions on the one hand and longer models on the other hand. An example worth mentioning is *quai de Seine* ('Seine bank') that illustrates the process of schwa deletion (of the word *de*) before a consonant (/s/ of *Seine*) after an open syllable (/kɛ/ of *quai*). Moreover the /d/ is assimilated to [t] due to the following unvoiced /s/ (cf. Snoeren, Hallé & Segui, 2006). In casual speech, the situation of complex combinations

of various reduction processes appear to be very common. Using large corpora therefore enables us to give a synthetic and exhaustive overview of the various reduction processes (cf. Schuppler, Ernestus, Scharenborg & Boves, 2008). As a first step in this direction, we propose to quantify temporal reductions using forced alignment and canonical pronunciations. This allows us to measure deviations from canonical temporal structures in terms of their phone segment duration distributions.

3. A corpus-based study using forced alignment

Forced speech alignment consists in linking a reference transcription to its acoustic speech signal using an ASR system. The resulting word and subword (typically phoneme) boundaries are determined with respect to the ASR system's configuration (acoustic phone models, model topology, word pronunciations, inter-word optional silences and so forth). Forced alignment is typically used to automatically label large manually transcribed speech corpora for acoustic model training. The resulting phone labels and boundaries are not necessarily in line with manual phonetic segmentation, nor completely compatible with different system configurations. However, previous studies have shown that major linguistic trends (e.g. vowel reduction and duration) can be consistently observed whilst using ASR systems developed independently by different research teams (Adda-Decker, Gendrot & Nguyen, 2008).

In a first series of explorations, we compared different speaking styles using phone segment duration distributions as obtained by forced alignment (Adda-Decker & Lamel, 2005). The questions we were interested in are the following. First, what is the effect of casual speech on the duration distribution as compared to careful speech? Second, how does the French data-set compare to the English data-set? Third, do the observed results hold for different types of casual speech? Finally, does the extent to which speech reduction occurs vary for vowels and for consonants? For this latter comparison between vowels and consonants, all segment duration distributions were examined using the following four duration classes:

short:		<=	40 ms
medium:	50	-	110 ms
long:	120	-	240 ms
very long:		>=	250 ms

Similar duration classes have proven to be useful in showing the impact of duration on F1/F2 formant values of oral vowels (see Gendrot & Adda-Decker, 2005). The *very long* duration class was mainly introduced to check the proportion of very long segments and includes silence and sounds other than speech (e.g., hesitations). Most temporal reductions of interest are assumed to be found in the duration class labelled *short*. It is important to point out that the tuning of 40 ms as upper duration limit should be considered as a permissive one, and by no means as a norm. On the basis of our speech data, it has been observed that a tighter limit of 30 ms would reduce the proportion of segments in this class by 50% for careful speech, and by 30% for casual speech. A further motivation for introducing these duration classes is to contribute to future improvements of acoustic pronunciation modeling by estimating specific acoustic models for each class.

3.1. Method

In the alignment system used here, each acoustic phone model corresponds to a three state left-to-right hidden Markov model (HMM). Each phoneme is associated to one (or several) context-dependent acoustic model(s) corresponding to the three state left-to-right HMMs. The three states are assumed to model phone segment onset (first state), middle (second state) and end parts (third state). In forced alignment, each state is associated to at least one acoustic vector of 10 ms. As there are three states, the minimum duration of a phone segment amounts to 30 ms. Figure 2 illustrates forced alignment with a canonical (citation form) word pronunciation in a temporal reduction situation. From an ASR speech modelling perspective, temporal speech reduction may be captured by sequential pronunciation variants with a smaller number of phonemes than the canonical pronunciation. Such productions, when aligned against canonical pronunciations, generally include one or several phone segments of a minimal duration (i.e. 30 ms here). These types of variants are considered to be the most problematic to ASR systems, since improper alignment results in poorer acoustic phone model accuracy.

Figure 3 gives an overview of how the ASR system can be used as an instrument for linguistic purposes. Starting with manually transcribed speech at the word-level that is not time-aligned with the audio files, pronunciations that deviate from an expected norm may be located. The expected norm may be a citation form pronunciation or simply an a priori fixed minimum word or phone segment duration. The idea is to select portions of speech

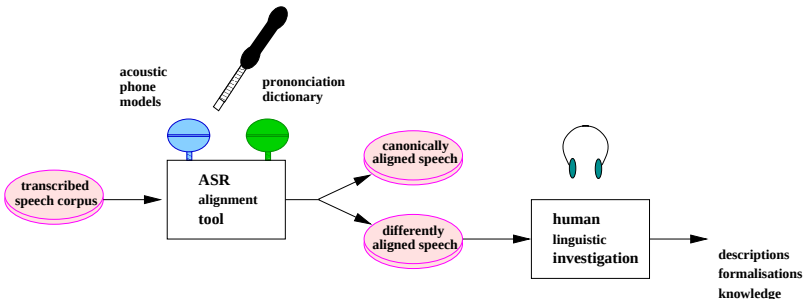

Figure 3: The automatic speech recognizer as an instrument to automatically select canonically and differently aligned subsets of speech deviating from expected representation. These subsets are of interest for more in-depth linguistic investigations.

that include a high proportion of such deviant forms (if they exist at all). Ideally, the adapted ASR tool might select the exact subset including all the deviant forms. In our case, the question of interest is temporal speech reduction, and therefore the number of deviant forms that are selected simply correlates with the number of low duration segments that is obtained by forced alignment. These subsets can then be studied more extensively by carrying out specific measurements. This approach may not only result in the improvement of the ASR model, but it may also enhance our knowledge of the linguistic phenomenon under investigation.

The proposed method aims at highlighting temporal reduction tendencies in large speech corpora via two different system configurations:

- In the first configuration, the alignment system makes use of a canonical pronunciation dictionary (see Table 4), which corresponds to the basic acoustic word model assumption of linear phoneme sequences as beads on a string. The approach consists of focusing on temporal structure mismatches

word	*canonical pronunciation*
le	lə
les	le lez(V)
maintenant	mɛ̃tənɑ̃

Table 4: Excerpt of the canonical pronunciation dictionary.

between the produced speech and the corresponding word models (i.e. HMM topologies of phonemic pronunciation models) as illustrated in Figure 2. The forced speech alignment technique matches the temporal structure of the

models onto the observed speech signal. Consequently, temporal reduction phenomena should correlate with higher rates of very short segments.

• A second configuration makes use of a pronunciation dictionary including optional schwas. Table 5 shows some lexical entries with a maximal number of pronunciations (including all possible phonemic segments as suggested either by the French writing conventions, or by pronunciation habits). The maximum length pronunciation thus includes all possible schwas, reflecting all mute-e occurrences within the word and potentially adding an epenthetic schwa at the end of a word-final closed syllable (without a graphemic presence of mute-e). Temporal reductions that result from schwa deletions can

Word	*Max. length pronunciation + variants*
Potential schwa: within word graphemic mute-e	
l$\underline{e}$	lə + l
c$\underline{e}$la	səla + sla
d$\underline{e}$v$\underline{e}$nu	dəvəny + dəvny dvəny dvny
Potential schwa: closed-syllable word end	
r$\underline{e}$vanch$\underline{e}$	ʁəvɑ̃ʃə + ʁəvɑ̃ʃ ʁvɑ̃ʃə ʁvɑ̃ʃ
d$\underline{e}$v$\underline{e}$nir$\underline{\#}$	dəvənirə + dəvənir dəvnirə...

Table 5: Excerpt of the pronunciation dictionary for the schwa study. All possible variants are summed (+ sign) and arise from within word or from word-final optional schwas.

be measured via schwa deletion rates.

3.2. Speech corpora

For the purpose of our study, several important speech corpora were used, among which broadcast news (BN) and conversational telephone speech (CTS) corpora. The careful speech data-set stems form French broadcast news (BN) and corresponds to 360 hours of various radio and TV shows that were used for the *Technolangue*-ESTER (Galliano et al., 2005) campaign. The casual speech data-set stems from the French telephone conversation (CTS) corpus and corresponds to 120 hours of LIMSI internal resources. French conversations from the corpora often took place between friends and/or family members, so the corpus therefore contains a highly casual speaking style. An additional French corpus (PFC) was used and provides different styles from the same set of users. The PFC (Phonologie du Français Contemporain, http://www.projet-pfc.net/) corpus is the result of an ambitious long-term project, initiated by French phonologists and phoneticians (cf. Durand,

Laks & Lyche, 2002, 2005). Several tens of varieties of the French language from different regions across the French-speaking areas in the world have been gathered. This amounts to the collection of data from hundreds of speakers to enable large scale studies on linguistic and sociolinguistic variation found in spoken French. For the present study, speech portions from ten regions have been used, corresponding to ten hours of speech. The speech portions were equally distributed between read speech and two sets of spontaneous speech: supervised conversation and free conversation (the latter having a slightly higher degree of casualness than the former one).

	French	
	# word tokens	duration
Careful	3600 k	360 h
Casual	1000 k	100 h
	English	
Careful	7200 k	720 h
Casual	25000 k	2300 h

Table 6: Corpus sizes for careful (BN) and casual (CTS) speech, for French (upper panel) and English (lower panel).

For comparison purposes, data from English corpora were added to the French data. The English careful speech data-set included hundreds of hours of broadcast news data for the DARPA *Rich Transcription 2004 Broadcast News* evaluation (Nguyen et al., 2005), distributed by LDC (Linguistic Data Consortium, http://www.ldc.upenn.edu/Catalog/). The casual speech data-set stems from the Switchboard corpus (Godfrey, Holliman & McDaniel, 1992) and the more recent Fisher data (see LDC) including thousands of hours of speech. In that corpus, 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.

3.3. Phone segment duration results

3.3.1. Casual versus careful speech styles

Figure 4 provides a line histogram of the segment proportions in the French corpus as a function of segment duration (expressed in ms). Results are broken down for prepared and conversational speech styles. The

results show that the majority of segments are part of segment durations up until about 150 ms. Concerning prepared speech, the duration bin with the largest number of segments (> 14%) corresponds to 60 ms. The casual speech distribution has by far the most segments (>18%) in the shortest duration bin of 30 ms, the total number of segments of 30 and 40 ms amounts up to more than 30% of the corpus. The distribution from the conversational speech corpus shows a somewhat flattened distribution, which suggest that casual speech is characterized by more fast and more slow speech. A Kolmogorov-Smirnov test confirmed the significant difference between the casual and careful distributions (D= 0.105, p<.0001) More detailed analyses on a speaker-per-speaker basis confirmed that the trend generally holds for each individual speaker and that the observed flattening is therefore not a result of mere averaging pools of fast and slow speakers. The lengthened segments may be partly due to prosodically stressed items (cf. Adda-Decker et al., 2008), and partly due to hesitation phenomena.

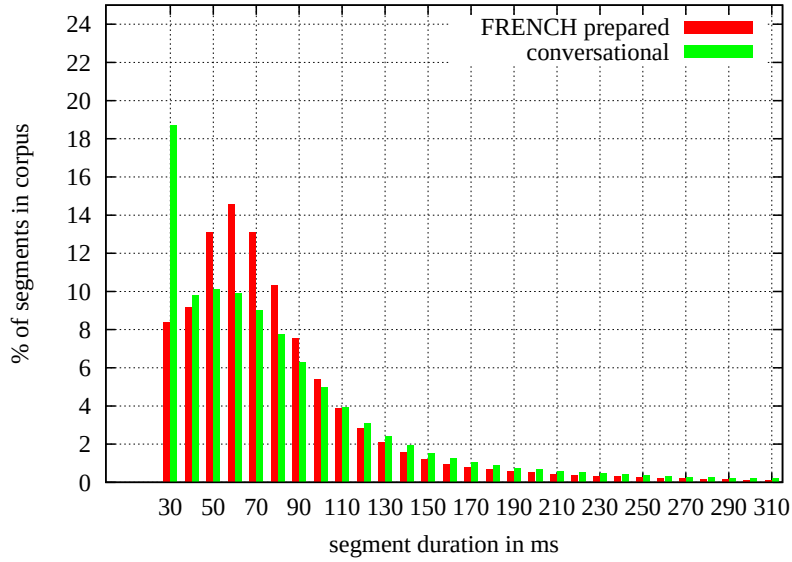

Figure 4: Line histogram of segment proportions from the corpus sets (cf. Table 6) as a function of segment durations for French. Journalistic prepared speech and conversational telephone speech styles are being compared.

3.3.2. French versus English

If casual speech generally produces a more flattened phone duration distribution, then it might be expected that this pattern should be observed irrespective of the corpus language. In other words, the observation should also hold for English. Figure 5 provides a line histogram of the segment proportions for English casual and careful speech as a function of segment duration.

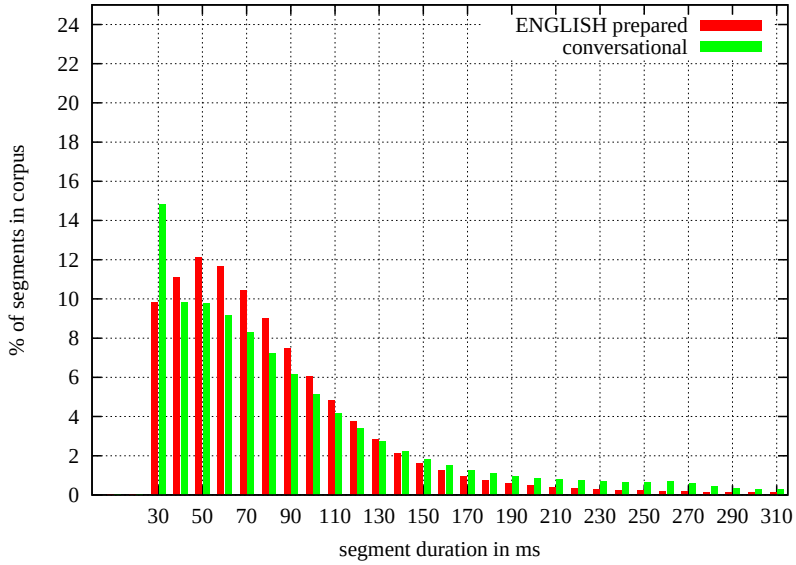

Figure 5: Line histogram of segment proportions from the corpus sets (cf. Table 6) as a function of segment durations for English. Journalistic prepared speech and conversational telephone speech are being compared.

Whereas the overall pattern slightly changes compared to the French results, the general tendency remains the same. There is a high proportion of segments in the minimum duration bin and a flattening of the conversational speech distribution, with more segments appearing in the longer segment durations (up until 310 ms). A Kolmogorov-Smirnov test confirmed the significant difference between the two speech styles ($D=0.084$, $p<.001$). Furthermore, it can be observed that the differences between careful and casual speech proportions up until 150 ms appear to be smaller in English than in French.

3.3.3. Comparing various casual speech styles

The comparisons between French and English were based on prepared journalistic speech and conversational speech styles taken from telephone conversations. To ascertain whether the tendencies found for spontaneous speech also hold for communicative settings other than telephone conversations, data from the French PFC corpus have been added. This corpus has the advantage of including a range of different speaking styles with read speech and face-to-face conversations, all produced by the same set of speakers.

Figure 6 shows the line histogram of segment proportions as a function of duration found for the PFC corpus (left-panel graph) and, for the reader's convenience, recalls the histogram for the previously shown French prepared and conversational corpus data (right-panel graph). The face-to-

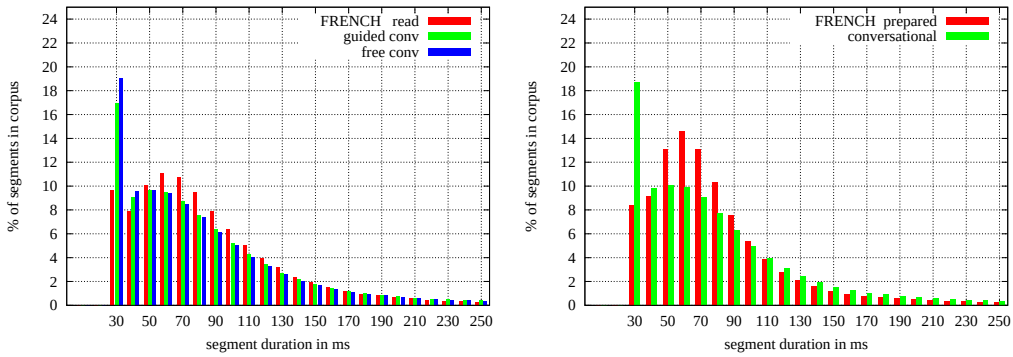

Figure 6: Segment duration proportions corresponding to read and conversational speech from the French PFC corpus (left) and journalistic prepared speech and conversational telephone speech (right).

face conversation distributions (guided and free) from the PFC corpus almost overlap with the conversational speech distribution, with quite similar casual speech effects. The read speech distribution was found to be significantly different from both the guided and free conversation distributions (Kolmogorov-Smirnov: $D = 0.089$, $p < .001$, and $D = 0.108$, $p < .001$ respectively). No significant difference was obtained between the PFC guided and free conversation distributions. Moreover, the read speech distribution exhibits a similar pattern as the prepared speech distribution on the right. However, it can be observed that there is a higher proportion of segments in the segment durations (up until 150 ms) for prepared journalistic speech as opposed to read PFC speech. This difference may very well be related

to time-pressure and speech rate performances for professional journalists, whereas it can be assumed that unprofessional readers speak more slowly.

3.3.4. Vowels versus consonants

The next question of interest is to know whether vowels or consonants exhibit similar duration patterns, and more specifically, which phonemes are the most prone to temporal reduction. As we have seen before, an obvious candidate for high temporal reduction rates is the schwa (e.g. in high frequency function words such as *le, de, ne...*, ('the', 'of', 'not'...)). Moreover, if it is true that function words are most prone to reduction, then the phonemes /l/ and /d/ should also be good candidates. Figure 7 shows overall proportions of consonants, vowels and sounds other than speech (this latter category includes real silences, but also breathing, hesitations and various other noises). The left-hand panel exhibits distributions taken from the prepared speech corpus, the right-hand one exhibits distributions taken from the spontaneous speech corpus. From the right-hand spontaneous speech distribution, it can be seen that vowels are only slightly more reduced and lengthened than consonants. This pattern holds for both careful and casual speaking styles.

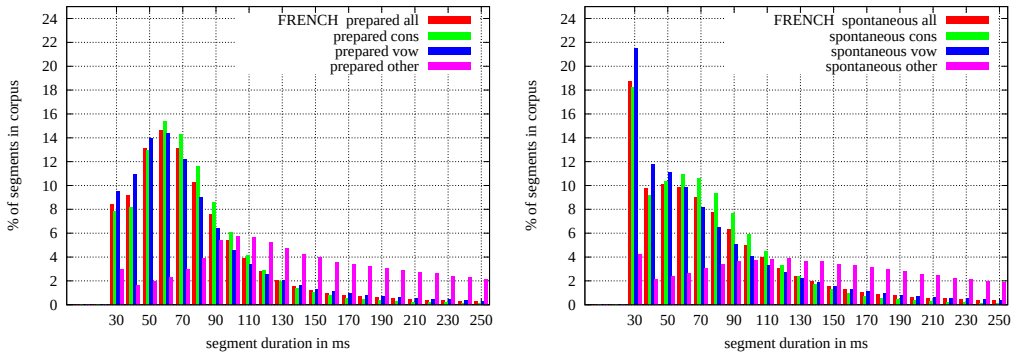

Figure 7: Segment duration distributions of vowels, consonants and other non-speech sounds (including silence) in prepared vs. spontaneous speech in French. For readability, each distribution sums up to 100%, even though the corpus is composed of 54% of consonants, 43% of vowels and 3% of other events (left) and of 50% of consonants, 43% of vowels and 7% of other events (right).

We will now be looking at duration phenomena for vowels and consonants separately. As was mentioned before, the segment proportions are represented as a function of the four duration classes, rather than continuous

segment durations. Turning to oral vowels, their duration distributions are shown in Figure 8. Given previous studies on the special status of the French schwa vowel mentioned earlier, it does not come as a surprise that the schwa vowel (the bar at the rightmost position) exhibits the highest figures in the short duration class, far above all the other vowels. Interestingly, the other French central vowel /ø/ (coded /eu/ in the figure, and near-homophone with a realized schwa) behaves in the opposite direction: it is actually one of the least reduced vowels. The duration patterns seen from the Figure, confirm

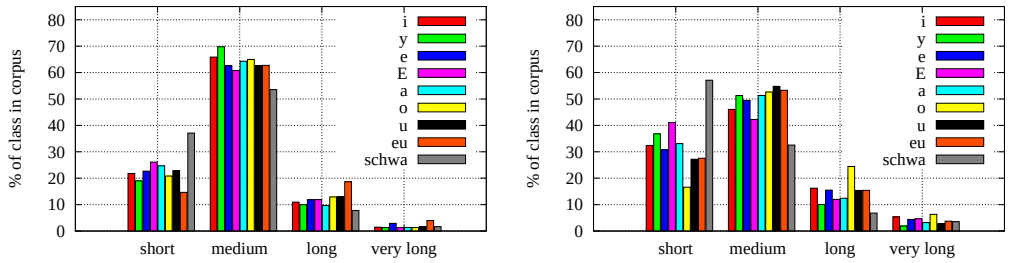

Figure 8: Proportions of short/medium/long/very long duration classes for French oral vowels. The left-hand panel shows the vowel distribution for careful speech distributions, the right-hand panel shows the vowel distribution for casual speech.

the special status of the French schwa vowel as an optionally (dis)appearing sound. Front and open vowels, such as /ɛ/ are good follow-up candidates in the temporal reduction hierarchy and tend to be temporally more reduced than back closed and rounded vowels.

We now move on to duration patterns for consonants. Figures 9 and 10 show duration distributions for the French voiceless and voiced fricatives respectively. Just like for the vowels, the results are shown for careful (left-hand panel) and casual speech (right-hand panel).

As can be seen from the Figures, the distribution patterns show that voiceless fricatives tend to be longer compared to the consonants' average duration (the bar at the rightmost position). In the *short* duration class the rates of the voiceless fricatives remain relatively low, particularly for /ʃ/.

Although it is well-established that voiced consonants tend to have shorter durations than their voiceless counterparts, the /v/ exhibits a somewhat atypical behavior with respect to /z/ and /ʒ/, in that a very high rate of /v/

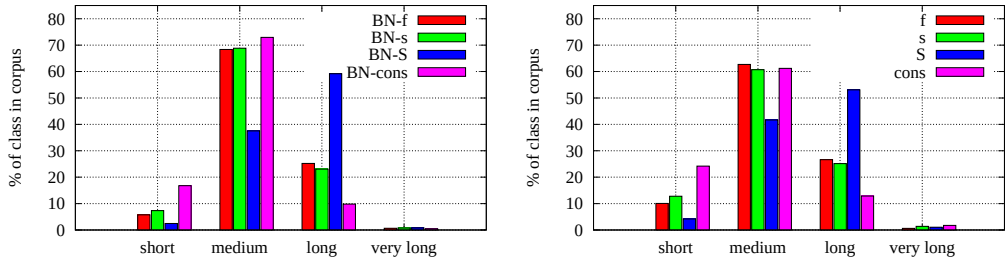

Figure 9: Proportions of short/medium/long/very long duration classes for French voiceless fricatives (/f/, /s/, /ʃ/) found in careful (left) and casual (right) speech styles.

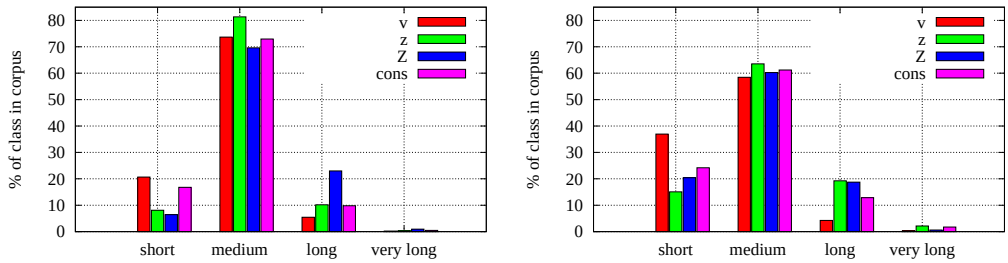

Figure 10: Proportions of short/medium/long/very long duration classes for French voiced fricatives (/v/, /z/, /ʒ/) found in careful (left) and casual (right) speech styles.

appear in the short duration class, with over 20% and 37% of segments for respectively careful and casual speech styles. In previous studies (see Adda-Decker et al., 2005), temporal reduction was already being observed for the function word *avec* /avɛk/ ('with'), that may be pronounced as something that is approximating [ɛg].

Finally, Figure 11 shows the results for the liquid and glide consonant classes in French. As opposed to the voiceless fricatives, liquids and glides tend to be rather short, given the increased proportions in the short duration class. The liquid /l/ which appears very often in frequent function words such as *le, la, les* ('the') appears to be most reduction-prone. For casual speech, the /ɥ/ glide also exhibits high reduction rates. These are

mainly stemming from the conversational specific words *suis* ('am') and *puis* ('then'). In particular the word sequence *je suis* ('I am') is frequently shortened to a pronunciation such as [ʃɥi] due to schwa-deletion and some complex consonantal assimilation processes.

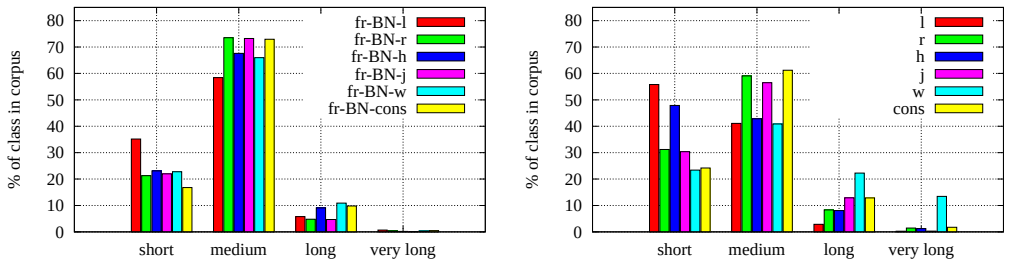

Figure 11: Proportions of short/medium/long/very long duration classes for French French liquids and glides found in careful (left) and casual (right) speech styles.

3.4. Optional schwa alignment

Instead of looking for evidence of temporal reductions in duration distributions from canonical alignments, the same objective can be pursued with a different tuning of the ASR model (see Adda-Decker, 2007, for more technical details). By adding sequential pronunciation variants to the pronunciation dictionary that includes optional segments (see Figure 5), the alignment system is free to keep or retrieve segments depending on their durations. It is clearly beyond the scope of the present article to provide an exhaustive overview of the implications of such adjustments to the model. However, we would like to give a flavour of the type of exploration one might conduct when focusing on French schwa realization and deletion.

Table 7 shows some schwa realization results concerning the French article *le* ('the') in the careful speech ESTER Corpus. Results highlight the contextual dependency of the schwa realization. Whereas the schwa is produced in almost 83% on average, there are left contexts where the rate approximates almost 100% (see, for instance, the rates for a left context of breathing (capital B in table 7) and in right unvoiced plosive /t/ context). On the other hand, a left vocalic context such as /u/ and the right voiced consonant /m/

context entails very low schwa production rates (14.3%). This specific context matches the idiom *tout le monde* [tuləmõdə] ('everyone'), which is most frequently produced as a bisyllabic word [tulmõd]. Globally, these results show that schwa production rates are about 90%, if the left context includes either breathing or a word ending with a closed syllable. A left vocalic context favours a schwa deletion. This is a good illustration of interesting temporal reduction phenomena such as can the case for the sequence *quai de Seine* ('dock of the Seine') that can be pronounced *quête saine* ('healthy quest').

cond.	%ə	%∅	*nb.*
#Cə	82.6	17.4	23950
B#Cə	92.7	7.3	2730
C#Cə	89.6	10.4	6480
V#Cə	76.7	23.3	12310
B#Cə#t	98.8	1.2	170
r#Cə#t	94.7	5.3	360
r#Cə#m	83.6	16.4	540
u#Cə#m	14.3	85.7	230

Table 7: Schwa realization (ə) and (complementary) deletion (∅) rates (expressed in %) of the French article *le* ('the') for the general pattern #Cə, and different left and right context conditions (#: word boundary, B: breathing, C: consonant, V: vowel). The total number of occurrences per condition (*nb*) are given in the last column.

4. Discussion

Whereas it may seem like a trivial matter to cite a number of examples of more or less severe reduction phenomena in day-to-day spoken language, the exact mechanisms that underly the processing of pronunciation variants that arise from these reductions are still relatively unknown. They are equally responsible for relatively high error rates in current ASR systems. More extensive descriptions are required to gain a better understanding of the type of pronunciation variation that both human listeners and automatic speech recognition systems are dealing with. By looking into transcription errors, it appears that a large number of ASR transcription errors in casual speech is due to a variety of "reduction phenomena" resulting in shorter productions. Given this observation, it seems straightforward to try and use

ASR systems as a tool to detect, quantify, and describe such phenomena using large spoken corpora. In the present paper, we focused on temporal patterns in French. Segmentation and labelling of speech regions that are prone to reduction phenomena were carried by using forced speech alignment in combination with canonical (full form) pronunciations. The forced speech alignment technique enabled us to localize and quantify phonemes that are particularly prone to temporal reduction. It must be borne in mind that upon employing forced alignment, an isolated occurrence of such a segment is not necessarily indicative of reduction. After all, an increase in duration of 50 ms will be relatively greater for a vowel that has a typical duration of 50 ms than for one that has a typical duration of 150 ms (see Campbell, 1992). Nevertheless, the larger the number of contiguous minimum duration segments, the stronger the hypothesis of an actual temporal reduction.

In a series of explorations in French, phone segment duration distributions were computed for a number of different speaking styles and speaker populations, ranging from careful speech (i.e., read speech, prepared journalistic speech)to casual speech (i.e., telephone conversations and face-to-face interviews). With regard to more refined comparisons between French vowels and consonants, the segment duration distributions were examined using four duration classes.

Our results showed that casual speech has a flatter duration distribution than careful speech, with an increase of more than 10% in the proportion of short segments, but also a larger number of longer segments. This general trend has been confirmed for English, even though the increase of short segments was limited to about 6%. This may be partly due to the fact that the corpora used for English included less casual telephone conversations. Future research should aim to investigate these measured differences with respect to known prosodic syllable/stress-timing differences between French and English (see, e.g., Cutler, Mehler, Norris & Segui, 1986). Moreover, observations also remain stable by shifting from prepared and telephone speech to the PFC corpus, where the same speaker population produced both read and spontaneous face-to-face speech. Next, duration patterns in vowels and consonants were examined. Vowels are slightly more shortened than consonants and, not surprisingly, the schwa vowel, that is notoriously known for being deleted in French, exhibits the highest figures of short segments, far above all other vowels. As for reduction patterns in consonants, it was shown that for voiced fricatives, the /v/ exhibits an un expected duration pattern, in that it is relatively short compared to the other voiced fricatives. The

liquid /l/ which appears very often in frequent function words is most prone to shortening. In casual speech, the /ɥ/ glide also exhibits high reduction rates due to conversation-specific words *suis* ('am'), *puis* ('then'). Finally, a schwa-specific exploration using forced alignment and an adapted pronunciation dictionary showed regularities of schwa production before pauses and breathing, and schwa deletions in idiomatic expressions such as *tout le monde* ('everyone').

By using forced alignment to quantify temporal reduction phenomena in French, we hope to have demonstrated how ASR systems may serve as a tool to systematically investigate variations that occur in the speech signal across different speaking styles. Hopefully, 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. 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 fine phonetic detail and features that the human listener seems to rely on in the face of ambiguity and noise. The perspectives available through an ASR approach are manifold. For researches working in the domain of ASR, the ultimate goal is to uncover the generic rules to generate pronunciation variants, even for rarely observed or unobserved words, for which variants cannot be estimated statistically. The framework developed should also help to describe and quantify more or less well known linguistic phenomena on phonemic and lexical levels, which is of relevance to linguists and cognitive scientists alike.

5. Acknowledgement

Parts of the research reported in this article have been funded by grants from the CNRS, ANR, Digiteo,and Quaero, awarded to the first author, and a grant from the Luxembourgish Fondation Nationale de la Recherche, awarded to the second author. We would like to thank Lori Lamel, Jean-Luc Gauvain, and Gilles Adda for their help and fruitful discussions during the preparation of this paper. We are greatly indebted to three anonymous reviewers for their constructive criticisms and suggestions on an earlier version of the paper.

References

Adda-Decker, Mhttp://ling.upenn.edu/phonetics/workshop., Gendrot C. & Nguyen N., (2008). Contributions du traitement automatique de la parole à l'étude des voyelles orales du français. *Traitement Automatique des Langues*, 49(3), 13-46.

Adda-Decker, M. (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, 211-216.

Adda-Decker, M., Boula de Mareüil, P., Adda, G., Lamel, L., (2005). Investigating syllabic structures and their variation in spontaneous French. *Speech Communication*, 46, 119-139.

Adda-Decker, M. & Lamel, L. (2005). Do speech recognizers prefer female speakers? In: *Proceedings of InterSpeech*, Lisbon.

Adda-Decker, M. & Lamel, L., (1999). Pronunciation variants across system configuration, language and speaking style. *Speech Communication*, 29,83-98.

Campbell, N. (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, R., Oshika, B., T., Noel, M., Lander, T., Fanty, M. (1994). Labeler Agreement in Phonetic Labeling of Continuous Speech. In: *Proceedings. of ICSLP*, 2, 2131-2134.

Cutler, A., Mehler, J., Norris, D., & Segui, J., 1986. The Syllable's Differing Role in the Segmentation of French and English. *Journal of Memory and Language*, 25, 385-400.

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

Dilley, L., Pitt, M. (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.

549 Duez, D., (2003). Modelling Aspects of Reduction and Assimilation in Spon-
550 taneous French Speech. In: *Proceedings of the IEEE-ISCA Workshop on*
551 *Sponrtaeous Speech Processing and Recognition*, 2003. Tokyo.

552 Durand, J., Laks B. & Lyche C. (2002). La phonologie du français contempo-
553 rain: usages, variétés et structure. In: C. Pusch & W. Raible (eds.) Roman-
554 istische Korpuslinguistik - Korpora und gesprochene Sprache / Romance
555 Corpus Linguistics - Corpora and Spoken Language. Tübingen: Gunter
556 Narr Verlag, pp. 93-106.

557 Durand, Jacques, Laks B. & Lyche C. (2005). Un corpus numérisé pour la
558 phonologie du français. In G. Williams (ed.) La linguistique de corpus.
559 Rennes: Presses Universitaires de Rennes. pp. 205-217.

560 Elman, J. & McClelland, J.M. (1988).Cognitive penetration of the mecha-
561 nisms of perception: Compensation for coarticulation of lexically restored
562 phonemes. *Journal of Memory and Language*, 27, 143-165.

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

566 Fougeron, C., Goldman, J.-P. & Frauenfelder, U.H., (2001). Liaison and
567 schwa deletion in French: an effect of lexical frequency and competition.
568 In: *Proceedings of Eurospeech* , Aalborg, 639-642.

569 Galliano et al. (2005). The Ester Phase II Evaluation Campaign for the Rich
570 Transcription of French Broadcast News. In: *Proceedings of InterSpeech.*

571 Ganong, W.F. (1980). Phonetic categorization in auditory word perception.
572 *Journal of Experimental Psychology: Human Perception & Performance*,
573 6, 110-125.

574 Gauvain J.-L., Lamel L.F., Adda G. & Adda-Decker M. (1994). Speaker-
575 Independent Continuous Speech Dictation. *Speech Communication*,15(*).

576 Gauvain, J.-L., Adda, G. , Adda-Decker, M., Allauzen, A., Gendner, V.,
577 Lamel, L., Schwenk, H. (2005). Where Are We in Transcribing French
578 Broadcast News? In *Proceedings of InterSpeech*, Lisbon.

Gendrot C. & Adda-Decker M., (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: *Proc. of InterSpeech*, Lisbon.

Godfrey J., Holliman E. & McDaniel J. (1992). Switchboard: telephone speech corpus for research and Development. In: *Proceedings of IEEE-Icassp.*

Greenberg, S. & Chang, S., 2000. Linguistic dissection of Switchboard-Corpus Automatic Speech Recognition Systems. In: *Proc. ISCA-ITRW Workshop on ASR*, Paris, pp. 195-202.

Greenberg, S., Carvey, H., Hitchcock, L., Chang, S. (2003). Temporal properties of spontaneous speech – a syllable-centric perspective. *Journal of Phonetics*, 31, 465-485.

Hosom, J.-P. (2009). Speaker-independent phoneme alignment using transition-dependent states. *Speech Communication*, 51(4), 352-368.

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

Lefèvre F., Gauvain J.-L., Lamel L.F., (2005). Genericity and portability for task-dependent speech recognition. *Computer Speech and Language*, 19:345-363, 2005.

Nakamura M., Furui S. & Iwano K. (2006). Acoustic and Linguistic Characterization of Spontaneous Speech Masanobu, *ISCA workshop on Speech Recognition and Intrinsic Variation*, Toulouse, France.

Nguyen L., Abdou S., Afify M., Makhoul J., Matsoukas S., Schwartz R., Xiang B. Lamel L., Gauvain J.L., Adda G., Schwenk H., Lefèvre F., (2005). The 2004 BBN/LIMSI 10XRT English Broadcast News Transcription System. In IEEE-Icassp 2005.

Prasad R., Matsoukas S., Kao C.L., Ma J., Xu D.X., Gauvain J.-L., Lamel L., Schwenk H., Adda G. & Lefèvre F., (2005). The 2004 BBN/LIMSI 20xRT English Conversational Telephone Speech Recognition System. In: *Proceedings of. InterSpeech*, Lisbon.

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

Schuppler B., Ernestus M., Scharenborg O. & Boves L. (2008). Corpus of Dutch Spontaneous Dialogues for Automatic Phonetic Analysis. In: *Proceedings of Interspeech*, Brisbane, pp. 1638-1641.

Snoeren, N., Hallé, P. & Segui, J. (2006). A voice for the voiceless : Production and perception of assimilated stops in French. *Journal of Phonetics*, 34, 241-268.

Shriberg, E. (1994). Preliminaries to a Theory of Speech Disfluencies. PhD thesis, University of California, Berkeley.

Strik H., Binnenpoorte D. & Cucchiarini C. (2005). Multiword Expressions in Spontaneous Speech: Do we really speak like that? In: *Proc. of InterSpeech*, Lisbon, pp.1161-1164.

Strik H., Elffers A., Bavcar D. & Cucchiarini C., (2006). Half a Word is Enough for Listeners, but Problematic for ASR. In: *Proceedings of ISCA workshop on Speech Recognition and Intrinsic Variation*, Toulouse, France.

Strik, H., Cucchiarini, C., (1999). Modelling pronunciation variation for ASR: A survey of the litterature. *Speech Communication*, 29,, 225-246.

Van Son, R.J.J.H., Pols & L.C.W., (2003). An Acoustic Model of Communicative Efficiency in Consonants and Vowels taking into Account Context Distinctiveness. In: *Proceedings ofthe 15th ICPhS*, Barcelona 2003, pp. 2141-2143.

Tseng S.-C., Features of Contracted Syllables of Spontaneous Mandarin, In: *Proceedings of InterSpeech*, Lisbon.

Verney Pleasants, J., (1956). Études sur l'*e* muet, timbre, durée, intensité, hauteur musicale. Klincksieck, Paris.