The Production and Perception of Signal-Based Cues to Word Boundaries

DISSERTATION

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Dahee Kim, B.A., M.A.

Graduate Program in Linguistics

The Ohio State University

2013

Dissertation Committee:

Dr. Cynthia G. Clopper, Advisor

Dr. Mark A. Pitt, Advisor

Dr. Shari R. Speer

Abstract

During speech comprehension, listeners must segment continuous speech into a series of discrete words. Previous studies of word segmentation have reported conflicting results as to whether talkers produce acoustic-phonetic cues demarcating word boundaries and whether the acoustic-phonetic details are sufficient to guarantee successful word segmentation by listeners. In this dissertation, we suggest that the conflicting results can be reconciled by considering acoustic-phonetic variation in the spoken language. Among the factors conditioning acoustic-phonetic variability, we focused on the influences of speech clarity and phonetic context on the production and perception of acoustic cues to word boundaries.

Forty native speakers of American English read aloud sentences containing word-boundary ambiguities (e.g., *collect<u>s</u> gulls* vs. *collect <u>s</u>kulls*) to three "listener" confederates, who were a young native, a young nonnative, and an older hearing impaired listener introduced by a short video clip. The word-boundary ambiguities involved consonant-vowel, /s/-consonant, and schwa-consonant sequences and the consonants were balanced for obstruents and sonorants within each sequence type. Talkers silently read two sentences that were visually presented and read aloud the one of the two sentences that flashed. For half of the talkers, the two sentences presented on a trial were unrelated, while for the other talkers, the two sentences contained word-boundary minimal pairs. Clarity of acoustic-phonetic cues to word boundaries was estimated by the

fit of logistic regression models predicting the location of word boundaries based on the durational properties of segments and pauses at and around the potential word boundaries. Statistical analyses of the model fit suggest that talkers adjust the extent to which they produce durational cues to word boundaries in that they enhance the clarity of word boundaries when speaking to listeners with foreseeable communication difficulties. The extent to which talkers clarified cues to word boundaries was constrained by the phonetic context surrounding word boundaries, as shown by the insufficiency of durational cues for distinguishing ambiguous schwa-initial sequences such as *a*l*ong* vs. *a* l*ong*, regardless of speech clarity.

Listening experiments tested whether and to what extent the talkers' clarity modulation of the acoustic-phonetic segmentation cues affected listeners' word segmentation. Listeners were shown to be more accurate and faster in determining the location of the word boundary when the talkers' speech was directed to an older hearing impaired or a young nonnative listener than when speech was directed to a young native listener. In addition, listeners were better at determining the location of the word boundaries when the stimuli were produced by talkers who read the target sentences after reading the sentence pairs containing word-boundary minimal pairs than the stimuli produced by talkers who read the target sentences after reading sentence pairs that did not highlight the word-boundary ambiguity. The extent to which listeners benefited from enhanced speech clarity due to the listener confederates' linguistic background or talkers' awareness of the word-boundary ambiguity differed depending on the phonetic context surrounding the potential word boundaries. The results contribute to our understanding of variation in speech production and word segmentation.

Vita

February 1999 .....Hanyoung Foreign Language High School

2002.....B.A. English, French, Yonsei University

2005.....M.A. English Linguistics, Yonsei University

2006 to present .....Graduate Teaching Associate, Department of Linguistics, The Ohio State University

Publications

Kim, D., Stephens, J. D., & Pitt, M. A. (2012). How does context play *a part* in splitting words *apart*? Production and perception of word boundaries in casual speech. *Journal of memory and language*, *66*(4), 509-529.

Kim, J., Kim, D. & Rhee, S-C. (2008). An analysis of Korean monophthongs produced by Korean native speakers and adult learners of Korean. *Malsori, the Journal of the Korean Society of Phonetic Sciences and Speech Technology*, *65*, 13-36.

Kim, D. (2005). Phonetic factors conditioning the release of English sentence-final stops. *Malsori, the Journal of the Korean Society of Phonetic Sciences and Speech Technology*, *53*, 1-16.

Fields of Study

Major Field: Linguistics

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

During speech communication, listeners may hear utterances that they have never heard before, because speakers can generate an infinite number of novel utterances. As a result, understanding spoken language must involve understanding the words that compose an utterance, rather than understanding an utterance as an indivisible whole. In order to understand individual words in spoken utterances, listeners have to find word boundaries and parse continuous speech into a series of words. The process of segmenting words from continuous speech is not trivial, especially given that a typical utterance that a speaker produces in a normal conversational situation is rarely a single word but composed of multiple words (Brent & Siskind, 2001; Aslin, Woodward, LaMendola, & Bever, 1996). Consequently, for most utterances that they hear and process, listeners have to segment continuous speech into a series of words, making the problem of word segmentation highly prevalent.

Word segmentation involves locating word boundaries in the speech signal and accessing the lexical items that talkers produced. Previous studies of word segmentation have sought to identify factors influencing listeners' perception of word boundary locations. "Minimal pairs" containing a word-boundary ambiguity—identical sequences of phonemes that differ only in the location or the presence/absence of a word boundary (e.g., *nitrate* vs. *night rate* vs. *Nye trait*)--have served as stimuli for a number of listening experiments. For instance, Lehiste (1960) presented listeners with word-boundary

minimal pairs produced by three talkers and had listeners report what they heard. Results showed that listeners were highly accurate in interpreting the word-boundary minimal pairs as intended by talkers. In a similar study, Hoard (1966) examined listeners' segmentation of connected speech produced by four talkers who read aloud a short story containing word-boundary minimal pairs. Listeners' segmentation accuracy was high (approximately 88%), suggesting that listeners may rely on acoustic-phonetic cues for word segmentation.

The finding that listeners are sensitive to the acoustic-phonetic markers of word boundaries has been replicated by studies using tasks other than the transcription tasks used in earlier studies. For instance, Gow and Gordon (1995) used a cross-modal lexical decision task and showed that *two lips* primed *kiss*, while its oronym *tulips* did not prime *kiss*, which suggests that the acoustic differences between the /l/ in *two lips* and the /l/ in *tulips* affected how listeners segmented the words. Davis, Marslen-Wilson and Gaskell (2002) confirmed the importance of acoustic-phonetic cues in segmentation by showing that listeners can disambiguate temporary word-boundary ambiguities. They had listeners perform a gating task, during which they provided written responses to successively presented sentence fragments generated from a longer sentence containing the syllable [kæp], which creates an ambiguity between *cap* as a single word versus [kæp] as the first syllable of *captain* (e.g., *The soldiers saluted the flag with his cap tucked under his arm* vs. *The soldier saluted the flag with his captain looking on*). Responses to intended shorter and longer word stimuli differed even before the stimuli diverged phonemically, suggesting that listeners use acoustic-phonetic cues to resolve word-boundary ambiguity.

Studies using on-line measures such as eye movements also confirmed listeners' sensitivity to the phonetic details marking word boundaries (Salverda, 2005; Shatzman & McQueen, 2006).

The finding that phonetic details influence listeners' perception of word boundaries has led researchers to investigate what acoustic properties correspond to listeners' perception of word boundaries. Lehiste (1960) performed detailed spectrographic analyses on naturally produced word-boundary minimal pairs and identified the following potential acoustic-phonetic cues to word boundaries in spoken English, among others: glottalization of word-initial vowels, lengthening of word-final vowels, flapping of word-medial coronal stops, allophonic variation of aspirated versus unaspirated stop consonants, more extreme formant structure of word-initial vowels, allophonic variation in clear /l/ and dark /l/ (see also Sproat & Fujimura, 1993; Smith & Hawkins, 2000), and lengthening of word-initial and -final segments and syllables as compared to those that occur word-medially (see also Lehiste, 1972). Though such acoustic-phonetic and allophonic cues might have assisted listeners in distinguishing the word-boundary minimal pairs, Lehiste's (1960) acoustic analyses revealed no evidence that every talker produces consistent acoustic markers to word boundaries equivalent to visual cues in written language such as white spaces. Similarly, Hoard (1996) pointed out that most allophonic variations occurring at word boundaries are segment-specific and thus may not function as cues to every word boundary in the speech stream. Umeda and Coker (1974) suggested that the discontinuity perceived at word boundaries may be attributable to word-initial consonants that are phonetically more consonantal than word-

medial or word-final consonants, potentially allowing listeners to perceive some discontinuity that they interpret as word boundaries. In summary, studies using natural speech have shown that talkers produce certain acoustic-phonetic markers of word boundaries, typically position-specific allophones.

Phonetic properties of naturally produced speech may co-vary, making it challenging to identify the precise acoustic correlates of perceived word boundaries. Thus, studies using synthetic stimuli have produced result to complement those obtained in studies using naturally produced speech. Nakatani and Dukes (1977) used synthetic stimuli created by splicing word-minimal pair productions (e.g., *no notion* and *known ocean*) into four subparts and then concatenating them. They found that the phonetic cues at and around a word boundary affected word segmentation and the strongest cues for word-boundary perception occurred at the beginning of the word, such as glottalization of word-initial vowels and allophonic variation of the liquids /r/ and /l/. Additionally, they distinguished qualitative (i.e., allophonic) and quantitative (i.e., durational, amplitude, and rate of formant transition) cues to word boundaries and suggested that the qualitative cues have a greater perceptual effect than the quantitative cues (Christie, 1974). Nakatani and Schaffer (1978) used reiterant speech composed of three syllables /ma.ma.ma/ and showed that listeners could parse the reiterant speech as /ma.ma # ma/ or /ma # ma.ma/ and listeners tended to base their decision on the durational properties of segments, but not on pitch or amplitude contour.

Another line of research that has shed light on the acoustic cues to word boundaries includes studies focusing on the acoustic markers of prosodic structure of

spoken languages. The discovery of utterance-final lengthening of segments and syllables (Joos, 1962; Oller, 1973; Klatt, 1976; White, 2002) prompted researchers to investigate what types of prosodic boundaries trigger phonetic variation and whether word boundaries also trigger the final lengthening of segments and syllables (Beckman & Edwards, 1990; Turk & Shattuck-Hufnagel, 2000; Smith, 2004, among others). Results from most studies have reported that talkers produce the duration of phonologically identical segments or syllables differently depending on where in a word the segments occur, in that syllables and segments that begin or end a word tend to be longer than syllables and segments that occur in the middle of a word (Lehiste, 1960; 1972; Oller, 1973; Harris & Umeda, 1974; Klatt, 1976; Port, 1981; Beckman & Edwards, 1990; Turk & Shattuck-Hufnagel, 2000; Smith, 2004). In addition to the phonetic processes at the end of prosodic domains, domain-initial (e.g., utterance-initial, phrase-initial, word-initial or syllable initial) phonetic processes have been shown to trigger phonetic variation, typically "strengthening." For instance, Byrd and Saltzman (1998) showed that articulation of a domain-initial consonant tended to be produced with more and longer constriction, resulting in longer acoustic duration of word-initial segments (Cooper, 1991; Fougeron, 2001), which may serve to signal the word boundary location as well.

In summary, the phonetics of word boundaries has extensively been studied so far from the perspectives of speech production and speech perception. Allophonic and durational properties at and surrounding word boundaries have been found to be the strongest factor determining listeners' word boundary perception. In addition, talkers were found to produce acoustic-phonetic variation conditioned by the location and

presence/absence of word boundaries, suggesting that talkers produce potential cues to word boundaries. However, rather surprisingly, the literature on spoken word recognition often states that speech contains no acoustic markers that are equivalent to the white spaces between words contained in the written language (Davis, 2000; McQueen, Cutler, Briscoe, & Norris, 1995; Cutler, 1996).

Proponents of the view that not all word boundaries are marked with certain acoustic events argue that the percept of discontinuity between words results from knowledge about the language. For instance, Cutler (1996) gave an example of hearing a language that the listeners do not know at all: listeners who are listening to some unfamiliar language would not hear speech of the unfamiliar language as an orderly sequence of discrete lexical units in part due to the lack of acoustic-phonetic cues demarcating word boundaries. Relatedly, Reddy (1976) addressed the question of the extent to which listeners can segment words without the help of syntactic and semantic context. He had four human listeners transcribe anomalous phrases such as *in mud eels are* or *in clays none are*. Responses generated from these two phrases showed that no listeners reported that they heard *mud* from the first phrase and only one listener reported that she heard *clay*. Instead, listeners reported that they heard *my deals*, *muddies*, and *model* for the intended *mud eels*. Results from Reddy's (1976) experiment confirm that listeners must rely on knowledge-based cues such as syntax and semantics to hear word boundaries as intended by talkers. Cole and Jakimik (1980) also pointed out that talkers may not always produce salient word boundary cues. They estimated how frequently word boundaries are acoustically demarcated and reported that at a conservative estimate,

fewer than 40% of word boundaries are marked by some acoustic events such as silence or allophonic variation and that word boundaries marked by salient acoustic events tend to coincide with larger prosodic boundaries. Though Cole and Jakimik (1980) did not report how they estimated the availability of word boundary cues, their finding suggests that talkers may not always produce acoustic-phonetic cues to word boundaries and listeners would need other information to compensate for the potential ambiguity caused by the lack of reliable word-boundary cues.

Models of word segmentation also acknowledge that the signal-based, acoustic-phonetic cues alone cannot guide word segmentation and have aimed at identifying the sources of knowledge-based cues that influence word boundary perception. Knowledge-based cues that have been shown to affect word boundary perception include generalizations over the lexicon such as phonotactics and metrical patterns (Norris, McQueen, Cutler, & Butterfield, 1997; McQueen, 1998), distal prosody (Dilley & McAuley, 2008), and higher-order knowledge such as syntactic, semantic, and contextual information (Mattys, White, & Melhorn, 2005; Mattys & Melhorn, 2007), among others. More importantly, recent studies have suggested that acoustic-phonetic and knowledge-based cues to word boundaries affect listeners' segmentation in combination and that the two types of cues are combined by a compensatory mechanism: the degree to which listeners rely one type of information depends on the availability and strength of the other types of cue (Mattys, 2004; Mattys, White, & Melhorn, 2005; Mattys & Melhorn, 2007).

For instance, Mattys et al. (2005) examined how acoustic cues and contextual cues to word boundaries are integrated by systematically matching different types of

segmentation cues against each other and comparing the cross-modal priming effects from lexical decision or word detection tasks. They had listeners perform a lexical decision task by responding to either *cremate* or *mate* after hearing "An alternative to traditional burial is to cremate the dead." The speaker deliberately produced a pause between the two syllables of the target word (e.g., *cremate*) such that the initial phoneme of the second syllable (e.g., /m/ in *cremate*) contained cues that would lead listeners to favor interpreting the second syllable as a single word (e.g., *mate*). This syllable then replaced its corresponding token in the two-syllable target word. Although one might expect the acoustic cues to cause mis-segmentation (i.e., listeners would hear *mate* instead of *cremate*), significant priming of the two-syllable target word *cremate* was found, suggesting that strong contextual information, when available, may override acoustic-phonetic cues in word segmentation. Results from subsequent studies (Mattys & Melhorn, 2007; Mattys, Melhorn, & White, 2007) revealed that the robust effect of strong contextual information on listeners' word boundary perception is constrained by the strength of the acoustic-phonetic cues to word boundaries. Mattys and Melhorn (2007) had listeners listen to word-boundary minimal pairs such as *plum pie* and *plump eye* in isolation and in a sentential context favoring one of the two interpretations in order to measure the strength of the signal-based word segmentation cues and examine how signal-based and knowledge-based cues are combined to affect listeners' word boundary perception. They found that the effect of context on listeners' word boundary perception was larger when the signal-based cues were mild than when they were strong.

In summary, the literature on the production and perception of word boundaries demonstrated that listeners' word boundary perception is affected by signal-based cues produced by talkers as well as knowledge-based cues that listeners recruit to resolve potential acoustic ambiguity. With regards to the availability and informativeness of signal-based cues to word boundaries, researchers have not reached a consensus, with some suggesting that talkers produce systematic phonetic variation signaling word-boundary location (Lehiste, 1960; Hoard, 1966; Turk & Shattuck-Hufnagel, 2000) and others suggesting that the speech signal lacks consistent acoustic markers of word boundaries (Klatt, 1976; Jakimik & Cole, 1980; Christiansen, Allen & Seidenberg, 1998) or that some acoustic-phonetic variation conditioned by word-boundary location is not robust enough to affect listeners' perception of word boundaries (Nakatani & Dukes, 1977; Christie, 1974; Nakatani & Schaffer, 1978). Studies aimed at clarifying the nature of the mechanism by which various word segmentation cues are combined have shown that the relative importance of knowledge-based cues depends on the availability and strength of signal-based cues (Mattys & Melhorn, 2007; Mattys, Melhorn, & White, 2007). Thus, in order to better understand word segmentation, one goal of this dissertation is to clarify the extent to which talkers provide signal-based cues to word boundaries and the extent to which listeners' segmentation can be guided solely by the signal-based segmentation cues.

Another goal of this dissertation is to clarify whether there is variation in the degree to which talkers produce signal-based segmentation cues. Since spoken language displays a vast amount of acoustic variability within and across talkers (Klatt, 1980), it

would not be unreasonable to hypothesize that the degree to which talkers produce signal-based segmentation cues will also display variability within and across talkers. In addition, studies aimed at describing the phonetic properties at and around word boundaries have reported that not all talkers use the same set of strategies to demarcate word boundaries (Lehiste, 1960; Hoard, 1966; Anderson & Port 1994; White, 2002; Smith, 2004). For instance, Lehiste (1960) found that the glottalization of word-initial vowels may serve to indicate the presence of a word boundary, but not all talkers glottalize word-initial vowels, suggesting potential inter-talker variation in the use of allophonic cues to word boundaries. Dilley, Shattuck-Hufnagel and Ostendorf's (1996) study of glottalization also found that talkers were more likely to glottalize utterance-initial vowels than phrase-initial or word-initial vowels, suggesting that using glottalized vowels for word segmentation will result in under-segmentation—some word-onset vowels that are not glottalized may be more confusable with word-medial or -final vowels than word-onset vowels that are glottalized. The example of glottalization at word boundaries suggests that the phonetics of word boundaries may display intra-talker variability as well as inter-talker variability and that the variability in the phonetics of word boundaries may influence listeners' perception of word boundaries.

Studying the variability of the phonetics of word-boundary cues is expected to inform models of word segmentation. Previous studies focusing on segmentation cue integration have typically used stimuli created by orthogonally manipulating the availability of signal-based and knowledge-based word segmentation cues from a single talker's speech (Mattys & Melhorn, 2007; Mattys, Melhorn, & White, 2007, among

others). The used of a single talker's speech was necessary to clarify cue integration processes. However, using one talker's speech somewhat simplifies the word segmentation problem by reducing acoustic-phonetic variability across talkers. In addition, individual talkers may differ with regards to intra-talker variability and the default and range of clarity with which they produce speech. The current study will examine intra-talker variation in the degree to which talkers produce signal-based segmentation cues across a range of different phonological and conversational contexts. The results will clarify the range of intra-talker variability in word-boundary productions, rendering a better estimate of the word segmentation problem that human listeners encounter when they process speech, by clarifying the extent to which word segmentation can be guided by signal-based cues.

In summary, despite extensive studies on the production and perception of word boundaries, it is still not entirely clear to what extent signal-based cues are produced by talkers, to what extent signal-based segmentation cues affect listeners' word boundary perception, and whether and to what extent the availability and informativeness of signal-based segmentation cues vary within and across talkers. Most importantly, previous studies have reported variable results with regards to the production, availability, and informativeness of signal-based segmentation cues. The apparent inconsistencies may be due to acoustic-phonetic variation within and across talkers. However, very little is known about variation in the production of word-boundary cues. This dissertation examines the production and perception of word boundaries, focusing on variability in the phonetics of word boundaries and clarifying the inconsistencies associated with the

phonetics of word boundaries. In particular, I focus on three factors that might have contributed to the inconsistencies in the literature: speech clarity, phonetic context surrounding the word boundaries, and the perceptual consequences of fine-grained acoustic variation induced by word boundaries.

Speech Clarity

One factor that may contribute to variation in the production of signal-based cues to word boundaries is speech clarity. It has been well established in the literature that talkers spend more articulatory energy and speak more clearly when speech clarity is demanded by communicative situations such as background noise (Lane & Tranel, 1971; Junqua, 1996; Summers, Pisoni, Bernacki, Pedlow, & Stokes, 1988; Lau, 2008) or listeners whose linguistic competence is suboptimal (Ferguson, 1975; Stern, Spieker, & McKain, 1982; Fernald & Simon, 1984; Sikveland, 2006, among others). Given that language is a means of communication, it makes intuitive sense that talkers need and have the ability to adjust the level of clarity with which they produce speech so that their listeners can understand them.

Talkers' ability to accommodate to listener needs and vary the ways in which they speak is best exemplified by the studies of clear speech. Clear speech refers to a goal-oriented speaking style aimed at facilitating speech comprehension and it can be elicited by instructing speakers to speak clearly and precisely or to speak in a hypothetical situation where listeners have potential difficulty in comprehension (e.g., "read the materials as if you were talking to someone who is hearing impaired, or not a native

speaker of your language"; Picheny, Durlach, & Braida, 1985; 1986). Clear speech is found to have some distinctive phonetic properties distinguishing it from plain or conversational speech. Phonetic properties that distinguish clear speech from other types of speech include, but are not limited to, slower rate and longer duration of segments, more frequent and longer pauses, more intermediate prosodic boundaries in an utterance, louder intensity, increased range of fundamental frequency, and more extreme articulatory movements which result in an expanded vowel space or more frequent release of English word-final stops (Picheny et al. 1986, 1989; Perkell, Zandipour, Matthies, & Lane, 2002; Ferguson & Kewley-Port 2002; Bradlow, Kraus, & Hayes, 2003; Krause & Braida 2004; Smiljanić & Bradlow 2005; 2008). Though individuals may differ in the ways in which they employ these strategies in order to make their clear speech "clear," perception experiments, typically using the paradigm of word or sentence recognition in noise, reveal a clear speech intelligibility benefit for listeners (Picheny et al. 1986, 1989; Bent & Bradlow, 2002; Bradlow et al., 2003; reviewed in Uchanski, 2005 and Smilianić & Bradlow, 2008). Taken together, the literature on clear speech suggests that talkers are able to adjust their speech clarity and when they speak "clearly" they produce acoustic-phonetic cues assisting listeners in understanding spoken language.

Based on talkers' ability to adjust the degree to which they produce acoustic cues assisting speech perception, Lindblom (1990) proposed that speech production can be conceptualized as talkers' maintaining a balance between speaking clearly to ensure listeners' successful speech comprehension and speaking unclearly to minimize the physical energy expended to produce spoken language. As a result, the H&H

(hypoarticulation and hyperarticulation) theory predicts that talkers produce speech with greater clarity when listeners are expected to experience difficulty in speech comprehension than no listener difficulty is expected.

One goal of the current project is to explore whether and to what extent talkers produce signal-based cues to word-boundaries across different speaking styles. Given that talkers are able to clarify their message by talking clearly and given that word segmentation is an integral part of spoken language processing, it seems intuitive to hypothesize that talkers would provide more signal-based segmentation cues when producing clear speech. However, it remains an empirical question whether and to what extent talkers vary the degree to which they produce acoustic cues to word boundaries as a function of speaking style. It is difficult to generate more specific hypotheses regarding this empirical question based on the H&H theory, since the theory does not prescribe the types of specific listener difficulty that talkers accommodate or what precise acoustic cues talkers provide to resolve such ambiguities. One interpretation of H&H theory would predict that talkers produce more signal-based segmentation cues if listeners need them, but a more conservative interpretation of H&H theory might generate a null hypothesis, at least without an estimate of the physical energy necessary to clearly demarcate word boundaries.

Only a few previous studies have examined the clarity of word boundary cues across speaking styles. In the clear speech literature, syllable, word, or sentence recognition tests in noise have frequently been used to measure intelligibility. In the case of sentence recognition tests in noise, the number of keywords that were accurately

transcribed by the listeners has typically been used as a measure of intelligibility (Van Engen & Bradlow, 2007; Bradlow & Bent, 2008, among others). Thus, listeners were required to segment speech into words for the sentence recognition tasks. However, in most previous studies, it is rarely reported to what extent recognition errors were due to mis-segmentation, indicating that word segmentation per se has not been focused on very much. While earlier studies of word segmentation have suggested that the clarity of word boundaries may vary across the methods used to elicit speech (Lehiste, 1960; Hoard, 1966) and even reported that word-boundary minimal pairs produced in connected speech are less likely to contain signal-based segmentation cues than those produced with no sentential context or in carrier phrases (Barry, 1981), studies comparing the production and perception of word boundaries in clear speech and plain speech are quite rare.

Cutler and Butterfield (1990a, 1990b) might be the earliest investigation of clarity variation in signal-based segmentation cues. They instructed participants that their speech would be distorted and heard by listeners in the next room, who would type out what they heard. Indeed, all participants were speakers and no real listeners participated in the experiment. For each trial, speakers received feedback, which they believed to be the listeners' responses to the distorted speech. After the feedback, speakers read aloud the sentences again. Feedback sentences were carefully constructed so that they contained possible mis-segmentation of the critical word boundaries. Talkers produced greater pre-boundary lengthening of syllables and longer pauses between words in the speech produced after the feedback than in the speech produced before the feedback, suggesting

that the degree to which talkers provide signal-based segmentation cues depends on listener need, in this case, prompted by the "feedback."

In a later study, White, Wiget, Rauch, and Mattys (2010) had speakers produce near word-boundary minimal pairs such as *great anchor* or *gray tanker* once in a map-task and then in a reading task. Tokens generated from the two styles were played to listeners who rated how ambiguous each token sounded on a 9-point scale, where 1 and 9 indicated high certainty with one of the two alternatives and 5 indicated a high degree of ambiguity. Though they did not find reliable differences in the listeners' ambiguous rating score between spontaneously produced tokens and read tokens, they found that word boundaries in spontaneous speech became more ambiguous with repetition, while the degree of word boundary ambiguity in read speech tokens was less affected by repetition. Instead of a rating task, White, Mattys, and Wiget (2012) used cross-modal identity priming to visual lexical decision to measure listeners' segmentation. Overall, lexical decision latencies were faster for read tokens than spontaneously produced tokens, suggesting potential differences in the informativeness in the signal-based segmentation cues across speaking styles.

The current project is aimed at clarifying whether and how talkers adjust the availability and strength of signal-based segmentation cues in different speaking styles. Based on the results from the Cutler and Butterfield (1990a, 1990b) and White et al. (2010, 2012) studies, it would be reasonable to hypothesize that speech clarity would have influences on the realization of a single acoustic parameter (i.e., duration of segments, number of pauses) and listeners' segmentation accuracy. However, previous

studies have rarely examined both production and perception. This dissertation will explore the effects of speech clarity on the talkers' production and listeners' perception of word boundaries through an acoustic analysis of speech elicited in a production study and the results of a perception study using the tokens generated from the production study as stimuli.

A novel contribution of this dissertation will be clarifying the extent to which the availability and informativeness of signal-based segmentation cues vary within and across talkers. Again, previous studies failed in reaching consensus regarding the presence of acoustic-phonetic cues to word boundaries. If the degree to which talkers produce signal-based cues to word boundaries varies depending on speech clarity, the apparent discrepancies in the literature regarding the presence of signal-based segmentation cues may be accounted for, in part, by the variation due to speech clarity. In other words, it may be the case that signal-based segmentation cues demarcate word boundaries in clear speech but not in plain speech.

Eliciting speech varying in clarity can be achieved in a number of ways. As summarized above, researchers have frequently used the methodology of instructing talkers to speak as if they are in a hypothetical communication situation that involves listeners with a specific linguistic background. Some recent studies used real interlocutors (see Scarborough et al., 2007 for a direct comparison of the degree to which speech clarity is modulated when talking to imaginary versus real listener confederates; White et al., 2010) and confirmed that talkers modulated speech clarity depending on listeners' linguistic profile, regardless of whether the interlocutors were present or imaginary.

Enhancement of speech clarity to the overall communicative situation is called "global" hyperarticulation since its effect is typically observed throughout the task (Oviatt, Levow, Moreton, & MacEachern, 1998a; 1998b).

In contrast to enhancing overall clarity by global hyperarticulation, talkers may enhance speech clarity of a highly localized region in the speech signal, which is called "focal" hyperarticulation. Oviatt et al. (1998a) observed that in computer-directed speech, talkers produce focal hyperarticulation of segments and syllables if they are given feedback regarding the specific types of failure that the recognition system experiences. Such feedback serves as a cue to talkers so that they can repair their pronunciation and provide specific cues to prevent specific recognition errors. Based on these findings and the findings by Cutler and Butterfield (1990a, 1990b), one may hypothesize that exposing word-boundary minimal pairs to talkers would lead them to produce focally hyperarticulated, signal-based cues to word boundaries.

In other words, hyperarticulation or clarity enhancement is indeed an umbrella term that can be used to refer to various adaptation strategies in speech production, including talking clearly to enhance overall intelligibility of speech for listeners (global hyperarticulation) and talking clearly to prevent specific types of confusion or misperception (focal hyperarticulation). Smiljanić & Bradlow (2005) drew a similar distinction between global hyperarticulation for the purpose of enhancing overall intelligibility of speech and local phonological enhancement for the purpose of enhancing meaningful distinctions between some linguistic elements, such as phonological contrasts.

Previous studies have identified multiple sources of hyperarticulation including listener need (Picheny et al. 1986, 1989; Ferguson & Kewley-Port 2002; Bradlow, Kraus, & Hayes, 2003; Krause & Braida 2004; Smiljanić & Bradlow 2005), prosodic phrasing (Fougeron & Keating, 1997; Fougeron, 2001; Cho & Keating, 2001), and lexical stress patterns (de Jong, 2004), among others. Cho, Lee, and Kim (2011) examined whether the differences in the sources of hyperarticulation cause different hyperarticulation strategies by comparing the acoustic realization of Korean stops and vowels produced in clear and casual speech and at different prosodic positions. Cho et al. (2011) showed that hyperarticulation driven by communicative needs and hyperarticulation driven by prosodic prominence did not always result in similar acoustic realizations, suggesting that different types of hyper-articulation are encoded separately in speech production. In addition, Cho et al. (2011) reported an interactive effect of communicatively-driven and prosodically-driven hyperarticulation on speech production by showing that the degree to which talkers lengthened IP-final vowels as compared to IP-medial vowels was greater in clear speech than in casual speech. Similar interactions were found by Baese-Berk and Goldrick (2009). They found that talkers produced longer voice onset times for voiceless word-initial stops for words that have a minimal pair neighbor distinguished by voicing of the word-initial stop (e.g., /k/ in *cod*, which has a minimal pair neighbor *god*) than for words that have no minimal pair neighbor (e.g., /k/ *cop* with no neighbor **gop*), confirming the effect of local hyperarticulation motivated by lexical neighborhoods. Baese-Berk and Goldrick (2009) also found that talkers produced even longer voice onset times when reading the target words that were visually presented with their minimal pair

neighbors (e.g., seeing *cod*, *god*, *yell* on the screen and reading *cod*; cf., seeing *cod*, *lamp*, *yell* on the screen and reading *cod*), suggesting that multiple sources of hyperarticulation—the presence of a minimal pair in the lexicon and the presence of a minimal pair in the visual context—may interact to affect speech production.

One goal of the current study is to examine the acoustic realization of hyperarticulation driven by listeners' linguistic profile and visual stimuli highlighting the word-boundary ambiguity. In other words, the current study tests whether talkers produce more signal-based segmentation cues depending on whom they direct their speech to and whether they are visually presented with sentences containing a word-boundary minimal pair. Talkers who produce speech towards a listeners whose speech comprehension is expected to be difficult (e.g., speaking to an older or hearing impaired listener; speaking to a non-native listener) and were presented with the visual stimuli illustrating the word-boundary ambiguity can be regarded as having a strong motivation to clarify word boundaries by signal-based segmentation cues. Productions generated in such a condition will shed light on what the upper limit is for the extent to which talkers can produce acoustic-phonetic cues to word boundaries.

Phonetic context surrounding the word boundaries

In real-life communication, potential word boundaries occur in diverse phonetic contexts. Thus, a comprehensive understanding of signal-based segmentation cues should examine the availability and strength of signal-based segmentation cues across diverse phonetic environments.

Early descriptive studies focusing on the phonetics of word boundaries investigated whether spoken language contains consistent acoustic markers to word boundaries. Lehiste (1960) identified allophonic cues to word boundaries such as glottalization of word-initial vowels and aspiration of word-initial voiceless stop consonants and listeners were shown to attend to such allophonic cues for word segmentation (Christie, 1974; Nakatani & Dukes, 1997). However, as pointed out by Lehiste (1960) and Hoard (1966), not all segments display allophonic variation conditioned by word-boundary location. Focusing on the differences between segments that display allophonic variation conditioned by word boundaries and segments that do not, Nakatani and Dukes (1977) suggested that the degree to which segments convey information about the word boundary location depends on the types of segments that are at and surround the word boundaries. For instance nasals (e.g., underlined n in *no n**o**tion* and *know**n** ocean*) do not display allophonic variation depending on whether they begin a word or not, unlike, for instance, liquids such as /l/, which display allophonic variation between clear /l/and dark /l/, suggesting that some segments have may be more informative than other segments with regards to the word-boundary location. Relatedly, Smith and Hawkins (2000), in a word-spotting experiment, confirmed that allophonic variation of voiceless stops and liquids facilitated listeners' perception of word boundaries, suggesting that some segments differ in their inherent informativeness for word boundary locations.

In addition to the inherent informativeness due to the presence or absence of allophonic variation, segments surrounding the word boundaries may also affect the

informativeness of word boundaries. In a sequence such as *the sky*, the possible word-boundary misperception due to low-informativeness (i.e., lack of allophonic variation) of the word-final vowel in *the* and word-initial /s/ can be compensated by /k/, a voiceless stop that shows allophonic variation by aspiration. The example above illustrates that the availability of signal-based cues is affected by two factors: segments at the word juncture position and the neighboring segments. In this study, both factors are manipulated to examine the production and perception of signal-based segmentation cues and listeners' word segmentation.

To summarize, the production and perception of word boundaries is predicted to vary depending on two factors: speech clarity and phonetic context at and surrounding the word boundaries. The two sources of variation have been shown to interact with each other in affecting the phonetic realizations of speech, thus they may interact in affecting the phonetic realization of word boundaries. For instance, Lindblom (1990) suggested that clarity modulation does not target all linguistic materials equally, since segments differ in their inherent articulatory and phonetic properties. As a result, segments differ in the degree of acoustic-phonetic variability as well. An empirical study by Yuan and Liberman (2008) examined acoustic variability across multiple American English speakers and revealed that acoustic realizations of /ʊ/ and /ŋ/ were more variable across talkers as compared to other vowels and nasal segments. Based on these findings, one may infer that the range of acoustic variability for each phoneme would vary due to its inherent phonetic properties. Studies exploring the effects of clarity modulation across different segments have shown that conversation-to-clear speaking style adjustments

target different sounds to different degrees. For instance, phonetically long segments such as tense vowels are lengthened more than phonetically short lax vowels by clarity modulation (Moon & Lindblom, 1994; Uchanski, 1988 and Uchanski, Millier, Reed, & Braida, 1992; Smiljanić & Bradlow, 2008). These findings and Yuan and Liberman's (2008) results confirm that some segments possess more room to phonetically vary and suggest that clarity modulation is more likely to target the segments that are inherently more variable.

Though researchers have acknowledged the potential interaction between speaking style modulation and the inherent phonetic variability of linguistic materials, not a lot of studies have examined how talkers tailor signal-based word boundary cues as a function of phonetic properties of the juncture segments, phonetic contexts surrounding the word boundaries, and speech clarity. In one of the few studies focusing on the interaction between phonological context of the word boundaries and speech clarity, Cutler and Butterfield (1990a) compared talkers' productions of English word boundaries before a strong syllable and before a weak syllable. Cutler and her colleagues focused on the distinction between strong and weak syllables because most bisyllabic content words in English begin with a strong syllable. They hypothesized that English speakers would use the metrical stress pattern for word segmentation and confirmed that English-speaking listeners were more likely to parse a word boundary upon hearing a strong syllable (Cutler, Mehler, Norris, & Segui, 1986; Cutler & Norris, 1988). Cutler and Butterfield (1990a) showed that talkers lengthened pre-boundary syllables more and produced longer pauses before a word-initial weak syllable (i.e., before a less typical

word onset) than before a word-initial strong syllable (i.e., before a more typical word onset), suggesting that talkers compensated for biases based on this generalization over the lexicon by producing stronger acoustic cues clarifying word boundaries. Most importantly, they showed that the effects of phonological context and speaking style (i.e., repetition) had an interactive effect on the pre-boundary segment duration and pause duration at the boundary, suggesting that talkers may clarify certain word boundaries more than other word boundaries.

If talkers clarify word boundaries selectively, in the sense that some phonetic environments may undergo more clarity modulation than others, a complete model of word boundary production should predict whether and how different segments undergo asymmetric clarity modulation at and around a word boundary. This dissertation investigates whether and how the juncture segment and the phonetic context surrounding the juncture segment affect talkers' productions of acoustic cues to word boundaries and listeners' perception of word boundaries.

The mapping between speech production and perception

Since the focus of the current project is variation in the degree to which talkers produce signal-based cues to word boundaries, it is necessary to estimate the availability of these signal-based cues. How do we know whether talkers produce or do not produce signal-based cues to word boundaries? The methodological question lies in defining "produce signal-based cues" and determining what constitutes valid ways of assessing the availability of signal-based cues to word boundaries.

Previous studies have typically used two methods to assess whether and to what extent talkers produce signal-based segmentation cues. The first is to examine whether talkers produce different acoustic patterns depending on whether a segment begins or does not begin a word. The second is to examine whether the signal-based cues are perceptible and used by human listeners for word segmentation.

It is well established that listeners may not exploit all acoustic-phonetic variation that talkers produce. For instance, Klatt (1976), based on acoustic analyses comparing word-initial, word-medial, and word-final segments as well as perception studies exploring the just-noticeable difference (JND) in segment duration, suggested that duration variation induced by word boundaries is typically below the JND level. Similarly, Kim, Stephens, and Pitt (2012) reported that talkers performing a recall-read-recall task, where they memorized two sentence fragments, combined the fragments and produced the complete sentence, do in fact produce acoustic differences between a single schwa-initial word (e.g., *along*) and a two-word phrase (e.g., *a long*). However, listeners failed in distinguishing them, indicating that the statistically significant differences across the two parses should not be interpreted as talkers producing acoustic cues to word boundary location that can be perceived and used by listeners for word segmentation.

To summarize, both Klatt (1976) and Kim et al. (2012) suggest that statistically significant differences in the acoustic measures depending on the word boundary location should be interpreted cautiously. For these reasons, the current study is composed of a production and a series of perception experiments to form a comprehensive understanding of how signal-based segmentation cues are produced and perceived.

Summary and overview of the dissertation

The central question of this dissertation is to clarify the production and perception of signal-based cues for word segmentation across speaking styles and phonetic contexts. In particular, this dissertation focuses on the impacts of speaking style modulation due to listeners' linguistic profile, speaking style modulation due to talkers' awareness of a specific ambiguity that needs to be resolved, and the phonetic context of word boundaries on the production of signal-based segmentation cues. Effects of these factors on speech production are examined by acoustic analyses as well as perception experiments using the productions as stimuli, in order to test whether signal-based cues to word boundaries influence listeners' perception of word boundaries.

The rest of the dissertation is organized as follows. Chapter 2 focuses on the talkers' production of acoustic cues to word boundaries. Statistical analyses will be used to examine the extent to which talkers provide signal-based cues to word boundaries. Chapter 3 focuses the listeners' word segmentation of the talkers' productions generated from Chapter 2. Listeners' segmentation accuracy will be used as a way to evaluate the degree to which talkers produced cues to word boundaries. Chapter 4 summarizes the findings and discusses the broader implications of the results with regards to speech production, perception, and word segmentation.

Chapter 2: Experiment 1: Production of acoustic cues to word boundaries

2.1. Introduction

Spoken utterances are typically composed of multiple words, requiring listeners to segment them into a sequence of words (Aslin, Woodward, LaMendola, & Bever, 1996). The prevalence of the word segmentation problem and the apparent ease with which human listeners segment words have led researchers to investigate whether and to what extent talkers produce acoustic cues to word boundaries. Despite extensive research focused on the phonetics of word boundaries, no consensus has been reached concerning whether and to what extent talkers produce acoustic-phonetic cues to word boundaries, with some suggesting that talkers produce systematic phonetic variation signaling word-boundary location (Lehiste, 1960; Hoard, 1966; Turk & Shattuck-Hufnagel, 2000) and others suggesting that the speech signal lacks consistent acoustic markers corresponding to word boundaries (Klatt, 1976; Jakimik & Cole, 1980; Christiansen, Allen & Seidenberg, 1998). The purpose of the current study is to resolve inconsistencies among studies on word segmentation, which have reported that talkers produce acoustic-phonetic markers at word boundaries (Lehiste, 1960; Hoard, 1966; Smith, 2004) and the studies suggesting that signal-based segmentation cues are insufficient to guarantee successful word segmentation (Davis, 2002; Cutler, 1987; Kim, Stephens, & Pitt, 2012).

The current study suggests that the conflicting results in the literature may be reconciled by considering the acoustic-phonetic variation that spoken language displays.

Acoustic-phonetic properties of spoken language display a high degree of variation (Klatt, 1980), and consequently, acoustic-phonetic cues to word boundaries could also be susceptible to a similar amount of variation. Variation in the acoustics of speech sounds results from a number of sources, such as individual (e.g., physiological and social) differences across talkers, diversity in the communicative situations, and linguistic properties of the message. Given that these factors contribute to the acoustic-phonetic variability, it is reasonable to hypothesize that they—between-talker differences, diverse communicative situations, and linguistic factors—could cause the acoustic-phonetic word-segmentation cues to vary as well. The current study is aimed at clarifying whether and how factors contributing to acoustic-phonetic variability affect talkers' production of signal-based cues to word boundaries.

The issue of variation has yet been in the focus of the word segmentation literature. However, it has been suggested or mentioned in passing that there might be individual differences among talkers, speaking styles, and phonetic contexts in the extent to which talkers produce acoustic-phonetic cues to word boundaries (Lehiste, 1960; Cooper & Paccia-Cooper, 1980; Quené, 1992, Smith & Hawkins, 2012; among others). Some talkers were found to produce more salient word-boundary cues such as glottalized word-initial vowels than others (Dilley, Shattuck-Hufnagel, & Ostendorf, 1996). In addition, listeners were found to be more accurate at segmenting speech produced by slower talkers than faster talkers (Schwab, Miller, Grosjean, & Mondini, 2008). These findings suggest that talkers differ in the extent to which they produce signal-based cues

to word boundaries, suggesting that an accurate picture of variation requires examining speech produced by multiple talkers.

The production of acoustic-phonetic segmentation cues may also vary even within a single talkers' speech. Extra-linguistic (i.e., stylistic or communicative; Bell, 1984) as well as linguistic factors have been hypothesized to cause variability in the signal-based cues to word boundaries. An extra-linguistic source of acoustic-phonetic variation within a single talker' production is speech clarity. It is well established that talkers accommodate to listener needs and tailor their speech clarity accordingly, resulting in intra-talker variability in the acoustic-phonetic realizations of speech sounds (Picheny, Durlach, & Braida, 1985; 1986; Lindblom, 1990). Given that talkers are able to choose to talk more or less clearly and that word segmentation is an integral part of speech comprehension, an individual talker could choose to produce more or less signal-based cues to word boundaries, for instance, by increasing the availability of signal-based segmentation cues when they produce clear speech than plain speech. Preliminary experimental evidence supports this hypothesis, for listeners' segmentation accuracy varied among the tokens spoken by a single talker depending on speaking rate (Barry, 1981). As well, talkers were found to produce greater pre-boundary lengthening of syllables and longer pauses between words as they were instructed to talk through signal distortion and to speak clearly to prevent listeners' missegmentation (e.g., *interviewer* for intended *in to view her*; Cutler & Butterfield, 1990a; 1990b). These findings reinforce the hypothesis that acoustic-phonetic cues to word boundaries would display intra-talker variability due to speech clarity.

Speech clarity is influenced by various factors, including listeners' linguistic background (Ferguson, 1975; Stern, Spieker, & McKain, 1982; Fernald & Simon, 1984; Sikveland, 2006) and talkers' awareness of specific type of ambiguity that they have to resolve for listeners (Oviatt, Levow, Moreton, & MacEachern, 1998; for the distinction between global versus focal hyperarticulation). One factor that affects speech clarity is the listener's listening ability in the language. The current study tested whether and to what extent talkers produce acoustic-phonetic word segmentation cues depending on the listeners' competence in the language. There is a consensus that speech production is influenced by generic listener need such as listeners' linguistic background and listeners' linguistic background has been used to elicit a "clear" speech aimed at facilitating speech comprehension (Picheny, Durlach, & Braida, 1985; 1986; Smilianić & Bradlow, 2009). Based on the findings of clear speech, it is hypothesized that talkers would produce more signal-based segmentation cues when talking to listeners whose listening ability is suboptimal.

While listeners' ability in the language comprehension yields a straightforward hypothesis regarding talkers' production of acoustic-phonetic cues to word boundaries, another factor that may affect speech clarity, awareness, is yet to be fully understood as to whether and to what extent it affects the production of signal-based cues to word boundaries. Studies focusing on talkers' awareness and speech production have typically focused on the production of words (Baese-Berk & Goldrick, 2009) or syntactically ambiguous sentences (Lehiste, 1976; Allbritton, McKoon, & Ratcliff, 1996; Snedeker & Trueswell, 2003, Schafer et al., 2000; Kraljic & Brennan, 2005; Speer, Warren, &

Schafer, 2011), leaving it an empirical question as to whether awareness has impact on the production of word boundaries. The current study fills this gap in the literature by manipulating talkers' awareness and examining its effects on the production of word boundaries. In addition, manipulating speech clarity by two factors (i.e., listeners' linguistic background and talkers' awareness of the ambiguity), acknowledging the multi-faceted nature of speech clarity, allows to examine how diverse motivations for clarity enhancement affect acoustic realizations of speech independently or in combination (Cho, Lee, & Kim, 2011; Baese-Berk & Goldrick, 2009).

An additional source of the apparent discrepancies regarding the production of signal-based segmentation cues is the phonetic context at and around word boundaries. While an investigation into the word-boundary ambiguity occurring between schwa and a following consonant (e.g., *along* vs. *a long*) suggested that talkers do not produce sufficient cues to word boundaries (Kim et al., 2012), studies focusing on phonetic environments other than the schwa-consonant sequences have identified some acoustic-phonetic properties corresponding to word boundaries (Lehiste, 1960; Hoard, 1966). Such discrepancy suggests that the production of signal-based cues to word boundaries is constrained by the phonetic context of the word boundaries, such as phonetic properties of segments at or around word boundaries. For instance, some segments display allophonic variation conditioned by the word-boundaries, such as aspiration of word-initial stop consonants. Such allophonic variation may serve as a salient word-boundary cue for listeners, while, for instance, nasals or fricatives do not display such allophonic variation (Nakatani & Dukes, 1977), suggesting that the phonetics of the segment at the

word boundary is an important source of variability in the availability of signal-based cues to word boundaries. In addition to the phonetic properties of segments of which the phonetic realization varies depending on whether they end or begin a word, phonetic context around potential word boundaries contribute to the informativeness of word boundaries. For instance, a word boundary "minimal pair" such as *this cars* vs. *the scar* can be disambiguated by allophonic variation, because /k/ is aspirated word-initially but not aspirated after a word-initial /s/. In contrast, /k/ occurring before a vowel, as in *make art* vs. *may cart*, may not be disambiguated by aspiration but can be disambiguated by optional glottalization of word-initial vowel in *make art*. After schwa (e.g., *acute* vs. *a cute*), neither aspiration nor optional glottalization of word-initial vowels resolves the word-boundary ambiguity, suggesting the informativeness of phonetic context in carrying word-boundary information depends on the phonetic properties of the segment occurring at the word juncture as well as by the phonetic context around the potential word boundaries.

To recapitulate, the current study is an empirical test of whether the apparent inconsistencies regarding the availability of signal-based cues to word boundaries can be understood as resulting from talker variability. Results are expected to provide a detailed picture of the production of word boundaries, and in particular, how factors conditioning talker variability in general affect the acoustic-phonetic realizations of signal-based cues to word boundaries. Primary contributors of talker variability, speech clarity and phonetic context at and around word boundaries, were systematically manipulated to investigate their influences on the production of word boundaries. To examine the influence of

phonetic context, productions of sentences containing word-boundary ambiguities occurring at diverse phonetic contexts were elicited. To examine how speech clarity affects the production of word boundaries, the materials were produced by talkers, who directed their speech to listeners with different linguistic backgrounds to encourage or discourage hyperarticulation. Additionally, talkers were exposed to different visual prompts where one set of prompts highlighted the word-boundary ambiguity while the other set did not to encourage the hyperarticulation of word boundaries.

If the lack of consensus concerning the availability of acoustic-phonetic cues to word boundaries can be accounted for by intra-talker variability, the extent to which talkers produce signal-based segmentation cues is predicted to differ as a function of speech clarity and phonetic context. A comparison across studies suggests that speech clarity should affect talkers' production of acoustic-phonetic cues to word boundaries. Listeners have been shown to be highly accurate in segmenting stimuli generated from a reading task (Lehiste, 1960; Hoard, 1966; Barry, 1981, among others), while listeners' segmentation accuracy was poor for stimuli generated from a task where talkers produced speech under memory load, which led talkers to speak less clearly (Kim et al., 2012; Brink, Wright, & Pisoni, 1998; Harnsberger & Pisoni, 1999). Phonetic context is also predicted to affect the extent to which talkers produce acoustic-phonetic cues to word boundaries, with sequences containing phonetic contexts conditioning allophonic variation (e.g., aspiration of word-initial stops or optional glottalization of word-initial vowels) containing more cues than sequences that do not contain such variation (Nakatani & Dukes, 1977; Smith & Hawkins, 2000). Speech clarity and phonetic context

may interact to affect the production of acoustic-phonetic cues to word boundaries, given that the extent to which speech clarity affects acoustics-phonetic realizations of segments depends on their inherent articulatory and phonetic properties (Lindblom, 1990).

2.2. Methods

Participants

40 participants (20 male and 20 female), who were recruited from the linguistics department and psychology department participant pools, participated for course credit. None of them reported speech or hearing related difficulties. Participants were randomly assigned to the unaware group or aware group. 20 talkers (10 male and 10 female) participated in each group. Each participant completed three 35-minute testing blocks separated by mandatory breaks to minimize fatigue.

Stimulus Materials

36 phrases containing word-boundary ambiguities were constructed. The sentences were created by modifying stimulus materials used by Lehiste (1960), Smith (2004), and Kim et al. (2012) and by creating some new sentences. The sequences can be categorized into three groups—/s/ + consonant (e.g., *collects # gulls* vs. *collect # skulls*), consonant + vowel (e.g., *beef # eater* vs. *bee # feeder*), and schwa + consonant (e.g., *along* vs. *a # long*). The sequences were selected from studies reporting the presence of acoustic-phonetic cues to word boundaries (/s/ + consonant sequences from Smith, 2004; consonant + vowel sequences from Lehiste, 1960) and Kim et al. (2012), who reported

that acoustic-phonetic cues to word boundaries were rarely produced by talker (schwa + consonant sequences). The three CV types also differ in whether they condition an allophonic variation that may provide cues to word boundaries and the kind of allophonic variation which they condition. For instance, after /s/, obstruent consonants display allophonic variation by the location of word boundaries, because word-medial obstruents are unaspirated, while word-initial obstruents are aspirated. Before vowels, consonants do not display an allophonic variation that can resolve word-boundary ambiguities, but word-initial vowels can undergo an optional glottalization. Finally, after schwa, neither consonants nor neighboring vowels displays allophonic variation. In order to evaluate how the production of signal-based segmentation cues is affected by phonetic contexts, consonant type was manipulated along with the CV type. Half of the sequences in each group had obstruent consonants after /s/, before the vowel, or after schwa, and the other half had sonorant consonants, in order to enable a comparison of the informativeness of consonant conditions and CV sequences in carrying word-boundary information.

These phrases were embedded in short sentences that had 6-10 words. Typically, the sentence pairs started highly similarly, but later disambiguated. An example pair of sentences is as follows (ambiguous region is underlined): *The gentleman <u>collects gulls</u> at the beach* vs. *The gentlemen <u>collect skulls</u> in the attic*. A complete set of target sentences is listed in Appendix A. Four additional filler phrases containing word-boundary ambiguity were constructed so that they could be used as practice stimuli for perception experiments (Chapter 3).

Procedure

Talkers wore a head-mounted microphone and performed the task in a sound attenuated room. They were told that they were going to do a web-chatting experiment with three people (i.e., "listener" confederates). The talkers received instructions, which did not contain any descriptions about the listeners but that the talkers would talk with three people. The "listener" confederates were a young native English speaker, an older hearing impaired native English speaker, and a young nonnative speaker of English. The conditions were constructed so that talkers would generate plain lab speech when talking to the young native speaker of English and clear speech when talking to the hearing impaired and nonnative listeners. The experiment was blocked by "listener" condition, and thus had three blocks separated by breaks.

To prevent confounding effects of hypo-articulation caused by repeated production of the sentences and listeners' linguistic background (i.e. listener being a young native English speaker), every talker had the young native listener block, during which plain lab speech (cf., clear speech) production was expected as the first block, during which no repetition-induced hypoarticulation is expected. After completing the young native listener block, talkers had the two other listener blocks during which clear speech production was expected. The order of the two clear speech conditions was counterbalanced so that half of the talkers had the hearing impaired listener condition as their second block, while the other half had the nonnative listener condition as their second block, in order to prevent confounds associated with repetition and the listeners' ability in language comprehension.

In the beginning of each block, talkers were introduced to their "listener" by watching a short prerecorded video clip of their listener introducing himself. The listeners used pseudonyms to introduce themselves. The young native listener called himself John, the older hearing impaired listener called himself Ed, and the young nonnative listener called himself Alan. The introduction was constructed so it included 1) his pseudonym, 2) how old he is, 3) where he is originally from, and 4) a short personal history about the listener.

For instance, for the young native listener condition, we wanted the talkers to perceive the listener as highly similar to them with respect to linguistic background to encourage the production of hypo-speech. Thus, we had the "listener" confederate say that he was 20 years old, originally from Columbus, Ohio, and went to OSU, worked at a library on campus, and did this experiment for course credit. For the older hearing impaired listener condition, we had the confederate talk about his hearing impairment history and his hearing aids not being perfect though quite helpful. For the nonnative listener condition, we had the confederate say that he was originally from China, just moved to Ohio "this fall," and was doing this experiment for an ESL class requirement. Data were collected during the fall and winter quarters.

For each trial, two sentences were presented on the computer monitor. After 3.5 seconds, one of the sentences changed color, and the change in color was the talkers' prompt to read the sentence that changed color into the microphone. Talkers were told that their "listeners" would decide which of the two sentences was read by the talker and they would receive occasional feedback regarding how well their listeners were doing.

For the participants who were in the unaware group, the two sentences presented during a single trial were highly dissimilar to each other, except for their length in characters. In contrast, participants who were in the aware group saw the two versions of the ambiguous sequences during a single trial. For example, the participants in the unaware group saw *The gentleman collects gulls at the beach* and *It took two hours to pick up all the trash*. The participants in the aware group saw *The gentleman collects gulls at the beach* and *The gentlemen collect skulls in the attic* on the screen and were then prompted to read one of the two sentences. Presenting the two sentences illustrating the ambiguity should cause talkers to be aware of the word-boundary ambiguity, which may then affect their production of speech in general or their production of the acoustic cues to word boundaries in particular. Talkers had six seconds to read aloud each sentence. There was a two-second interval between trials.

For both groups, there were no listeners in action during the experiment, but the feedback served as a reminder to maintain the speaking style that we wanted the talkers to adopt. For the young native listener block, feedback sentences were highly positive (e.g., "John is doing very well. Keep doing what you are doing."), to encourage hypoarticulation. For the other two blocks, the feedback sentences were neutral (e.g., "Alan is doing okay. Keep doing what you are doing.") or negative (e.g., "Alan is having trouble understanding you. Be sure to talk clearly" or "Alan missed the last trial. Remember he is a nonnative speaker of English."), to ensure that talkers did not regress to hypoarticulation. The feedback sentences for the older hearing impaired listener

condition and nonnative speaker condition were kept almost identical, except for the description about the listener.

After they completed the experiment, the talkers were interviewed. During the post-experiment interview, they were asked to give a number from 0% to 100%, indicating how well their listeners were doing and to rank order the listeners according to how well they understood the talkers. Participants were also asked whether they noticed the word-boundary ambiguity during the experiment and if they did, at what point of the experiment they noticed the ambiguity. Finally, talkers received debriefing information which included information about deception and their right to withdraw their data from analysis. None requested to withdraw their data from analysis.

Measurements and Analysis

Acoustic analysis was performed on the production data (20 talkers × 2 groups × 40 sequences (including fillers) × 2 word-boundary conditions × 3 listener conditions = 9,600 sound files; 20 talkers × 2 groups × 36 sequences × 2 word-boundary conditions × 3 listener conditions = 8,640 sound files excluding filler items). From each token, including filler items, duration of the entire utterance, duration of the ambiguous target sequence, RMS amplitude of the non-silent portions, and the range of fundamental frequency were measured in order to examine whether listener conditions led talkers to produce different speaking styles. This confirmatory test was necessary, since the procedure of the current study differs from those that have been widely used in previous studies in that we used prerecorded introductions of listener confederates. As pointed out

in the review of clear speech research by Smilianić & Bradlow (2009), research methods that have typically been used to elicit clear speech include giving talkers explicit instructions such as asking them to talk "clearly and precisely" or giving talkers descriptions about imaginary interlocutors, such as instructing the talkers to talk as if they were talking to someone who is hearing impaired and/or not a native speaker of their language (Picheny et al., 1985). Some recent studies used confederates as real interlocutors (see Scarborough et al., 2007 for a direct comparison of the degree to which speech clarity is modulated when talking to imaginary versus real listener confederates; White et al., 2010) and confirmed talkers' clarity modulation regardless of whether talkers were imaginary or real.

Given the robustness of clarity modulation reported in the literature, we predicted that typical phonetic properties found in clear speech—longer utterance duration, longer duration of the ambiguous sequence, greater RMS amplitude of non-silent portions, and larger dynamic ranges of f0 (Picheny et al. 1986; Bradlow et al. 2003; Krause & Braida 2004)—would be observed when speech was directed to an older hearing impaired or a nonnative listener than when speech was directed to a young native listener. The confirmatory test of the "listener" manipulation is also expected shed new light on the methods of eliciting clear speech.

The second set of analyses tested whether and to what extent the production of segmentation cues differed across the four conditions of interest: listener condition, awareness condition, CV type surrounding the word boundary, and whether the consonant at the word boundary was a sonorant or an obstruent. Among the acoustic

measures corresponding to word boundaries, we focused on the duration of segments at and around the ambiguous target word boundaries, because duration is a generic measure that can be obtained from all segments and that is strongly correlated with listeners' perception of word boundaries (Kim et al., 2012).

Availability of segmentation cues was estimated by the fit of a logistic regression model, which predicted whether the critical juncture segment (e.g., /s/ in *it*s *praise* or *it* *s**prays* is regarded as the critical juncture segment since the two phrases are distinguished by whether it begins or ends a word; in cases of schwa + consonant sequences, the consonant after schwa, /l/ in *a**l**ong* vs. *a* *l**ong*, was considered as the critical juncture segment) begins or does not begin a word (i.e., word-medial or word-final). The acoustic measures that served as predictor variables were as follows: the duration of the critical juncture segment and the duration of segments surrounding the critical juncture segments (e.g., duration of /t/, /s/, /p/ in *it*s *praise* or *it* *s**prays*, where the underlined /s/ is the critical juncture segment).

The logistic regression model categorizes the tokens based on their durational properties. If a token has durational properties that are distinct from the durational properties of its word boundary "minimal pair," the logistic regression model should accurately categorize that particular token. On the contrary, if a token lacks salient duration cues to word boundaries, it will not be accurately categorized by the logistic regression model with respect to the location of the intended word boundary.

The fitted model values indicate whether or not the model can accurately categorize the intended segmentation of a given sound file. Model accuracy was

aggregated over the tokens that were produced by the same talker in the same listener, awareness, CV type, and consonant conditions. Aggregated means were interpreted as the estimated availability of acoustic segmentation cues produced by speakers, where higher means indicated tokens having more salient word segmentation cues and lower means indicated the relative paucity of segmentation cues. Aggregated means of regression model accuracy were compared across conditions, in order to test whether the availability of acoustic segmentation cues differs depending on the listener condition, awareness condition, CV type, and consonant condition. For instance, a main effect of CV type on the mean of the regression model accuracy would suggest that the availability of durational cues to word boundaries differs depending on the CV type and that some CV combinations may have inherently clearer segmentation cues than others.

2.3. Results

2.3.1. Effect of listener condition on global clarity modulation

In order to confirm whether talkers generated speech with different clarity across listener conditions, phonetic properties that have been shown to depend on speech clarity—rate measured by entire utterance duration and duration of ambiguous target sequences, RMS amplitude of non-silent portions, and range of fundamental frequency--were compared across listener conditions.

Out of the 9,600 tokens, 426 tokens (4.44%) were excluded from the analysis of the entire utterance duration, because they were produced with errors. 110 tokens (1.15%) were also excluded from the analysis of the entire utterance duration, since it

took longer than the allotted recording time (6 seconds) for talkers to produce the sentence and thus it was impossible to accurately measure how long the talkers took to produce a sentence. Similarly, 216 tokens (2.25%) were excluded from the analysis of the ambiguous target sequence duration and from the analysis of fundamental frequency range, since the ambiguous target phrase in these tokens was produced with errors or included disfluencies.

A mixed-effects linear regression model tested whether the listener condition had a statistically significant effect on the acoustic measures. Statistical significance was evaluated using the pvals.fnc function of the languageR package (Baayen, 2008) in R. Talkers and items were treated as random effects and listener condition and block order were treated as fixed effects. Means and standard deviations (in parentheses) of the acoustic measures across listener conditions are presented in Table 1.

	Listener Condition		
Measurement	Young Native	Older Hearing Impaired	Young Nonnative
Entire utterance Duration	2567 ms (529 ms)	3155 ms (758 ms)	3203 ms (796 ms)
Ambiguous target sequence duration	576 ms (184 ms)	738 ms (279 ms)	738 ms (279 ms)
RMS amplitude of non-silent portions	68 dB (7 dB)	70 dB (6 dB)	70 dB (6 dB)
Range of fundamental frequency	55 Hz (47 Hz)	67 Hz (53 Hz)	68 Hz (52 Hz)

Table 1. Means and standard deviations (in parentheses) of the acoustic measures across listener conditions

As can be seen in Table 1, when they spoke to the older hearing impaired or the young nonnative listeners than the young native listener, talkers produced slower and louder speech with a larger fundamental frequency range. The fixed effects of listener condition on the four acoustic measurements were statistically significant at $\alpha = 0.05$, suggesting that the manipulation of listener condition affected listeners' speech clarity. A complete model output is provided in Appendix B. Although talkers showed a tendency to speak slowest during the third block (Means of the entire utterance duration for the second and the third blocks were 3069 ms (SD = 729 ms) and 3289 ms (SD = 807 ms), respectively; means of the ambiguous target sequence duration were 772 ms (SD = 254 ms) for the second block and 774 ms (SD = 297 ms) for the third block), the fixed effects

of block order were not robust enough to reach a statistical significance. Instead, block order and listener condition had interactive effects on the entire utterance duration and ambiguous target sequence duration. No fixed effects or interactions were found to be statistically significant of block reached statistical significance for the analyses of amplitude and fundamental frequency range.

Findings suggest that talkers produced speech that differed in speaking styles between the young native listener condition and the other two listener conditions, as indicated by the statistical differences in the acoustic properties—rate, amplitude, and range of fundamental frequency—between speech directed to a young native listener and those of the speech directed to an older hearing impaired or a young nonnative listener. More importantly, these results confirm that the elicitation method used in Experiment 1, namely introducing the "listener" confederates by a prerecorded video including information about their linguistic background, had talkers produce speech with different degrees of clarity.

2.3.2. Availability of segmentation cues

Duration of a segment systematically varies depending on its location at or around a word boundary: word-initial segments tend to be longer than word-medial or word-final segments (Lehiste, 1960; 1972; Oller, 1973; Harris & Umeda, 1974; Klatt, 1976; Beckman & Edwards, 1990; Turk & Shattuck-Hufnagel, 2000; Smith, 2004, among other). As a consequence, variation in segment duration can inform the extent to which talkers produce acoustic-phonetic cues to word boundaries. If talkers vary the extent to

which they produce acoustic-phonetic cues to word boundaries according to speech clarity and phonetic context, the extent to which durational cues predict the location of word boundaries should differ as a function of phonetic speech clarity and phonetic context. In order to estimate the extent to which durational cues predict the location of word boundaries, a series of logistic regression analyses were conducted, of which the independent variables were durational properties at and around word boundaries and the dependent variables was word-boundary location. Fitted accuracy of the logistic regression models were compared across speech clarity and phonetic context conditions in order to examined whether and to what extent to which speech clarity and phonetic context contributed to the variability in the production of signal-based cues to word boundaries. Two sets of logistic regression models were used to estimate the extent to which talkers produced durational cues to word boundaries.

The first model is highly conservative in that it predicts word boundary location solely based on the duration of a single segment at word juncture, which distinguishes the two phrases depending on whether the segment begins a word or ends a word for the /s/ + consonant and consonant + vowel conditions (e.g., /s/ in *it<u>s</u> praise* vs. *it <u>s</u>prays*) and whether the segment begins a word or not for the schwa + consonant condition (e.g., /l/ in *a<u>l</u>ong* vs. *a <u>l</u>ong*). If talkers consistently produce durational differences between word-initial versus non-initial segments, duration of the juncture segment should predict the location of a word boundary. If talkers vary the extent to which they produce durational cues to word boundaries as a function of speech clarity and phonetic context, means of

logistic regression model accuracy should differ across the speech clarity and phonetic context conditions.

A mixed-effects logistic regression model estimated the extent to which the duration of the juncture segment predicted the location of the word boundary as intended by the talker. Talkers and items were random effects. Results suggested that the duration of the critical segment is a statistically significant predictor of word boundary location ($\beta = 0.02$, $z = 30.39$, $p < 0.001$). Figure 1 shows the probability of the regression model to classify a given target segment as it begins a word, as a function of critical segment duration in milliseconds.

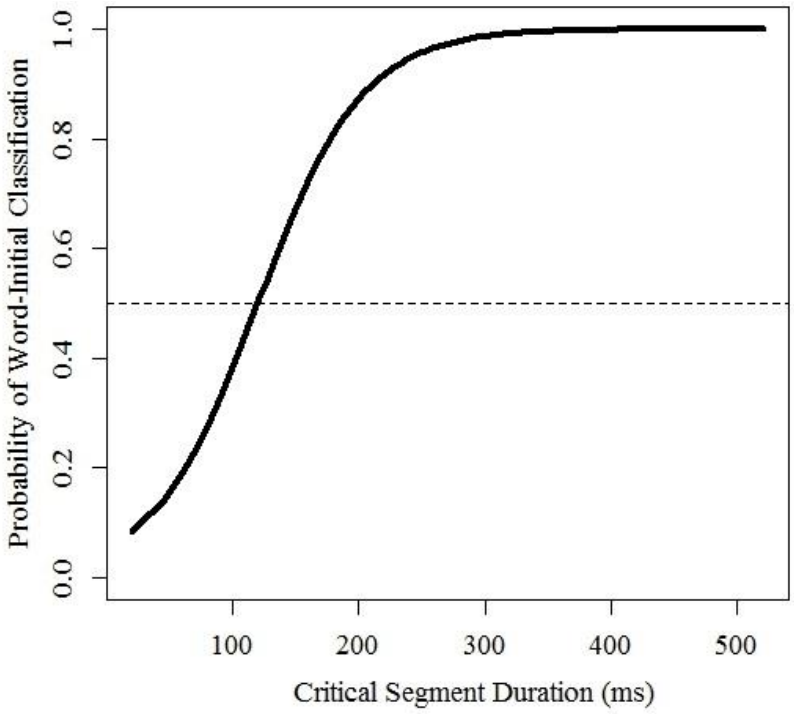

Figure 1. Probability of the regression model to classify a given target segment as beginning a word as a function of critical segment duration in milliseconds. The dashed line represents the chance level (0.5).

The fitted values of the regression model accurately categorized the intended segmentation of 66.88% of the tokens, suggesting that overall, model accuracy was higher than the chance level. In order to test whether model accuracy differs depending on speech clarity and phonetic context, model accuracy was aggregated over the tokens that were produced by the same talker and in the same listener, awareness, CV type, and consonant conditions. A series of repeated measures analyses of variance were performed on the average model accuracy. For the by-subject analysis, listener condition, CV type, and consonant type were treated as within-subject independent variables and awareness was a between-subject independent variable. For the by-item analysis, listener and awareness conditions were within-item independent variables, while CV and consonant type were between-item independent variables. Figure 2 below shows model accuracy that significantly differed across conditions.

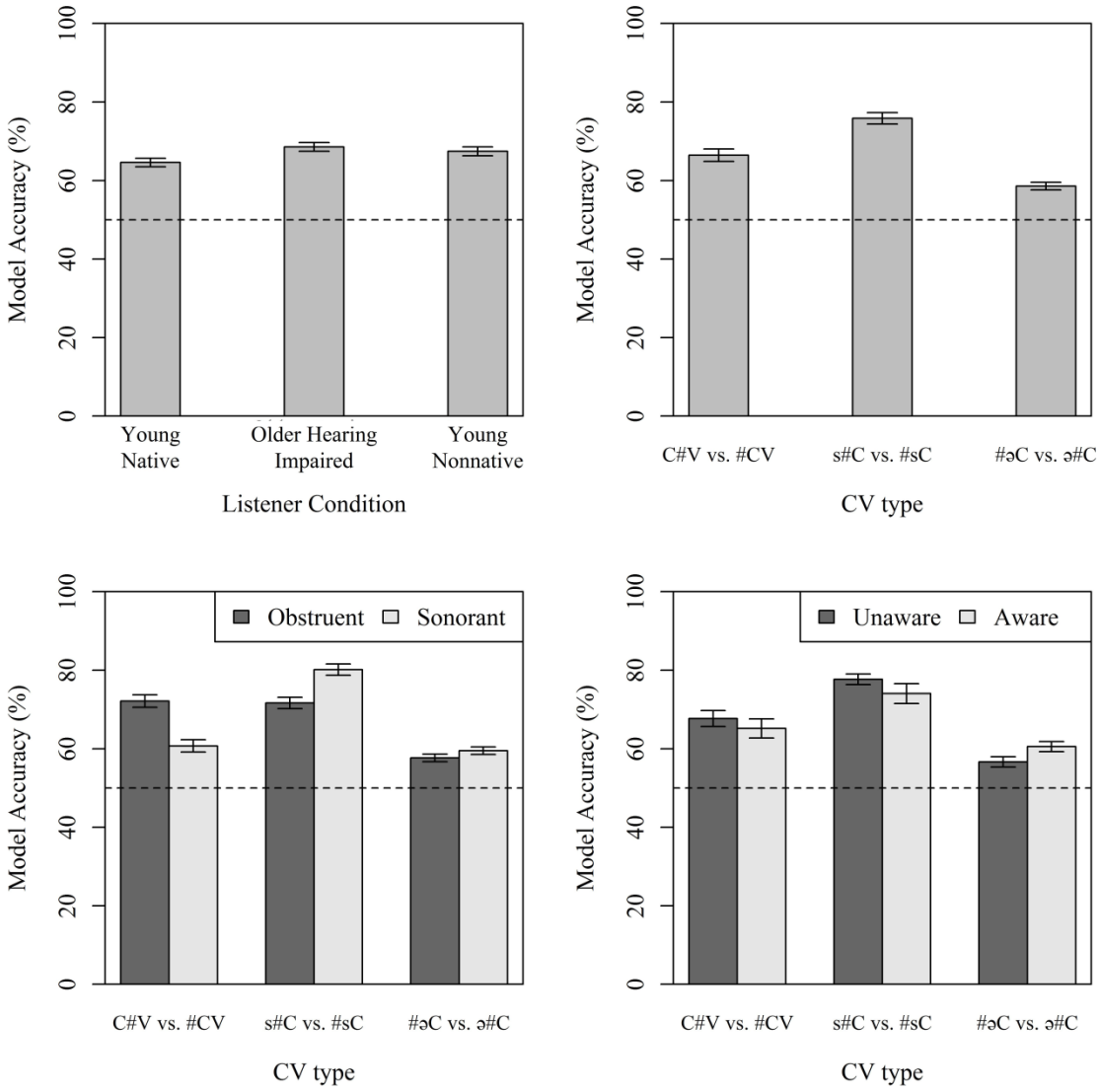

Figure 2. Model accuracy across listener conditions (top left), CV types (top right), CV type by consonant conditions (bottom left), and CV type by awareness conditions (bottom right). Error bars represent subject standard errors.

As can be seen from the top left panel of Figure 2, listener condition affected talkers' production of acoustic-phonetic cues to word boundaries. Model accuracy was higher for the tokens spoken to the older hearing impaired listener (68.56%) and the

nonnative listener (67.45%) than those spoken to the young native listener (64.61%), as predicted by research on clear speech (Picheny, Durlach, & Braida, 1985; 1986). The main effect of listener condition on model accuracy was statistically significant ($F1(2, 38) = 5.02$, $p < 0.01$; $F2(2, 34) = 9.54$, $p < 0.001$). The finding that talkers produced more signal-based segmentation cues as they spoke to listeners who would benefit from a higher degree of speech clarity suggests that speech clarity is a potential source of variability in the extent to which talkers produce signal-based segmentation cues. Post-hoc comparisons revealed statistically significant differences in the model accuracy between the young native and the older hearing impaired listener conditions ($t_{subj}(39) = -3.04$, $p < 0.01$; $t_{item}(35) = -3.46$, $p < 0.01$) and between the young native and the young nonnative listener conditions ($t_{subj}(39) = -2.22$, $p < 0.05$; $t_{item}(35) = -3.19$, $p < 0.01$). Between the two clear speech conditions, however, the difference in the model accuracy was not statistically significant ($t_{subj}(39) = 0.97$, *ns*; $t_{item}(35) = 1.33$, *ns*).

Shown in the top right panel of Figure 2 is model accuracy across CV types. Model accuracy differed across CV types ($F1(2, 38) = 48.04$, $p < 0.001$; $F2(2, 34) = 9.16$, $p < 0.001$), suggesting that talkers produced more segmentation cues for certain CV types than others and that stimulus choice could have resulted in the variability in the production of signal-based cues to word boundaries. The predictive model accurately classified 66.44% of the tokens containing a word boundary ambiguity between a consonant and a vowel (e.g., *beef eater* vs. *bee feeder*; juncture segments are underlined), 75.86% of the tokens containing a word-boundary ambiguity between /s/ and a consonant (e.g., *collects gulls* vs. *collect skulls*), and 58.49% of the tokens containing a word

boundary ambiguity after schwa (e.g., *a*l*ong* vs. *a* l*ong*). Post-hoc comparisons on the subject means revealed statistically significant differences between all CV types (t (39) = -5.63, $p < 0.001$ for the difference in model accuracy between the consonant + vowel and /s/ + consonant conditions; t (39) = 3.98, $p < 0.001$ for the difference between the consonant + vowel and schwa + consonant conditions; t (39) = 10.11, $p < 0.001$ for the difference between the /s/ + consonant and schwa + consonant conditions). In contrast, post-hoc comparisons on the item means revealed statistically significant differences in model accuracy only between the /s/ + consonant and the schwa + consonant conditions (t (22) = 5.12, $p < 0.001$), in part because the CV condition was a between-item variable and thus had a less statistical power in the items analysis than in the subject analysis.

The bottom left panel of Figure 2 shows the CV type by consonant type interaction on model accuracy, which was marginally significant ($F1$(2, 38) = 33.15, $p < 0.001$; $F2$(2, 34) = 3.13, $p = 0.06$). The CV type by consonant type interaction suggests that the extent to which talkers produced durational cues to word boundaries differed depending on the phonetic context at the word boundary (consonant type) as well as the phonetic context around the word boundaries (CV type). When the potential word boundary occurred between a consonant and a vowel (e.g., *beef eater* vs. *bee feeder*), model accuracy was higher for the tokens that contained an obstruent consonant (72.14%) than a sonorant consonant (60.72%). In contrast, for the two other CV types, model accuracy was higher if a sonorant consonant occurred at the potential word boundary than when an obstruent occurred at the potential word boundary (71.64% vs. 80.13% accuracy for the CV type where a potential word boundary occurs between /s/

and a consonant, 57.67% vs. 59.51% accuracy for the CV type where a potential word boundary occurs after schwa and a consonant). Post-hoc comparisons on the subject means revealed that the differences in the model accuracy between consonant types were reliable only in two of the three CV type conditions, where the potential word boundary was between a consonant and a vowel (t (39) = 6.14, $p < 0.001$) and where the potential word boundary was between /s/ and a consonant (t (39) = 4.93, $p < 0.001$). The item mean comparisons did not reveal any statistically significant differences.

Lastly, shown in the bottom right panel of Figure 2 is an interaction between talkers' awareness of the word-boundary ambiguity and CV type on model accuracy, which was significant only in the items analysis ($F1$(2, 38) = 2.72, $p = 0.07$; $F2$(2, 34) = 4.31, $p < 0.05$). The interaction between awareness and CV type on model accuracy suggests that talkers' production of signal-based segmentation cues is affected by speech clarity as well as phonetic context. The extent to which speech clarity affected the production of durational cues to word boundaries differed across CV types. To elaborate, talkers' awareness of the word-boundary ambiguity resulted in a significantly higher model accuracy only in the schwa-initial CV type (56.65% accurate for ambiguous schwa-initial tokens produced by talkers who were unaware of the word-boundary ambiguity vs. 60.54% for ambiguous schwa-initial tokens produced by talkers who were aware of the word-boundary ambiguity; t_{subj} (19) = -2.12, $p < 0.05$; t_{item} (11) = -4.59, $p < 0.001$). For the other CV types, talkers' awareness of the word-boundary ambiguity resulted in a statistically insignificantly lower model accuracy (mean unaware = 77.67% vs. mean aware = 74.04% for a potential word boundary between /s/ and a consonant;

mean unaware = 67.70% vs. mean aware = 65.18% for a potential word boundary between a consonant and a vowel). Similarly, the influence of awareness differed across phonetic contexts (i.e., CV type by consonant conditions). There was a marginally significant three-way interaction among awareness, CV type, and consonant condition, which was significant in the by-subject analysis but not in the by-item analysis (*F1*(2, 38) = 3.40, $p < 0.05$; *F2*(2, 34) =2.86, $p = 0.07$). Post-hoc comparisons revealed that awareness decreased model accuracy (75.00% vs. 68.25% accuracy for unaware and aware conditions, respectively) only when the potential boundary occurred between /s/ and a following obstruent (t_{subj} (38) = 2.15, $p < 0.05$; t_{item} (5) = 3.04, $p < 0.05$), contrary to the prediction that awareness would lead talkers to enhance the clarity of word boundaries. Post-hoc comparisons for the other conditions did not reveal any statistically significant differences.

A second model was constructed that was less restrictive than the first model, in the sense that its predictors included durations of segments surrounding the word boundary (e.g., duration of /t/, /s/, /p/ in *it<u>s p</u>raise* or *i<u>t sp</u>rays*), as well as optional pauses produced by the talkers. The methods used to perform the statistical analyses were identical to those used to perform analyses in the first model. Figure 3 shows main effects that were revealed by the ANOVAs on the average model accuracy.

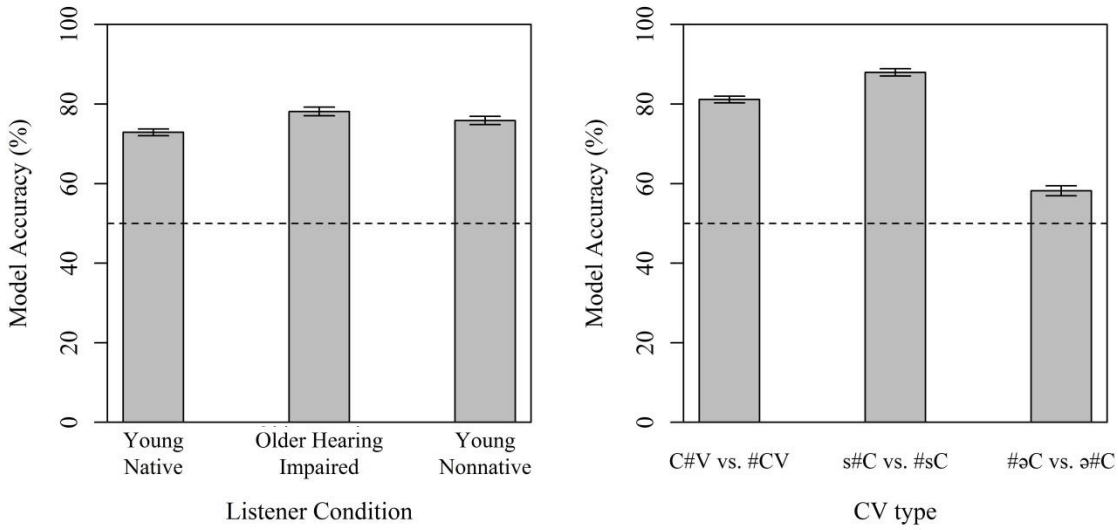

Figure 3. Less restrictive model accuracy across listener conditions (left) and CV types (right)

Shown in the left panel of Figure 3 is accuracy of the less restrictive model across listener conditions. As was found in the conservative analyses, listener condition had a main effect on model accuracy ($F1(2, 38) = 9.94$, $p < 0.001$; $F2(2, 34) = 10.79$, $p < 0.001$), suggesting that speech clarity is one source of variation in the extent to which talkers produce signal-based cues to word boundaries. Overall, the less restrictive model accurately categorized 75.64% of the tokens (cf., 66.88% accuracy for the conservative model). Model accuracy was lowest (72.89%) for the tokens spoken toward a young native listener and higher for the tokens spoken toward the two other listeners (cf., 78.13% and 75.87% for the tokens spoken to an older hearing impaired and a young nonnative listener, respectively). Post-hoc comparisons revealed statistically significant differences in model accuracy across all three listener conditions ($t_{subj}(39) = -4.09$, $p < 0.001$; $t_{item}(35) = -4.35$, $p < 0.001$ between the young native and the older hearing

impaired listener conditions; t_{subj} (39) = -2.43, $p < 0.05$; t_{item} (35) = -2.27, $p < 0.05$ between the young native and the young nonnative listener conditions; t_{subj} (39) = 2.16, $p < 0.05$; t_{item} (35) = 2.28, $p < 0.05$ between the older hearing impaired and the young nonnative listener conditions).

The right panel of Figure 3 shows the main effect of CV type on model accuracy (*F1*(2, 38) = 339.22, $p < 0.001$; *F2*(2, 34) = 42.34, $p < 0.001$), suggesting the availability of signal-based cues to word boundaries depends on the phonetic context around the word boundaries. Consistently to the findings from the conservative analyses, model accuracy was lowest for the ambiguous schwa-initial sequences (58.20%) and highest for word boundaries between /s/ and a consonant (87.97%; 81.14% accuracy for the boundary between a consonant and a vowel). Post-hoc pairwise comparisons revealed that the model accuracy for ambiguous schwa-initial sequences was significantly lower than the two other CV types (t_{subj} (39) = 21.89, $p < 0.001$; t_{item} (22) = 4.63, $p < 0.001$ between /s/-consonant and schwa-consonant conditions; t_{subj} (39) = 19.97, $p < 0.001$; t_{item} (22) = 4.37, $p < 0.001$ between consonant-vowel and schwa-consonant conditions). Difference in model accuracy between the consonant-vowel and the /s/-consonant conditions was significant only in the subject mean comparison (t_{subj} (39) = -6.00, $p < 0.001$; t_{item} (22) = 1.25, *ns*).

In addition to the main effects of listener condition and CV type, there was a significant interaction among listener condition, CV type, and consonant condition on the accuracy of the less restrictive model (*F1*(4, 34) = 5.84, $p < 0.001$; *F2*(4, 32) = 3.30, $p < 0.05$). This three-way interaction, which is shown in Figure 4, and the main effects of

listener condition and CV types suggest that speech clarity and phonetic context affects the production of signal-based segmentation cues independently and in combination.

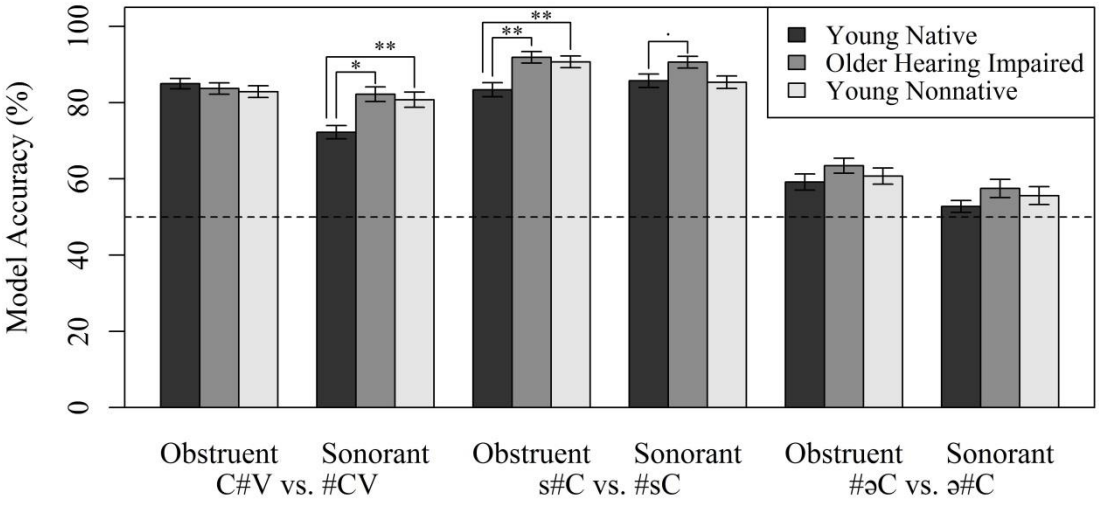

Figure 4. Model accuracy across listener condition, CV type, and consonant condition. ** indicates $p < 0.01$; * indicates $p < 0.05$; and . indicates $p = 0.05$ for the post-hoc comparisons

As can be seen in Figure 4, across most CV type by consonant conditions, model accuracy was highest for the older hearing impaired listener condition and lowest for the young native listener condition. However, the extent to which talkers enhanced the clarity of word boundaries across listener conditions depended on phonetic context (i.e., CV type and consonant condition), confirming and extending the finding that clarity modulation does not target all linguistic elements equally (Lindblom, 1990) to a new linguistic environment--phonetic contexts at and around potential word boundaries.

A series of post-hoc pairwise comparisons were performed to determine the phonetic contexts that displayed significant differences in model accuracy across listener

conditions. Between the young native and the older hearing impaired listener conditions, model accuracy differed in two of the six CV type by consonant conditions: when the potential word boundary occurred between a sonorant consonant and a vowel (t_{subj} (39) = -4.06, $p < 0.001$; t_{item} (5) = -3.80, $p < 0.05$) and between /s/ and an obstruent (t_{subj} (39) = -5.22, $p < 0.001$; t_{item} (5) = -5.99, $p < 0.01$). For the tokens where the potential word boundary occurred between /s/ and a sonorant, model accuracy difference between the young native and the older hearing impaired listener conditions was significant (t_{subj} (39) = -2.12, $p < 0.05$; t_{item} (5) = -2.57, $p = 0.05$). Likewise, when the word boundary ambiguity occurred between a sonorant consonant and a vowel (t_{subj} (39) = -3.70, $p < 0.001$; t_{item} (5) = -4.18, $p < 0.01$) and between /s/ and an obstruent (t_{subj} (39) = -3.94, $p < 0.001$; t_{item} (5) = -5.06, $p < 0.01$), model accuracy significantly differed between the young native and the young nonnative listener conditions. No differences in model accuracy between the older hearing impaired and the young nonnative listener conditions were revealed to be statistically significant by post-hoc comparisons.

For the schwa + consonant condition, regardless of whether the consonant following schwa was an obstruent or a sonorant, differences in the model accuracy among listener conditions failed in reaching statistical significance. This finding suggests that talkers were not only least likely to produce durational cues to word boundaries but also least likely to enhance the clarity of potential word boundaries that occurred between schwa and a consonant. The finding that ambiguous schwa-initial sequences can hardly be disambiguated by talkers' production of signal-based segmentation cues is consistent

with Kim et al. (2012) and suggests that phonetic context affects the extent to which talkers produce acoustic-phonetic cues to word boundaries.

Although model accuracy was, in general, higher for the clear speech conditions (i.e., the older hearing impaired and the young nonnative listener conditions) than the plain speech condition (i.e., the young native listener condition), the obstruent consonant + vowel (e.g., *beef eater* vs. *bee feeder*) condition was found to be an exception to the pattern. One possible account for this exception is variability among the items that belong to the phonetic context condition. For three out of the six ambiguous sequences that contained the word-boundary ambiguity between an obstruent consonant and a vowel, model accuracy was lower for the clear speech conditions (i.e., the older hearing impaired and the young nonnative listener conditions) than plain speech condition (i.e., the young native listener condition). Such sequences had fricatives or affricates as critical junctures segments (e.g., *beef eater* vs. *bee feeder*; *buys ink* vs. *buy zinc*; *launch air* vs. *lawn chair*) followed by a vowel. For the rest of the items, which had a stop consonant followed by a vowel, model accuracy was higher for the clear speech conditions than plain speech conditions, as predicted.

Finally, as was found in the conservative analyses, talkers' awareness of the word-boundary ambiguity did not have a significant effect on model accuracy. None of the interactions with awareness was statistically significant, either.

2.3.3. Inter-talker variation in the range of model accuracy across listener conditions

The main focus of the current study was to examine whether and to what extent speech clarity and phonetic context condition the variation in the signal-based segmentation cues. The design of the current study enables to investigate the variation in the extent to which talkers produced signal-based segmentation cues as a function of listener condition within a single talker. In addition, the large number of tokens produced by each of the 40 talkers enables to investigate the inter-variability in the extent to which talkers enhanced the clarity of acoustic-phonetic cues to word boundaries.

In order to examine the inter-talker variability in the extent to which talkers varied the availability of durational cues to word boundaries across listener conditions, model accuracy was averaged over the tokens produced in each of the three listener conditions by each of the 40 talkers. The range of intra-talker variation was computed for each talker, by subtracting the lowest listener condition mean of model accuracy from the highest listener condition mean of model accuracy. For 25 out of the 40 talkers (62.50% of the talkers), average model accuracy was lowest for the young native listener condition. For 24 out of the 40 talkers (60.00%), average model accuracy was highest for the older hearing impaired listener condition. Figure 5 shows the distribution of the ranges of listener condition variability for each talker.

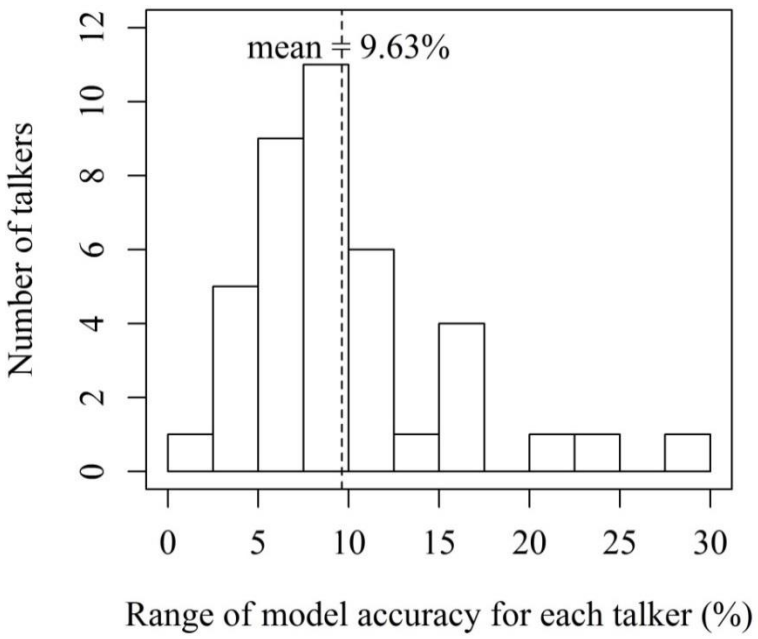

Figure 5. A histogram of the range of model accuracy across listener conditions for each talker

As can be seen in Figure 5, the extent to which talkers varied the clarity of word-boundary cues differed across talkers, as the range of model accuracy varied between 2.45% to 27.78%. On average, talkers' range of intra-talker variation in the model accuracy was 9.63%. On average, female talkers had a wider range (10.83%) than male talkers (8.43%). However, the range of model accuracy did not correlate with the gender of the talker, talkers' awareness of the ambiguity, or talkers' reports about the listeners' performance. Although it is left for future research to explore what factors determine the degree to which talkers vary the availability of durational cues to word boundaries, the inter-talker variability in the extent to which talkers vary the clarity of word-boundary cues suggests that individual talker differences might have contributed to the

inconsistencies in the previous literature regarding whether or not talker produce acoustic-phonetic cues to word boundaries.

2.3.4. Durational cues to word boundaries across CV types

One of the major findings of the current study is that phonetic context around the potential word boundaries (i.e., CV type) affects talkers' production of durational cues to word boundaries. This finding suggests that phonetic context around the potential word-boundaries is a potential source of variation in the extent to which talkers produce signal-based segmentation cues. In particular, word-boundary ambiguities between schwa and a following consonant (e.g., *acute* vs. *a cute*) were rarely disambiguated by durational cues to word boundaries (see also, Kim et al., 2012). The lack of durational cues resolving the word-boundary ambiguities occurring between schwa and consonant was suggested by the analyses of predictive model accuracy, regardless of whether the predictor variables of the predictive model was duration of a single segment (i.e., the conservative morel) or duration of segments and pauses (i.e., the less restrictive model). In order to investigate why ambiguous schwa-initial sequences were hardly disambiguated by the predictive models, the distribution of juncture segment duration as compared across the location of the word boundary (i.e., word-initial or non-initial positions) and the three CV type conditions, which is shown in Figure 6.

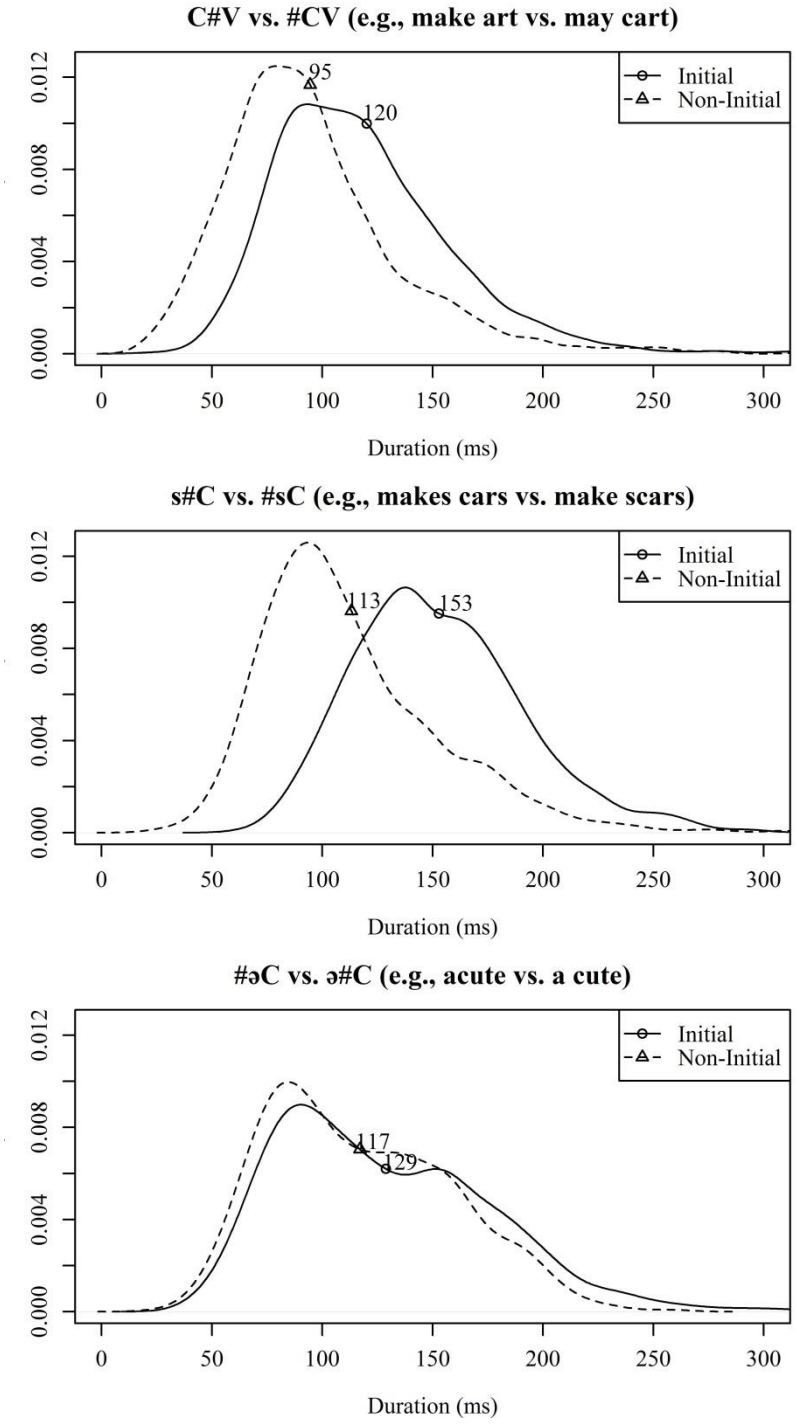

Figure 6. Distribution of word-initial and non-initial critical juncture segment duration across CV types

As can be seen in Figure 6, the distributions of word-initial and non-initial (i.e., word-medial or word-final) juncture segment duration overlapped most in the ambiguous schwa-initial sequences. This high degree of overlap creates a higher degree of ambiguity as compared to the two other CV types. Also noticeable from Figure 6 is that the non-initial juncture segment after the schwa vowel (e.g., /k/ in *a<u>c</u>ute*; mean duration = 117 ms) is relatively longer than the non-initial juncture segment that occurs in the consonant + vowel condition (95 ms). Such "lengthening" of non-initial segments after schwa is caused by the stress pattern of the schwa-initial ambiguous sequences, where the second syllable is stressed (White, 2002). Due to the lengthening caused by stress, a consonant that occurs after schwa undergoes lengthening which may create a confounding cue for the logistic regression model and potentially for the human listeners.

2.3.5. Awareness effects

In the current study, speech clarity was manipulated by two factors, listeners' ability in the language and talkers' awareness of word-boundary ambiguity. The analysis of predictive model accuracy, regardless of whether the predictive model was conservative or less restrictive, consistently revealed significant main effects of listener condition, suggesting that variation in speech clarity resulting from listener accommodation affects talkers' production of signal-based cues to word boundaries. In contrast, the analysis of model accuracy did not reveal statistically significant main effects or interactions of talkers' awareness of the word-boundary ambiguity. Although the ANOVAs on model accuracy did not reveal any statistically significant main effects

or interactions of awareness, it could still be the case that talkers' awareness of the word-boundary affected the production of durational cues to word boundaries.

In order to investigate whether and to what extent awareness affected durational properties of talkers' productions, a series of duration measures were compared across awareness conditions. Means and standard deviations (in parentheses) of the duration measures are presented in Table 2.

	Awareness condition	
Measurement	Unaware	Aware
Entire utterance duration	2934 ms (723 ms)	3020 ms (797 ms)
Ambiguous target sequence duration	657 ms (237 ms)	711 ms (284 ms)
Ratio of ambiguous target sequence duration to the entire utterance duration	22.79% (7.49 %)	23.64% (7.71 %)
Average duration of silent gap	39 ms (59 ms)	51 ms (81 ms)

Table 2. Means and standard deviations (in parentheses) of the durational measures between awareness conditions

As can be seen in Table 2, talkers who were presented with a visual prompt highlighting the word-boundary ambiguity (i.e., talkers in the aware group) tended to produce slower speech than those who were presented with a visual prompt that did not highlight the word-boundary ambiguity (i.e., talkers in the unaware group). In addition, the ratio of ambiguous target sequence duration to entire utterance duration was slightly

(~ 1%) longer for the utterances spoken by talkers who were in the aware group than those who were in the unaware group. Finally, talkers who were in the aware group tended to produce longer silent gaps than talkers who were in the unaware group. Not only that the overall silent gap duration was longer for the aware group talkers than unaware group talkers, talkers who were in the aware group produced pauses more frequently (847 instances) than those who were in the unaware group (262 instances). Although the visual prompt highlighting the word-boundary ambiguity led talkers to produce slower speech with more frequent and longer pauses demarcating word boundaries as compared to the visual prompt that did not highlight the word-boundary ambiguity, these temporal adjustment did not result in statistically significant differences in the model accuracy.

One possible explanation for the lack of awareness effect on model accuracy is that the design of the current study, which involves reading ambiguous target sequences multiple times, might have led all talkers, regardless of whether or not the visual prompt highlighted the word-boundary ambiguity, to notice the word-boundary ambiguity. If all talkers, even those who were in the unaware group, eventually became aware of the word-boundary ambiguity, all talkers might have compensated for the word-boundary ambiguity, resulting in the awareness effects to be leveled out. In order to address whether repetition decreased the potential effects of awareness, analyses were performed on the tokens produced during the first block of the experiment. The analysis of the data from the first block of the experiment revealed a significant main effect of CV type ($F1(2, 38) = 147.79$, $p < 0.001$; $F2(2, 34) = 36.61$, $p < 0.001$) and a marginally significant main

effect of consonant condition ($F1(1, 39) = 26.12$, $p < 0.001$; $F2(1, 35) = 3.84$, $p = 0.06$). Model accuracy for the tokens produced during the first block was lower for the ambiguous schwa-initial sequences (55.97%) than the other two CV types (78.64% for the consonant-vowel condition; 84.54% for the /s/-consonant condition), which is consistent with the results from the analysis of the tokens produced during all three blocks of the experiment. In addition, model accuracy was higher for the tokens containing an obstruent consonant (75.76%) than the tokens containing a sonorant consonant (69.98%). The analysis of the data from the all three blocks of the experiment did not reveal a significant main effect of consonant condition, but revealed an interaction among listener condition, CV type, and consonant condition on model accuracy (pp. 55-58 and Figure 4). Most importantly, no main effects or interactions with awareness condition reached statistical significance even in the analyses of the tokens produced during the first block of the experiment, reinforcing that the lack of awareness effect is not attributable to repetition. Instead, it is suggested that presenting talkers with a visual prompt that highlighted the word-boundary ambiguity did not lead them to produce clearer durational cues to word boundaries even in the first block of the experiment.

2.4. Discussion

The production of acoustic-phonetic cues to word boundaries has extensively been examined in the literature (Lehiste, 1960; Hoard, 1966; Turk & Shattuck-Hufnagel, 2000; Klatt, 1976; Jakimik & Cole, 1980; Kim et al., 2012, among others). However, there has yet been a consensus as to whether talkers produce signal-based cues to word

boundaries or not. Focusing on the lack of consensus regarding the production of signal-based cues to word boundaries and the acoustic-phonetic variation that spoken language displays, the current study tested whether and to what extent talkers' production of signal-based cues to word boundaries depends on factors conditioning phonetic variation. Among the factors that influence phonetic realizations of speech, the current study focused on speech clarity and phonetic context that have been shown to be major contributors of phonetic variability and that differed across previous studies focusing on the phonetics of word boundaries.

Results from the current study confirmed that the extent to which talkers produced signal-based cues to word boundaries depended on speech clarity, phonetic context, and the interaction between the two. Talkers' adjustment in the extent to which they produced durational cues to word boundaries was found from the analyses of the predictive model accuracy, regardless of whether the model was highly conservative or less restrictive. As predicted, model accuracy was higher when talkers spoke to the listeners whose linguistic ability implied communicative difficulties than when they spoke to a young native listener. These findings confirm that talkers clarify potential acoustic-phonetic markers of word boundaries in order to accommodate to listener needs (Cutler & Butterfield, 1990a; 1990b) and shed light on studies of clear speech by suggesting that enhanced word boundaries may constitute a phonetic property of clear speech.

More importantly, the current study suggested that the extent to which talkers produced signal-based cues to word boundaries systematically varied as a function of speech clarity. Since speech clarity is a source of variation in the production of word-

boundary cues, it is suggested that the apparent discrepancies in the literature regarding the production of word boundary cues can be understood as resulting from differences in speech clarity. To elaborate, studies that have examined the production and perception of read speech reported that talkers produce acoustic-phonetic markers at word boundaries (Lehiste, 1960; Hoard, 1966), while Kim et al. (2012), who examined speech produced under memory load thus less clearly spoken than read speech, argued that talkers rarely produce sufficient signal-based segmentation cues. The current study suggests that speech clarity may account for why previous studies have failed in reaching consensus as to whether talkers produce signal-based cues to word boundaries.

Results from the current study also suggest that phonetic context surrounding the word boundaries (CV type) contribute to the variability in the production of signal-based segmentation cues. Talkers produced the weakest durational cues to word boundaries when potential word boundaries occur between schwa and a consonant, as compared to the potential word boundaries between a consonant and a vowel or /s/ and a consonant. The distributions of word-initial and word-medial or final juncture segment duration overlapped most in the ambiguous schwa-initial sequences, possibly causing the accuracy of the predictive model to be lowest among the three CV types. In addition, the finding that talkers rarely produced durational cues to word boundaries when potential word boundaries occur between schwa and a consonant can be one potential explanation for why results of Kim et al. (2012), who solely focused on the word-boundary ambiguity of schwa-initial sequences, were discrepant from the results of other studies (Lehiste, 1960; Hoard, 1966; Smith & Hawkins, 2012) that have used the stimuli including consonant +

vowel or /s/ + consonant sequences and suggested that talkers produce signal-based cues to word boundaries. Although it is still left for future research to examine why ambiguous schwa-initial sequences were rarely disambiguated by talkers' production of signal-based segmentation cues, it has been suggested that stress (White, 2002) and syntactic properties of indefinite article (Turk & Shattuck-Hufnagel, 2000) might blur potential word boundaries occurring between schwa and a following consonant. That phonetic context around the word boundaries may affect the extent to which talkers produce signal-based segmentation cues also suggests that an accurate picture of word-boundary production requires examining word boundaries occurring in a wide variety of phonetic contexts.

The production of signal-based cues to word boundaries was found to be constrained by the interaction between factors pertinent to the phonetic context of a word boundary (i.e., CV type and consonant condition), factors pertinent to speech clarity (i.e., listener condition). The robust effects of listener accommodation were not found in certain phonetic contexts, such as potential word boundaries between a fricative consonant and a vowel or between a schwa and a consonant. The finding that phonetic context at and around word boundaries constrains the extent to which talkers can produce and enhance the clarity of durational cues to word boundaries is inconsistent with Cutler's (1987) hypothesis that constraints pertinent to speech perception do not affect talkers' production of acoustic-phonetic cues to word boundaries. Instead, our findings suggest that the effects of constraints pertinent to speech perception are modulated by constraints pertinent to speech production, such as inherent phonetic properties at and around

potential word boundaries, replicating that adjustments in speech clarity does not target all sounds to an equal degree (Moon & Lindblom, 1994; Uchanski, 1988 and Uchanski et al., 1992).

Although speech clarity and phonetic context were shown to affect the extent to which talkers produced signal-based segmentation cues, not all factors pertinent to speech clarity and phonetic context affected talkers' production of word boundaries. While listener condition and CV type had robust effects on the extent to which talkers produced durational cues to word boundaries, consonant condition and awareness did not have such robust effects on model accuracy. Consonant condition affected the production of word boundary cues by interacting with other factors. Awareness effects were not robust enough to reach statistical significance. The design of the current study, where speech clarity was manipulated by two factors, acknowledges that speech clarity is multi-faceted and that diverse motivations for speech clarity enhancement may affect speech production differently. Likewise, phonetic context was manipulated by CV type and consonant condition, which allowed for a detailed investigation into the effects of phonetic context on the production of word boundaries.

The discussion above speaks to Lindblom's (1990) theory of Hypo- and Hyper-articulation (henceforth H&H theory), which conceptualized speech production as a balancing act between constraints pertinent to speech production and perception. The two contributors to talkers variability investigated by the current study, speech clarity and phonetic context, are comparable to the perception and production constraints of H&H theory. In addition, the effects and interactions of listener condition and CV type confirm

the H&H theoretic hypothesis that talkers tailor their speech depending on listener needs (perception constraints), phonetic properties of message being conveyed (production constraints), and the interaction between the two. In addition, the finding that not all perception constraints affect the production of word-segmentation cues equally suggests that the H&H theory can be refined by orthogonally manipulating multiple perception constraints and investigating how diverse motivations for perception constraints affect speech production independently as well as in combination.

Another new finding of the current study was that talkers' awareness did not affect the availability of word-boundary cues. This finding speaks to and potentially extends the conclusions drawn from studies focusing on the role of talkers' awareness of a particular ambiguity on speech production. There has been a long-standing debate as to whether awareness is a requirement for talkers' producing acoustic cues resolving ambiguities. In particular, a great amount of research has been devoted to the investigation of whether awareness of ambiguity is a requirement for talkers to produce prosodic cues resolving syntactic ambiguity (Lehiste, 1976; Allbritton, McKoon, and Ratcliff, 1996; Snedeker & Trueswell, 2003, among others). For instance, Lehiste (1976) and Allbritton et al. (1996) showed that talkers who were informed of the syntactic ambiguity produced prosodic cues resolving syntactic ambiguity when they read written sentences, suggesting that talkers produce prosodic cues resolving syntactic ambiguity when they are aware of the ambiguity. Snedeker and Trueswell (2003) compared talkers' productions of syntactically ambiguous sentences produced in an experimental setup where the ambiguity resolution was necessary for successful communication (e.g., when the talkers instructed listeners to

tap the frog with the flower, where the referential context included a flower, a frog holding a flower, and a frog who did not have a flower) and in an experiment setup where the experimental setup resolved syntactic ambiguity (e.g., for the sentence *tap the frog with the flower*, the referential context included only one frog, either holding or not holding a flower), and suggested that talkers used disambiguating prosody when the situation required them to do so for successful communication and when they were aware of the structural ambiguity. The aforementioned studies regarded talkers' production of prosodic boundary cues as accommodation to listener needs.

However, not all studies have shown that awareness is a necessary condition for the production of prosodic disambiguation. For instance, a number of recent studies using interactive tasks between talkers and listeners reported that talkers produce disambiguating prosody consistently and spontaneously (Schafer et al., 2000; Kraljic & Brennan, 2005; Speer, Warren, & Schafer, 2011, among others). Talkers' production of cues to prosodic boundaries seems almost redundant, in the sense that prosodic cues demarcating syntactic boundaries were produced even when there were other constraints (e.g., situational context) that would resolve the syntactic ambiguity (Krajic & Brennan, 2005; Speer et al, 2011) or even when the sentences were not ambiguous (Watson & Gibson, 2005). These findings suggest that the production of prosodic cues is contingent upon constraints of speech production.

The current study extended the aforementioned studies investigating the relationship between awareness and speech production by studying a new type of ambiguity, word boundary ambiguity. We also confirmed that talkers' awareness of the

word-boundary ambiguity did not cause them to enhance clarity of word boundaries, suggesting that awareness was not a requirement for talkers' producing durational cues resolving potential word-boundary ambiguities. Results from the current study also appear to contradict Baese-Berk and Goldrick's (2009) results, which revealed an interaction between multiple sources of hyperarticulation—the presence of a minimal pair in the lexicon and the presence of a minimal pair in the visual context, which would lead talkers to be aware of the minimal pairs—and additive effects of awareness on the production of voice onset times. Instead, we found no effects of visual context and no interaction between listener condition and awareness. This discrepancy might be due to the difference in the linguistic unit of investigation (i.e., contrast between words vs. word-boundary ambiguity).

Results from the current study may also shed light on models of word segmentation. Mattys et al.'s (2005; 2007) studies have suggested that word segmentation involves integrating signal-based and knowledge-based cues according to their relative strength. Given that the availability of signal-based cues differs across listener condition and phonetic contexts, the extent to which knowledge-based cues are necessary for word segmentation may vary depending on whom the talkers speak to and phonetic context at and around word boundaries. In order to obtain a precise estimate regarding the extent to which knowledge-based cues are necessary to resolve a potential word-boundary ambiguity, it would be necessary to study the extent to which signal-based cues can guide listeners' word segmentation. An investigation into whether and the extent to which listeners' segmentation accuracy is affected by the contributors of variation in the talkers'

production of signal-based segmentation cues is expected to clarify another puzzle in the literature regarding whether signal-based or knowledge-based cues should take precedence in word segmentation.

In this study, statistical analyses were used to estimate the extent to which talkers produced word boundary cues, and thus the availability of acoustic cues to word boundaries for listeners. However, it is still an open question whether and how the estimated availability of acoustic segmentation cues correlates with listeners' segmentation accuracy (i.e., whether listeners interpret the ambiguous sequences as the talker intended). In addition, it is an empirical question whether listeners' segmentation accuracy also varies as a function of the production and perception constraints which were found to affect the degree to which talkers produced durational cues to word boundaries. For instance, there might be some phonetic variation that is perceptually salient but does not strongly correlate with duration, which would result in the regression models in this chapter underestimating the availability of acoustic cues to word boundaries. Alternatively, listeners might not be sensitive to the fine-grained duration differences that talkers produced. For instance, Klatt (1976), based on studies examining Just-Noticeable Differences (JNDs) in speech perception, suggested that variation in segment duration due to word boundary location is below the JND level and thus listeners might not benefit from talkers' production of fine-grained duration differences.

In order to address these empirical questions, in Chapter 3, I present three perception experiments testing the perceptual consequences of acoustic-phonetic variation at and around word boundaries, focusing on whether listeners' segmentation

accuracy changes as a function of whom the talkers spoke to ("listener" condition), talkers' awareness of the word-boundary ambiguity (awareness), phonetic context surrounding the word boundary (CV type) and at the word boundary (consonant condition).

Chapter 3: Perception of acoustic cues to word boundaries

3.1. Introduction

Spoken language is a continuous stream of speech sounds. In order to comprehend spoken language, listeners must segment such continuous speech into a series of discrete units of meaning, typically words. A substantial literature has investigated how listeners segment the speech signal into words, focusing on identifying the factors that influence listeners' perception of word boundaries.

It has widely been suggested that acoustic-phonetic events that co-occur with word boundaries affect listeners' perception of word boundaries. Listeners have been shown to distinguish word-boundary "minimal pairs," identical sequences of phonemes that differ only in the location or the presence/absence of a word boundary (e.g., *nitrate* vs. *night rate* vs. *Nye trait*; Lehiste, 1960; Hoard, 1966, among others), considerably above the chance level. The finding that listeners are highly accurate in interpreting word-boundary minimal pairs as intended by talkers suggests that listeners may exploit sub-phonemic, acoustic-phonetic details for word segmentation. The acoustic-phonetic details that have been shown to correlate with listeners' perception of word boundaries include, but are not limited to, glottalization of word-initial vowels, lengthening of word-final vowels, lengthening of word-initial segments, flapping of word-medial coronal stops, allophonic variation such as aspiration of voiceless stops and clear versus dark allophones

of lateral /l/ (Lehiste, 1960; Christie, 1974; Nakatani & Dukes, 1977; Nakatani & Schaffer, 1978; Sproat & Fujimura, 1993; Smith & Hawkins, 2000).

Additional evidence suggests that listeners are highly efficient in using the acoustic-phonetic details for word segmentation. For instance, listeners were shown to be primed for *kiss* after hearing *two lips*, but not after hearing *tulips*, suggesting that they can attend to the acoustic differences between the /l/ in *two lips* and the /l/ in *tulips* (Gow & Gordon, 1995). Listeners' early sensitivity to the acoustic-phonetic markers to word boundaries has also been confirmed by studies focusing on how listeners resolve a temporary word-boundary ambiguity due to embedded words (e.g., *cap*, which can be a single monosyllabic words as well as the first syllable of a disyllabic word, *captain*). Listeners were shown to distinguish a short monosyllabic word (i.e., *cap* as a single word as in *the soldier saluted the flag with his cap tucked under his arm*) and the first syllable of a longer word (i.e., *cap* as the first syllable of *captain* as in *the soldier saluted the flag with his captain looking on*) even before the stimuli diverged phonemically, suggesting that listeners are efficient in using acoustic-phonetic cues to resolve word-boundary ambiguity (Davis, Marslen-Wilson & Gaskell, 2002; Salverda, 2005; Shatzman & McQueen, 2006, among others).

In summary, it has been suggested that the speech signal may contain acoustic-phonetic markers corresponding to listeners' perception of word boundaries and that listeners are highly efficient in using such acoustic-phonetic cues for word segmentation in that they readily attend to the acoustic-phonetic markers and parse the signal with high accuracy. If human speakers consistently produce acoustic-phonetic markers to word

boundaries and listeners attend to the acoustic-phonetic cues, a purely bottom-up mechanism should guarantee successful word segmentation. However, in real-life communication, word-boundary misperception is not infrequent, with some studies reporting that word-boundary misperceptions make up 18% of spontaneously occurring errors in speech perception (Garnes and Bond, 1975; 1980), suggesting that signal-based acoustic information alone may not guarantee successful word segmentation. In addition, in a study aimed at estimating the extent to which acoustic-phonetic cues alone (i.e., without syntactic or semantic context) guide word segmentation, listeners were found to be highly inaccurate in perceiving word boundaries as intended by talkers (Reddy, 1976; Kim et al., 2012). These findings suggest that the speech signal may not contain sufficient cues for listeners' word segmentation and thus listeners must rely on cues other than acoustic-phonetic information, in order to perceive word boundaries as intended by talkers.

Taken together, there has not yet been a consensus concerning whether and to what extent signal-based cues guide listeners' word segmentation, with some studies suggesting that acoustic-phonetic information plays an instrumental role in listeners' word segmentation (Lehiste, 1960; Hoard, 1966; Gow & Gordon, 1995; Davis, Marslen-Wilson & Gaskell, 2002; Salverda, 2005; Shatzman & McQueen, 2006, among others) and other studies suggesting that acoustic-phonetic details are not sufficient to guarantee successful word segmentation (Reddy, 1976; Kim et al., 2012) and that listeners must attend to and use information other than what is in the speech signal (Norris, McQueen, Cutler, & Butterfield, 1997; McQueen, 1998; Dilley & McAuley, 2008; Mattys, White, &

Melhorn, 2005; Mattys & Melhorn, 2007, among others). In the current study, we suggest that talker variability may account for the apparent lack of consensus concerning the role of signal-based cues to word boundaries.

The acoustic-phonetic cues to word boundaries are produced by talkers and consequently, the extent to which listeners can rely on the acoustic-phonetic cues to word boundaries should be contingent upon factors conditioning talker variability. Results of Experiment 1 suggested that the extent to which talkers produce signal-based segmentation cues varies as a function of speech clarity and phonetic context based on the fit of statistical models of which the independent variables were durational properties around the word boundaries and of which the dependent variable was the location of word boundaries.

If talkers vary the extent to which they produce acoustic-phonetic cues to word boundaries and if listeners attend to acoustic-phonetic cues for word segmentation, listeners' perception of word boundaries should depend on speech clarity and phonetic context surrounding the word boundaries. The current study is an empirical test of whether the variation in the production of word boundaries as a function of speech clarity and phonetic context, which was found in Experiment 1, has perceptual consequences.

It is well established in the literature that talkers tailor their speech according to listener needs and that clarity modulation has perceptual consequences, as demonstrated by a higher intelligibility of clear speech than plain speech by listeners (Perkell, Zandipour, Matthies, & Lane, 2002; Ferguson & Kewley-Port 2002; Bradlow, Kraus, & Hayes, 2003; Krause & Braida 2004; Smiljanić & Bradlow 2005). Experiment 1

replicated the findings of the clear speech literature by showing that the accuracy of the statistical models predicting the location of word boundaries based on the duration of segments and optional pauses around the word boundaries was higher when talkers spoke to listeners whose linguistic background implies communicative difficulties and thus produced clearer speech than when talkers spoke to a young native listener of the language being spoken.

Based on the results of Experiment 1 and the findings from the studies reporting clear speech intelligibility benefit, it is hypothesized that the talkers' modulation of acoustic-phonetic cues to word boundaries will have perceptual consequences, such as enhancing listeners' segmentation accuracy, enhancing overall intelligibility of speech, among others. Although it seems straightforward to hypothesize that the findings from acoustic analyses and statistical analyses should generalize to listeners' perception, it is still an empirical question whether and to what extent human listeners benefit from such modulation in the acoustic-phonetic properties, especially given that some acoustic-phonetic variation due to speech clarity may be below the JND level and thus listeners might not benefit from talkers' production of fine-grained acoustic-phonetic differences (Klatt, 1976). In addition, results from acoustic analyses and perception experiments might diverge, since listeners' segmentation may be affected by some acoustic-phonetic properties that were not measured or used as predictor variables for Experiment 1. Similarly to Experiment 1, modulation in speech clarity was manipulated by two variables, listener condition and talkers' awareness of the word-boundary ambiguity.

Results from empirical studies specifically focusing on speech clarity and listeners' perception of word boundary cues reinforce the hypothesis that the extent to which listeners exploit acoustic-phonetic cues for word segmentation varies as a function of speech clarity. Listeners were shown to be 3% more accurate in segmenting word-boundary minimal pairs produced at a comfortable and careful speed than a fast speed (Barry, 1981), suggesting that speech rate and corresponding speech clarity might affect the availability of signal-based segmentation cues. Relatedly, listeners who heard primes containing a word-boundary ambiguity (e.g., *great anchor* vs. *gray tanker*) were faster in performing a lexical decision task for read primes than spontaneously produced primes (White et al., 2010; White et al., 2012), suggesting potential differences in the informativeness in the signal-based segmentation cues across speaking styles and corresponding speech clarity.

The preceding discussion suggests that listeners' perception of word boundaries may be affected by talkers' speech rate modulation (Barry, 1981) or differences in the elicitation methods, which impact speech clarity (White et al., 2010; White et al., 2012). Additionally, the high segmentation accuracy for stimuli generated from a reading task (Lehiste, 1960; Hoard, 1966; Barry, 1981, among others), in comparison to the poor segmentation accuracy for stimuli generated from a task where talkers produced speech under memory load (Kim et al., 2012; Brink, Wright, & Pisoni, 1998; Harnsberger & Pisoni, 1999) also suggests that speech clarity might influence listeners' word-segmentation accuracy. Based on these findings, it was hypothesized that listeners' word boundary perception will be affected by speech clarity.

An additional source of variation that can lead to variation in the availability of acoustic-phonetic cues influencing listeners' perception of word boundaries is phonetic context at and around potential word boundaries. It has well been documented that some segments display qualitative (i.e., allophonic) variation depending on the whether they begin or end a word, while other segments display quantitative (e.g., durational) variation (Nakatani & Dukes, 1977). Since the availability of allophonic cues is specific to individual segments, informativeness of acoustic-phonetic cues to word boundaries is expected to vary across segments. Listeners were shown to be better at spotting word boundaries when segments that show allophonic variation as a function of word position, such as /p, t, k/ and /l/, occurred around a word boundary than spotting word boundaries around segments without allophonic variation. These findings suggest that phonetic contexts may differ in the degree to which they inform listeners of the location or presence/absence of word boundaries. In addition, the high segmentation accuracy for the word-boundary minimal pairs containing a word-boundary ambiguity between /s/ and a following consonant or between a consonant and a vowel (Lehiste, 1960; Hoard, 1966; Barry, 1981), as compared to the low segmentation accuracy for the word-boundary minimal pairs containing a word-boundary ambiguity between schwa and a following consonant (Kim et al., 2012) suggests that the phonetic context surrounding word-boundaries will also affect the extent to which listeners can exploit acoustic-phonetic cues for word segmentation. Comparisons across previous studies and the results of Experiment 1, which revealed that phonetic contexts surrounding the word boundaries (CV type) affected the production of acoustic-phonetic cues to word boundaries, leads to

the hypothesis that the availability of signal-based segmentation cues and corresponding segmentation accuracy should depend on the phonetic contexts around the word boundaries.

To summarize, the goal of the current study was to explore the variability in the extent to which signal-based segmentation cues guide listeners' word segmentation. The current study tested the hypothesis that the availability and informativeness of signal-based cues to word boundaries, which were measured by listeners' perception of word boundaries, depend on speech clarity and phonetic context at and around the word boundaries. Results are expected to confirm that the results from Experiment 1 reflect listeners' perception of word boundaries as well as shed light on the discrepancy in the previous studies concerning the extent to which acoustic-phonetic cues affect listeners' word segmentation.

Some methodological concerns remain as to how to measure listeners' word segmentation while avoiding a ceiling effect masking the potential differences due to speech clarity. The small effect of talkers' speech rate on listeners' segmentation accuracy (3.23%) as well as overall high accuracy of word segmentation (83.05%) reported by Barry (1981) suggest that it is generally easy for listeners to segment words from read speech and the small slow-speech benefit on word segmentation might be due to a ceiling effect. Similarly, Schwab, Miller, Grosjean, and Mondini (2008), who examined variation in listeners' segmentation accuracy for ambiguous phrases containing a four-way word boundary ambiguity (e.g., *great eyes*, *gray ties*, *great ties*, and *gray eyes*) depending on the talkers' speech rate, reported that listeners were 95.6% accurate in word segmentation,

with 2.5% higher accuracy for the slower speech condition than the faster speech condition, again suggesting a potential ceiling effect. Given these findings and the overall accuracy of the predictive model (Experiment 1, 75.64% accurate), which was considerably above the chance level, we predicted that listeners would be, in general, fairly accurate in segmenting words from the word-boundary minimal pairs generated from Experiment 1. In other words, it might be the case that all sound files are clear enough for listeners to accurately perceive word boundaries as intended by the talkers, especially when the task is a forced-choice task between the possible parses, as in the tasks used by Barry (1981) and Schwab et al. (2008).

For these reasons, in the current study, multiple dependent measures were used by having listeners perform the following three tasks: forced-choice decision regarding which alternative they heard, clarity rating, and an open-set transcription task, in order to measure listeners' word segmentation. For Experiments 2a and 3, listeners performed a forced-choice task followed by a clarity rating task. Tokens produced by the talkers who were in the "unaware" group served as stimuli for Experiment 2a. Tokens produced by the talkers who were in the "aware" group served as stimuli for Experiment 3. Upon hearing each stimulus, listeners made a response indicating which alternative they heard as quickly as possible so that response time data could be used to explore the potential effects of speech clarity and phonetic context on the speed of listeners' word segmentation even if a ceiling effect was found in the accuracy analysis.

After performing the speeded forced-choice task, listeners performed a clarity rating task. Clarity rating scores should provide information that complements the

accuracy and response time analyses. In addition, an investigation into listeners' subjective clarity rating is expected to extend previous findings regarding the correlation between perceived clarity and speech intelligibility. Ferguson and Kerr (2009) suggested that subjective clarity rating tasks may offer a good estimate of speech intelligibility, by reporting a high correlation between subjective clarity rating and vowel intelligibility in noise as well as a high correlation between subjective clarity rating and vowel perimeter (i.e., the sum of the Euclidean distances between /i/ and /ae/, /ae/ and /a/, /a/ and /u/, and /u/ and /i/ in *F1* by *F2* space). By using a forced-choice task as well as a clarity rating task, we could examine whether the high correlation between clarity rating and intelligibility found for the perception of vowels is also found for the perception of word boundaries.

For Experiment 2b, we had listeners perform an open-set transcription task on the same set of stimuli used for Experiment 2a, in order to test whether the forced-choice experiment (Experiment 2a) accurately reflected what the listeners would hear without having the alternatives to choose from. In addition, the use of the open-set task is expected to reduce the possibility of a ceiling effect, since listeners' recognition accuracy tends to be higher for closed-set tests with a very small number of options to choose from than for open-set tasks (Clopper, Pisoni, & Tierney, 2006).

In summary, the current study tested whether and to what extent listeners' perception of word boundaries varied depending on the major sources of talker variability, speech clarity and phonetic contexts, in order to clarify the role of signal-based segmentation cues on listeners' word segmentation. Listeners' word segmentation was

measured by multiple dependent measures in order to avoid a ceiling effect. The organization of the remainder of this chapter is as follows. Experiments 2a, 2b, and 3 are empirical tests of listeners' perception of word boundaries across the speech clarity and phonetic context conditions, so that we can confirm the perceptual consequences of the talker variability in the acoustic-phonetic cues to word boundaries.

3.2. Experiment 2a: two-alternative forced choice and clarity rating for stimuli produced by talkers who were unaware of the word-boundary ambiguity

3.2.1. Methods

Participants

36 new participants who did not take part in the production experiment were recruited from the same pools or invited for paid participation. Participants recruited from the linguistics and psychology participant pool received course credit, and paid participants received $8 for participation. None of them reported speech or hearing related difficulties.

Stimulus materials

Stimuli were created by excising the ambiguous phrases from a sample of the productions generated in Experiment 1. For experiment 2a, only the tokens that were produced by talkers who were in the "unaware" group were used. Ten (five male and five female) talkers were selected from the twenty talkers who participated in the "unaware" condition of Experiment 1. In order to ascertain that the selected ten talkers were good

representatives of the entire set of talkers, talkers were chosen by pairing talkers who were similar with respect to speech rate and the logistic regression model fit predicting the location of the word boundary based on the acoustic cues, and randomly choosing one talker from each of the ten pairs. To include productions from these ten talkers and every item, productions of half (18) of each talker's 36 target items were sampled, so that the stimulus set contained productions of all three CV types and both of the segment conditions from each talker. If a talker made a mistake in one of the six utterances in a set (i.e., 2 word-boundary conditions by 3 listener conditions) of an ambiguous sequence, the set was not included for the perception experiment for that talker.

Three lists were constructed so that each listener heard all six instances of an ambiguous sequence produced by a single talker. Each list contained at least one instance of the 36 target items produced by one of the ten talkers. A typical list was composed of 52 or 53 sets of six utterances. Each listener heard stimuli from one of the three lists. Stimuli were presented in a semi-random order, so that listeners heard the same voice or same item no more than three times in a row. The differences in the number of trials across lists were kept minimal. Listeners heard a maximum of 316 trials per a session, and the entire session took approximately 35-40 minutes.

A short practice session preceded the main part of the experiment. 32 sound files, which were selected from the productions for the four fillers items and were spoken by the talkers whose productions were not included in the stimulus set for the main part of the experiment, served as stimuli for the practice session.

Procedures

Each trial in Experiment 2a began with a fixation mark presented in the middle of the computer screen for 750 ms. 250ms after the fixation mark was erased, listeners heard an audio stimulus. At the offset of the audio stimulus, two possible parses were printed on the screen, one on the left-hand side of the screen, the other on the right-hand side of the screen.

The listeners' first task was to decide which of the two alternatives the talker said as quickly and as accurately as possible. They had 3.5 seconds to respond and their responses were timed. After the listeners responded to the first task, they rated how clearly the sequence was spoken on a scale from 1 (spoken very clearly) to 5 (spoken very unclearly).

3.2.2. Results

3.2.2.1. Segmentation Accuracy

Participants' segmentation accuracy was compared across the following three conditions: whom the talkers spoke to (i.e., listener condition), phonological context surrounding the word boundary (i.e., whether the potential boundary occurred between a consonant and a vowel (e.g., *beef eater* vs. *bee feeder*), between /s/ and a consonant (e.g., *collects gulls* vs. *collect skulls*), or between schwa and a consonant (e.g., *along* vs. *a long*; referred to as CV type), and whether the consonant at the word boundary was an obstruent or a sonorant (referred to as consonant condition).

A series of repeated measures analyses of variance were performed on the mean of participants' segmentation accuracy. Listener condition, CV type, and consonant condition were treated as within-participant independent variables; listener condition was also a within-item independent variable, while CV type and consonant condition were between-item independent variables.

Overall, participants' segmentation accuracy in the two-alternative forced choice task was 77.05%, which was higher than chance level (50%) and slightly higher than the accuracy of the less restrictive model reported in Experiment 1 (75.64%). Participants' segmentation accuracy differed depending on who the talkers spoke to (i.e., listener condition; left panel of Figure 7), consistently with the findings of Experiment 1.

Segmentation accuracy was lowest when talkers directed their speech to a young native listener (73.31%) and highest when talkers directed their speech to an older hearing impaired listener (79.97%). The accuracy in the young nonnative listener condition was in between the other two conditions (77.86%). Statistical analysis revealed a significant main effect of listener condition on average segmentation accuracy ($F1(2, 34) = 27.93$, $p < 0.001$; $F2(2, 34) = 91.06$, $p < 0.001$). Post-hoc comparisons revealed that listeners' segmentation accuracy in the plain speech condition was significantly lower than for the clear speech conditions ($t_{subj}(35) = -7.22$, $p < 0.001$; $t_{item}(35) = -5.50$, $p < 0.001$ for comparisons between the young native and the older hearing impaired listener conditions; $t_{subj}(35) = -4.99$, $p < 0.001$; $t_{item}(35) = -3.67$, $p < 0.001$ for comparisons between the young native and the young nonnative listener conditions). The difference in

the listeners' accuracy between the two clear speech conditions was significant by subjects (t_{subj} (35) = 2.53, $p < 0.05$), but not by items (t_{item} (35) = 1.26, *ns*).

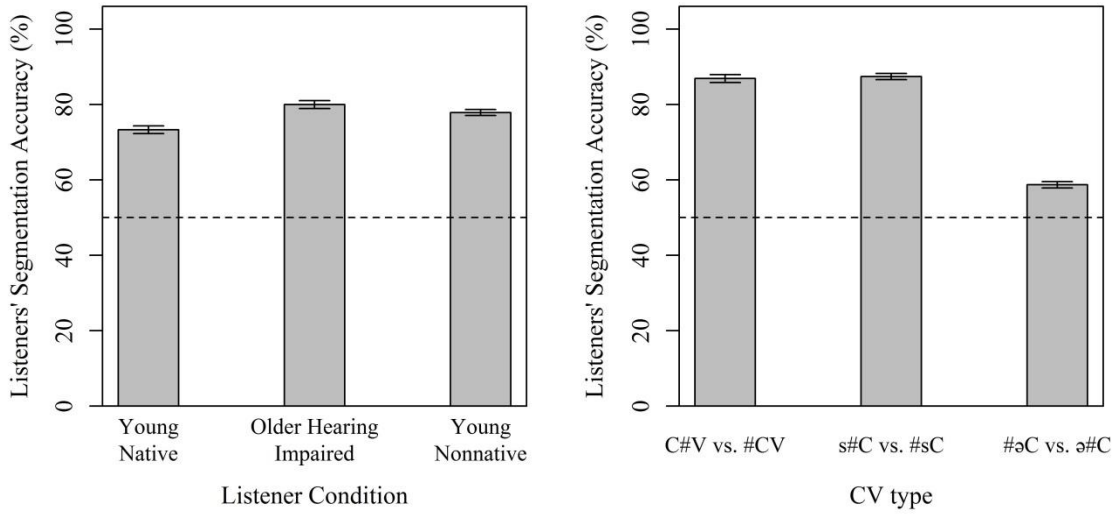

Figure 7. Listeners' segmentation accuracy by listener condition (left) and CV type (right). Error bars represent subject standard errors.

As can be seen from the right panel of Figure 7, listeners were more accurate in segmenting certain CV types than others, consistently with the findings of Experiment 1. Listeners were much more accurate (86.88%) in segmenting a potential word-boundary between a consonant and a vowel (e.g., *beef eater* vs. *bee feeder*) or the boundary between /s/ and a following consonant (e.g., *collects gulls* vs. *collect skulls*; 87.40%) as compared to the boundary between schwa and a following consonant (e.g., *along* vs. *a long*; 58.71%). The main effect of CV type on the participants' segmentation accuracy was statistically significant ($F1(2, 34) = 27.93$, $p < 0.001$; $F2(2, 34) = 91.06$, $p < 0.001$). Post-hoc pairwise comparisons on listeners' segmentation accuracy revealed that

segmentation accuracy for the ambiguous schwa-initial sequences was significantly lower than the two other CV types (t_{subj} (35) = 36.31, $p < 0.001$; t_{item} (22) = 11.35, $p < 0.001$ for comparisons between /s/-consonant and schwa-consonant conditions; t_{subj} (35) = 45.03, $p < 0.001$; t_{item} (22) = 16.86, $p < 0.001$ for comparisons between consonant-vowel and schwa-consonant conditions). The comparisons between the consonant-vowel and the /s/-consonant conditions was not statistically significant (t_{subj} (35) = -0.65, *ns*; t_{item} (22) = 0.10, *ns*). The main effect of consonant condition was significant in the subject analysis but not in the item analysis (*F1*(1, 35) = 5.59, $p < 0.05$; *F2*(1, 35) = 0.79, *ns*). Listeners showed a tendency to hear more word boundaries as intended by talkers when the word-boundary minimal pair contained an obstruent consonant at the word boundary (77.55%) than a sonorant consonant (76.52%).

No interactions reached statistical significance at ($\alpha = 0.05$) in both subject and item analyses, but a marginally significant interaction between listener condition and CV type on listeners' segmentation accuracy (*F1*(4, 32) = 5.94, $p < 0.001$; *F2*(4, 32) = 2.43, $p = 0.06$). Figure 8 shows the listener condition by CV type interaction on listeners' segmentation accuracy.

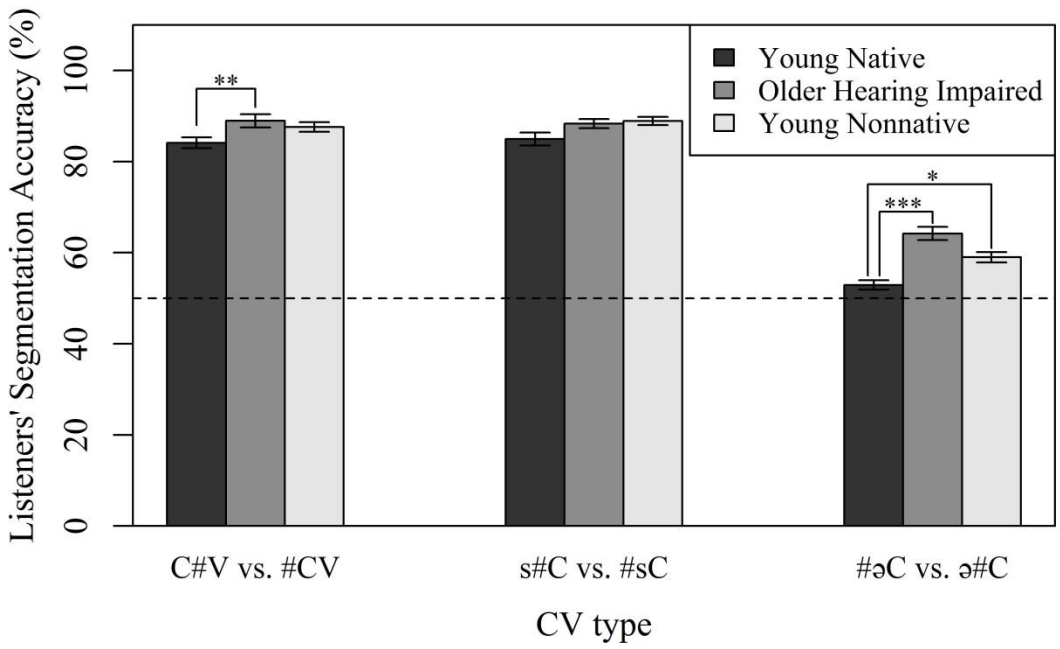

Figure 8. Listener condition by CV type interaction on listeners' word segmentation accuracy. *** indicates $p < 0.001$; ** indicates $p < 0.01$; and * indicates $p < 0.05$ for the post-hoc comparisons

As can be seen in Figure 8, the degree to which listener condition and corresponding speech clarity affected listeners' segmentation accuracy depended on the CV type. Post-hoc comparisons revealed that listeners' segmentation accuracy significantly differed between the young native and the older hearing impaired listener conditions in two of the three CV type conditions: when the word boundary ambiguity occurred between a consonant and a vowel (t_{subj} (35) = -4.31, $p < 0.001$; t_{item} (11) = -3.61, $p < 0.01$) and when the word boundary ambiguity occurred between schwa and a following consonant (t_{subj} (35) = -7.42, $p < 0.001$; t_{item} (11) = -5.52, $p < 0.001$). Post-hoc comparisons also revealed that listeners' segmentation accuracy differed between the young native and the young nonnative listener conditions when the word boundary

ambiguity occurred between schwa and a consonant (t_{subj} (35) = -4.48, $p < 0.001$; t_{item} (11) = -2.88, $p < 0.05$). No differences in the listeners' segmentation accuracy between the older hearing impaired and the young nonnative listener conditions were revealed to be statistically significant.

3.2.2.2. Response times in the two-alternative forced choice task

Overall, we found that listeners' segmentation accuracy was high and although the segmentation accuracy differences across speaking styles were significant, they were relatively small (6.66% difference in the listeners' segmentation accuracy between the young native and older hearing impaired listener conditions; 4.55% difference in the listeners' segmentation accuracy between the young native and young nonnative listener conditions). The high segmentation accuracy as well as the small effect of listener condition might suggest that all of the stimuli were clear enough for listeners to accurately parse word boundaries, and the accuracy analysis might have displayed a ceiling effect, missing some more fine-grained differences across conditions. If this is the case, an investigation into the response time data may reveal some differences across the listener, CV type, and consonant conditions, shedding light on the extent to which talkers produced acoustic cues to word boundaries that listeners can exploit for word segmentation. A repeated measures ANOVA was performed on the mean response time for the two-alternative forced choice task. Listener condition, CV type, and consonant condition were within-subject independent variables; listener condition was a within-item

independent variable and CV type and consonant condition were between-item independent variables.

The ANOVA on the mean response time for the two-alternative forced choice task revealed a statistically significant main effect of listener condition ($F1(2, 34) = 22.30$, $p < 0.001$; $F2(2, 34) = 23.70$, $p < 0.001$). Listeners' response times were faster for the tokens spoken towards an older hearing impaired (1271 ms) or a young nonnative (1315 ms) listener than the tokens spoken towards a young native listener (1352 ms). Post-hoc comparisons revealed that the response times in the forced-choice task differed across all listener conditions. ($t_{subj}(35) = 6.00$, $p < 0.001$; $t_{item}(35) = 6.57$, $p < 0.001$ for comparisons between the young native and the older hearing impaired listener conditions; $t_{subj}(35) = 3.25$, $p < 0.01$; $t_{item}(35) = 2.79$, $p < 0.01$ for comparisons between the young native and the young nonnative listener conditions; $t_{subj}(35) = -3.80$, $p < 0.001$; $t_{item}(35) = -4.25$, $p < 0.001$ for comparisons between the older hearing impaired and the young nonnative listener conditions). Faster reaction times in the forced-choice task for the clear speech conditions suggest that listeners benefited from talkers' enhancing the clarity of acoustic-phonetic cues demarcating word boundaries. In contrast to the analysis on listeners' segmentation accuracy, in which accuracy did not differ significantly for the two clear speech conditions, we found that response times significantly differed across all three listener conditions, as was found in the analysis of the model accuracy in Experiment 1.

The main effects and interactions of the other independent variables—CV type and consonant condition —failed in reaching significance in both subject and item

analyses. In the subject analysis, the repeated measures ANOVA revealed a significant main effect of CV type ($F1(2, 34) = 4.16$, $p < 0.05$; $F2(2, 34) = 2.03$, *ns*), a main effect of consonant condition ($F1(1, 35) = 9.70$, $p < 0.01$; $F2(1, 35) = 1.26$, *ns*), and an interaction between CV type and consonant condition ($F1(4, 32) = 4.16$, $p < 0.05$; $F2(4, 32) = 2.03$, *ns*). Listeners tended to respond faster to the tokens containing the word-boundary ambiguity between a consonant and a vowel (e.g., *beef eater* vs. *bee feeder*; 1275 ms) than the tokens containing a word-boundary ambiguity between /s/ and a following consonant (e.g., *collects gulls* vs. *collect skulls*; 1320 ms) or between schwa and a following consonant (e.g., *along* vs. *a long*; 1343 ms). With regards to the consonant condition, listeners showed a tendency to respond faster to the tokens containing an obstruent at the word boundary (1300 ms) than to the tokens containing a sonorant (1328 ms). Finally, the difference in the response times between the consonant conditions was largest for the word boundaries between /s/ and a following segment (1279 ms for obstruents vs. 1362 ms for sonorants; 1275 ms for obstruents vs. 1275 ms for sonorants for the word-boundary ambiguities between a consonant and a vowel; 1338 ms for obstruents vs. 1349 ms for sonorants for the word-boundary ambiguities between schwa and a consonant). This interaction is most likely due to the robust perceptual effects of the allophonic variation in aspirated versus unaspirated stop consonants following /s/, but this difference failed in reaching statistical significance in the item analysis, in part because the CV and consonant conditions were between-item variables and thus had less statistical power in the item analysis than the subject analysis.

3.2.2.3. Clarity rating

After responding to the two-alternative forced choice task, listeners rated how clearly each sound file was spoken on a scale from 1 (spoken very clearly) to 5 (spoken very unclearly). By performing a correlation analysis, we tested whether the high correlation between clarity rating and intelligibility found by Ferguson and Kerr (2009) for the perception of vowels was also found for the perception of word boundaries. For each sound file, accuracy and clarity rating score were averaged over participants. A mixed-effects linear regression analysis was performed to test whether and to what extent average clarity rating predicts average accuracy in the two-alternative forced choice task. Talkers and items were treated as random factors.

The mixed-effect model revealed that the average clarity rating score is a statistically significant predictor of average forced-choice accuracy (β = -12.70; t = -11.10; pseudo r^2 = 0.40), suggesting that tokens that were rated as clearly spoken tended to be accurately segmented. The finding that clarity rating is a significant predictor of listeners' forced-choice segmentation accuracy suggests that subjective clarity rating may offer a good estimate of segmentation accuracy and extends Ferguson and Kerr's (2009) finding that subjective clarity rating and vowel intelligibility are highly correlated with each other to another linguistic context, word boundaries.

Although we found a high correlation between clarity rating and listeners' segmentation accuracy in the forced-choice task, this high correlation does not necessarily mean that factors that influence listeners' segmentation accuracy also influence listeners' clarity rating scores. To determine which factors affected listeners'

clarity rating scores, we performed a series of repeated measures ANOVAs on average clarity rating across listener conditions, CV types and consonant conditions. The repeated measures ANOVAs on average clarity rating score revealed significant main effects of listener condition ($F1(2, 34) = 103.60$, $p < 0.001$; $F2(2, 34) = 57.82$, $p < 0.001$), CV type ($F1(2, 34) = 47.41$, $p < 0.001$; $F2(2, 34) = 15.32$, $p < 0.001$), and consonant condition ($F1(1, 35) = 37.53$, $p < 0.001$; $F2(1, 35) = 4.53$, $p < 0.05$).

Participants rated tokens spoken to an older hearing impaired (2.14) or a nonnative (2.25) listener as being more clearly spoken than tokens spoken to a young native listener (2.55). Post-hoc pairwise comparisons revealed statistically significant differences in the clarity rating score across all listener conditions ($t_{subj}(35) = 12.35$, $p < 0.001$; $t_{item}(35) = 11.57$, $p < 0.001$ for comparisons between the young native and the older hearing impaired listener conditions; $t_{subj}(35) = 9.30$, $p < 0.001$; $t_{item}(35) = 7.68$, $p < 0.001$ for comparisons between the young native and the young nonnative listener conditions; $t_{subj}(35) = -5.40$, $p < 0.001$; $t_{item}(35) = -2.93$, $p < 0.001$ for comparisons between the older hearing impaired and the young nonnative listener conditions).

With regards to CV types, schwa-initial ambiguous sequences, which were least likely to be accurately segmented by talkers, were perceived as least clear (2.61; cf., mean rating scores of 2.11 for potential word-boundaries between a consonant and a vowel; 2.18 for potential word-boundaries between /s/ and a consonant). Post-hoc pairwise comparisons revealed that clarity rating scores for the tokens containing potential word boundaries between schwa and a consonant were lower than the other two conditions ($t_{subj}(35) = -7.69$, $p < 0.001$; $t_{item}(22) = -5.43$, $p < 0.001$ for comparisons

between the potential word boundaries between a consonant and a vowel and potential word boundaries between schwa and a consonant; t_{subj} (35) = -7.21, $p < 0.001$; t_{item} (22) = -4.17, $p < 0.001$ for comparisons between the potential word boundaries between /s/ and a consonant and word boundaries between schwa and a consonant). The differences in the clarity rating between potential word boundaries between a consonant and a vowel and potential word-boundaries between /s/ and a consonant were not statistically significant.

Also revealed by the ANOVAs was a significant main effect of consonant type on listeners' clarity rating (*F1*(1, 35) = 37.53, $p < 0.001$; *F2*(1, 35) = 4.53, $p < 0.05$). Listeners perceived tokens containing an obstruent consonant at the juncture as more clearly produced (2.23) than tokens containing a sonorant consonant at the juncture (2.40), probably because of allophonic variation between aspirated and unaspirated stop consonants.

No statistically significant interactions were revealed by the ANOVAs on average clarity rating in both subject and items analyses. The subject analysis revealed a significant interaction between CV type and consonant condition on listeners' clarity rating, but this interaction failed in reaching statistical significance in the item analysis (*F1*(2, 34) = 24.48, $p < 0.001$; *F2*(2, 34) = 2.11, *ns*). Though statistically insignificant by items, the extent to which tokens containing an obstruent juncture segment were perceived as being more clearly produced than tokens containing a sonorant juncture segment was largest for the /s/-consonant condition (0.40 difference in the clarity rating score between the tokens containing /s/ and an obstruent following /s/ (1.98); cf. mean

clarity score of 2.38 for the tokens containing /s/ and a sonorant following /s/) and smallest (0.03) for the schwa-consonant condition (average clarity rating score of 2.60 for the tokens containing schwa and an obstruent following schwa; cf. 2.63 for the tokens containing schwa and a sonorant following schwa). The larger difference in the perceived clarity between consonant conditions when the tokens contained potential word boundaries between /s/ and a consonant might be due to allophonic variation between aspirated word initial stops and unaspirated word-medial stops after /s/.

3.2.2.4. Summary of results, Experiment 2a

Experiment 2a tested whether listeners' perception of word boundaries differed depending on speech clarity and phonetic context in order to confirm whether the findings from acoustic analyses (Experiment 1) generalizes to listeners' perception and to suggest that talker variability is a potential explanation for the discrepancies in the literature regarding the sufficiency of signal-based cues to word-boundaries. As was hypothesized in Experiment 1, speech clarity and phonetic context at and around word boundaries were hypothesized to affect listeners' word segmentation. In order to measure listeners' perception of word boundaries, three sets of dependent measures—1) segmentation accuracy and 2) response times in the two-alternative forced-choice task, and 3) clarity ratings on a five-point scale--were obtained and compared across speech clarity (i.e., listener condition) and phonetic context conditions (i.e., CV type and consonant condition).

Speech clarity, which was manipulated by the listener condition, had a statistically significant main effect on the dependent measures. Speech directed to an older hearing impaired listener or a nonnative listener were more likely to be accurately segmented, more quickly responded to in the forced-choice task, and rated as being more clearly spoken than speech directed to a young native listener. This finding suggests that talkers produced more perceptible acoustic cues to word boundaries as they spoke to listeners who might experience difficulty in speech comprehension and that the participants of the current study, who were young native listeners, benefited from these clearer acoustic cues demarcating word boundaries. The effect of listener condition on the listeners' segmentation accuracy is consistent with the results from Experiment 1, where we found a statistically significant main effect of listener condition on the segmentation accuracy of statistical models predicting the location of word boundaries based on the durational cues at and around the word boundaries. As predicted by the statistical model performance, listeners were better at segmenting words from speech directed towards "listeners" whose linguistic competence is suboptimal. In addition, the finding that speech clarity affects the availability and informativeness of acoustic-phonetic cues to word boundaries suggests that discrepancies in the literature regarding the role of signal-based segmentation on the perception of word boundaries may be accounted for by differences in speech clarity with which talkers produced stimuli.

Phonetic context surrounding the potential word boundaries (i.e., CV type) also had a significant main effect on listeners' segment accuracy and clarity rating, mainly because schwa-initial sequences were the least likely to be accurately segmented and the

least likely to be heard as being clearly spoken. These findings are consistent with the findings from Experiment 1, which showed that ambiguous schwa-initial sequences were the least accurately categorized by the model. The effect of CV type on model accuracy and listeners' segmentation accuracy suggest that the differences in the clarity of word boundaries across CV types were not only statistically significant but also perceptible by human listeners. The finding that CV type is a source of variation in the availability and informativeness of signal-based segmentation cues also suggests that stimulus choice could have contributed to the inconsistencies associated with the production and perception of word boundaries.

The analyses of listeners' segmentation accuracy revealed a marginally significant interaction of listener condition and CV type, suggesting that the extent to which listeners benefited from talkers' enhancing speech clarity differed depending on the phonetic context surrounding the potential word boundaries. Talkers' clarity enhancement resulted in a statistically significantly higher accuracy for the tokens containing potential word-boundaries between a consonant and a vowel or between schwa and a consonant, but not for the tokens containing word-boundaries between /s/ and a following consonant. Similar interactions between speech clarity and phonetic context have been revealed by the analyses of predictive model accuracy (Experiment 1), such as the three-way interaction among listener condition, CV type, and consonant condition on the accuracy of the less restrictive model (shown in Figure 4). Although the results from both experiments suggest that the speech clarity and phonetic context interact to affect the availability and informativeness of signal-based cues to word boundaries, the locus of

interaction differed across experiments. In the analyses of model accuracy, predictive model accuracy for tokens containing a word-boundary between schwa and a following consonant did not differ between listener conditions. In contrast, listeners' segmentation accuracy for tokens containing a word-boundary between schwa and a consonant differed across listener conditions. The differences in the locus of interaction across experiments suggest that listeners might have attended to signal-based cues to word boundaries other than durational properties that were used to estimate the extent to which talkers produced word-boundary cues for Experiment 1.

Finally, consonant condition--whether the segment at the word juncture was an obstruent or a sonorant—affected clarity rating. Listeners rated the tokens containing an obstruent at the potential word boundaries as more clearly produced than the tokens containing a sonorant at the potential word boundaries, probably because of allophonic variation between aspirated and unaspirated stops. The consonant condition, however, did not have statistically significant effects on the other dependent measures.

In summary, results from Experiment 2a confirmed that the availability and informativeness of signal-based segmentation cues, which were estimated by listeners' segmentation accuracy and response times in the two-alternative task and subjective clarity rating, differ depending on speech clarity and phonetic context surrounding the potential word boundaries. The finding that speech clarity and phonetic context surrounding the word boundaries contribute to the variability in the extent to which signal-based cues can guide listeners' word segmentation is consistent with the findings from Experiment 1 that speech clarity and phonetic context affect the extent to which

talkers produce durational cues to word boundaries. The converging evidence between the two experiments suggest that the extent to which talkers produce and listeners rely on signal-based segmentation cues vary depending on speech clarity and phonetic context of the potential word boundary. The findings also suggest that the apparent discrepancies in the literature regarding whether talkers produce sufficient acoustic-phonetic cues to word boundaries cues may result from talker variability conditioned by speech clarity and phonetic context of word boundaries.

3.3. Experiment 2b: Open-set transcription

In order to measure listeners' segmentation accuracy, Experiment 2a used a two-alternative forced choice task. Experiment 2b was conducted in order to test whether the forced-choice experiment accurately reflected what the listeners would hear without having the alternatives to choose from.

3.3.1. Methods

Participants

36 new participants who did not take part in any of the previous experiments were recruited from the same pools or invited for paid participation ($8).

Stimulus Materials

The stimulus materials and lists used for Experiment 2b were identical to the stimulus materials and lists used in Experiment 2a.

Procedures

For each trial in Experiment 2b, listeners heard an ambiguous phrase and typed out what they heard (cf. listeners heard an ambiguous phrase and performed a two-alternative forced-choice task followed by a clarity rating task in Experiment 2a).

Analysis

Listeners' responses were scored in two ways. First, each typed response was scored as accurate if the listener's typed response contained every segment that was present in the visual stimulus that the talkers read and contained a space indicating a word boundary at the location where the visual stimulus contained a word boundary. For instance, for the intended *he dyed*, written responses such as *he dyed* or *he died* were scored as accurate. However, written responses such as *hid eyed* or *he'd died* were scored as inaccurate. An additional, more lenient scoring method only focusing on the match between intended and reported location of the word boundary was used to measure the extent to which listeners heard the location of a word boundary, regardless of precisely what they heard. For instance, for the intended *he'd eyed*, a written response such as *hid I'd* or *he'd eye* were scored as accurately segmented.

3.3.2. Results

Listeners' typed responses contained all segments that the talkers intended to produce with the accurately perceived location of a word-boundary 59.28% of the time. 69.94% of the time, listeners accurately perceived the location of the word boundary. Regardless of how listeners' responses were scored, listeners' accuracy in the open-response transcription task was lower than listeners' accuracy in the forced-choice task (Experiment 2a, 77.05% accurate).

In order to test whether the forced-choice experiment accurately reflected what the listeners would hear without having the alternatives to choose from, a series of correlation analyses were performed. For each sound file in the stimulus set, the following measures were computed: listeners' mean accuracy in the forced-choice task (Experiment 2a), the mean accuracy in the transcription task using the strict scoring method, and the mean accuracy in the transcription task using the more permissive scoring method.

A mixed-effects linear regression model tested whether and to what extent the strict transcription accuracy was predicted by the accuracy in the forced-choice task. Talkers and items were treated as random effects. Listeners' accuracy in the forced choice task and listeners' accuracy in the open-response task showed a high correlation. The mixed-effects linear regression model revealed that the forced-choice accuracy is a statistically significant predictor of strictly scored transcription accuracy ($\beta = 0.82$; $t = 25.12$; pseudo $r^2 = 0.56$), suggesting that listeners' accuracy in the forced-choice task and open-set task are strongly correlated with each other. The strong correlation confirms that

the responses to the forced-choice task reflected what the listeners heard. Similarly, the forced-choice accuracy was revealed to be a statistically significant predictor of permissively scored open-set segmentation accuracy ($\beta = 0.79$; $t = 25.94$; pseudo $r^2 = 0.52$).

Although we found a high correlation between listeners' segmentation accuracy in the open-set transcription task and listeners' segmentation accuracy in the forced-choice task, this high correlation does not necessarily mean that both measures were affected by the same set of factors. In order to determine which factors influenced listeners' segmentation accuracy in the open-set transcription task, a series of repeated measures ANOVAs were performed on the listeners' accuracy in the transcription task.

Similar to the findings of Experiment 2a, listeners were more accurate in transcribing what they heard (i.e., accuracy using the strict scoring method) when the talkers directed their speech to the older hearing impaired (61.97% accurate) or the nonnative (59.47%) listener than when their speech was directed to the young native listener (56.41% accurate; $F1(2, 34) = 28.16$, $p < 0.001$; $F2(2, 34) = 16.06$, $p < 0.001$). Post-hoc pairwise comparisons revealed statistically significant differences in the listeners' strictly scored accuracy across all listener conditions ($t_{subj}(35) = -7.63$, $p < 0.001$; $t_{item}(35) = -6.12$, $p < 0.001$ for comparisons between the young native and the older hearing impaired listener conditions; $t_{subj}(35) = -4.48$, $p < 0.001$; $t_{item}(35) = -3.02$, $p < 0.01$ for comparisons between the young native and the young nonnative listener conditions; $t_{subj}(35) = 3.41$, $p < 0.01$; $t_{item}(35) = 2.20$, $p < 0.05$ for the comparisons between the older hearing impaired and the young nonnative listener conditions).

As was found in the analyses of forced-choice accuracy, listeners were least accurate in transcribing the tokens containing the potential word boundary between schwa and a following consonant (42.44% accurate; cf. 67.24% for the potential word-boundary between a consonant and a vowel; 69.93% for the potential word-boundary between /s/ and a consonant). The ANOVA on listeners' transcription accuracy revealed a main effect of CV type ($F1(2, 34) = 125.50$, $p < 0.001$; $F2(2, 34) = 15.27$, $p < 0.001$). Post-hoc pairwise comparisons revealed that listeners' transcription accuracy for the tokens containing potential word boundaries between schwa and a consonant was lower than the other two conditions ($t_{subj}(35) = 12.35$, $p < 0.001$; $t_{item}(22) = 4.68$, $p < 0.001$ for comparisons between the potential word boundaries between a consonant and a vowel and the potential word boundaries between a schwa and a consonant; $t_{subj}(35) = 12.83$, $p < 0.001$; $t_{item}(22) = 5.64$, $p < 0.001$ for comparisons between the potential word boundaries between /s/ and a consonant and the potential word boundaries between schwa and a consonant). The differences in the listeners' transcription accuracy between potential word boundaries between a consonant and a vowel and potential word boundaries between /s/ and a consonant were not statistically significant ($t_{subj}(35) = -1.70$, *ns*; $t_{item}(22) = -0.20$, *ns*).

Finally, the ANOVA on listeners' transcription accuracy revealed a three-way interaction among listener condition, CV type, and consonant condition ($F1(4, 32) = 5.94$, $p < 0.001$; $F2(4, 32) = 2.60$, $p < 0.05$). Figure 9 shows this three-way interaction among listener condition, CV type, and consonant condition on listeners' accuracy in the transcription task (strictly scored).

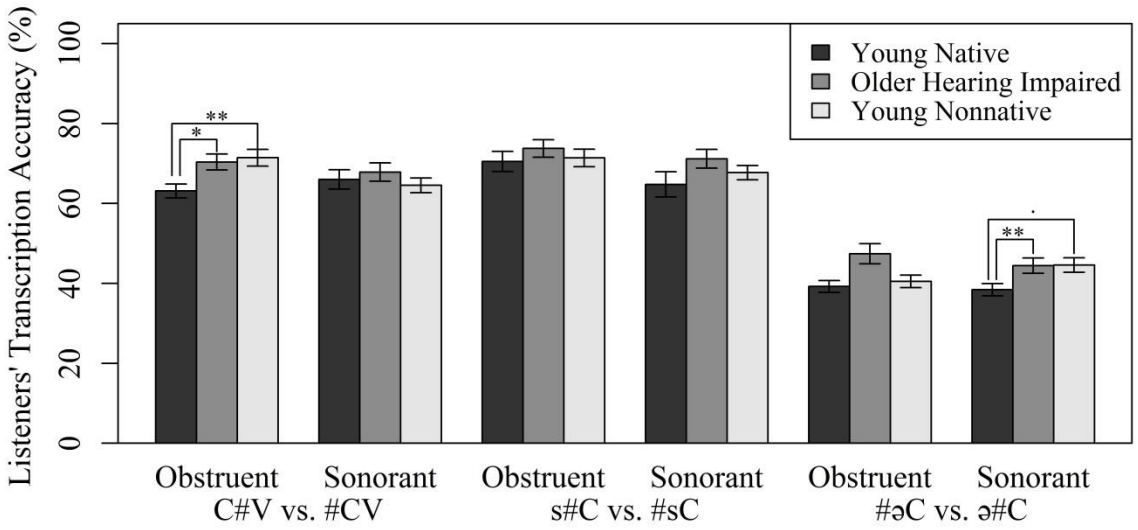

Figure 9. Listeners' accuracy in the transcription task by listener condition, CV type, and juncture consonant condition.

As can be seen in Figure 9, the extent to which listeners benefited from talkers' modulation of speech clarity due to listener condition (i.e., whom the talkers spoke to) differed depending on the phonetic context at (i.e., the consonant condition) and around (i.e., the CV type condition) the potential word boundaries. Post-hoc comparisons revealed that listeners' transcription accuracy significantly differed between the young native and the older hearing impaired listener conditions in two of the six CV type by consonant conditions: when the word boundary ambiguity occurred between a consonant and a vowel and the juncture segment was an obstruent (t_{subj} (35) = -4.92, $p < 0.001$; t_{item} (5) = -3.88, $p < 0.05$) and when the word boundary ambiguity occurred between schwa and a sonorant (t_{subj} (35) = -3.14, $p < 0.01$; t_{item} (5) = -4.27, $p < 0.01$). Post-hoc comparisons also revealed that listeners' transcription accuracy differed between the

young native and the young nonnative listener conditions when the word boundary ambiguity occurred between a consonant and a vowel and the juncture segment was an obstruent (t_{subj} (35) = -5.75, $p < 0.001$; t (5) = -5.39, $p < 0.01$). The difference in the listeners' accuracy in the transcription task between the young native and the young nonnative listener conditions marginally significant when the word boundary ambiguity occurred between schwa and a sonorant (t_{subj} (35) = -3.24, $p < 0.01$; t_{item} (5) = -2.44, $p = 0.06$). No differences in transcription accuracy between the older hearing impaired and the young nonnative listener conditions were revealed to be statistically significant by post-hoc comparisons.

For completeness, we performed a comparable set of analyses on the segmentation accuracy which is based on the more permissive scoring method only focusing on the location of reported word boundaries. As was found from the strictly scored accuracy, analyses on the permissively scored segmentation accuracy revealed statistically significant effects of listener condition ($F1$(2, 34) = 20.62, $p < 0.001$; $F2$(2, 34) = 8.86, $p < 0.001$) and CV type ($F1$(2, 34) = 114.30, $p < 0.001$; $F2$(2, 34) = 14.17, $p < 0.001$). Listeners' accuracy in the transcription task, when the responses were permissively scored, statistically differed across all listener conditions (t $_{subj}$ (35) = -6.47, $p < 0.001$; t_{item} (35) = -3.98, $p < 0.001$ for comparisons between the young native (67.11% accurate) and the older hearing impaired listener conditions (71.91%); t_{subj} (35) = -4.29, $p < 0.001$; t_{item} (35) = -2.06, $p < 0.01$ for comparisons between the young native (67.11%) and the young nonnative listener conditions (69.88%); t_{subj} (35) = 2.57, $p <$

0.05; t_{item} (35) = 2.16, $p < 0.05$ for comparison between the older hearing impaired and the young nonnative listener conditions).

Analyses of the permissively scored transcription accuracy revealed that listeners' segmentation accuracy for the ambiguous schwa-initial sequences was significantly lower than the two other CV types (t_{subj} (35) = 13.86, $p < 0.001$; t_{item} (22) = 5.54, $p < 0.001$ for comparisons between consonant-vowel and schwa-consonant conditions; t_{subj} (35) = 12.50, $p < 0.001$ for subject mean comparison; t (22) = 5.21, $p < 0.001$ for item mean comparison between /s/-consonant and schwa-consonant conditions). The comparisons between the consonant-vowel and the /s/-consonant conditions were not statistically significant.

The ANOVAs on the permissively scored transcription accuracy also revealed a marginally significant interaction between listener condition and CV type ($F1$(4, 32) = 4.81, $p < 0.01$; $F2$(4, 32) = 2.44, $p = 0.06$). As was found from the forced-choice task accuracy analyses of Experiment 2a (Figure 8), the extent to which listeners benefited from clarity modulation due to the talkers' directing their speech to listeners whose linguistic competence is suboptimal (i.e., older hearing impaired listener or nonnative listener conditions) depended on the phonetic context surrounding the word boundaries (i.e., CV type). The interaction for the permissive scoring in the transcription task is shown in Figure 10.

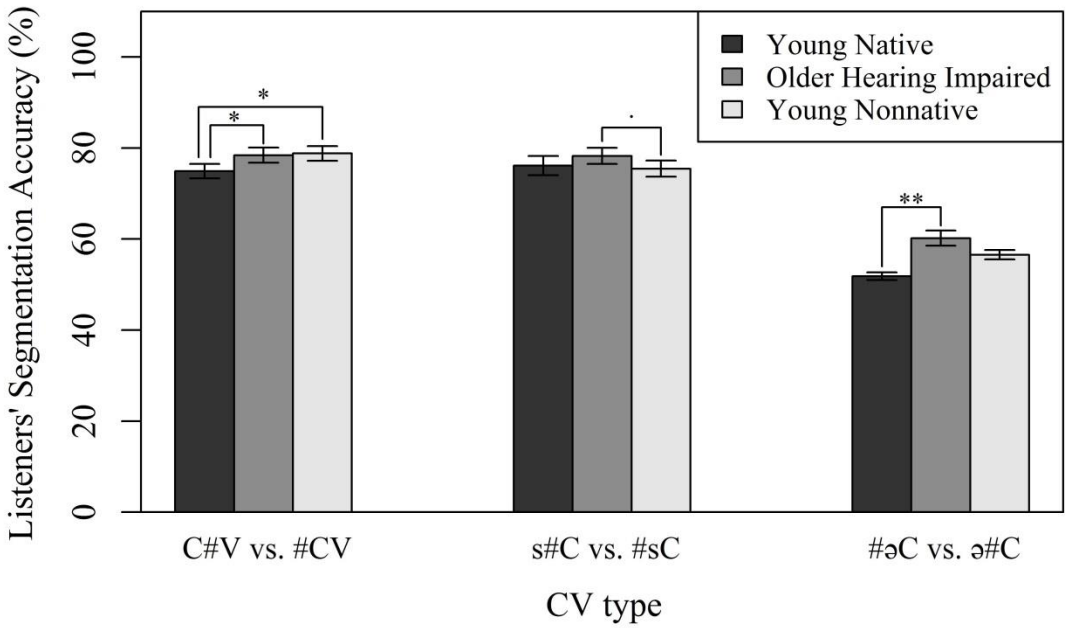

Figure 10. Listener condition by CV type interaction on listeners' word segmentation accuracy in the transcription task. ** indicates $p < 0.01$; and * indicates $p < 0.05$, and . indicates $p < 0.06$ for the post-hoc comparisons

Post-hoc pairwise comparisons on the permissively scored segmentation accuracy across listener conditions revealed that the extent to which listeners benefited from the talkers' clarity modulation depended on the phonetic context surrounding the potential word boundaries. More specifically, listeners' segmentation accuracy was significantly higher in the older hearing impaired listener condition than the young native listener condition when the potential word boundaries occurred between a consonant and a vowel (t_{subj} (35) = -3.86, $p < 0.001$; t_{item} (11) = -2.29, $p < 0.05$) and when the potential word boundaries occurred between schwa and a consonant (t_{subj} (35) = -4.93, $p < 0.001$; t_{item} (11) = -3.67, $p < 0.01$). In addition, listeners' segmentation accuracy was significantly higher in the young nonnative listener condition than the young native listener condition,

when the potential word boundaries occurred between a consonant and a vowel (t_{subj} (35) = -3.78, $p < 0.001$; t_{item} (11) = -2.29, $p < 0.05$). The differences in the listeners' segmentation accuracy between the older hearing impaired listener condition and the young nonnative listener condition were not statistically significant, except when the word-boundary occurred between /s/ and a consonant, where the accuracy differences between these two listener conditions was marginally significant (t_{subj} (35) = 2.28, $p < 0.05$; t_{item} (11) = 2.18, $p = 0.05$). Findings from Experiment 2a and 2b regarding the interactive effects of listener condition and CV type on listeners' segmentation accuracy suggest that listeners were more likely to benefit from clarity modulation for the tokens containing potential word boundaries between a consonant and a vowel or between schwa and a consonant than for the tokens containing potential word boundaries between /s/ and a consonant.

Finally, the ANOVAs on listeners' segmentation accuracy for transcription responses revealed that the interaction between CV type and consonant condition was significant by subjects but not by items ($F1(2, 34) = 4.26$, $p < 0.05$; $F2(2, 34) = 0.24$, *ns*). Listeners tended to be more accurate in reporting the intended word boundaries in their transcription responses when the token contained a sonorant segment than an obstruent segment and the word boundary occurred between a consonant and a vowel (79.17% accurate for tokens containing potential word boundaries between a sonorant and a vowel; cf. 75.67% accurate for tokens containing potential word boundaries between an obstruent and a vowel). In contrast, the reverse tendency was found for word-boundaries after /s/, where /s/-obstruent sequences were more likely to be accurately segmented

(77.28%) than /s/-sonorant sequences (75.93%), and for word-boundaries after schwa, where schwa-obstruent sequences were more likely to be accurately segmented (56.70%) than schwa-sonorant sequences (55.66%).

3.3.3. Summary of Experiments 2a and 2b

Results from Experiments 2a and 2b confirm that, regardless of the dependent measures or experimental paradigm, listener condition and phonetic context around potential word boundaries (i.e., CV type) had robust effects on how accurate listeners were in identifying the word boundaries intended by the talkers. The robust effects of listener condition and CV type are consistent with the findings from Experiment 1, which showed that the accuracy of statistical models predicting the word boundary location based on the durational properties of segments and pauses at and around the boundaries were higher when the "listener" condition suggests communicative difficulties and the word boundary occurred between schwa and a following consonant. Taken together, findings from Experiments 1, 2a, and 2b suggest that the availability and informativeness of signal-based cues to word boundaries depend on constraints pertaining to speech perception (i.e., listener needs depending on to whom the talkers spoke) and constraints pertaining to speech production (i.e., the inherent informativeness of the phonetic context surrounding word boundaries).

Statistical analyses of the accuracy measures (cf. reaction times or clarity rating scores) also revealed interactions between listener condition and phonetic context (i.e., CV type and / or consonant conditions), suggesting that the extent to which listeners

benefited from enhanced speech clarity was constrained by the phonetic context at and surrounding the potential word boundaries. Listeners' segmentation accuracy benefited from enhanced speech clarity, mainly for the tokens containing the potential word boundaries between a consonant and a vowel or between schwa and a consonant, but not for the tokens containing potential word boundaries between /s/ and a consonant, suggesting that some phonetic contexts undergo more clarity modulation than others.

A comparison between listeners' accuracy in the closed-set and the open-set transcription task revealed that listeners were on average 17.76% more accurate in the closed-set task than the open-set task, when the responses to the open-set transcription task were scored based on the strict criteria. Repeated measures ANOVAs on listeners' accuracy, with task type as a between-subject independent variable and listener condition, CV type, and consonant condition as within-subject independent variables revealed statistically significant main effects of task type ($F1(1, 71) = 179.50$, $p < 0.001$; $F2(1, 35) = 94.61$, $p < 0.001$), listener condition ($F1(2, 70) = 55.10$, $p < 0.001$; $F2(2, 34) = 20.23$, $p < 0.001$) and CV type ($F1(2, 70) = 470.02$, $p < 0.001$; $F2(2, 34) = 39.71$, $p < 0.001$).

When the responses to the open-set transcription task were scored based on the permissive criteria, the ANOVAs revealed a significant main effect of task type (7.41% higher accuracy in the forced-choice than the transcription task, $F1(1, 71) = 30.99$, $p < 0.001$; $F2(1, 35) = 31.77$, $p < 0.001$) as well as significant main effects of listener condition ($F1(2, 70) = 48.33$, $p < 0.001$; $F2(2, 34) = 14.86$, $p < 0.001$) and CV type ($F1(2, 70) = 520.54$, $p < 0.001$; $F2(2, 34) = 39.34$, $p < 0.001$). In addition, the ANOVAs on the forced-choice accuracy and permissively scored transcription accuracy revealed

significant interactions between listener condition and CV type ($F1(4, 68) = 9.02$, $p < 0.001$; $F2(4, 32) = 2.73$, $p < 0.05$), suggesting that the extent to which listeners benefited from enhanced speech clarity depended on the phonetic context around the word boundaries. Also revealed by the ANOVAs on listeners' accuracy was a statistically significant interaction between task type and CV type ($F1(2, 70) = 12.27$, $p < 0.001$; $F2(2, 34) = 4.18$, $p < 0.05$), suggesting that the extent to which listeners benefited from having alternatives to choose from (i.e., relative easiness of the task) depended on the phonetic context around the word boundaries. Listeners were 9.50% and 10.78% more accurate in the forced-choice than the transcription task (permissively scored) for the consonant + vowel and /s/ + consonant conditions. In contrast, for the tokens containing the potential word-boundary between schwa and a consonant, forced-choice accuracy was 2.50% higher than the transcription accuracy.

Finally, the mixed-effects linear regression analyses predicting listeners' segmentation accuracy in the transcription task based on listeners' accuracy in the forced-choice task revealed that listeners' accuracy in the forced-choice task was a statistically significant predictor of listeners' accuracy in the transcription task, regardless of whether the transcription responses were scored using the strict criteria (i.e., focusing on the recognition of every segment and the word boundaries) or the permissive criteria (i.e., focusing solely on the match between the intended and the perceived word boundaries). Since the results from Experiments 2a and 2b confirmed that listeners' responses to the forced-choice task reflected what they heard, in Experiment 3, we had listeners perform only the forced-choice task followed by the clarity rating task in order to investigate

factors affecting listeners' segmentation accuracy for the tokens produced in the aware condition in Experiment 1.

3.4. Experiment 3: two-alternative forced choice and clarity rating for stimuli produced by talkers who were aware of the word-boundary ambiguity

3.4.1. Methods

Participants

36 new participants who did not take part in any of the previous experiments were recruited from the same pools or invited for paid participation ($8).

Stimulus Materials

The stimulus materials and lists used for Experiment 3 were comparable to the stimulus materials and lists used for Experiment 2a, but the stimuli were sampled from the production data generated by the talkers from the "aware" group in Experiment 1.

Procedures

The procedures used for Experiment 3 were identical to the procedures used in Experiment 2a.

3.4.2. Results

3.4.2.1. Segmentation Accuracy

For the stimuli produced by talkers who saw a pair of sentences containing a word-boundary "minimal pair" on each trial, participants' segmentation accuracy in the two-alternative forced choice task was 81.39%, indicating that listeners' segmentation accuracy was higher than chance level (50%), than the accuracy of the less restrictive model reported in Experiment 1 (75.64%), and than the accuracy of listeners who heard stimuli produced by talkers who saw a target sentence containing an ambiguous target phrase and a sentence that is not related to the target sentence (Experiment 2a, 77.05%; $t_{subj}(71) = -4.04$; $p < 0.001$; $t_{item}(35) = -4.18$; $p < 0.001$).

Participants' segmentation accuracy differed depending on the listener condition and CV type. Segmentation accuracy was lowest when talkers directed their speech to a young native listener (77.83%) and highest when talkers directed their speech to an older hearing impaired listener (82.98%) or a young nonnative listener (83.39%). Statistical analysis revealed a significant main effect of listener condition on average segmentation accuracy ($F1(2, 34) = 34.81$, $p < 0.001$; $F2(2, 34) = 12.21$, $p < 0.001$). Post-hoc pairwise comparisons revealed statistically significant differences in listeners' accuracy between the young native listener condition and the clear speech conditions ($t_{subj}(35) = -6.61$, $p < 0.001$; $t_{item}(35) = -4.15$, $p < 0.001$ for comparisons between the young native listener condition and the older hearing impaired listener condition; $t_{subj}(35) = -6.75$, $p < 0.001$; $t_{item}(35) = -3.93$, $p < 0.001$ for comparisons between the young native and the young nonnative listener conditions). Post-hoc pairwise comparisons did not reveal a significant

difference in the segmentation accuracy between the older hearing impaired and the young nonnative listener conditions (t_{subj} (35) = -0.74, *ns*; t_{item} (35) = -0.35, *ns*). The effect of listener condition on listeners' accuracy in the forced-choice task was also revealed by the results of Experiment 2a.

In addition, listeners were much more accurate in segmenting the potential word boundary between a consonant and a vowel (e.g., *beef eater* vs. *bee feeder*; 89.81% accurate) or the boundary between /s/ and a following segment (e.g., *collects gulls* vs. *collect skulls*; 88.80%) than the boundary between schwa and a following consonant (e.g., *along* vs. *a long*; 67.27%). The main effect of CV type on the participants' segmentation accuracy was statistically significant (*F1*(2, 34) = 261.80, $p < 0.001$; *F2*(2, 34) = 47.97, $p < 0.001$). Post-hoc pairwise comparisons on listeners' segmentation accuracy revealed that segmentation accuracy for the ambiguous schwa-initial sequences was significantly lower than the two other CV types (t_{subj} (35) = 12.50, $p < 0.001$; t_{item} (22) = 8.83, $p < 0.001$ for comparisons between /s/-consonant and schwa-consonant conditions; t_{subj} (35) = 13.86, $p < 0.001$; t_{item} (22) = 8.24, $p < 0.001$ for comparisons between consonant-vowel and schwa-consonant conditions). The comparisons between the consonant-vowel and the /s/-consonant conditions were not statistically significant (t_{subj} (35) = -1.06, *ns*; t_{item} (22) = 0.47, *ns*). The findings that CV type had a main effect on listeners' accuracy and that schwa-initial sequences were most likely to be heard inaccurately are consistent with the findings from Experiment 2a.

No interactions reached significance in both subject and item analyses. The CV type by consonant condition interaction was significant only by subjects (*F1*(2, 34) =

1.06, $p < 0.01$; $F2(2, 34) = 2.00$, *ns*). The three-way interaction among listener condition, CV type, and consonant condition was also significant by subjects, but not by items ($F1(4, 32) = 2.48$, $p < 0.05$; $F2(4, 32) = 1.01$, *ns*; cf. a marginally significant interaction between listener condition and CV type on listeners' forced-choice accuracy, Experiment 2a, reported in section 3.2.2.1.).

3.4.2.2. Response times in the two-alternative forced choice task

Listeners' response times were faster when the stimuli were spoken to an older hearing impaired (1214 ms) or a nonnative (1219 ms) listener than when the stimuli were spoken to a young native listener (1258 ms; $F1(2, 34) = 12.18$, $p < 0.001$; $F2(2, 34) = 6.58$, $p < 0.01$). Post-hoc comparisons revealed that the response times in the forced-choice task differed between the plain speech condition (i.e., the young native listener condition) and the clear speech conditions (i.e., the older hearing impaired and the young nonnative listener conditions; $t_{subj}(35) = 4.60$, $p < 0.001$; $t_{item}(35) = 3.18$, $p < 0.01$ for comparisons between the young native and the older hearing impaired listener conditions; $t_{subj}(35) = 3.86$, $p < 0.001$; $t_{item}(35) = 2.59$, $p < 0.05$ for comparisons between the young native and the young nonnative listener conditions). However, post-hoc pairwise comparisons did not reveal a significant difference in listeners' response times between the older hearing impaired and the young nonnative listener conditions ($t_{subj}(35) = -0.43$, *ns*; $t_{item}(35) = -0.28$, *ns*).

The finding that tokens spoken towards the older hearing impaired listener or the young nonnative listener were responded to faster than tokens spoken towards the young

native listener in the forced-choice task is consistent with the findings from Experiment 2a (reported in section 3.2.2.2.): listeners benefited from talkers' enhancing the clarity of acoustic-phonetic cues demarcating word boundaries according to listener needs, regardless of whether the talkers were aware of the word-boundary ambiguity or not. The main effects and interactions of other independent variables—CV type and consonant condition —failed in reaching significance in both subject and items analyses. The three-way interaction among listener condition, CV type, and consonant condition was significant by subjects, but not by items ($F1(4, 32) = 4.58$, $p < 0.01$; $F2(4, 32) = 1.84$, *ns*).

3.4.2.3. Clarity Rating

As in Experiment 2a, for each sound file, accuracy and clarity rating score were averaged over participants. A mixed-effects linear regression analysis was performed to test whether and to what extent average clarity rating predicts average accuracy in the two-alternative forced choice task. Talkers and items were treated as random factors. The mixed-effects linear regression analysis revealed that the average clarity rating score is a statistically significant predictor of average forced-choice accuracy ($\beta = -15.48$; $t = -12.25$; pseudo $r^2 = 0.41$), suggesting that tokens that were rated as clearly spoken were more likely to be accurately segmented. This finding is consistent with the findings from Experiment 2a (section 3.2.2.3.) and suggests that subjective clarity rating may offer a good estimate of segmentation accuracy.

A repeated measures ANOVA on average clarity rating score revealed significant main effects of listener condition ($F1(2, 34) = 59.88$, $p < 0.001$; $F2(2, 34) = 36.89$, $p < 0.001$) and CV type ($F1(2, 34) = 17.44$, $p < 0.001$; $F2(2, 34) = 12.56$, $p < 0.001$). Participants rated tokens spoken to an older hearing impaired (mean clarity rating score = 1.97) or a nonnative (1.98) listener as being more clearly spoken than tokens spoken to a young native listener (2.23). Post-hoc pairwise comparisons revealed statistically significant differences in the clarity rating score between the young native listener condition and the two clear speech conditions ($t_{subj}(35) = 8.88$, $p < 0.001$; $t_{item}(35) = 7.99$, $p < 0.001$ for comparisons between the young native and the older hearing impaired listener conditions; $t_{subj}(35) = 7.85$, $p < 0.001$; $t_{item}(35) = 6.79$, $p < 0.001$ for comparisons between the young native and the young nonnative listener conditions). The difference in the clarity rating between the two clear speech conditions was not statistically significant ($t_{subj}(35) = -0.57$, *ns*; $t_{item}(35) = -0.48$, *ns*).

With regards to CV type, schwa-initial ambiguous sequences, which were least likely to be accurately segmented by talkers, were perceived as least clear (2.24; cf., mean rating scores of 1.89 for potential word boundaries between a consonant and a vowel; 2.03 for potential word boundaries between /s/ and a consonant). Post-hoc pairwise comparisons revealed that the tokens containing potential word boundaries between schwa and a consonant were rated as less clearly spoken than the other two conditions ($t_{subj}(35) = -5.41$, $p < 0.001$; $t_{item}(22) = -6.19$, $p < 0.001$ for comparisons between the potential word boundaries between a consonant and a vowel and potential word boundaries between schwa and a consonant; $t_{subj}(35) = -2.91$, $p < 0.01$; $t_{item}(22) =$

-3.03, $p < 0.01$ for comparisons between the potential word boundaries between /s/ and a consonant and word boundaries between schwa and a consonant). The differences in the clarity rating between potential word boundaries between a consonant and a vowel and potential word boundaries between /s/ and a consonant were statistically significant by subjects but not by items (t_{subj} (35) = -4.32, $p < 0.001$; t_{item} (22) = -1.60, *ns*).

The findings that listeners' clarity rating depended on listener condition and CV type are consistent with the findings from Experiment 2a. However, contrary to what was found from Experiment 2a, the ANOVAs did not reveal statistically significant differences in the clarity rating scores between consonant conditions. The main effect of consonant condition failed in reaching statistical significance by items ($F1$(1, 35) = 18.07, $p < 0.001$; $F2$(1, 35) = 1.36, *ns*). Likewise, the interaction between listener condition and CV type ($F1$(4, 32) = 3.66, $p < 0.001$; $F2$(4, 32) = 1.50, *ns*) and the interaction among listener condition, CV type, and consonant condition ($F1$(4, 32) = 2.56, $p < 0.05$; $F2$(4, 32) = 0.62, *ns*) on listeners' clarity score were significant in the subject analyses but not the items analyses.

3.4.3. Comparison of Experiments 2a and 3

As was found in Experiment 2a, listener condition had robust effects on all three dependent measures (i.e., accuracy, response time, and clarity rating score) and CV type also had robust effects on the accuracy and clarity rating scores. In addition to these similarities across experiments, the high correlation between clarity rating scores and forced-choice accuracy from Experiment 2a was also replicated.

In the current study, awareness was a between-subject variable and thus the effect of awareness on listeners' perception of word boundaries could only be estimated by comparing participants across Experiments 2a and 3. A series of repeated measures ANOVAs on listeners' segmentation accuracy, with awareness as a between-subject independent variable and listener condition, CV type, and consonant condition as within-subject independent variables, revealed significant main effects of talkers' awareness of word-boundary ambiguities, listener condition, and CV type. Listeners were more likely to accurately segment the stimuli produced by talkers who were aware of the word-boundary ambiguity (81.40% accurate) than the stimuli produced by talkers who were unaware of the word-boundary ambiguity (77.05%; $F1(1, 71) = 16.92$, $p < 0.001$; $F2(1, 35) = 19.15$, $p < 0.001$).

For completeness, ANOVAs comparable to the preceding analyses were performed on response times and clarity rating score. Listeners were faster in responding to the stimuli produced by talkers who were aware of the word-boundary ambiguity (1230 ms) than the stimuli produced by talkers who were unaware of the word-boundary ambiguity (1313 ms), but the difference was not statistically significant in the subject analysis ($F1(1, 71) = 1.24$, *ns*; $F2(1, 35) = 80.70$, $p < 0.001$), probably because awareness was a between-subject variable thus had less statistical power. Listeners rated the stimuli produced by talkers who were aware of the word-boundary ambiguity as being more clearly spoken (2.06) than the stimuli produced by talkers who were unaware of the word-boundary ambiguity (2.31), and the difference was statistically significant ($F1(1, 71) = 5.89$, $p < 0.05$; $F2(1, 35) = 67.26$, $p < 0.001$).

Due to the design of the current study, where awareness was a between-listener variable, it is possible that the differences in listeners' segmentation accuracy between the aware and unaware conditions are due to differences in the participants. In addition, the criteria used to sample stimuli from the production data were goodness of regression model fit (i.e., % accurate classification of the less restrictive model) and rate of speech. Though these two measures are highly correlated with human listeners' segmentation accuracy ($\beta = 0.54$, $r^2 = 0.52$ for the correlation between the listeners' segmentation accuracy in the forced-choice task and the less restrictive model accuracy for each talker; $\beta = 1291.8$, $r^2 = 0.43$ for the correlation between listeners' segmentation accuracy and average duration of the ambiguous phrase), they may not correspond exactly to listeners' word segmentation accuracy. Future research should scrutinize the effect of awareness by using more talkers' productions as stimuli and treating awareness as a within-participant variable.

That awareness had a statistically significant effect on listeners' word segmentation, especially on the segmentation accuracy, appears to contradict the findings from Experiment 1, where awareness was found to have no effect on the predictive model accuracy, regardless of whether the model was conservative or less restrictive. This apparent discrepancy might be due to the choice of predictor variables in Experiment 1. The predictor variables of the statistical model were durations of segments and optional pauses at and surrounding the potential word boundaries. The finding that awareness did not affect model accuracy but affected listeners' segmentation accuracy might suggest

that awareness modulated the production of signal-based segmentation cues that cannot be captured by variation in duration and that have perceptual consequences.

3.5. Discussion

There has been a long-standing debate concerning whether and to what extent the speech signal influences listeners' perception of word boundaries. While previous studies reporting listeners' high segmentation accuracy typically suggested that the speech signal is rich in the acoustic-phonetic markers to word boundaries (Lehiste, 1960; Hoard, 1966; Gow & Gordon, 1995; Davis, Marslen-Wilson & Gaskell, 2002; Salverda, 2005; Shatzman & McQueen, 2006), studies reporting that successful word segmentation cannot be guided solely by the acoustic-phonetic signal have suggested that the speech signal may not be a reliable source of information regarding word boundaries (Reddy, 1976; Kim et al., 2012; Norris, McQueen, Cutler, & Butterfield, 1997; McQueen, 1998; Dilley & McAuley, 2008).

Focusing on the lack of consensus regarding the extent to which signal-based cues affect listeners' perception of word boundaries and the fact that signal-based cues are produced by talkers and therefore should be susceptible to vary according to the factors conditioning the acoustic-phonetic variation, the current study hypothesized that the availability and informativeness of signal-based segmentation cues may vary depending on speech clarity and phonetic context at and around the word boundaries. The availability and informativeness of signal-based segmentation cues were estimated by

listeners' perception of word boundaries and was compared across speech clarity and phonetic context conditions. Results of the perception studies are summarized in Table 3.

As can be seen from Table 3, listeners were more accurate and faster in segmenting speech directed to listeners whose linguistic background suggests communicative difficulties. This finding is consistent with the findings of Experiment 1, where the availability of signal-based segmentation cues was estimated by the accuracy of predictive models predicting the word boundary locations based on durational properties of segments and pauses at and around the word boundaries. These findings are consistent with the findings of clear speech studies in that talkers' enhancing speech clarity facilitates speech comprehension (Perkell, Zandipour, Matthies, & Lane, 2002; Ferguson & Kewley-Port 2002; Bradlow, Kraus, & Hayes, 2003; Krause & Braida 2004; Smiljanić & Bradlow 2005). Given that word segmentation is an integral part of speech comprehension, speech produced with enhance clarity is expected to be segmented more accurately and easily, and these predictions were confirmed by the results of the perception experiments. In addition, results of the current study extend studies of clear speech by focusing on the perception of word boundaries and suggesting that one perceptual advantage that clear speech has is containing more salient signal-based cues demarcating word boundaries.

			Main Effects			Interactions			
Experiment	Dependent Measure	Tests	Listener	CV Type	Segment Type	Listener by CV type	Listener by segment type	CV by segment type	Listener by CV type by Segment type
Experiment 2a	Accuracy	Subj	✓	✓	✓	✓			
		Items	✓	✓		$p = 0.06$			
	Response Time	Subj	✓	✓	✓			✓	
		Items	✓						
	Clarity Rating	Subj	✓	✓	✓			✓	
		Items	✓	✓	✓				
Experiment 2b	Accuracy (strict)	Subj	✓	✓					✓
		Items	✓	✓					✓
	Segmentation accuracy	Subj	✓	✓		✓		✓	
		Items	✓	✓		$p = 0.06$			
Experiment 3	Accuracy	Subj	✓	✓			✓		✓
		Items	✓	✓					
	Response Time	Subj	✓						✓
		Items	✓						
	Clarity Rating	Subj	✓	✓	✓	✓			✓
		Items	✓	✓					

Table 3. Results from the statistical tests on the data from perception experiments. ✓ indicates statistical significance at $\alpha = 0.05$

The finding that speech clarity is one source of variation in the availability and informativeness of acoustic-phonetic cues to word boundaries also suggests that speech clarity might account for the inconsistencies in the literature regarding whether or not signal-based cues are sufficient to guide word segmentation. For instance, perception studies using stimuli generated from a reading task (Lehiste, 1960; Hoard, 1966; Barry, 1981, among others) have typically reported that listeners distinguish word-boundary minimal pairs. In contrast, studies using stimuli generated from a task where talkers produced speech under memory load (Kim et al., 2012) reported that talkers do not produce sufficient cues for word segmentation. Such apparent discrepancy might suggest that speech clarity is one source of variation in the extent to which signal-based cues can guide listeners' word segmentation.

Also found from the perception experiments is that the informativeness of signal-based cues differ depending on phonetic context of the word boundaries. Listeners' segmentation accuracy and perceived clarity was affected by CV type, mainly because schwa-initial sequences, unlike other CV types, were segmented with low accuracy. These findings are consistent with the findings of Experiment 1, where model accuracy for the ambiguous schwa-initial sequences was much lower than model accuracy for the other CV types. The difficulty with segmenting schwa-initial sequences is due to the relative lack of durational cues marking the presence or absence of a word boundary between schwa and the following segment. As shown in Experiment 1, the non-initial juncture segment after the schwa vowel (i.e., /l/ in *a*l*ong*) is long, relative to the non-initial segments that occur after /s/ or after a vowel. Such "lengthening" of non-initial

juncture segments after schwa is caused by the stress pattern of the schwa-initial word, where the second syllable is stressed and thus the onset of the stressed syllable undergoes lengthening (White, 2002). The lengthening due to stress pattern within a single word and the lengthening due to the presence of a word boundary may result in listeners' missegmentation.

The finding that phonetic context surrounding a word boundary affects the availability and informativeness of signal-based segmentation cues suggest that phonetic context might have contributed to the discrepancies in the literature regarding the extent to which word segmentation can be guided by acoustic-phonetic cues to word boundaries. As summarized in the methods section of Experiment 1, the stimuli used for the current study were selected from those that were used in the previous studies: consonant + vowel sequences from Lehiste (1960), /s/ + consonant sequences from Smith (2004), who suggested the presence and importance of signal-based segmentation cues on word segmentation, and schwa + consonant sequences from Kim et al. (2012), who suggested that talkers rarely produce signal-based segmentation cues that listeners can attend to and use to hear word boundaries. Taken together, the apparent discrepancies in the literature regarding the production and perception of signal-based segmentation cues may indeed be variation in the availability and informativeness of signal-based segmentation cues conditioned by phonetic contexts of the word boundaries.

The main effects of listener condition and CV type indicate that the extent to which talkers produce perceptible cues to word boundaries is constrained by constraints pertaining to speech perception (i.e., who the listeners were) as well as constraints

pertinent to production (i.e., the phonological context surrounding the word boundary). The main effects of listener condition and CV type were found in the analyses on accuracy of predictive models (Experiment 1) as well as listeners' segmentation accuracy (Experiments 2a, 2b, and 3), providing converging evidence. However, results of the two sets of analyses did not always converge. For instance, analyses on model accuracy did not reveal effects of talkers' awareness, while those on listeners' segmentation accuracy did. One possible explanation for this discrepancy across analyses could be because signal-based cues were more narrowly defined in the analyses of model accuracy. The predictor variables for the predictive model were durational measures, which might affect listeners' perception of word boundaries, but listeners might have relied on acoustic-phonetic properties other than durational properties. Similarly, both the analyses of model accuracy and the analyses of listeners' segmentation accuracy revealed that listener condition and CV type interact to affect the availability and informativeness of signal-based segmentation cues, but the locus of interaction differed. In order to explore reasons why and how the predictive models and listeners' word boundary perception differ from each other, it would be necessary to examine the precise mapping between acoustic details and the perception of word boundaries, which is left for future research.

In Experiments 2a and 2b (i.e., the experiments using stimuli produced by talkers who were unaware of the word-boundary ambiguity), we found a small but consistent effect of listener condition as well as a marginally significant interaction between listener condition and CV type on listeners' segmentation accuracy. For the potential word boundaries between a consonant and a vowel and for the potential word boundaries

between /s/ and a consonant, listeners' segmentation accuracy benefited relatively little from talkers' producing clear speech: 4.14% and 3.68% more accurate in the clear speech conditions, respectively. However, for the word boundaries between schwa and a following consonant, listeners' segmentation accuracy improved by 8.71% (52.90% vs. 61.61%) from the young native listener condition to the clear speech conditions. The small difference between speaking styles (i.e., listener conditions) and disproportional clear speech benefit across CV conditions might suggest that even in the plain speech condition, acoustic cues to word segmentation were ample, except for the ambiguous schwa-initial sequences. Thus, in order to estimate the availability of signal-based word segmentation cues, we need to consider the phonological contexts surrounding word boundaries and the degree to which talkers may hypoarticulate cues to word boundaries in certain contexts.

Unlike the CV type, consonant condition was not shown to affect listeners' segmentation accuracy. For instance, the juncture consonant condition —whether the segment at the word juncture was an obstruent or a sonorant—did not have a significant effect on segmentation accuracy or response times. However, listeners rated sequences containing an obstruent juncture segment as more clearly spoken. The tendency to hear ambiguous sequences containing an obstruent juncture segment as being more clearly spoken than ambiguous sequences containing a sonorant segment may be attributable to allophonic variation between aspirated and unaspirated stop consonants, which may provide listeners with useful cues to word boundaries. The lack of a significant effect of the consonant condition, especially in the items analyses, might also be due to the

inherent nature of the consonant condition variable, which is a property of an item and thus is a between-item variable. In order to draw firm conclusions regarding the effect of juncture consonant condition on the availability of acoustic word segmentation cues, it would be necessary to conduct a follow-up study with more items in each condition.

To summarize, the current study suggested that speech clarity and phonetic context are potential sources of variation in the extent to which talkers produce signal-based cues to word boundaries, and consequently, sources of variation in the extent to which listeners rely on the speech signal for word segmentation. Results suggested that signal-based segmentation systematically varies depending on speech clarity and phonetic context, helping to understand the apparent discrepancies in the literature regarding listeners' use of signal-based cues for word segmentation as resulting from systematic talker variability.

Chapter 4: Discussion

4.1. Summary of the research

This dissertation investigated the production and perception of signal-based cues to word boundaries. Although the production and perception of word boundaries have extensively been investigated in the literature, no consensus has been reached concerning the availability and informativeness of signal-based cues to word boundaries. Focusing on the fact that signal-based segmentation cues are produced by talkers, it was hypothesized that the acoustic-phonetic realizations of signal-based segmentation cues would display systematic intra-talker variability and that the variation might have contributed in the discrepancies in the past literature. Among the potential sources of talker variability, the current study focused on two sources of intra-talker variability, speech clarity and phonetic context, and aimed to provide a detailed picture of the variation in the production and perception of signal-based segmentation cues.

Results from the production experiment revealed that talkers adjust the extent to which they produce durational cues distinguishing word-boundary minimal pairs depending on their listeners' linguistic background, the phonetic context surrounding the word boundaries, and the interaction between these two factors.

Results from the perception experiments revealed that the talkers' clarity modulation in the acoustic-phonetic cues to word boundaries has perceptual consequences. Listeners were more accurate and faster in determining the location of the

word boundary when the talkers' speech was directed to an older hearing impaired or a young nonnative listener than when speech was directed to a young native listener. In addition, listeners were better at determining the location of the word boundaries when the stimuli were produced by talkers who read the target sentences after reading the sentence pairs containing word-boundary minimal pairs. The extent to which listeners benefited from enhanced speech clarity due to the listener confederates' linguistic background or talkers' awareness of the word-boundary ambiguity differed depending on the phonetic context surrounding the potential word boundaries.

Taken together, the results from the current study suggest that the production of signal-based cues to word boundaries is contingent upon constraints pertinent to speech perception, such as listeners' linguistic background or knowledge of the specific ambiguity that listeners would face, as well as constraints pertinent to speech production, such as inherent phonetic properties of segments or sequencing of speech sounds surrounding potential word boundaries. The findings that speech clarity and the phonetic contexts of potential word boundaries lead to variation in the extent to which signal-based segmentation cues are produced by talkers and used by listeners help understand the apparent inconsistencies in the literature regarding the production and perception of signal-based segmentation cues as resulting from systematic talker variability.

New findings from the current study shed light on existing models of speech production. In addition, results from the current study suggest that models of word segmentation should take into account the variation in the availability and informativeness of acoustic-phonetic cues to word boundaries.

4.2. Implications for speech production

Results from the current study are consistent with Lindblom's (1990) theory of Hypo- and Hyper- articulation (henceforth H&H theory). In H&H theory, speech production is conceptualized as a balancing act between talkers' minimizing articulatory effort and still maintaining successful communication for their listeners. To elaborate, H&H theory suggests that constraints pertaining to speech perception lead talkers to produce speech at a certain level of clarity so that their speech contains sufficient acoustic cues for listeners to ensure successful speech comprehension. At the same time, production constraints lead talkers to conserve effort so that speech production is not too energy-consuming.

Results of the production experiment conforms to the H&H theoretic predictions that constraints pertinent to speech production and constraints pertinent to speech perception affect the acoustic-phonetic properties of speech sounds independently and in combination. In addition, results of the production study suggest that the H&H theory can be refined and expanded to accommodate the new findings. In the current study, perception constraints were manipulated by two factors, listeners' linguistic profile and talkers' awareness of the word-boundary ambiguity (i.e., the specific difficulty that listeners would face). The motivation for this manipulation was to acknowledge the multi-faceted nature of speech clarity and to test how the diverse motivations for perception constraints result in clarity modulation. The current study found that listeners' linguistic profile had a significant main effect on the availability of durational cues to

word boundaries, while awareness of the word-boundary ambiguity did not modulate the production of durational cues to word boundaries. This finding suggests that perception constraints may differ in the phonetic property that they modulate and/or the extent to which they affect speech production. It is also suggested that the H&H theory can be refined by orthogonally manipulating multiple types of perception constraints and investigating how the perception constraints affect speech production independently and in combination.

In addition, we manipulated production constraints by two factors: phonetic context surrounding the word boundaries (i.e., CV type) and whether the juncture segment was an obstruent or a sonorant (i.e., consonant condition). Results suggested that some segments are more informative than other segments (e.g., stops were inherently more informative than fricatives), and informativeness of segments regarding the location of a word boundary depends on the segments surrounding the word boundaries. Based on these findings, we suggest that an accurate model of connected speech production should acknowledge the finding that the extent to which acoustic-phonetic details indicate discontinuity between lexical items depends on the phonetic properties of segments surrounding potential word boundaries.

Also observed in the production study was the interaction between phonetic context and speech clarity, which suggests that perception constraints interact with production constraints. This finding is consistent with both Lindblom's (1990) proposal and empirical studies demonstrating that speech clarity adjustments target different sounds to different degrees (Moon & Lindblom, 1994; Uchanski, 1988 and Uchanski et

al., 1992; for discussion of the conversation-to-clear speaking style adjustments of English lax versus tense vowels). Taken together, the current study extends previous studies of talker and listener constraints on variability in word and segment productions (typically vowel productions) to connected speech (i.e., sequences of words), which is understudied relative to smaller linguistic units.

4.3. Implications for speech perception and word segmentation

Results from the current study also shed light on models of word segmentation. Mattys et al.'s (2005; 2007) studies have suggested that word segmentation involves integrating signal-based and knowledge-based cues according to their relative strength. Given that the availability of signal-based cues differs across speech clarity and phonetic contexts, the extent to which knowledge-based cues are necessary for word segmentation may vary depending on speech clarity and phonetic context at and around word boundaries.

Among the three CV types that this study investigated, schwa-initial ambiguous sequences differed from the other CV types in that the word-boundary ambiguity was least likely to be resolved by signal-based cues. However, although CV type affects the extent to which talkers produce signal-based segmentation cues and listeners rely on acoustic-phonetic cues for word segmentation, it is unknown which CV type is a better representation of connected speech. In addition, if the availability and informativeness of signal-based cues to word boundaries vary depending on the segments that occur at and around word boundaries, the availability of signal-based segmentation cues should

fluctuate as speech unfolds. Given that the availability and strength of signal-based cues to word boundaries vary, how should a model of word segmentation handle such variation? Ideally, a good word segmentation model should be robust against acoustic-phonetic variation, including the variation in the extent to which signal-based segmentation cues are produced by talkers.

One solution to this problem involves positing a model that is robust against variation in the availability of signal-based cues to word boundaries, so that words can be segmented and recognized even in the absence of signal-based segmentation cues. Mattys' hierarchical model (2005), which assumes that knowledge-based segmentation cues always dominate signal-based cues, might be one example of a model that is robust against talker variability. However, given that knowledge-based cues such as semantic context are, in fact, sequences of accurately segmented words, and that knowledge-based cues typically builds up as communicative exchanges proceed, a model that relies too heavily on the knowledge-based cues would not be ideal.

An alternative solution would be to posit a segmentation model that weighs the relative importance of signal-based and knowledge-based cues in a flexible manner. In other words, the extent to which the model recruits different types of segmentation cues would depend on the relative strength of different types of segmentation cues (Mattys et al., 2007). Such a model would recruit more knowledge-based information as it detects that signal-based cues are insufficient in the speech signal, for instance, because there was a potential word-boundary ambiguity between schwa and a following consonant. Although it can avoid problems such as the lack or insufficiency of knowledge-based

cues, such a model might involve complicated computation continuously evaluating relative strength of signal- versus knowledge-based cues to word boundaries.

Yet another alternative solution is to posit a model that does not require ambiguity resolution at every word boundary, by not immediately ruling out other ways to parse word boundaries from connected speech unless there are very strong acoustic cues signaling word boundaries are present in the speech signal (Gow & Gordon, 1995). Although the current study does not provide sufficient evidence to claim that one model is better than the other, findings of the current study helps identify potential problems that an ideal word-segmentation model is required to handle.

The findings from the current study regarding the availability of signal-based segmentation cues across phonetic contexts are still limited, in that only a handful of phonetic contexts were investigated. In order to obtain a precise estimate regarding the extent to which knowledge-based cues are necessary to resolve a potential word-boundary ambiguity, it would be necessary to study more diverse phonetic contexts.

In this study, we used two methods to estimate the extent to which signal-based cues to word boundaries are made available: statistical models of the production data as well as perception experiments. The two sets of analyses produced converging evidence suggesting the robust effects of listener condition and CV type on the availability and informativeness of acoustic-phonetic cues to word boundaries. However, with regards to the effects of awareness, the statistical models suggested the lack of significant effects of awareness on the availability of durational cues to word boundaries, while the perception experiments suggested that awareness had a small yet statistically significant effect on

listeners' segmentation accuracy. Due to this discrepancy, it is inconclusive as to whether awareness affected the production of word boundaries. At most, what can be concluded is that awareness might have modulated non-durational cues or acoustic-phonetic properties that were not measured and used as predictor variables of the statistical model. Several explanations may be offered to account for the discrepancy or mismatch between results from production and perception experiments. First, as also suggested by previous studies, duration is not the only acoustic-phonetic property corresponding to listeners' perception of word boundaries, though duration is highly correlated with a number of acoustic-phonetic properties that co-occur with word boundaries. Second, the statistical model could have underestimated the perceptual consequences of some durational cues that occur rather sporadically, such as optional pauses between word boundaries. Relatedly, it could be the case that some small durational differences might be perceptually very salient, which would lead the statistical model to diverge from human listeners' perception of word boundaries. The precise mapping between acoustic details and the perception of word boundaries is left for future research.

4.4. Implications for methods of eliciting clear speech

The current study also has methodological implications, since it used an elicitation method that facilitates talkers' estimation of listeners' linguistic competence. Many previous studies have used explicit instructions to elicit clear speech (see Smiljanić & Bradlow, 2009, for a review), such as instructing the talkers to imagine a nonnative or a hearing impaired listener. Participants in the current study watched video clips of their

hypothetical listeners, which might have assisted them in estimating listener needs. During the post-experiment interview, all talkers could select which listener confederate had most and least trouble comprehending their message, although listener needs reported by talkers did not strongly correlate with the predictive model accuracy, listeners' segmentation accuracy, or speech rate. However, talkers unanimously agreed that the young native listener understood their message best and were most likely to produce hypo-speech (estimated by the model and listeners' segmentation accuracy and speech rate) for the young native listener condition.

Since the elicitation method was novel, analyses were first performed to confirm whether the method was successful in eliciting speech produced with different degrees of clarity. Measurements and analyses of the phonetic properties that have been shown to depend on speech clarity revealed that talkers produced speech with different degrees of clarity depending on who they spoke to (i.e., the video clip they watched before each block of the production experiment started), suggesting that the method used in the current study led talkers to generate speech with different degrees of clarity.

Interestingly, the analyses on the global measures of clarity modulation, such as modulation of overall speech rate or loudness, suggested that talkers typically distinguished two levels of speech clarity: plain (i.e., spoken towards the young native listener) versus clear (i.e., spoken towards the older hearing impaired listener or the young nonnative listeners). Contrary to the measurements pertaining to overall clarity, analyses of the production regression model fit, listeners' response times, clarity ratings, and transcription accuracy revealed that the availability and informativeness of signal-

based cues to word boundaries differed across all three listener conditions, suggesting that talkers might have produced qualitatively different speech depending on whether they spoke to the older hearing impaired listener or the young nonnative listener. In the post-experiment questionnaire, 15 talkers (out of 40) said that the young nonnative listener had the most trouble understanding them, while 23 talkers said that the older hearing impaired listener had the most trouble understanding them. The remaining two talkers said that both listeners had similar degree of communicative difficulty.

In the current study, we obtained a large number of speech productions (9,600 tokens from 40 talkers), which made it challenging to obtain perception data from all tokens by a reasonably high number of listeners. For these reasons, a subset of the production data was sample to serve as stimuli for the perception experiments. In order to include as much variability induced by talkers and items as possible, we constructed the stimulus set so that the stimuli includes ten talkers' voices (out of twenty talkers who were in the "unaware" and "aware" groups, respectively), half (18) of the items produced by any one talker, and all 36 target items. However, any discrepancy between the results from the production experiment and the results from the perception experiments might have been due to the specific sampling method that we used for the current study.

The preceding discussion leads to a methodological question regarding what constitutes a valid method to sample a subset of speech materials for perception experiments. Given that testing time and participants are not unlimited resources, having a gold standard of sampling stimulus subsets from a large production corpus would be highly desirable. Logically, if we can ensure that talkers are similar to each other,

reducing the number of talkers in favor of preserving item-variability would be desired. In contrast, if we can ensure that findings from a subset of items are generalizable to other items, reducing the items and preserving talker variability would be desired. We are currently conducting a study focusing on comparing multiple subsets of the production data in order to determine a sampling method that would generate a sample that can successfully represent the population (i.e., results from perception experiments where every listener responds to every token produced by every talker).

4.5. Conclusions

The current study extends the previous research focusing on the production and perception of word boundaries and talker variability, by studying variability in the degree to which talkers produce acoustic-phonetic cues to word boundaries and in the degree to which signal-based segmentation cues guide listeners' perception of word boundaries. We found that the production of acoustic-phonetic cues to word boundaries was no exception to the general principles of speech production in that the extent to which talkers produce word-segmentation cues depends on speech clarity (one type of "perception constraints," Lindblom, 1990) and phonetic contexts at and around word boundaries ("production constraints"). The finding that speech clarity and phonetic context condition the extent to which signal-based segmentation cues are produced suggests that the apparent discrepancies in the literature regarding whether or not talkers produce sufficient acoustic-phonetic cues to word boundaries for listeners can be understood as resulting from talker variability. By orthogonally manipulating multiple types of perception

constraints and production constraints and by investigating how linguistic environment affect the production of word-segmentation cues, which has rarely been explored with regards to acoustic-phonetic variability, we sought to obtain a better understanding of the production of signal-based cues to word boundaries.

A series of perception experiments tested how listeners perceived word boundaries from speech productions that differ in speech clarity and phonetic context surrounding word boundaries. Results from the perception experiments confirmed that talkers produced more perceptible acoustic cues to word boundaries as they spoke to listeners who might experience difficulty in speech comprehension and that the participants in the current study, who were young native listeners, benefitted from these clearer acoustic cues demarcating word boundaries. Listeners were more accurate in distinguishing the word-boundary minimal pairs that contain specific sequences of vowels and consonants (e.g., /s/-stop sequences), suggesting that the extent to which listeners can rely on acoustic cues for word segmentation depends on the inherent phonetic properties of segments and sequences of segments surrounding word boundaries.

By providing a detailed and more accurate picture of how word boundaries and produced and perceived, the current study sheds light on the ways in which models of speech production can be refined. As well, results of the current study can inform the desired properties of word-segmentation models.

References

Allbritton, D.W., McKoon, G., & Ratcliff, R. (1996). Reliability of prosodic cues for resolving syntactic ambiguity. Journal of Experimental Psychology: Learning, Memory, and Cognition, 22, 714-735.

Altmann, G. T. M. (2001). The mechanics of language: Psycholinguistics in review. The British Journal of Psychology, 92, 129-170.

Anderson, S., & Port, R. (1994). Evidence for syllable structure, stress and juncture from segmental durations. Journal of Phonetics, 22(3), 283-315.

Aslin, R. N., Woodward, J. Z., LaMendola, N. P., & Bever, T. G. (1996). Models of word segmentation in fluent maternal speech to infants. Signal to syntax: Bootstrapping from speech to grammar in early acquisition, 117-134.

Aylett, M. & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. Language and Speech 47 (1), 31–56.

Aylett, M. & Turk, A. (2006). Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. Journal of Acoustical Society of America 119 (5), 3048–3058.

Baese-Berk, M. & M. Goldrick. (2009). Mechanisms of interaction in speech production. Language and Cognitive Processes 24 (4), 527 – 554.

Barry, W. J. (1981). Internal juncture and speech communication. In W. J. Barry & K. J. Kohler (Eds), Beitrage zur experimentalen und angewandten phonetik, Kiehl, Germany: AIPUK.

Bard, E. G., Anderson, A. H., Sotillo, C., Aylett, M. Doherty-Sneddon, G., & Newlands, A. (2000). Controlling the intelligibility of referring expressions in dialogue. Journal of Memory and Language, 42 (1), 1-22.

Bard, E. C. &Aylett, M. P. (2005). Referential form, word duration, and modeling the listener in spoken dialogue. In J. C. Trueswell and M. K. Tanenhaus (Eds.), Approaches to studying world-situated language use: Bridging the language-as-

product and languageas-action traditions, pp. 173 – 191. Cambridge, Massachusetts: MIT Press.

Beckman, M. E. & Edwards, J. (1990). Lengthenings and shortenings and the nature of prosodic constituency. In J. Kingston & M. E. Beckman, eds., Papers in laboratory phonology I: Between the grammar and the physics of speech, pp. 152-178. Cambridge University Press.

Bell, A. (1984). Language style as audience design. Language in society, 13(2), 145-204.

Bradlow, A. R., G. Torretta, & D. Pisoni. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. Speech Communication 20 (3–4), 255–272.

Bradlow, A. R., Kraus, N., & Hayes, E. (2003). Speaking clearly for learning-impaired children: Sentence perception in noise. Journal of Speech, Language, and Hearing Research, 46, 80-97.

Bradlow, A. R. (2002). Confluent talker- and listener-related forces in clear speech production. In Gussenhoven, C. & Warner, N. (Eds.) Laboratory Phonology 7. Berlin & New York: Mouton de Gruyter. pp. 241-273.

Bradlow, A. R. & Bent, T. (2008) Perceptual adaptation to non-native speech. Cognition, 106, 707-729.

Brent, M. R., & Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. Cognition, 81(2), B33-B44.

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

Brown, P., & Dell, G. (1987). Adapting production to comprehension -- the explicit mention of instruments. Cognitive Psychology 19, 441-472.

Burnham, D., Kitamura, C., & Vollmer-Conna, U. (2002). What's new, pussycat? On talking to babies and animals. Science 296, 1435.

Byrd, D., & Saltzman, E. (1998). Intragestural dynamics of multiple prosodic boundaries. Journal of Phonetics, 26(2), 173-199.

Cho, T., Lee, Y., & Kim, S. (2011). Communicatively driven versus prosodically driven hyper-articulation in Korean. Journal of Phonetics, 39(3), 344-361.

Christie, W. M. 1974: Some cues for syllable juncture perception in English. Journal of the Acoustical Society of America 55, 819–21.

Christiansen, M. H., Allen, J., & Seidenberg, M. S. (1998). Learning to segment speech using multiple cues: A connectionist model. Language and cognitive processes, 13(2-3), 221-268.

Christiansen, M. H., Conway, C.M. & Curtin, S. (2005). Multiple-cue integration in language acquisition: A connectionist model of speech segmentation and rule-like behavior. In J. W. Minett & W.S.-Y. Wang (Eds.), Language acquisition, change and emergence: Essay in evolutionary linguistics, pp. 205-249. Hong Kong: City University of Hong Kong Press.

Clark, H. & C. Marshall. (1981). Definite reference and mutual knowledge. In A. Joshe, B. Webber, and I. Sag (Eds.), Elements of Discourse Understanding, pp. 10–63. Cambridge: Cambridge University Press.

Cole, R. A., & Jakimik, J. (1980). A model of speech perception. In R. Cole (Ed.), Perception and production of fluent speech. Hillsdale, NJ: Erlbaum.

Cooper, A. (1991). Laryngeal and oral gestures in English /p,t,k/. In Proceedings of the 12th International Congress of Phonetic Sciences. 2: 50-53.

Cooper, W. E. & Paccia-Cooper, J., 1980. Syntax and Speech. Harvard Univ. Press, Cambridge, MA.

Cutler, A. (1987). Speaking for listening. In A. Allport, D.G. MacKay, W. Prinz & E. Sheerer (Eds.), Language Perception and Production: Relationships between Listening, Speaking, Reading and Writing (pp.23-40). London: Academic Press Ltd.

Cutler, A. (1996). Prosody and the word boundary problem. In Morgan, J.L. and Demuth, K. (eds.), Signal to Syntax. Mahwah: Erlbaum. 87-99.

Cutler, A., & Butterfield, S. (1990a). Durational cues to word boundaries in clear speech. Speech Communication, 9, 485-495.

Cutler, A., & Butterfield, S. (1990b). Syllabic lengthening as a word boundary cue. Proceedings of the 3rd Australian International Conference on Speech Science and Technology, 324-328.

Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. Journal of Experimental Psychology: Human Perception and Performance, 41, 113-121.

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.

Davis, M. H. (2000). Lexical segmentation in spoken word recognition. Unpublished PhD thesis, Birkbeck College, University of London.

Davis, M. H., Marslen-Wilson, W. D., & Gaskell, M. G. (2002). Leading up the lexical garden path: Segmentation and ambiguity in spoken word recognition. Journal of Experimental Psychology: Human Perception and Performance, 28(1), 218-244.

Dell, G., & Brown, P. (1991). Mechanisms for listener-adaptation in language production: Limiting the role of the "model of the listener." In D. J. Napoli & J. A. Kegl (eds.), Bridges Between Psychology and Linguistics. Hillsdale: Erlbaum.

Dilley, L., Shattuck-Hufnagel, S., & Ostendorf, M. (1996). Glottalization of word-initial vowels as a function of prosodic structure. Journal of Phonetics, 24(4), 423-444.

Flemming, E. (2010). Modeling listeners. In C. Fougeron, B. Kühnert, M. D'Imperio and N. Vallée (Eds) Laboratory Phonology 10, pp. 587-606, Berlin: De Gruyter Mouton.

Ferguson, S.H. & Kerr, E.E. (2009). Use of subjective ratings for perceptual assessment of clear and conversational speech. Journal of the Academy of Rehabilitative Audiology, XLII, 51-66.

Ferguson, S. H., & Kewley-Port, D. (2002). Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners. The Journal of the Acoustical Society of America, 112, 259.

Fernald, A., & Simon, T. (1984). Expanded intonation contours in mothers' speech to newborns. Developmental psychology, 20(1), 104-113.

Fougeron, C. (2001). Articulatory properties of initial segments in several prosodic constituents in French. Journal of Phonetics 29: 109-135.

Gay, T., Ushijima, T., Hirose, H., & Cooper, F.S. (1974). Effect of speaking rate on labial consonant-vowel articulation. Journal of Phonetics 2, 46–63.

Godfrey, J. J., & Holliman, E. (1997). Switchboard-1 Release 2 Linguistic Data Consortium, Philadelphia.

Gow, D. W., & Gordon, P. C. (1995). Lexical and prelexical influences on word segmentation: Evidence from priming. Journal of Experimental Psychology: Human Perception and Performance, 21(2), 344-359.

Harris, M. S., & Umeda, N. (1974). Effect of speaking mode on temporal factors in speech: vowel duration. The Journal of the Acoustical Society of America, 56, 1016-1018.

Hoard, J. E. (1966). Juncture and syllable structure in English. Phonetica, 15(2), 96-109.

Horton, W. & B. Keysar (1996). When do speakers take into account common ground? Cognition 59 (1), 91–117.

Johnson, K. (2004) Massive reduction in conversational American English. In: K. Yoneyama & K. Maekawa (Eds.) Spontaneous Speech: Data and Analysis. Proceedings of the 1st Session of the 10th International Symposium. Tokyo: The National Institute for Japanese Language.

Joos, M . 1962. The five clocks. International Journal of American Linguistics 28, 9-62.

Junqua, J. C. (1996). The influence of acoustics on speech production: A noise-induced stress phenomenon known as the Lombard reflex. Speech Communication, 20(1), 13-22.

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

Keysar, B., & Henly, A. S. (2002). Speakers' overestimation of their effectiveness. Psychological Science, 13, 207–212.

Kim, D., Stephens, J. D. W., & Pitt, M. (In Press) How does context play a part in splitting words apart? Production and perception of word boundaries in casual speech. Journal of Memory and Language.

Klatt, D. (1976). Linguistics uses of segmental duration in English: Acoustic and perceptual evidence. Journal of the Acoustical Society of America, 59, 1208–1221.

Klatt, D. H. (1980). Speech perception: A model of acoustic-phonetic analysis and lexical access. In R. A. Cole (Ed.), Perception and production of fluent speech (pp. 243-288). Hillsdale, NJ: Erlbaum.

Kraljic, T. & Brennan, S. E. (2005). Prosodic disambiguation of syntactic structure: For the speaker or for the addressee? Cognitive Psychology, 50, 194–231.

Krause, J. C. & L. D. Braida. (2002). Investigating alternative forms of clear speech: The effects of speaking rate and speaking mode on intelligibility. The Journal of the Acoustical Society of America, 112 (5), 2165 – 2172.

Krause J. C. and L. D. Braida. (2004) Acoustic properties of naturally produced clear speech at normal speaking rates. The Journal of the Acoustical Society of America, 115:362–378.

Landauer, T. & L. Streeter (1973). Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition. Journal of Verbal Learning and Verbal Behavior 12 (2), 119–131.

Lehiste, I. (1960). An acoustic–phonetic study of internal open juncture. Phonetica, 5(Suppl. 5), 5–54.

Lehiste, I. (1972). The timing of utterances and linguistic boundaries. The Journal of the Acoustical Society of America, 51, 2018-2024.

Lane, H., & Tranel, B. (1971). The Lombard sign and the role of hearing in speech. Journal of Speech, Language and Hearing Research 14:677-709.

Lau, P. (2008). The lombard effect as a communicative phenomenon. UC Berkeley Phonology Lab Report, 1-9.

Lieberman, P. (1963). Some effects of semantic and grammatical context on the production and perception of speech. Language and Speech 6 (3), 172–187.

Lindblom, B. (1963). Spectrographic study of vowel reduction. The Journal of the Acoustical Society of America 35 (5), 783.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H and H theory. In W. J. Hardcastle and A. Marchal (Eds.), Speech production and speech modelling, pp. 403–439. Dordrecht, The Netherlands: Kluwer.

Mattys, S. L. & Melhorn, J. F. (2007). Sentential, lexical, and acoustic effects on the perception of word boundaries. Journal of the Acoustical Society of America, 122, 554-567.

Mattys, S. L., White, L., & Melhorn, J. F. (2005). Integration of multiple speech segmentation cues: A hierarchical framework. Journal of Experimental Psychology: General, 134, 477-500.

Mattys, S. L. (2004). Stress versus coarticulation: Towards an integrated approach to explicit speech segmentation. Journal of Experimental Psychology: Human Perception and Performance, 30, 397-408.

McQueen, J. M. (1998). Segmentation of continuous speech using phonotactics. Journal of Memory and Language, 39:21–46.

McQueen, J. M., Cutler, A., Briscoe, T., & Norris, D. (1995). Models of continuous speech recognition and the contents of the vocabulary. Language and Cognitive Processes, 10(3-4), 309-331.

Moon, S. & Lindblom, B. (1994). Interaction between duration, context, and speaking style in English stressed vowels. Journal of Acoustical Society of America 96 (1), 40–55.

Munson, B. & Solomon, N. P. (2004). The influence of phonological neighborhood density on vowel articulation. Journal of Speech, Language, and Hearing Research 47 (2), 1048-1058.

Nakatani, L. H., & Dukes, K. D. (1977). Locus of segmental cues for word juncture. Journal of the Acoustical Society of America, 62(3), 715-719.

Nakatani, L. H., & Schaffer, J. A. (1978). Hearing''words''without words: Prosodic cues for word perception. The Journal of the Acoustical Society of America, 63, 234-245.

Norris, D., McQueen, J. M., Cutler, A., & Butterfield, S. (1997). The possible-word constraint in the segmentation of continuous speech. Cognitive Psychology, 34, 191-243.

Oller, D. K. (1973). The effect of position‐in‐utterance on speech segment duration in English. Journal of the Acoustical Society of America, 54, 1235‐1247.

Oviatt, S., MacEachern, M., & Levow, G. A. (1998a). Predicting hyperarticulate speech during human-computer error resolution. Speech Communication, 24(2), 87-110.

Oviatt, S., Levow, G. A., Moreton, E., & MacEachern, M. (1998b). Modeling global and focal hyperarticulation during human–computer error resolution. The Journal of the Acoustical Society of America, 104, 3080-3098.

Perkell, J. S., Zandipour, M., Matthies, M. L., & Lane, H. (2002). Economy of effort in different speaking conditions. I. A preliminary study of intersubject differences

and modeling issues. The Journal of the Acoustical Society of America, 112, 1627-1641.

Picheny, M. A., Durlach, N. I., & Braida, L. D. (1986). Speaking clearly for the hard of hearing. II. Acoustic characteristics of clear and conversational speech. Journal of Speech and Hearing Research, 29, 434–446.

Picheny, M. A., Durlach, N. I., & Braida, L. D. (1985). Speaking clearly for the hard of hearing. I. Intelligibility differences between clear and conversational speech, Journal of Speech and Hearing Research. 28, 96–103.

Port, R. F. (1981). Linguistic timing factors in combination. The Journal of the Acoustical Society of America, 69, 262-274.

Reddy, D. R. (1976). Speech recognition by machine: A review. Proceedings of the IEEE, 64(4), 501-531.

Quené, H. (1992). Integration of acoustic-phonetic cues in word segmentation. The auditory processing of speech: From sounds to words, 349-356.

Quené, H. (1993). Segment Durations and Accent as Cues to Word Segmentation in Dutch. Journal of the Acoustical Society of America, 94(4), 2027-2035.

Salverda, A. P. (2005). Prosodically-conditioned detail in the recognition of spoken-words. Doctoral dissertation, Radboud University Nijmegen (MPI Series in Psycholinguistics, Vol. 33). Wageningen: Ponsen & Looijen.

Scarborough, R., Brenier, J., Zhao, Y., Hall-Lew, L., & Dmitrieva, O. (2007) An Acoustic Study of Real and Imagined Foreigner-Directed Speech. Proceedings of the International Congress of Phonetic Sciences, 2007.

Scarborough, R. (2004). Coarticulation and the structure of the lexicon. Unpublished Ph.D. dissertation, UCLA.

Schafer, A. J., Speer, S., Warren, P., & White, S. D. (2000). Intonational disambiguation in sentence production and comprehension. Journal of Psycholinguistic Research, 29(2), 169–182.

Schwab, S., Miller, J. L., Grosjean, F. & Mondini, M. (2008). Effect of speaking rate on the identification of word boundaries, Phonetica, 65, 173-186.

Shatzman, K. B. & McQueen, J. M. (2006). Segment duration as a cue to word boundaries in spoken-word recognition. Perception & Psychophysics, 68(1), 1-16.

Sikveland, R.O., 2006. How do we speak to foreigners?—phonetic analyses of speech communication between L1 and L2 speakers of Norwegian. Proc. Fonetik 2006. Centre for Language and Literature, Lund University, Lund, Sweden, pp. 109–112.

Smiljanić, R. & Bradlow, A. R. (2008). Temporal organization of English clear and plain speech. Journal of the Acoustical Society of America, 124(5), 3171-3182.

Smiljanić, R. & Bradlow, A. R. (2009). Speaking and hearing clearly: Talker and listener factors in speaking style changes. Linguistics and Language Compass, 3(1): 236–264.

Smith, R., & Hawkins, S. (2012). Production and perception of speaker-specific phonetic detail at word boundaries. Journal of Phonetics, 40(2), 213-233.

Smith, R. & Hawkins S. (2000) "Allophonic influences on word spotting experiments," Proceedings of the ISCA Workshop on Spoken Word Access Processes, Nijmegen, The Netherlands, 139-142.

Smith, R. H. (2004). The role of fine phonetic detail in word segmentation. Unpublished Ph.D. thesis, University of Cambridge.

Snedeker, J., & Trueswell, J. (2003). Using prosody to avoid ambiguity: Effects of speaker awareness and referential context. Journal of Memory and Language, 48, 103–130.

Speer, S. R., Warren, P., & Schafer, A. J. (2011). Situationally independent prosodic phrasing. Laboratory Phonology, 2, 35-98.

Sproat, R. & Fujimura, O. (1993) Allophonic variation in English /l/ and its implications for phonetic implementation. Journal of Phonetics, 21, 291–311.

Stern, D. N., Spieker, S. & MacKain, K. (1982). Intonation contours as signals in maternal speech to prelinguistic infants. Developmental Psychology, 18, 727-735.

Turk, A., & Shattuck-Hufnagel, S. (2000). Word boundary-related duration patterns in English. Journal of Phonetics, 28, 397-440.

Uchanski, R. M., Millier, K. M., Reed, C. M., & Braida, L. D. (1992). Effects of Token Variability on Resolution for Vowel Sounds, in The Auditory Processing of Speech: From Sounds to Words, M. E. H. Schouten, Ed., Mouton de Gruyter, Berlin, pp. 291-302.

Uchanski, R. M. (1988). Spectral and temporal contributions to speech clarity for hearing impaired listeners. Unpublished Doctoral dissertation, Massachusetts Institute of Technology.

Umeda, N., & Coker, C. H. (1974). Allophonic variation in American English. Journal of Phonetics, 2:1-5

Van Engen, K. & Bradlow, A. R. (2007). Sentence recognition in native- and foreign-language multi-talker background noise. Journal of the Acoustical Society of America, 121(1), 519-526.

Van Son, R. & Pols, L. (2003). How efficient is speech. In Proceedings of the Institute of Phonetic Sciences, Volume 25, pp. 171–184.

Van Summers, W., Pisoni, D. B., Bernacki, R. H., Pedlow, R. I., & Stokes, M. A. (1988). Effects of noise on speech production: Acoustic and perceptual analyses. The Journal of the Acoustical Society of America, 84(3), 917-928.

Watson, D., & Gibson, E. (2004). The relationship between intonational phrasing and syntactic structure in language production. Language and Cognitive Processes, 19. 713–755.

Watson, D, & Gibson, E. (2005). Intonational phrasing and constituency in language production and comprehension. Studia Linguistica, 59. 279–300.

White, L.S. (2002). English speech timing: a domain and locus approach. Unpublished Ph.D. dissertation, University of Edinburgh.

White, L., Wiget, L., Rauch, O., & Mattys, S. L. (2010). Segmentation cues in spontaneous and read speech. In Proceedings of the Fifth Conference on Speech Prosody, 2010, Chicago.

White, L., Mattys, S.L., & Wiget, L. (2012). Segmentation cues in conversational speech: Robust semantics and fragile phonotactics. Frontiers in Psychology, 3, Article 375, 1-9.

Wright, R. (1997). Lexical competition and reduction in speech: A preliminary report. In Research on Speech Perception Progress Report No. 21, pp. 471–485. Bloomington, Indiana: Speech Research Laboratory, Psychology Department, Indiana University.

Wright, R. (2004). Factors of lexical competition in vowel articulation. In J. Local, R. Ogden, and R. Temple (Eds.), Papers in Laboratory Phonology VI, pp. 26–50. Cambridge: Cambridge University Press.

Yao, Y. (2011). The Effects of Phonological Neighborhoods on Pronunciation Variation in Conversational Speech. Unpulished PhD dissertation. Department of Linguistics, UC Berkeley.

Yuan, J. & Liberman, M. (2008). Speaker identification on the SCOTUS corpus. In Proceedings of Acoustics 2008, 5687-5690.

Appendix A: Stimulus Materials

Item Number	Sentence
1	She loves her parrot because its wings are black and white
	She loves her cradle because it swings back and forth
2	If its limbs are down, you'll want to repot the tree
	If it slims you down, you'll want to workout again
3	I bet he gets lower response rates than before
	I bet we get slower response rates than before
4	The shepherd's sheep likes leaping in the field
	The shepherd's sheep like sleeping at night
5	That's why the airman eats wheat products
	That's why the airmen eat sweet pastries
6	The gentleman collects gulls at the beach
	The gentlemen collect skulls in the attic
7	The clock keeps ticking under the bed
	The clocks keep sticking at three o'clock
8	I wonder why it's praise but not worship
	I wonder why it sprays but doesn't sprinkle
9	This game character takes beers from a fridge to the bar
	These game characters take spears to fight one another
10 (filler)	He was sure this guy was not cheating
	He was sure the sky was not falling
11	I wish I knew what makes cars drive faster
	I wish I knew how to make scars disappear faster
12	Emily didn't know that the cook's truck hit me
	Emily didn't know that the cook struck him

13	The scientist takes oil samples from the marsh
	The scientists take soil samples from the swamp
14	The explorers did sail east into the Pacific Ocean
	The explorers did say least about what they found
15	I believe the earth has the oldest known ocean in the space
	I believe the world nowadays has no notion of justice at all
16	Dr Miller's project started while earning his first PhD
	Dr Miller's project is about why learning is life long
17	He didn't seem able to run the hospital
	He didn't see Mable touring the hospital
18	Come with your friends and find a team at any time
	Come with your friends and get a tea mat for free
19 (filler)	The kids saw an iceman from the bedroom
	The kids saw a nice man from the roof top
20	I couldn't think of an aim for this new proposal
	I couldn't think of a name for this new project
21 (filler)	He picked the shirt he'd eyed earlier
	He picked the shirt he dyed white
22	Many people make art for various purposes
	Some people may cart things that aren't too heavy
23	They wanted to buy a new lawn chair for next summer
	They were planning to launch air quality monitor stations
24 (filler)	He is not a huge fan of Grade A maple syrup
	He is not a huge fan of a gray day during football season
25	The zoo has an old great ape but nobody wants to see it
	My aunt has an old grey tape but doesn't know what's on it
26	Every Monday, we buy zinc for the experiment
	Every Monday, he buys ink for his office staff
27	There are too many bee feeders in Utah
	There are too many beef eaters in the US

28	Then he'd iced the cakes
	Then he diced the carrots
29	The teenager came to a maze in the spacious field
	The teenager came to amaze all the spectators
30	The man went to a line there on the floor
	The man went to align them on the floor
31	The angry crowd came to a rest at the courthouse
	The angry crowd came to arrest the criminal
32	Steve's dog was a way to avoid loneliness
	Steve's dog was away ten days at the kennel
33	Megan was ready for a round of our kickboxing class
	Megan was ready for around an hour of kickboxing
34	I think Jane has a loud sewing machine
	I think Jane has allowed Sue to go there
35	Lauren didn't know what a fair deal would have been
	Lauren didn't know what affair Dave was involved in
36	The servant came to a door that leads to the hall
	The servant came to adore that little boy in the hall
37	They have been a part of our school for years
	They have been apart ever since the argument
38	People often claim that a tax lawyer would help the city
	People often claim that attacks largely happen at night
39	The young girl had a cute kitten in her arms
	The young girl had acute kidney disease
40	They said it was a cross they saw in the church
	They said it was across the street from the church

Appendix B: Output of the Mixed-effects Linear Regression Models Testing the Effects of Listener Condition on the Measures of Global Clarity Modulation

Note: Block order 1 refers to the order of young native, older hearing impaired, and young nonnative listener conditions. Block order 2 refers to the order of young native, young nonnative, and older hearing impaired listener conditions.

Appendix B-1. Entire utterance Duration

	Estimate	**t-value**	**Pr(>|t|)**
Intercept Young Native Listener, Block order 1	2554.68	21.88	< 0.001
Older Hearing Impaired Listener	546.67	36.64	< 0.001
Young Nonnative Listener	814.76	54.58	< 0.001
Block Order 2	28.74	0.21	0.84
Older Hearing Impaired, Block Order 2	122.98	5.84	< 0.001
Young Nonnative, Block Order 2	-334.47	-15.86	< 0.001

Appendix B-2. Ambiguous target sequence duration

	Estimate	**t-value**	**Pr(>|t|)**
Intercept Young Native Listener, Block order 1	577.13	15.4	< 0.001
Older Hearing Impaired Listener	130.89	24.69	< 0.001
Young Nonnative Listener	203.34	38.37	< 0.001
Block Order 2	1.44	0.04	0.97
Older Hearing Impaired, Block Order 2	61.02	8.15	< 0.001
Young Nonnative, Block Order 2	-83.61	-11.15	< 0.001

Appendix B-3. RMS amplitude of non-silent portions

	Estimate	**t-value**	**Pr(>|t|)**
Intercept Young Native Listener, Block order 1	69.59	56.98	< 0.001
Older Hearing Impaired Listener	1.91	16.24	< 0.001
Young Nonnative Listener	2.34	19.95	< 0.001
Block Order 2	-3.17	-1.85	0.06
Older Hearing Impaired, Block Order 2	1.16	7.00	< 0.001
Young Nonnative, Block Order 2	-0.20	-11.15	< 0.001

Appendix B-4. Range of fundamental frequency

	Estimate	**t-value**	**Pr(>|t|)**
Intercept Young Native Listener, Block order 1	56.53	6.70	< 0.001
Older Hearing Impaired Listener	9.65	7.46	< 0.001
Young Nonnative Listener	14.51	11.21	< 0.001
Block Order 2	-3.75	-0.34	0.06
Older Hearing Impaired, Block Order 2	4.53	2.48	< 0.001
Young Nonnative, Block Order 2	-3.14	-1.72	0.23