Introduction, or: why rethink reduction?

Margaret Zellers, Barbara Schuppler, Meghan Clayards

Abstract

In phonetic reduction, segments may be shorter, less clearly articulated, or absent, compared to “canonical” or dictionary forms. While traditionally considered “slurred” or deficient, recent work has shown that reduction phenomena are highly complex. This chapter takes a historical and multidisciplinary approach, describing how views of reduced speech have evolved over time in the domains of phonetics, speech perception and automatic speech recognition. We bring these perspectives together to raise several questions: Are reduced forms really inferior to canonical forms? Is the scope of reduction best described at the level of the feature, syllable, or larger unit? Does reduction generally occur in places where the content is more predictable? The volume’s chapters are then contextualized with regard to these questions.

Author affiliation

Margaret Zellers, Stuttgart University, Stuttgart
Barbara Schuppler, Graz University of Technology, Graz, Austria
Meghan Clayards, McGill University, Montreal, Canada

1.1 Introduction

This volume is focused around the phenomenon of phonetic reduction in speech. In phonetic reduction, segments may be shorter, less clearly articulated, or absent, compared to “canonical” or dictionary forms. A classical view on reduction is given by Jakobson and Halle (1956):

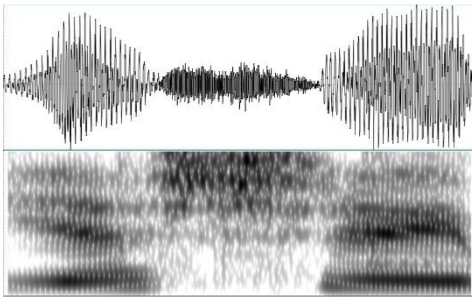
> Since in various circumstances the distinctive load of the phonemes is actually reduced for the listener, the speaker, in his turn, is relieved of executing all the sound distinctions in his message: the number of effaced features, omitted phonemes and simplified sequences may be considerable in a blurred and rapid style of speaking. The sound shape of speech may be no less elliptic than its syntactic composition…. But, once the necessity arises, speech that is elliptic on the semantic or feature level, is readily translated by the utterer into an explicit form which, if needed, is apprehended by the listener in all its explicitness.
>
> The slurred fashion of pronunciation is but an abbreviated derivative from the explicit clear-speech form which carries the highest amount of information…. When analyzing the patterns of phonemes and distinctive features composing them, one must resort to the fullest, optimal code at the command of the given speakers.
>
> Jakobson & Halle (1956: 6)

Many early studies of reduction take a similar attitude to Jakobson and Halle’s characterization of reduced forms as “slurred”, “slovenly”, or otherwise deficient. However, more recent work has shown clearly that reduction is much more complex. For example, while reduction is primarily thought of as a casual speech phenomenon, it also occurs in read speech, and indeed can make read speech easier to listen to and process; speech synthesis for the blind adopts reduction phenomena to make texts easier to listen to, especially over longer stretches of time (Jande 2003). Furthermore, an increasing amount of evidence suggests that “canonical” and “reduced” forms are not simply categorical oppositions, but may rather fall along a spectrum of pronunciation variants that are more or less clearly articulated (Nolan 1992). An acoustically “absent” segment may still leave prosodic and/or articulatory traces (Kohler & Niebuhr, 2011; Niebuhr & Kohler 2011; Torreira & Ernestus 2011); and conversely, segments may be hyperarticulated to a point that, while their pronunciation may be “super-canonical”, it does not reflect the typical production of that segment (Clayards & Knowles 2015). Schuppler et al. (2012) for instance found, that in conversational Dutch only 11.7% of the tokens show canonical realizations of word-final /t/ (i.e., a voiceless closure followed by one strong burst, produced at an alveolar place of articulation).

While a great deal of current research addresses the kinds of difficulties posed by dealing with reduced forms and spontaneous speech, cf. the recent special issue of the Journal of Phonetics (39[3], 2011) addressing this very topic, this volume will ask the question of whether the ways we think about “reduction” are helpful, and how as researchers we could potentially shift our paradigms and methodologies, leading to greater understanding of this kind of variation in phonetic forms. Thus the current volume brings together work from a variety of research and language backgrounds aimed at widening our understanding of what reduction is and how we as language users make use of it.

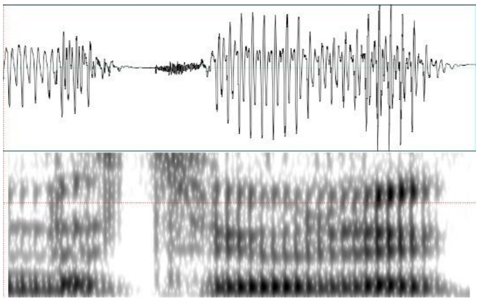
1.2 Examples of reduction

Reduction phenomena can be highly variable, particularly across languages. A few examples are provided here to illustrate some of the possibilities.

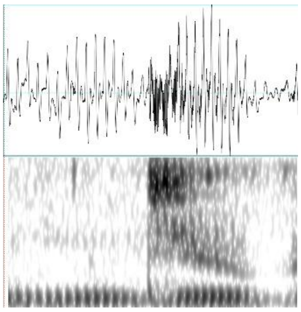


(1.1) yesterday, realised as [’jeʃa͡i]; Buckeye corpus

This example is taken from the utterance “we were supposed to see yesterday, but he felt really bad.” For native speakers, it is easily understandable, even though it is realized with fewer segments, with only two instead of three syllables, and with different segments than the canonical form. The remaining segments are the initial consonant, the stressed vowels, and the fricative. As frequently shown in the literature, unstressed vowels and plosive closures are deleted completely. With a segment based approach this word would be hard to annotate and segment, as boundaries between the segments are overlapping. (However cf. Xu & Liu 2013, who propose that segments, like prosody, are produced as a series of approximations to dynamic targets, and that segmentation is thus preferable on the basis of gesture onsets in the context of syllable structure. Thus, in the current example, the onset of the fricative occurs before the first vowel gesture is complete, as can be seen by the continuation of voicing as well as by the clear formant structure visible within the frication.)



(1.2) *natuurlijk*, canonical, Ernestus Corpus of Spontaneous Dutch



(1.3) *natuurlijk* [nt'yk]

These examples of *natuurlijk* ("naturally") are taken from a Dutch corpus of spontaneous dialogues (Ernestus, 2000). In Dutch, *natuurlijk* can be used with different functions; it can be an adjective, as in "natural languages", or it can have the discourse-oriented meaning "of course". If presented in isolation, listeners are not able to understand the reduced form (Van de Ven et al. 2012), even though the canonical form used as "of course" never occurs in the corpus (Schuppler et al. 2011). When used as an adjective, however, the canonical form is observed frequently. Thus, the same sequence of phones is reduced differently for two different functions.

1.3 A historical and multidisciplinary perspective on reduction

1.3.1 Phonetics

Like Jakobson & Halle (1956) above, most early views on reduction coincided in considering reduced forms as lacking in some way. Jakobson and Halle thus argue that reduced forms are not worth studying compared to canonical forms, which contain the "fullest, optimal" information. However, not all contemporary researchers held the same view, and a number of phonetic studies touched on topics relating to reduction, particularly in the context of research on lexical stress. Fry (1955, 1958), for example, reports that lexical stress in American English is marked by duration and intensity ratios between syllables in disyllabic words; that is, reduced duration and intensity are associated with the unstressed syllable. Similarly, Lieberman (1960) reports that stressed syllables in American English have higher fundamental frequency (F0), higher amplitude, and longer duration than unstressed syllables. Research on Swedish by Fant (1962) demonstrates that the formants in unstressed vowels move closer to those typical of schwa (i.e. 500, 1500, and 2500Hz for the first three formants). Fónagy (1966) reports for Hungarian that increased articulatory effort (as in stressed syllables), measured by EEG, is associated with higher formant amplitude and broader bandwidth.

Lindblom (1963) addresses reduction phenomena directly by asserting the existence of "physiologically invariant" vowel targets, which are more or less closely approached based on the articulatory context. He treats reduction as a phenomenon resulting from limits on articulatory speed, and claims that timing is more important in influencing reduction than lexical stress; that is, targets are undershot on the basis of decreased duration, not on the basis of lexical stress alone.

As phonetic research on reduction began to move beyond analysis of lexical stress alone, a large body of work also arose in the phonological community, providing contextual analyses that provided rules for certain types of reduction (e.g. schwa allophones of vowels, or elision of consonants). Kohler (1974), for example, reports rules for the common elision of schwa from the German suffix *-en*. Gimson (1977), Elgin (1979), and Lass (1984) provide detailed grammars including elision rules. For Lass, deletion is phonologically defined as a segment merging with a null segment, though he indirectly addresses the phenomenon of reduction by stating that deletion is often the last stage of a lenition process (1984: 187). Since he deals with a phonological process rather than a phonetic one, however, the kinds of deletions he reports are not synchronically variable (in contrast with, for example, Dutch *natuurlijk*, as discussed in section 1.2 above).

Reduction was also addressed from a sociolinguistic perspective in this time period. Labov (1972) reports on, among other phenomena, elision of word-final alveolar plosives in New York Black American English, and social distribution of syllable-final /r/ in department store workers. He does not study reduction as a set of individual phonetic cases, but rather as a way in which people modify their speech to demonstrate group membership. Ohala (1981) studies reduction in the context of explaining sound change over time. His work is related to Lindblom's on timing, arguing that articulatory timing is constrained by physical characteristics of articulators.

A turning point for the analysis of reduction came with Lisker's (1984) argument that invariance is not actually something to be expected of the speech signal. He points out that, while invariance was considered as important as an explanation for why listeners are able to constantly identify sequences of a limited set of sounds from the continuous speech signal, in fact, "phonemes... are not perceptual constants" (1984: 1201). This argument grew in part out of contemporaneous phonological analyses, in which articulatorily and acoustically different sounds appeared as allophones for the same phoneme. In fact, the argument that contrastiveness, rather than invariance, is essential for speech processing, is fundamental for modern work on reduction.

Following this, more phonetic studies of reduction phenomena began to appear. Dalby (1986) characterizes a number of reduction phenomena occurring in fast American English speech. Kohler (1990), moving beyond his earlier phonological analysis, argues that phonology-based characterizations of reduction are too restrictive, and that reduction rules should instead be based on phonetic features, generating a larger variety of alternatives. Kohler specifically argues that phonology, being an abstraction, cannot account for the physical and/or phonetic processes that lead to reduction, and that therefore it is an insufficient basis for analyzing reduction. Instead, he refers back to "motor economy", as proposed by Lindblom (1963; see discussion above) and others, as a fundamental consideration.

Lindblom (1990) himself also argues against the "problem" of invariance. His theory of Hyper- and Hypo-articulation (H&H Theory) states that articulation is influenced by both constraints on the production system (i.e. a preference for minimal expenditure of effort) as well as constraints on the output (i.e. the need to make oneself understood). Since segmental production must take both of these constraints into account, reduction is not a problem but simply a response to a particular set of system settings. Again, the goal is to create sufficient contrast (and thus understandability) rather than to connect to something invariant.

Articulatory Phonology (Browman and Goldstein 1990) provides a somewhat different analysis of reduction, while staying within a similar tradition. In this theory, variation in forms occurs on the basis of varying articulation rate and thus varying degrees of gestural overlap. At extreme degrees of overlap gestures can be "hidden" so that they are not acoustically/perceptually available. For instance, Browman and Goldstein's measurements showed that word-final /t/ in word combinations like perfect memory could be absent in the acoustic signal, even though the tongue clearly moved to the alveolar ridge. Since all form variations can be accounted for in this model by increased gestural overlap and decreased gestural magnitude (i.e. gestures are never added, changed, or deleted), this model necessarily asserts that all reduction is gradient. Johnson, Flemming & Wright (1993) follow up on the idea of a single gradient system for speech sound forms: "...fortitions are more accurately seen as descriptions of the pronunciation of phonetic targets in the absence of lenitions, and hence it is apparently the case that for every lenition there is an equal and opposite fortition." (524).

More recently, new insights about reduced words and the conditions under which they occur have been achieved by two parallel growing interests: 1) in conversational speech and the collection of large conversational speech corpora, e.g., for English, Pitt, Johnson, Hume, Kiesling, and Raymond (2005); Godfrey, Holliman, and McDaniel (1992); for Dutch, Ernestus (2000); for German, Peters (2005); and for French Torreira et al. (2010); and 2) in creating automatic tools for phonetic annotation, e.g., forced alignment, Adda-Decker and Lamel (2000), Adda-Decker and Snoeren (2011). Studies based on such large, automatically annotated corpora obviously do not focus on the detailed pronunciation of words. However, this quantitative approach can identify trends for certain speaking styles, and allows for the calculation of sophisticated statistical models in order to learn about the conditions under which certain reductions are likely to occur. For instance, Schuppler et al. (2012) found that the realisation of word-final /t/ in Dutch spontaneous dialogues is conditions by word frequency, bigram frequency, segmental context, morphological structure and phrase position. For English, Yuan and Liberman (2009) investigated conditions for /l/ variation in English. Chapters 4 (Adda-Decker et al.) and 7 (Cuttugno et al.) of the current volume present studies carried out with such quantitative, corpus based approaches.

1.3.2 Speech perception

The "lack of invariance" problem (Lisker 1984) was a central issue for speech perception studies as much as for acoustic phonetic studies. That is, the acoustic cues that signal a phone are very context-dependent, even in carefully produced citation forms such as one encounters in lab speech. Since then the sources and varieties of variation that have been considered have been greatly multiplied and include casual speech phenomena like assimilations, flapping or deletions as well as changes in speaking style and speaker. The traditional approach has been to assume that these variations pose a problem for the (canonically-based) speech recognition system and the search has been for processes or representations that can accommodate this variation such as talker normalization (see Johnson 2005 for summary); general auditory processes that normalize coarticulation and assimilation (e.g. Diehl, Lotto, & Holt 2004; Fowler 1986; Gow 2003; Mitterer, Csépe, & Blomert 2006); and statistical processes that infer canonical forms from variable data (e.g. Gaskell & Marslen-Wilson 1996; McMurray & Jongman 2011; Snoeren 2011; Sonderegger & Yu 2010). Other approaches explicitly move away from the the idea of a phone-based canonical form as the mental representation (e.g. Hawkins & Smith 2001; Port 2007; Pisoni 1997; Goldinger 1998; Johnson 1997, 2006; Clayards 2010).

Thus a central set of questions in the perception literature has been what level of processing deals with pronunciation variation (pre-lexical, lexical, context) and what kind of units or representations best accommodate or even incorporate variation. Another important question has been the role of the canonical form versus other forms in representation. These questions are mirrored in the phonetic and ASR literature as discussed in sections 1.4.1 and 1.4.2 below. Despite the central role of representation and processing of phonetic variation of all kinds to the speech perception literature, perception studies have overwhelmingly focused on non-spontaneous and non-conversational speech. Some notable exceptions focusing on reduction include Pickett and Pollack (1963), Ernestus and Baayen (2007), Janse and Ernestus (2011), Kohler and Niebuhr (2011), Van der Ven, Schreuder, and Ernestus (2012), and Brouwer,Mitterer and Huettig (2013). Examining perception of spontaneous, and especially conversationally produced, speech is clearly an important direction for future research in order to enrich our understanding of speech perception more generally. Chapter 6 (Cole and Shattuck-Hufnagle) takes a novel approach to examining perception of spontaneous speech through the use of imitation.

1.3.3 Automatic Speech Recognition

Since the first ASR experiments in the 1970s mainly focused on the recognition of isolated words, there was no need to investigate methods to deal with reduced pronunciations. At that time, researchers thought that speech recognition was soon to be a solved problem. However, as new applications continued to be proposed for ASR systems, ASR research needed to move beyond recognizing only isolated words. When researchers began using connected words and read speech, the need arose to find ways of incorporating coarticulation and pronunciation variation. In the ensuing decades, interest progressed more and more towards spontaneous and conversational speech, and recently to conversations between more than two people. Since the frequency of reduction phenomena is highest in conversations and lowest in read words, the importance of dealing with reduction has steadily increased alongside the increasing focus on conversational speech. For instance, a recent study on Austrian German database GRASS (Schuppler et al. 2014a) shows that while in read texts only 33.1% of the words are produced with a pronunciation different from the canonical form, in conversational speech this number rises to 63.2% (Schuppler et al. 2014b). The performance of ASR systems drops in proportion to the degree of spontaneity of the speech in question , as shown by Adda-Decker et al. (2013): an ASR system which was trained on read speech reaches nearly 96% word accuracy on read speech, whereas only 27% on spontaneous conversations involving the same speakers (without further adaptations and/or additional pronunciation modeling). It is thus clear that in order to make ASR systems work, pronunciation modelling has to be taken into account.

In general, reduction has been dealt with in the speech recognition community using the same methods as other sources of pronunciation variation (e.g. regional and social variation, variation due to anatomical differences of speakers, emotional status, etc.; for an overview of different methods used see Strik and Cucchiarini 1999). In a typical ASR system, the basic unit chosen is the phone; alternative approaches which are assumed to deal better with reduced words are the use of syllable (see Chapter 7, Cutugno et al.) or phonetic features. Apart from the choice of the basic unit, an additional question arising is which component of the ASR system needs to be adapted in order to make it robust to reduced speech; this is discussed further in section 1.4.3.

1.4 Some open questions about reduction

1.4.1 What do “canonical” and “reduced” actually mean?

When we talk about “reduced” forms, we are usually implying that they are reduced compared to something else. This something else is generally a “canonical” form, which might also be called a full or citation form, or a dictionary form. It can be defined as the form of a lexical item that is used in clear speech, with all underlying phonemes having a phonetic realization; indeed, this is the traditional definition that researchers like Jakobson and Halle (1956) assume. However, recent research shows that such a simple, binary contrast is not sufficient to characterize the many different facets of reduction that can be observed in speech, and that some parts of a phonemic (or other underlying) structure may be more necessary for conveying information than others.

In the first place, the language surrounding canonical and reduced forms may be problematic in that it carries with it an implicit assumption that reduced items are deficient compared to canonical forms - indeed, the implication is that phonetic information is somehow missing from the reduced form, taken away from the “full” canonical form. Even in the Phonetics of Talk-in-Interaction, where the potential meaningfulness of all phonetic forms is a basic tenet, the terms phonetic “upgrade” and “downgrade” may imply a hierarchical relationship between these kinds of forms, with downgraded forms being somehow “less” than upgraded forms (Traci Walker, personal communication). Lindblom’s influential H&H theory (cf. section 1.3.1 above) relies on the assumption that reduced forms are lacking by suggesting that talkers reduce effort and therefore reduce phonetic information when the needs of the listener are minimal. Chapter 2 of the current volume questions the premise that reduction and listener needs are as tightly linked as proposed in H&H theory. However, it still shares the assumption that reduced forms are less useful for extracting lexical information from the signal.

One problem inherent in this loaded value comparison between canonical and reduced forms is the assumption about what linguistic forms are primary. Despite criticisms of “sloppy” conversational speech, researchers like Abercrombie (1969) have long been pointing out that “spoken prose”, the focus of most traditional linguistic analyses, is derived from conversation, rather than vice-versa. If we consider, following Abercrombie, that conversation is primary, it becomes increasingly difficult to make the leap to saying that reduced forms are less privileged compared to citation/canonical forms. Why then have “canonical forms” become so prominent? And how do we reconcile this with psycholinguistic findings that lend at least some support to the idea of canonical forms being privileged in perception, even outside of contexts where they would be produced naturally.

In basic word recognition tasks, canonical forms do seem to have an advantage. They are recognized faster (Ernestus & Baayen 2007; Janse 2004; Janse, Nooteboom, & Quené 2007; Ranbom & Connine 2007; Tucker 2011) and more easily (Pitt, Dilley & Tat 2011); they prime more effectively (Andruski et al. 1994; Ranbom, Connine & Rudman 2009), and exhibit stronger lexical biases on perception (e.g. Pitt, 2009), than corresponding reduced forms. In cases where a pronunciation variant is extremely frequent, such as the flap variant of word internal intervocalic d/t in American English, it may behave similarly to the canonical form (Connine 2004; Pitt, Dilley & Tat 2011).

One limitation of findings related to the so-called “canonical advantage” is that they are often tested in the context of single spoken words, where canonical forms are very much more expected than reduced forms (e.g. Pitt 2009; Tucker 2011). In fact research has shown that the perception of reduced words is very much dependent on aspects of the context, including the speaking rate (Dilley & Pitt 2010). Other studies such as those by Ranbom, Connine and Rudman (2009) and Viehbahn, Ernestus & McQueen (2015) include the canonical and reduced forms in full sentences which should favour reduced forms to a greater extent. Ranbom, Connine and Rudman (2009) further vary whether the prosodic environment favors the reduced or canonical variant (presence vs.

absence of a prosodic break) and found a canonical advantage even in the environment favoring flapping. Viehbahn, Ernestus & McQueen (2015) varied the predictability of the words, with more predictable environments favoring the reduced variant, and again found a consistent advantage for canonical forms. However, even for these studies, because they involved deliberately and not spontaneously produced reduction, it is unclear whether the speaking style truly favored the canonical or reduced variant. Sumner (2013) found that when reduced forms are produced in a casual speaking style they are recognized just as well as canonical forms produced in a careful speaking style. Tucker (2011) also found that the canonical advantage depended on word frequency, such that very high frequency words were processed more quickly with reduced variants. Thus, some of the processing advantage observed for canonical productions may be due to how well they match contextual expectations. Furthermore, the phonetic details of deliberately and spontaneously produced variation may not be the same. Gow (2002, 2003) has shown that spontaneous nasal place assimilations are phonetically distinct from deliberately produced ones, and that these phonetic differences may facilitate speech perception.

On the other hand, research on the perception of spontaneously produced reduced speech continues to find that recognition of reduced words is difficult. When the perception of severely reduced words from spontaneous productions is tested, recognizing reduced forms in isolation is very difficult (Ernestus, Baayen, and Schreuder 2002; Janse & Ernestus 2011), though recognition goes up when the context is provided. Furthermore, Brouwer et al. (2013) present both reduced and clearer pronunciations of words in their original contexts, and find that canonical forms still have an advantage.

Several of the psycholinguistic studies discussed above raise the question of whether reduction can be treated as a simple application of rules (as in the Viehban et al. study, in which “reduced” variants were carefully produced for an experiment) or whether the influence of a larger context is strictly necessary to bring about a correct form. In ASR systems, the most typical component where reductions are incorporated is the Pronunciation Dictionary. When the basic unit chosen is the phone, pronunciation variants are typically incorporated in form of deletions, substitutions and insertions of segments to the canonical form. Starting in the early seventies, pronunciation variants were created automatically by formulating phonological rules and applying them to the canonical forms in terms of “re-write rules.” Barnett (1974), for instance, developed a phonological rule compiler for American English, which took the phonetic features of the sounds into account (manner and place of articulation, voiced vs. voiceless), as well as the stress pattern of the word and the position of the segment within the word. These phonological rules mostly deal with coarticulation (degemination, flap generation, homorganic stop insertion, etc.) and include few reductions for conversational speech, since they were not necessary for the data in question.

Since then new approaches continue to be developed, all of them with the aim of creating a lexicon, but not all of them based on the development of overt, human-defined rules. Besides the rule-based approach (mentioned above; cf. also Van Bael et al. 2007), the increasing number of transcribed speech databases has allowed for the development of data-driven approaches (e.g., Hämäläinen et al. 2008; Kessens et al. 2003). A set of variants derived with a data-driven approach is specific to the database from which the variants were extracted and tends to contain fewer pronunciation variants for most words than a lexicon created with the knowledge-based approach. Not all plausible variants will be present for all word types, especially for words with a low frequency of occurrence. Since low frequency words occur seldom by definition, correspondingly few variants of those words will appear in the speech material. For highly frequent words, however, the data-driven approach yields a good set of pronunciation variants. In order to compensate for the disadvantages, knowledge- and data- driven approaches have been combined (Wester et al. 2002, Schuppler et al. 2011, Schuppler et al. 2014b; for a broader overview see Barry & van Dommelen 2005 and Hain 2005).

The question of the relationship between canonical and reduced forms is further complicated by research results from within Conversation Analysis/Phonetics of Talk-in-Interaction demonstrating that the context in which canonical versus reduced (or upgraded versus downgraded) forms may occur is not necessarily simply determined. Curl and colleagues (Curl 2002, 2005; Curl, Local & Walker 2006) have pursued a number of lines of study shedding light on the reduced (or not) properties of repeated elements in conversation, responding to a body of literature arguing that first mentions of lexical items tend to be more "canonical" in form, while mentions of something that is already present in the conversation tend to be more reduced (cf. Fowler & Housum 1987; Fowler 1988; Bard et al. 1989; Jurafsky et al. 1998; Bell et al. 1999 *inter alia*; also discussion in section 1.4.3 about informativeness). While Ohala (1994) reports that creating a context of mishearing leads to speakers repeating their productions with more "canonical" features, Curl (2002) finds that not all repetitions as part of repair sequences, i.e., those in which speakers resolve some kind of misunderstanding or incorrect information, are produced the same. Instead, these repetitions are sensitive to how the repair process fits in with the rest of the conversation. If it is well-fitted to the current location, the repaired item is repeated with a phonetically "upgraded" form: louder, expanded pitch range, increased duration, as well as differences in articulatory form. If the repair turn is not well-fitted in its current location, however, the phonetic form employed in the repetition tends to closely resemble the phonetic form in the earlier production. Both of these findings contrast with the finding in experimental contexts and corpus studies that second mentions tend to be produced in a reduced form compared to the first mention (e.g. Baker & Bradlow 2009). They also highlight that a range of discourse factors play a role in determining the form a word takes. Curl also points out that since these turns are all produced clearly enough to obtain a display of understanding from the interlocutor, that appropriateness in context is not necessarily the same thing as optimally clear speech.

Several of the papers in the current volume address the question of the identity of "canonical" and "reduced" forms, and how they relate to one another. For example, Ernestus & Smith (chapter 5) propose that the essence of a word (similar to Niebuhr and Kohler's term "phonetic essence", Niebuhr & Kohler 2011; Kohler & Niebuhr 2011) is something other than a canonical form, and that we should perhaps be more interested in what may not be reduced away than in what is present when the "full form" is produced. Cole and Shattuck-Hufnagel (chapter 6) similarly propose that the essence of a word may be captured in the form of "landmarks", which reduced forms still aim at. Espy-Wilson et al.'s research in articulation during spontaneous speech (chapter 8) gives evidence for an underlying remnant of articulatory form, as would be predicted in the Articulatory Phonology framework, even when acoustic evidence of a segment is absent. Van Dommelen (chapter 3) addresses the issue of exposure to canonical and reduced forms in second-language learning, and the extent to which this influences L2 speakers' production behavior.

1.4.2 Units and scope of reduction phenomena

As discussed in section1.3.1, the bulk of the body of research into phonetic reduction has been focused at the level of the segment. However, the scope of reduction phenomena can also be considered to be much larger. Work in ASR fields, in particular, has investigated some larger domains for reduction, such as syllables or longer-range acoustic features.

1.4.2.1 Syllable

Greenberg (1999) presents a quantitative analysis of pronunciation variation in conversational speech taken from the Switchboard corpus (Godfrey et al. 1992), and shows that variation is systematic if analyzed at the level of the syllable. This study concludes that syllabic onsets are realized in their canonical form much more frequently than nuclei or codas, and that word stress has systematic effects on the pronunciation of syllables. Greenberg (1999) concludes that the syllable is a more suitable unit than the phone for describing the variation occurring in spontaneous speech.

Since Greenberg's quantitative phonetic analyses, many studies in the field of speech technology have investigated whether a syllable-based speech recognition system may have advantages, especially for the case of spontaneous speech. The benefit of using the syllable, however, is not always straightforward. Hämäläinen et al. (2008), for instance, compare context-independent single-path and multi-path syllable models with context-dependent phone models. In contrast to their original hypothesis, single-path syllable models and context-dependent phone models outperform multi-path syllable models. Their analysis shows that word recognition is mostly conditioned by syllabic context and lexical confusability. Their results suggest that multi-path syllable models are only beneficial to an ASR system if the pronunciation variation described at the syllable level of pronunciation can be linked with the word level in the language model.

The aforementioned study by Hämäläinen et al. can be seen as involving a full syllabic system, since the basic units of the acoustic models are syllables. The disadvantage of this approach is that there are more different syllables and syllabic contexts than phones and phonetic contexts (e.g., for triphones), and thus more data are necessary in order to have enough material to train the syllable models. Promising methods, however, have been found by combining the training of acoustic phone models with information about their positions within the syllabic structure. For example, Shafran and Ostendorf (2003) incorporate syllabic structures into acoustic model clustering. They find that phone model clustering on the basis of syllabic structures outperforms traditionally trained pentaphones in a recognition task on the spontaneous speech material from Switchboard.

Whereas the above methods deal with reductions implicitly (i.e., the acoustic models are trained on both reduced and fully realized segments in the training material), there are also explicit ways to incorporate reductions specific to certain syllabic structures and/or properties into the pronunciation modelling component of the ASR system. Schuppler et al. (2011), for instance, created pronunciation variants by applying reduction rules, which were dependent on syllabic structure and stress patterns, to the canonical pronunciations of words. Thus, different reduction rules apply to nuclei of stressed vs. unstressed syllables, to onset vs. coda consonants and consonant-clusters.

1.4.2.2 Acoustic phonetic features

Even though the syllable is a larger unit than the word, it is still treated as linear in most analyses. One problem with segment-based approaches in ASR is that deletions are seen as "complete" deletions, and that there are no ways to capture "traces of segments left" in surrounding segments (i.e. overlapping features), as in the "yesterday" example in section 1.2. Two chapters in this volume further explore the relationship between features and reduction in terms of acoustic phonetics. Ernestus & Smith (chapter 5) and Cole & Shattuck-Huffnagel (chapter 6), both show that certain acoustic features are more or less likely to be reduced than others and that features that are left behind are often blended together. Researchers following a Firthian tradition (cf. e.g. Firth 1948; Ogden & Walker 2001), have begun to investigate segments in terms of their parallels to prosody; cf. also Xu & Liu's (2013) extension of the Target Approximation model to account for segment dynamics. The approach of Articulatory Phonology (Browman & Goldstein 1990; 1992), discussed in section 1.3.1 above, seeks to model speech as a set of overlapping and dynamic gestures which are represented directly by both the speaker and the listener and is able to capture many reduction phenomena.

A study by Ostendorf (1999) proposes moving beyond the traditional ‘‘Beads-on-a-String’’ model of speech by using representations of speech based on acoustic features (AFs), which strongly resemble the articulatory gestures that Articulatory Phonology considers as primitives). Such a representation could consist of several layers: for instance, one for manner of articulation, one for place of articulation, one for voicing and one for nasality. Boundaries on different layers are placed independently of each other, and are thus capable of capturing the asynchronous gestures of the articulators, i.e., the acoustic correlates of the articulatory gestures. Thus, AFs seem to offer a natural way for representing (semi-) continuous articulatory gestures and the ensuing acoustic characteristics of speech signals (e.g., Frankel et al. 2007).

Promising results using AFs as the basic unit instead of phones have been forthcoming since the early 1990s. Deng and Erler (1992) compare multidimensional (or multivalued) feature representation of speech with a phone-based representation in an HMM recognition framework. They show that due to the high degree of data sharing, training data can be used well, and the resulting models are very capable of capturing coarticulatory effects such as feature spreading. For the task of stop consonant discrimination, they show performance gains for the AF-based system in comparison with the phone-based system.

Since the development of these early AF classifiers, automatic AF classification has been continuously further developed and used for speech recognition in adverse conditions (e.g. Kirchhoff 1999; Kirchhoff et al. 2002; Schutte and Glass 2005), to build language-independent phone recognizers (Siniscalchi et al. 2008, Siniscalchi and Lee 2014), and in computational models of human spoken-word recognition (Scharenborg 2010).

1.4.2.3 Prosody

Since many studies have shown that acoustic reduction is conditioned by the prosodic properties of a word, the potential benefit of incorporating prosodic information into ASR systems has also been investigated. For these purposes, a large number of prosodic features are usually automatically extracted from the speech signal, including fundamental frequency (F0), energy and rhythm features such as timing, durations, and silent pauses. According to Ostendorf et al. (2003), “A major problem in computational modeling of prosody is that these acoustic correlates provide cues to many different types of information associated with different time scales, from segmental to phrasal to speaker characteristics.” (148)

Prosody is incorporated into speech technology systems for various purposes: e.g., the detection of utterance endings or possible places for the realization of backchannels in dialogue systems, and the detection of paralinguistic characteristics such as emotions and speaker uncertainty. It has also been demonstrated that the incorporation of prosodic information into acoustic modelling and pronunciation modelling shows benefits, as for instance Ostendorf et al. (2003) do for a recognition task on the conversational speech material of the American English Switchboard corpus (Greenberg 1995). Ostendorf et al.’s model makes use of both intermediate symbolic representations and acoustic correlates of prosody. For the incorporation of prosody into large-vocabulary speech recognition, however, some obstacles are still to be overcome. Whereas acoustic properties are relatively easily extracted automatically, symbolic representations of prosody in conversational, spontaneous speech can still only be created by hand, a very time-consuming project, despite its great promise. One attempt at addressing this problem is the development of “silver standard” corpora, which involve high-quality automatic segmentation and labelling (cf. Mahlow et al. 2014).

1.4.3 Role of Predictability

It has long been observed that there is a relationship between reduced speech and predictable speech, and one common assumption has been that words that are harder to perceive or produce (because they are less predictable) are produced more clearly and this is sometimes explained with appeals to audience design (e.g. Lindblom’s H&H theory). This issue is addressed by Clopper and Turnbull (chapter 2), who review the evidence in favor of the conclusion that more predictable and higher frequency words and segments are on average shorter, and vowels in them are more centralized. However, when one looks at very reduced words, they are often not especially predictable in their immediate context (Brouwer, Mitterer, & Huettig, 2013). This means that in online production of speech, there is a lot of variability in the degree of reduction that is not accounted for by predictability. Clopper and Turnbull (chapter 2) also note a number of complications to the relationship. It is also clear from much of the perception literature that more reduced words are harder to recognize, even (or especially) when they are spontaneous productions heard in their original context (Brouwer, Mitterer, & Huettig, 2013). This again calls into question the idea that talkers reduce on-line where they can “get away with it” because the context makes up for their lack of clarity. Seyfarth (2014) argues that word durations are explained in part by mean predictability and not just contextual predictability. That is, words that are on average more predictable will tend to be reduced even when they occur in an unpredictable context. This result is more in line with a representational account of reduction rather than an on-line or listener oriented account (cf. Clopper and Turnbull, chapter 2).

The phonetic literature on Talk-in-Interaction has also noted the discrepancy between accounts of reduction related to predictability. For example, Local, Kelly & Wells (1986; see also Niebhurn, Görs & Gaupe 2013), in their discussion of turn-taking in Tyneside English, include more centralized vowel quality, i.e. a reduced production, as one feature of turn-constructional unit (TCU) ends which are followed by speaker transition. The centralized vowels in these turn-final locations need not be part of repetitions of previous lexical material, although they may be. Instead, they are produced in a reduced form (in concert with other cues) in order to indicate a speaker's intention to stop speaking. Local et al. (1986) do not find evidence that lexical material produced in these locations is less informative than in other locations. Thus the use of "reduced" forms must be context-appropriate on some basis other than it simply being easier to access. Curl and colleagues’ observations about repetitions in the context of repair (discussed in section 1.4.1) also show that some repairs are not clearer than the original productions, again challenging the notion that reduction occurs in places where the listener requires less information to recognize the word.

In light of this discussion it is also of interest to consider how the structure of a typical ASR system can be compared with the H&H model: a top-down language model (LM) makes hypotheses about how likely different words are given a specific n-gram context, while a bottom-up acoustic model (AM) plus a lexicon makes the same hypotheses based on the acoustic signal. If the H&H model were strictly true, i.e. if loss of information in the bottom-up signal occurs in locations where the top-down model should make strong predictions, these two should balance each other out. However, the fact that ASR models struggle with reduction seems to challenge this assumption. One factor may be that in a traditional ASR system, the LM, AM and the Pronunciation Model (i.e. the lexicon) are trained independently. In order to make ASR systems more robust for conversational speech (which has a high number of reduced words), methods have been developed to combine the modelling of the components. For instance, this can be achieved by using the variants themselves (instead of the underlying words) to calculate the N-grams of the LM. First, this would easily allow incorporation of cross-word processes and reductions at the word boundaries. Second, variants which tend to occur together (for instance in multi-word expressions, highly frequent bigrams and trigrams such as “I have done”, “I don’t know”, etc.) are also modelled together. This approach seems simple, but runs into problems of data-coverage. Whereas for regular LMs all text sources available can be used to train the N-grams, for such a combined LM phonetically transcribed speech material is required. One way to get around this “lack of data” issue is presented by Kessens et al. (1999), who train a LM to which pronunciation variants which are created from phonological rules are added for each word, including both within-word and cross-word processes. With this approach, the system’s WER improved by 8.8%.

1.5 Chapter summaries

The chapters presented in this volume each address some aspect of the issues we have raised surrounding our understanding of reduction phenomena. The chapter by **Clopper and Turnbull** addresses sources of reduction, including listener-oriented, talker-oriented, and structural/evolutionary accounts of reduction. They find evidence for a complex relationship between these different sources.

The next two chapters examine reduction from a cross-linguistic perspective. The chapter by **van Dommelen** investigates influences on reduction for second-language speakers of a language, who often have more access to canonical productions in their target language. He finds evidence that second-language learners may reduce in similar ways to native speakers, though to differing degrees. The chapter by **Adda-Decker and Lamel** uses ASR tools to conduct automatic analysis of reduction phenomena across different languages and speaking styles, taking advantage of canonical word dictionaries to identify areas with reduced phonetic production.

The last four chapters question the common assumption of phone-based reduction. The chapter by **Ernestus and Smith** reports factors conditioning the reduction of *eigenlijk* in Dutch, and raises the issue of the automaticity of reduction phenomena, as well as that of invariant landmarks occurring even in the most reduced productions of a word. The chapter by **Cole and Shattuck-Hufnagel** further investigates phonetic landmarks in imitations of reduced forms, reporting on which kinds of landmarks are most stable across speakers and contexts. The chapter by **Cutugno, Origlia and Schettino** addresses mismatches between expected and observed production in speech on the syllable level. They use automatic methods to identify syllables with reduced forms, and propose a rule-based relationship between reduced and non-reduced forms. The chapter by **Espy-Wilson, Tiede, Mitra, Sivaraman, Saltzman and Goldstein** uses a speech inversion system to identify areas of overlapping gesture, even where acoustic evidence for reduced segments may be lacking, with the aim of improving ASR systems.

Finally, the **conclusion** considers reduction from the perspective of its role in the history of the field(s) and draws together some of the implications of the studies reported in this volume for future conceptions of and research on phonetic reduction.

Works Cited

Adda-Decker, M. and Lamel, L. (2000). Modeling reduced pronunciations in German. Phonus 5, Institute of Phonetics, University of the Saarland, pp. 129–143.

Adda-Decker, M. and Snoeren, N.D. (2011). Quantifying temporal speech reduction in French using forced speech alignment," Journal of Phonetics 39(3): 261-270.

Andruski, J.E., Blumstein, S., and Burton, M. (1994). The effect of subphonetic differences on lexical access. Cognition 52: 163–187.

Baker, R.E., & Bradlow, A.R. (2009). Variability in word duration as a function of probability, speech style, and prosody. Language & Speech 52(4): 391-413.

Barnett, J., (1974). A phonological rule compiler. In: Erman., L. (Ed.), Proceedings of the IEEE Symposium on Speech Recognition, Carnegie-Mellon University, Pittsburgh, PA, pp. 188-192.

Bard, E.G., Lowe, A.J. & Altmann, G.T.M. (1989). The effect of repetition on words in recorded dictations. In Eurospeech '89: Proceedings of the European Conference on Speech Communication and Technology, vol. 2, pp. 573–576.

Barry, W.J. & van Dommelen, W.A. (2005) The Integration of Phonetic Knowledge in Speech Technology. Dordrecht, the Netherlands: Springer. ISBN 1-4020-2635-8.

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C. & Gildea, D. (1999). Forms of English function words – effects of disfluencies, turn position, age and sex, and predictability. In Proceedings of ICPhS, San Francisco, USA.

Brouwer, Susanne, Holger Mitterer & Falk Huettig. (2013). Discourse context and the recognition of reduced and canonical spoken words. *Applied Psycholinguistics* 34. 519–539. doi:10.1017/s0142716411000853.

Browman, C.P., & Goldstein, L. (1990). Tiers in articulatory phonology, with some implications for casual speech. Papers in laboratory phonology I: Between the grammar and physics of speech, 341-376.

Browman, C.P., & Goldstein, L. (1992). Articulatory phonology: An overview. Phonetica, 49(3-4), 155-180.

Clayards, M. (2010). Using probability distributions to account for recognition of canonical and reduced word forms. In LSA Annual Meeting Extended Abstracts (Vol. 1, 4 pages). http://dx.doi.org/10.3765/exabs.v0i0.529

Clayards, M., & Knowles, T. (2015). Prominence enhances voiceless-ness and not place distinctions in English voiceless sibilants. Proceedings of ICPhS 2015, Glasgow, Scotland.

Connine, C. (2004). It's not what you hear but how often you hear it: On the neglected role of phonological variant frequency in auditory word recognition. Psychonomic Bulletin & Review, 11(6): 1084–1089.

Curl, T.S. (2002). *The phonetics of sequence organization: an investigation of lexical repetition in other-initiated repair sequences in American English*. Ph.D. dissertation, University of Colorado.

Curl, T.S. (2005). Practices in other-initiated repair resolution: the phonetic differentiation of 'repetitions'. Discourse Processes 39(1): 1-43.

Curl, T.S., Local, J. & Walker, G. (2006). Repetition and the prosody-pragmatics interface. Journal of Pragmatics 38(10): 1721-1751.

Deng L., & Erler, K. (1992). Structural design of HMM speech recognizer using multi-valued phonetic features: comparison with segmental speech units. Journal of the Acoustical Society of America 92: 3058- 3067.

DiCanio, C., Nam, H., Amith, J.D., Castillo García, R., & Whalen, D.H. (2015). Vowel variability in elicited versus spontaneous speech: Evidence from Mixtec. Journal of Phonetics 48: 45-59.

Diehl, R.L., Lotto, A.J., & Holt, L.L. (2004). Speech perception. Annual Review of Psychology 55: 149–179. doi:10.1146/annurev.psych.55. 090902.142028

Dilley, L.C., & Pitt, M.A. (2010). Altering context speech rate can cause words to appear or disappear. Psychological Science 21(11): 1664-1670. 10.1177/0956797610384743

Elgin, S.H. (1979). *What is Linguistics.* Second Edition. Englewood Cliffs, New Jersey: Prentice-Hall.

Ernestus, M. (2000). *Voice assimilation and segment reduction in casual Dutch. A corpus-based study of the phonology-phonetics interface*. Ph.D. dissertation, LOT, Vrije Universiteit Amsterdam, The Netherlands.

Ernestus, M., & Baayen, R.H. (2007). The comprehension of acoustically reduced morphologically complex words: The roles of deletion, duration and frequency of occurrence. In Proceedings of the 16th International Congress of Phonetic Sciences, Saarbrücken, Germany, pp. 773-776.

Ernestus, M., Baayen, H., & Schreuder, R. (2002). The recognition of reduced word forms. Brain and Language 81: 162–173.

Fant, G. (1962). *Den akustika fonetikens grunder*. Kungliga Tekniska Högskolan, Taltransmissionslab. 2nd printing. Stockholm: Royal Institute of Technology, no. 7.

Firth, J.R. (1948). Sounds and Prosodies. Reprinted in Palmer (ed.) (1970) *Prosodic Analysis*. London: Oxford University Press, 1-26.

Fónagy, I. (1966). Electrophysical and acoustic correlates of stress and stress perception. Journal of Speech and Hearing Research 9:231-244.

Fowler, C.A. (1986). An event approach to the study of speech perception from a direct-realist perspective. Journal of Phonetics 1: 3–28.

Fowler, C.A. (1988). Differential shortening of repeated content words produced in various communicative contexts. Language & Speech 31(4): 307–20.

Fowler, C.A. & Housum, J. (1987). Talkers signaling of ‘new and ‘old’ words in speech and listeners’ perception and use of the distinction. Journal of Memory and Language 26: 489–504.

Frankel, J., Wester, M., King, S., (2007). Articulatory feature recognition using dynamic Bayesian networks. Computer Speech and Language 21(4): 620–640.

Fry, D.B. (1955). Duration and intensity as physical correlates of linguistic stress. Journal of the Acoustical Society of America 27:4, 765-768.

Fry, D.B. (1958). Experiments in the perception of stress. Language & Speech 1(2): 126-152.

Gaskell, M.G., & Marslen-Wilson, W.D. (1996). Phonological variation and inference in lexical access. Journal of Experimental Psychology: Human Perception and Performance 22: 144–158. doi:10.1037/ 0096-1523.22.1.144

Gimson, A.C. (1977). *An introduction to English pronunciation*. London: Edward Arnold.

Godfrey, J.J., Holliman, E.C. and McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In Proceedings of ICASSP, vol. 1, pp. 517–520.

Goldinger, S D. (1998). Echoes of echoes? An episodic theory of lexical access. Psychological Review 105(2): 251.

Gow, D.W., Jr. (2002). Does English coronal place assimilation create lexical ambiguity? Journal of Experimental Psychology: Human Perception and Performance 28: 163–179.

Gow, D.W., Jr. (2003). Feature parsing: Feature cue mapping in spoken word recognition. Perception & Psychophysics 65: 575–590. doi: 10.3758/BF03194584

Greenberg, S. (1999). Speaking in shorthand. A syllable-centric perspective for understanding pronunciation variation. Speech Communication 29: 159-176.

Hain, T. (2005). Implicit modelling of pronunciation variation in automatic speech recognition. Speech Communication 46: 171–188.

Hawkins, S., & Smith, R. (2001). Polysp: A polysystemic, phonetically-rich approach to speech understanding. Italian Journal of Linguistics 13: 99-188.

Hämäläinen, A., Ten Bosch, L., Boves, L. (2008). Modelling pronunciation variation with single-path and multi-path syllable models: Issues to consider. Speech Communication 51(2): 130-150.

IPDS (1997). CD-ROM: The Kiel Corpus of Spontaneous Speech, vol i- vol iii. Corpus description available http://www.ipds.uni- kiel.de/forschung/kielcorpus.de.html (viewed 25/04/2011).

Jakobson, R. & Halle, M. (1956). *Fundamentals of language*. The Hague: Mouton & Co.

Jande, P.-A. (2003). Evaluating rules for phonological reduction in Swedish. PHONUM 9: 149-152.

Janse, E. (2004). Word perception in fast speech: Artificially time-compressed vs. naturally produced fast speech. Speech Communication 42: 155–173.

Janse, E., Nooteboom, S., & Quené, H. (2007). Coping with gradient forms of /t/- deletion and lexical ambiguity in spoken word recognition. Language and Cognitive Processes 22: 161–200.

Janse, E., & Ernestus, M. (2011). The roles of bottom-up and top-down information in the recognition of reduced speech: evidence from listeners with normal and impaired hearing. Journal of Phonetics 39(3): 330-343.

Johnson, K. (1997). Speech perception without speaker normalization. In K. Johnson & Mullenix, J. W. (Eds.), *Talker variability in speech processing*. San Diego: Academic Press, pp. 145–165.

Johnson, K. (2005). Speaker normalization. In R. Remez, & D. B. Pisoni (Eds.), *The Handbook of Speech Perception*. Oxford: Blackwell.

Johnson, K. (2006). Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. Journal of Phonetics 34(4): 485-499.

Johnson, K., Flemming, E. & Wright, R. (1993). The hyperspace effect: phonetic targets are hyperarticulated. Language 69(3): 505-528.

Jurafsky, D., Bell, A., Fosler-Lussier, E., Girand, C. & Raymond, W. (1998). Reduction of English function words in switchboard. In Proceedings of ICSLP, vol. 7, Sydney, Australia, pp. 3111–3114.

Kessens, J.M., Wester, M., Strik, H., (1999). Improving the performance of a Dutch CSR by modelling within-word and cross-word pronunciation variation. Speech Communication 29(2-4): 193-207.

Kessens, J.M., Cucchiarini, C., and Strik, H. (2003). A data-driven method for modeling pronunciation variation. Speech Communication (40), pp. 517–534.

Kirchhoff, K., (1999). *Robust speech recognition using articulatory information*. Ph.D. dissertation, University of Bielefield.

Kirchhoff, K., Fink, G.A., Sagerer, G., (2002). Combining acoustic and articulatory feature information for robust speech recognition. Speech Communication 37: 303–319.

Kohler, K.J. (1974). Koartikulation und Steuerung im Deutschen. In Sprachsystem und Sprachgebrauch: Festschrift für Hugo Moser, Teil 1. Düsseldorf: Schwann, pp. 172-192.

Kohler, K.J. (1990). Segmental reduction in connected speech in German: phonological facts and phonetic explanations. In W.J. Hardcastle, & A. Marchal (Eds.), Speech production and speech modelling. Dordrecht: Kluwer Academic Publishers, pp. 69-92.

Kohler, K.J. & Niebuhr, O. (2011). On the role of articulatory prosodies in German message decoding. Phonetica 68:1-31.

Lass, R. (1984). *Phonology: an introduction to basic concepts*. Cambridge, UK: Cambridge University Press.

Lieberman, P. (1960). Some acoustic correlates of word stress in American English. Journal of the Acoustical Society of America 32(4): 451-454.

Lindblom, B. (1963). Spectrographic study of vowel reduction. Journal of the Acoustical Society of America 35(11): 1773-1781.

Lisker, L. (1984). The pursuit of invariance in speech signals. Journal of the Acoustical Society of America 77(3):1199-1202.

Local, J. (2003). Variable domains and variable relevance: interpreting phonetic exponents. Journal of Phonetics 31: 321-339.

Local, J., Kelly, J. and Wells, W. (1986). Towards a phonology for conversation: turn-taking in Tyneside English. Journal of Linguistics 22: 411–437.

Mahlow, C., Eckart, K., Stegmann, J., Blessing, A., Thiele, G., Gärtner, M. and Kuhn, J. (2014). Resources, tools and applications at the CLARIN center Stuttgart. Proceedings Konvens 2014, Hildesheim, Germany.

McMurray, B., & Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. Psychological Review 118(2): 219.

Mitterer, H., Csépe, V., & Blomert, L. (2006). The role of perceptual integration in the recognition of assimilated word forms. Quarterly Journal of Experimental Psychology 59: 1395–1424. doi:10.1080/17470210500198726

Niebuhr, O., K. Görs & Graupe, E. (2013). Speech reduction, intensity, and F0 shape are cues to turn-taking. Proceedings of the 14th Annual SigDial Meeting on Discourse and Dialogue, Metz, France, pp. 261-269.

Niebuhr, O., & Kohler, K.J. (2011). Perception of phonetic detail in the identification of highly reduced words. Journal of Phonetics 39: 319-329.

Nolan, F. (1992). The descriptive role of segments: evidence from assimilation. In G. Docherty and D.R. Ladd (Eds.), Laboratory Phonology 2. Cambridge: Cambridge University Press, pp. 261-280.

Ogden, R. & Walker, G. (2001). 'We speak prosodies and we listen to them.' Symposium on Prosody and Interaction, Uppsala, Sweden.

Ohala, J. J. (1994). Acoustic study of clear speech: a test of the contrastive hypothesis. In International Symposium on Prosody, Vol. 18, pp. 75-89.

Ostendorf, M., (1999). Moving beyond the ‘‘Beads-on-a-String’’ model of speech. In: Proceedings of 1999 IEEE Workshop on Automatic Speech Recognition and Understanding, Vol. 1. pp. 79–83.

Ostendorf, M., Shaffran, I. and Bates, R. (2003). Prosody models for conversational speech recognition. In Proceedings of the 2nd Plenary Meeting and Symposium on Prosody and Speech Processing, pp. 147-154.

Peters, B. (2005). The database - The Kiel Corpus of Spontaneous Speech. Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel (AIPUK), 35a, 1-6.

Pickett, J. M., & Pollack, I. (1963). Intelligibility of excerpts from fluent speech: Effects of rate of utterance and duration of excerpt. Language & Speech 3: 151–164.

Pisoni, D. B. (1997a). Some thoughts on ‘‘normalization’’ in speech perception. In K. Johnson & Mullenix, J. W. (Eds.), Talker variability in speech processing. San Diego: Academic Press, pp. 9–32.

Pitt, M. A. (2009). The strength and time course of lexical activation of pronunciation variants. Journal of Experimental Psychology: Human Perception and Performance, 35, 896–910.

Pitt, M. A., Dilley, L., & Tat, M. (2011). Exploring the role of exposure frequency in recognizing pronunciation variants. Journal of Phonetics 39(3): 304-311.

Pitt, M.A., Johnson, K., Hume, E., Kiesling, S. and Raymond, W.D. (2005). The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability,” Speech Communication 45: 89–95.

Port, R. (2007). How are words stored in memory? Beyond phones and phonemes. New Ideas in Psychology 25(2): 143-170.

Ranbom, L., & Connine, C. (2007). Lexical representation of phonological variation in spoken word recognition. Journal of Memory and Language 57(2): 273-298.

Ranbom, L. J., Connine, C. M., & Yudman, E. M. (2009). Is phonological context always used to recognize variant forms in spoken word recognition? The role of variant frequency and context distribution. Journal of Experimental Psychology: Human Perception and Performance 35(4): 1205-1220.

Sarac ̧lar, M., Nock, H.J., Khudanpur, S. (2000). Pronunciation modelling by sharing Gaussian densities across phonetic models. Computer Speech and Language 14: 137–160.

Scharenborg, O. (2010). Modeling the use of durational information in human spoken-word recognition. Journal of the Acoustical Society of America 127(6): 3758–3770.

Schegloff, E. A., Jefferson, G., & Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. Language 53: 361–382.

Schuppler, B., Adda-Decker M., & Morales-Cordovilla J. A. (2014b). Pronunciation variation in read and conversational Austrian German. Proceedings of Interspeech 2014, pp. 1453-1457.

Schuppler, B., Ernestus M., Scharenborg O., & Boves L. (2011). Acoustic reduction in conversational Dutch: A quantitative analysis based on automatically generated segmental transcriptions. Journal of Phonetics 39: 96 - 109

Schuppler, B., Hagmüller M., Morales-Cordovilla J. A., & Pessentheiner H. (2014a). GRASS: The Graz Corpus of Read and Spontaneous Speech. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pp. 1465-1470.

Schuppler, B., van Dommelen, W., Koreman, J. and Ernestus, M. (2012). “How linguistic and probabilistic properties of a word affect the realization of its final /t/: Studies at the phonemic and sub-phonemic level,” Journal of Phonetics 40: 595–607.

Schutte, K. and Glass, J. (2005). Robust detection of sonorant landmarks. In Proceedings of 6th Interspeech, Lisbon, Portugal, pp. 1005–1008.

Seyfarth, S. (2014). Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. Cognition 133(1): 140-155.

Shafran, I. and Ostendorf, M. (2003). Acoustic Model Clustering Based on Syllable Structure. Computer Speech & Language 17(4): 311–328.

Siniscalchi, S. M., Lee, C.-H., (2014). An attribute detection based approach to automatic speech recognition. Loquens 1(1), e005.doi:http://dx.doi.org/10.3989/loquens.2014.005.

Siniscalchi, S. M., Svendsen, T., Lee, C.-H. (2007). Towards bottom-up continuous phone recognition. In Proceedings of IEEE ASRU Workshop, pp. 566–569.

Snoeren, N. D. (2011). Psycholinguistique cognitive de la parole assimilée. Saarbrücken, Germany: Editions Universitaires Européennes.

Sonderegger, M., & Yu, A. C. L. (2010). A rational account of perceptual compensation for coarticulation. In S. Ohlsson & R. Catrambone (Eds.), Proceedings of the 32nd Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society, pp. 375–380.

Strik, H. and Cucchiarini, C. (1999). Modeling pronunciation variation for ASR: a survey of the literature. Speech Communication 29 (2-4): 225-246.

Sumner, M. (2013). A phonetic explanation of pronunciation variant effects. Journal of the Acoustical Society of America 134(1): EL26-EL32.

Torreira, F., Adda-Decker, M. and Ernestus, M. (2010). The Nijmegen Corpus of Casual French. Speech Communication 52(3): 201–212.

Torreira, F., & Ernestus, M. (2011). Vowel elision in casual French: The case of vowel /e/ in the word c’était. Journal of Phonetics 39(1): 50 -58.

Tucker, B. V. (2011). The effect of reduction on the processing of flaps and/g/in isolated words. Journal of Phonetics 39(3): 312-318.

Van Bael, C., Boves, L., van den Heuvel, H., and Strik, H. (2007). Automatic phonetic transcription of large speech corpora. Computer Speech and Language 21: 652–668.

Van de Ven, M., Ernestus, M., & Schreuder, R. (2012). Predicting acoustically reduced words in spontaneous speech: The role of semantic/syntactic and acoustic cues in context. Laboratory Phonology 3: 455-481.

Viebahn, M. C., Ernestus, M., & McQueen, J. M. (2015). Syntactic predictability in the recognition of carefully and casually produced speech. Journal of Experimental Psychology: Learning, Memory, and Cognition 41(6): 1684-1702.

Wester, M. (2002). *Pronunciation Variation Modeling for Dutch Automatic Speech Recognition*. PhD dissertation, Radboud University Nijmegen, The Netherlands.

Xu, Y. & Liu, F. (2013). Intrinsic coherence of prosodic and segmental aspects of speech. In Niebuhr, O. & Pfitzinger, H.R. (Eds.) Prosodies: context, function, and communication. Berlin/New York: de Gruyter, pp. 1-26.

Yuan, J. and Liberman, M. (2009). Investigating /l/ variation in English through forced alignment. Proceedings of Interspeech 2009, pp. 2215–2218.