

23c. Methods for Studying Spontaneous Speech

Natasha Warner, University of Arizona

23c.1. Introduction/terminology

The overwhelming majority of all research in phonetics and laboratory phonology has used "careful" speech, but interest in "spontaneous," non-careful speech is now surging. This could lead to a very different understanding of how speech and communication work. Spontaneous speech often includes sequences with such strong reduction phenomena that one could never have predicted them, and is rather surprised to see them when one examines the spectrogram (e.g. Fig. 1, with multiple deletions, reduction of stops to fricatives, and changes to vowel qualities). Yet these sequences usually sound intelligible and normal, at least to native listeners. But what types of speech are at issue? This section will offer a brief discussion of terminology (see also Warner submitted).

PLEASE INSERT FIGURE 1 ABOUT HERE

One could establish a continuum of carefulness or naturalness. On one end might be vowels or nonsense monosyllables read in isolation (perhaps while wearing an airflow mask). At the other end might be informal conversation among family or friends, perhaps at home with no microphone present. Several terms would fall along this continuum: careful or laboratory speech (near the careful end), any non-read speech (including responses to prompts), connected speech (anything in a longer utterance, read or not), spontaneous speech (nothing read, but including monologues and structured speech such

as Map Task dialogues), and conversational speech (even with interviewers). The "clear speech" that Bradlow's group has studied (e.g. Bradlow and Bent 2002, Smiljanić and Bradlow 2009), for addressing hearing-impaired or L2 listeners, is near the careful end of the continuum, more careful than typical read lab speech. The term "natural speech" could be taken as the other end of this continuum, but researchers of differing backgrounds use "natural" very differently. "Natural speech" can mean anything produced by a human vocal tract (not synthesized), or in linguistic anthropology it can set strict requirements on the interactional setting. Therefore, "natural speech" will be left undefined.

Sociolinguists have put considerable effort into defining various ways in which speakers vary their speech style, some of which overlap with the carefulness dimension delineated here. Schilling-Estes (2002) provides a clear overview of sociolinguistic approaches to speech style, and of shortcomings of simple explanations for why speakers vary style as they do. The carefulness continuum here does not claim any of the particular explanations Schilling-Estes (2002) discusses, but is simply a description of a continuum along which several types of speech fall. Speakers of course vary their speech style in many other ways not covered by "carefulness," for example in order to show affiliation with a variety of groups, their attitude toward interlocutors' utterances, etc.

The terms "reduced" and "fast" speech are not on the same continuum as carefulness. Careful, read speech and casual conversation can both be fast or slow, and speech rate can be measured acoustically, unlike spontaneity. I take "reduced speech" to refer to any speech exhibiting reduction from the canonical, careful pronunciation, e.g. speech with segments or syllables deleted, with expected stops realized as approximants,

with vowels approaching the center of the vowel space, with incomplete tongue closures, etc. That is, reduction is defined by the results (acoustic or articulatory), not by the circumstances under which it is recorded. One finds reductions even in isolated word list reading, but not as often as in conversation. Figure 2 shows a schematic representation of carefulness, speech rate, and degree of reduction as separate continua.

PLEASE INSERT FIGURE 2 ABOUT HERE

With the terminology defined, the rest of this chapter will turn first to methods, then to theory. Regarding methods, the major issues for reduced speech are how to obtain recordings of reduced or less-careful speech (23c.2), how to obtain or create stimuli for perception experiments on such speech (23c.3), and how to analyze the resulting data (23c.4). Section 23c.5 discusses implications of reduced speech for phonetic, phonological, and psycholinguistic theories.

23c.2. Methods for eliciting less-careful speech (for audio recordings)

This section will address recording methodologies. The main purposes are acoustic phonetic analysis, obtaining stimuli for perception studies, or development of Automatic Speech Recognition (ASR) systems. None of the methods are optimal, and they are very much under exploration.

One common method to balance control and naturalness is to record subjects conversing with an interviewer/experimenter, as in the Buckeye Corpus (Pitt et al. 2005). Such speech can be rather natural, but still have very good acoustic conditions. The

interviewer can attempt to make the interactions casual, and less like a formal interview. The interviewer can keep the subject talking, and can steer the conversation to include topics that are likely to elicit target words, to obtain some matched words across speakers. There are a few disadvantages: the subject does not know the interviewer, making speech more formal than in conversation with friends or family. The subject's and interviewer's voices are likely to overlap sometimes, although the interviewer can attempt to avoid overlaps. This can be avoided if the interviewer is outside the sound-protected booth and the subject inside, and they hear each others' speech over headphones, but that creates an unfamiliar and less natural setting. (See Scobbie and Stuart-Smith, this volume, on the overall topic of the effect of presence of experimenter or recording equipment on naturalness of speech.)

The most obvious way to record conversational speech might be to put two subjects who know each other well in a sound booth together, with head-mounted directional microphones that pick up little of the other speaker's voice, and have them converse. I do not know of a study that has done exactly this, perhaps because most recording booths are small. However, Ernestus and colleagues use a clever variant on this method (Torreira et al., submitted), by seating two subjects who know each other and a confederate in the booth, so that the experimenter/confederate can help get the conversation going well (making the subjects comfortable), and can steer the conversation toward certain topics. The confederate then leaves the booth on the pretense of switching out a broken microphone, leaving the subjects to converse. This method has some danger of overlapping speech in the recordings, and requires a rather large sound booth. Speakers might also be distracted by a head-mounted microphone on their

interlocutor's face, reminding them of the unusual conversational setting. However, Scobbie and Stuart-Smith (this volume) find that even obvious ultrasound equipment has little negative effect on naturalness if speakers speak casually with an interlocutor who is a peer, so an unobtrusive microphone may not be a problem.

Recording telephone conversations is another method to obtain very spontaneous speech. It avoids several problems of recording two speakers conversing in person: overlapping speech will be recorded separately, and speakers need not sit together in a sound booth. Speakers are also very comfortable with conversing casually on the phone, so speech is very natural. The Switchboard, CallHome, and CallFriend corpora all exemplify this approach (e.g. Canavan and Zipperlen 1996, or Switchboard as analyzed by Bell et al. 2009). Conversations can be between acquaintances (e.g. CallHome, CallFriend) or between two volunteers introduced for the phone call (Godfrey and Holliman 1997). The naturalness of such recordings is a clear advantage to this method, particularly when the speakers know each other well. However, the recordings retain only telephone speech bandwidth (500-3500 Hz), and speakers call from locations with highly variable, sometimes extremely loud background noise. It may be difficult to collect detailed information about speakers' language and dialect backgrounds. There is, of course, no control whatsoever over what the speakers say. One can take this approach even further and simply attach a recording device to the speaker and leave it recording while they carry on with their daily-life activities, without the researcher present (Mehl and Pennebaker 2003, Podesva 2006). Podesva uses this method for sociophonetic analysis, while Mehl and Pennebaker use it to obtain social psychology data. Although

Podesva's data allowed for detailed phonetic analysis of unusually natural speech, this method has a clear danger of failure to obtain high acoustic quality in recordings.

The favored method in my own lab is to have a speaker sit in a sound booth and talk on the telephone to a close friend or family member, while wearing a head-mounted microphone over the opposite ear from the telephone. The recording only includes one side of the conversation (losing discourse information), but the acoustics are excellent, and the speech is extremely natural and casual. Speakers rapidly become comfortable with the sound booth, and begin animatedly discussing informal topics (e.g. gossiping about one's boyfriend to one's best friend). This retains all advantages of recording telephone speech (except losing the interlocutor's side), but provides high-quality acoustic recordings as well. One might be able to recover discourse information about the interlocutor's utterances (although not phonetic information) from a weak signal the microphone might pick up from the telephone. However, if the microphone picks up enough of the interlocutor's speech to be intelligible, this might require that the interlocutor also be a consented human subject (depending on local regulations), which would present logistical problems.

A method one step less natural is to record spontaneous monologues (e.g. "now please tell us about yourself") over the telephone or in the lab. The OGI corpora for various languages use this method (Muthusamy et al. 1992). This is easier to set up than conversations: subjects call a toll-free number and hear recorded prompts, so no pairs of subjects need be arranged. This method gives spontaneous but not conversational speech. Some speakers find it difficult to speak with no interlocutor, or to speak naturally to an

answering machine, but surprisingly many subjects do quite well at this task (Warner and Arai 2001).

The Map Task (Bard et al. 2001, Shattuck-Hufnagel & Veilleux 2007, among others, and see Warren and Hay, this volume) elicits relatively spontaneous, conversational speech while maintaining considerable control over target words. In this method, two speakers look at non-identical maps. One speaker directs the other on how to go from one location to another on the map. Because the maps differ, the listener is likely to ask for clarification. Neither speaker is reading a script, although some features on the map might be labeled in order to induce speakers to use specific target words (the labeled items) that contain phonological features of interest. This method leads to conversational but relatively formal speech.

Moving further toward controlled speech, one can record speakers reading a very large quantity of connected texts. ATR (Kyoto, Japan) in the development of their speech synthesis program might commonly record a speaker reading a newspaper out loud for an hour (Campbell 1992, 1999). This is not spontaneous speech, but when reading for so long, speakers are likely to speak less carefully. This method gives control over the content and some control over likely intonational patterns. The speech is less variable than conversation, where speakers shift rapidly from enthusiastic speech to slow, tired-sounding utterances.

Recently, it has become possible to obtain large quantities of relatively natural speech over the internet, even for a variety of languages. (Kim (2004) provides just one example of work answering a question about connected speech with such material.) Radio and television broadcasts, often available for download, can provide huge publicly

available, pre-recorded corpora. One step in using such recordings is to classify the types of speech and speaker, since the material includes both professional newscasters and non-professional speakers (e.g. in interviews). Some speech may be scripted and some spontaneous, and one cannot necessarily tell which. Background noise (recorded in studio vs. on site), background music, dialect, topic, and genre may all vary. Files may be compressed, in a variety of only partially predictable ways, which may make some more detailed acoustic analyses impossible. Language background information is likely unavailable. However, with the number of broadcasts available over the internet rapidly increasing, this provides an exciting opportunity to study relatively natural speech, particularly for languages where large recording experiments might be impossible.

Going a final step toward controlled speech, one can simply manipulate speech rate by instructing speakers to read target sentences quickly, normally, or slowly. If speakers succeed in reading quickly, they are likely to produce some reductions. Research on topics other than reduction uses this method (Ladd et al. 1999, Hirata et al. 2007, and Adank and Janse 2009 provide a few examples), and speakers can vary their speech rates, although this may not be the same as what they do in natural speech. However, speech rate is not the same thing as speech style, spontaneity, or casualness. Overtly asking speakers to vary their speech rate should not be the main method for recording reduction.

23c.3. Current methods for obtaining stimuli for reduction perception studies

Section II summarized methods for obtaining acoustic recordings, but perception of spontaneous speech may be even more interesting. Perception experiments on reduced

speech require stimuli containing reduction, which are even harder to obtain than good spontaneous acoustic recordings. Research on perception and psycholinguistic processing of reduced or spontaneous speech was almost non-existent until a few years ago, with intriguing exceptions such as Mehta and Cutler (1988) and Koopmans-van Beinum (1980). Sociolinguists have long used relatively natural speech in perception experiments (e.g. Labov 1989), but these studies are often for the purpose of studying perception across dialectal varieties, not perception of reduced speech of one's own variety. (See Warren and Hay, this volume, on the importance of perception studies for sociophonetic topics, as well.) Researchers are now developing an array of methods for obtaining stimuli, varying in naturalness of the source speech.

The most direct method is to extract stimuli from large, relatively natural corpora that have been collected using the most spontaneous and conversational speech methods above. For example, one can record a conversation and extract stimuli from only the parts without overlapping speech, or record one side of a telephone conversation and extract stimuli from that (e.g. Ernestus et al. 2002, Warner et al. 2009, Brouwer et al. submitted, cf. also Labov 1989). This has the advantage that stimuli definitely represent what real speakers produce in conversation, and what listeners hear in their daily lives. However, the stimuli are highly variable and uncontrolled. One cannot make a target word list in advance, but must be able to use a wide variety of words, of varied numbers of syllables, spoken with any intonational pattern, in any context, etc. Items will not match across conditions, either. One can record a long conversation and use only portions that meet criteria as stimuli (e.g. only content words in a particular intonational context), but the materials will still vary widely. For example, in one study in my lab, utterances for the

target "he's" include "Well, because he's turning 24 and he hasn't accomplished anything in his life," and "He's like, 'I just...'" despite their different lengths. In another study, target content words for use in a cross-modal priming lexical decision task included both "kindergarten" and "free." In psycholinguistic studies not on reduction, all targets often contain the same number of syllables, are controlled for some of the phonemes, and are recorded in isolation or in a consistent frame sentence. For example, Gaskell and Marslen-Wilson (1996), although they go considerably further toward connected speech materials than many studies, use prime words such as "broad, cloud, crowd, bread" etc., with the most varied items being "concede, horrid, wicked" etc. Not surprisingly, one may not obtain significant results with the more variable stimuli one takes from open conversation. However, with tasks or questions for which varied stimuli can work (e.g. Ernestus et al. 2002), this method may be optimal.

A related method is to record spontaneous speech, extract usable stimuli, then bring the same speaker back to read those word strings again as careful speech, out of context. One can thus compare listeners' reactions to spontaneous vs. careful speech using the same targets, with words and voice controlled. This requires two recording sessions with each speaker though, and speakers may read casually because the phrases are from their own spontaneous conversations, minimizing style effects. Intonation may differ unpredictably between the spontaneous and read utterances. Mehta and Cutler (1988) use this method successfully with a phoneme-monitoring task, and my own lab has attempted this recording method. However, we were unable to obtain even a priming effect using cross-modal identity priming with such varied stimuli.

To obtain controlled reduced stimuli, one can record spontaneous or careful speech and resynthesize or splice to manipulate acoustic characteristics one sees in reductions (with PSOLA, LPC, or intensity resynthesis). One can also instruct the speaker to say the words over and over, sometimes "in a sloppy way" and sometimes "normally," to obtain reduced and careful tokens of each target. Examples include Niebuhr (2008), Warner et al. (2009), and Mitterer and Ernestus (2006). Sociolinguists have also used these methods to study listeners' perception of sociolinguistically marked variables. Campbell-Kibler (2008), for example, uses splicing between [ɪŋ] and [ən] versions of the English *-ing* suffix, then uses resynthesis to match duration, intensity, and pitch to a target pronunciation. The degree of control with resynthesis is a clear advantage: one can know that only one acoustic aspect of reduction varies at a time (Fig. 3). However, one can never know whether the stimuli are truly representative of spontaneous speech, although one can resynthesize beginning from both a careful and a reduced production, or from a variety of productions (Warner et al. 2009), for example, to determine whether the perceptual effect holds despite other cues that may be present. One could also use parametric synthesis from scratch and vary specific acoustic characteristics that mimic what one sees in natural reductions. This provides even more control, but departs from the external validity of spontaneous speech stimuli.

PLEASE INSERT FIGURE 3 ABOUT HERE

A final method is to record a phonetician intentionally producing reduced and unreduced (careful) forms of target items. One obtains well-matched stimuli in a

consistent voice, without synthesis, and the stimuli may represent what naïve speakers do in daily life speech, but one cannot be absolutely sure of this. Acoustic measurements can partially confirm that stimuli match natural reductions (Tucker 2007).

None of these methods is perfect, but all can contribute -- there is no clear way to study perception of reduced speech under controlled circumstances. Therefore, rather than discarding methods as flawed, or worse yet avoiding the entire topic of reduction, we should use multiple methods and look for convergent evidence. Since the field of phonetics has been dominated by careful speech, and acoustics is ahead of perception for reduction studies, it is no surprise that the methods are just being explored. A wide variety of methods in flux may signal an innovative research area.

23c.4. Spontaneous speech analysis methods

After collecting data, one must determine how to analyze it. For acoustic work, researchers are developing some novel methods in order to measure the variable and unexpected segments of reduced speech. With controlled wordlists, one knows what segments to expect and can define specific measurement criteria, e.g. offset of voicing for voiceless stops vs. offset of F2 for voiced ones. With reduced speech, though, one can rarely predict what segments will be present, or what manner of articulation they will have. This forces flexibility in measurements.

One relatively common method (Greenberg 1999, Johnson 2004, Shattuck-Hufnagel and Veilleux 2007) is to transcribe a corpus at phonetic and word levels, then compare the segments in the phonetic transcription to the segments given for the same words in a searchable dictionary. One can then tally deletions and substitutions relative to

the canonical form as it is transcribed in an electronic dictionary. This method can answer the overall, descriptive question of how much reduction is happening, and it can be applied regardless of what words speakers use and what segments are realized how. However, this method is heavily influenced by transcription conventions, and it assumes that the signal can be transcribed as distinct, categorical segments. In reduced speech, one often hears a segment that one can identify as vocalic, but one cannot say if it is a segment of the language, let alone which one. The heavy coarticulation of reduced speech makes counts of transcription mismatches suspect, even if the segments do seem to be identifiable. Current theories do not all assume that each word has a single, invariant lexical entry, making comparison of the realization to the single form in the electronic lexicon less meaningful, but this is a theoretical issue that goes beyond the methodological point. A related method is to use automatic speech recognition (ASR) to (help) produce a transcription or locate segment boundaries (Pluymaekers et al. 2006).

One can avoid transcription problems by using a more global measure such as syllable count (surface perceived syllable count vs. underlying expected count) or overall speech rate (in underlying syllables per second, for example). ASR can also offer innovative global measures of reduction (Nakamura et al. 2007).

Alternatively, one can go to a more detailed, rather than global, method by using traditional phonetic measures such as duration, intensity, etc. to measure particular segments in detail. However, one must define criteria that cover any manner of articulation speakers might produce. In my lab's work, we study reduction of stops and flaps, and define criteria conditionally depending on whether the target is realized as a

stop/flap (voiced or voiceless, with or without burst), an approximant (with or without weakening of formants), or is deleted.

For perception studies, however one obtains stimuli, one can use them in standard phonetic and psycholinguistic perception tasks (e.g. phonetic identification, discrimination, phoneme/word monitoring, lexical decision, priming). Some methods require filler non-words, which can be challenging to make from spontaneous conversation, but most methods are possible. Thus, for production studies, the methods issues include both how to obtain and how to analyze data, but for perception, the primary issue is how to obtain stimuli, not how to analyze results.

23c.5. What theoretical questions can the methods answer? What have we learned from them? Do the methods presuppose any theoretical commitments?

This section offers a brief list of theories for which spontaneous speech methods may be relevant. See also Coetzee, Warren and Hay, and Scobbie and Stuart-Smith (all this volume), and Warner (submitted). The most obvious theoretical relations are with articulatory topics, such as Articulatory Phonology and task dynamics, because gestures describe reductions conveniently. Reduction also has clear relevance for Lindblom's H&H model (1990) and the idea of competing constraints in OT favoring ease of articulation vs. perceptibility (critiqued by Hale and Reiss 2000). Reduced speech could also impact most phonetics-phonology interface questions, such as what about language is gradient vs. categorical, and what is conditioned vs. random variability. So much varies unexpectedly in spontaneous speech, it provides an excellent theoretical testing ground regarding what is under the control of an abstract, categorical phonology. For all theories

of speech production (phonetic, phonological, and psycholinguistic (see Bell et al. 2009)), reduction tests whether a theory generalizes to daily-life speech, since theories are nearly always developed based on careful pronunciations. However, Warren and Hay (this volume) point out that not all questions can be answered using daily-life speech, and some theoretical questions should be addressed using controlled, targeted laboratory speech.

Reduced speech could lead to a change in theories of formal phonology: just the idea that there are many possible realizations of a given word, which are not entirely predictable from a single underlying form, is problematic for most theories. For example, how would one derive [wiçõ] for "weekend" (Fig. 1), or [p^herĩ] for "apparently" (Johnson 2004), while still being able to derive the more canonical forms of the words and the many other possible pronunciations? Phonological theories usually generate a single surface form, not the tens of distinct pronunciations Greenberg (1999) documents (e.g. 117 for "that," with the most common, [ðæ], representing only 11% of tokens).

Turning to theories of perception, reduced speech clearly impacts theories of spoken word recognition, such as TRACE, SHORTLIST, Merge, etc. Recognition of reduced words in their many surface forms is problematic for the same reason as the multiple forms are problematic for phonological (production) theories: the multiplicity of forms are not systematically derivable from the underlying form (Ernestus et al. 2002). If "that" has at least 117 distinct pronunciations, must multiple forms, or all forms, be stored in the lexicon? This can shade into an exemplar model (Johnson 1997, 2006, Pierrehumbert 2001, 2002). One might also expect reduced speech to be relevant for articulatory theories of speech perception (e.g. the Motor Theory and Direct Realism),

although this topic has not been well developed yet. Overall, reduced speech is relevant to testing any theory of speech or word perception, because any cues present in it differ so radically from the kinds of perception stimuli that are typically studied.

Detailed findings will be left for other works, but a few overall findings from spontaneous speech research methods can be summarized here. It is clear that individual words are realized with a wide variety of forms (Greenberg 1999), and that listeners can recognize these forms well in context, but at best poorly out of context (Arai 1999, Ernestus et al. 2002). Listeners recognize unreduced forms more easily than reduced forms (Ernestus et al. 2002, Ranbom and Connine 2007, Tucker 2007), even if the reduced form is more common. One thing we have certainly learned from spontaneous speech is that the real speech we all produce and process every day is far, far more variable than one would ever expect based on more controlled methods. Furthermore, we have learned that however listeners perceive speech and recognize words, they must be able to handle far more variability than most theories address.

Some methods include theoretical assumptions, and may not be useful for testing anything if those assumptions are not true. The methods for working on reduced speech make only one minimal assumption about theory: that variability well below the level of the phoneme is interesting. If reduced speech, fast speech, casual speech reduction, etc. are all relegated to "phonetic implementation" and considered external to the grammar, and the grammar is the topic of study, then reduced speech is of no interest. However, if any aspect of reduction is language-specific rather than caused by universal biological constraints on articulation, then speakers would need to know the language-specific aspects of reduction as part of the grammar (cf. Keating 1985, 1990 and Kingston and

Diehl 1994 on language-specific phonetic detail in general, regardless of reduction, and Barry and Andreeva 2001 on cross-linguistic patterns of reduction). Many recent phonological theories extend the realm of interest to include low-level gradient variability (Coetzee, this volume).

Another part of how methods relate to theory is that some may feel it is better science to test theories on controlled data, rather than on spontaneous data. When investigating a new topic, about which little is understood (e.g. intonation in a language for which it has never been studied), one should probably begin with controlled, stable data, such as matched target items in frame sentences. However, when studying a topic with extensive past literature, the field may be ready to move to data that is more representative of daily life speech. Warner and Arai (2001) argue this for a study of Japanese mora-rhythm using spontaneous speech.

23c.6. Conclusions

There are problems with all methods of obtaining and analyzing spontaneous speech and stimuli. Researchers are exploring a wide variety of methods. While this may seem chaotic, it is exciting. As large speech corpora have appeared, spontaneous speech research has increased rapidly. For example, LabPhon X (Paris, 2006) and ICPhS 2007 (Saarbrücken) both had a proliferation of papers using large speech corpora or investigating speech style. 2008 saw the First Nijmegen Speech Reduction Workshop (program at <http://www.u.arizona.edu/~nwarner/>).

Perception studies on reduction have lagged behind production studies, perhaps because of the methodological challenge of obtaining stimuli, but are now leading to

fascinating studies. Moving beyond the core areas of native adult production and perception, there has been only the most tentative exploration into the relationship of spontaneous speech to L1 or L2 acquisition (Bleses 2008, Shockey 2008), cross-linguistic and cross-dialectal language use, or disordered speech (dysarthria, Mattys and Liss (2008)). We can expect development into these areas soon. Returning to theory, current phonetic theories (and even more so formal phonological ones) have had little development for handling massive reduction phenomena. We can expect, or work toward, an impact of spontaneous speech on many theories in upcoming years.

ACKNOWLEDGEMENTS

The author would like to thank Mirjam Ernestus, Ben Tucker, Anne Cutler, Holger Mitterer, and Rob Podesva for helpful discussion and feedback on the issues in this article. All errors, of course, are the author's own.

References

- Adank, Patti and Janse, Esther (2009). 'Perceptual learning of time-compressed and natural fast speech'. *Journal of the Acoustical Society of America*, 126: 2649-2659.
- Arai, Takayuki (1999). 'A case study of spontaneous speech in Japanese'. *Proceedings of the International Congress of Phonetic Sciences (ICPhS), San Francisco*, 1: 615-618.
- Bard, Ellen Gurman, Sotillo, Catherine, Kelly, M. Louise, and Aylett, Matthew P. (2001). 'Taking the hit: Leaving some lexical competition to be resolved post-lexically'. *Language and Cognitive Processes*, 16: 731-737.
- Barry, William and Andreeva, Bistra (2001). 'Cross-language similarities and differences in spontaneous speech patterns'. *Journal of the International Phonetic Association*, 31: 51-66.
- Bell, Alan, Brenier, Jason, Gregory, Michelle, Girand, Cynthia, and Jurafsky, Dan (2009). 'Predictability effects on durations of content and function words in conversational English'. *Journal of Memory and Language*, 60: 92-111.
- Bleses, Dorte (2008). 'The struggle of Danish word-learning babies: The role of sound structure in word learning in a cross-linguistic framework'. Presentation at the First Nijmegen Speech Reduction Workshop, MPI, Nijmegen, the Netherlands.
- Bradlow, Ann R. and Bent, Tessa (2002). 'The clear speech effect for non-native listeners'. *Journal of the Acoustical Society of America*, 112: 272-284.
- Brouwer, Susanne, Mitterer, Holger, and Huettig, Falk. Submitted. 'Phonological competition during the recognition of spontaneous speech: Effects of speech style and reduction'.

- Campbell, Nick (1992). 'Segmental elasticity and timing in Japanese speech', in Tohkura, Y., Vatikiotis-Bateson, E., and Sagisaka, Y. (eds.), *Speech Perception, Production, and Linguistic Structure*. Tokyo: Ohmsha.
- Campbell, Nick (1999). 'Data-driven speech synthesis'. *Journal of the Acoustical Society of America*, 105: 1029-1030.
- Campbell-Kibler, Kathryn (2008). 'I'll be the judge of that: Diversity in social perceptions of (ING)'. *Language in Society* 37: 637-659.
- Canavan, Alexandra, and Zipperlen, George (1996). *CALLHOME Japanese Speech*. Philadelphia: Linguistic Data Consortium.
- Coetzee, Andries W. (Submitted). 'Variation: Where laboratory and theoretical phonology meet'. This volume.
- Ernestus, Mirjam, Baayen, R. Harald, and Schreuder, Rob (2002). 'The recognition of reduced word forms'. *Brain and Language*, 81: 162-173.
- John J. Godfrey and Edward Holliman (1997). *Switchboard-1 Release 2*. Philadelphia: Linguistic Data Consortium.
- Gaskell, M. Gareth and Marslen-Wilson, William D. (1996). 'Phonological variation and inference in lexical access'. *Journal of Experimental Psychology: Human Perception and Performance*, 22: 144-158.
- Greenberg, Steven (1999). 'Speaking in shorthand - A syllable-centric perspective for understanding pronunciation variation'. *Speech Communication*, 29: 159-176.
- Hale, Mark, and Reiss, Charles (2000). 'Substance abuse and dysfunctionalism: Current trends in phonology'. *Linguistic Inquiry*, 31: 157-169.

- Hirata, Yukari, Whitehurst, Elizabeth, and Cullings, Emily (2007). 'Training native English speakers to identify Japanese vowel length contrasts with sentences at varied speaking rates'. *Journal of the Acoustical Society of America*, 121: 3837-3845.
- Johnson, Keith (1997). 'Speech perception without speaker normalization: An exemplar model', in Johnson, K., and Mullennix, J. (eds.), *Talker Variability in Speech Processing*. San Diego: Academic Press. pp. 145-165.
- Johnson, Keith (2004). 'Massive reduction in conversational American English', in Yoneyama, K. and Maekawa, K. (eds.), *Spontaneous Speech: Data and Analysis. Proceedings of the 1st Session of the 10th International Symposium*. Tokyo: The National International Institute for Japanese Language, pp. 29-54.
- Johnson, Keith (2006) 'Resonance in an exemplar-based lexicon: The emergence of social identity and phonology'. *Journal of Phonetics*, 34: 485-499.
- Keating, Patricia A. (1985). 'Universal phonetics and the organization of grammars', in Fromkin, V. (ed.), *Phonetic Linguistics*. Academic Press, pp. 115-132.
- Keating, Patricia A. (1990). 'Phonetic representations in a generative grammar'. *Journal of Phonetics*, 18: 321-334.
- Kim, Sahyang (2004). *The role of prosodic phrasing in Korean word segmentation*. (Doctoral dissertation, UCLA, Department of Linguistics).
- Kingston, John and Diehl, Randy L. (1994). 'Phonetic knowledge'. *Language*, 70: 419-454.
- Koopmans-van Beinum, Florian J. (1980). *Vowel Contrast Reduction: An Acoustic and Perceptual Study of Dutch Vowels in Various Speech Conditions*. Amsterdam: Academische Pers B.V.

- Ladd, D. Robert, Faulkner, Dan, Faulkner, Hanneke, and Schepman, Astrid 1999.
Constant "segmental anchoring" of F0 movements under changes in speech rate.
JASA 106:1543-1554.
- Labov, William (1989). 'The limitations of context: Evidence from misunderstandings in Chicago', in *Papers from the 25th Annual Regional Meeting of the Chicago Linguistic Society, Part 2: Parasession on Language in Context*. Chicago: Chicago Linguistic Society, pp. 171-200.
- Lindblom, Björn. (1990) 'Explaining phonetic variation: a sketch of the H&H theory', in Hardcastle, W.J. and Marchal, A. (eds.), *Speech Production and Speech Modeling*. Netherlands: Kluwer Academic Publishers, pp. 403- 439.
- Mattys, Sven L., and Liss, Julie M. (2008). 'On building models of spoken-word recognition: When there is as much to learn from natural "oddities" as artificial normality'. *Perception and Psychophysics*, 70: 1235-1242.
- Mehl, Matthias R., & Pennebaker, James W. (2003). 'The sounds of social life: A psychometric analysis of students' daily social environments and natural conversations'. *Journal of Personality and Social Psychology*, 84: 857-870.
- Mehta, Gita, and Cutler, Anne (1988). 'Detection of target phonemes in spontaneous and read speech'. *Language and Speech*, 31, 135-156.
- Mitterer, Holger, and Ernestus, Mirjam (2006). 'Listeners recover /t/s that speakers reduce: Evidence from /t/-lenition in Dutch'. *Journal of Phonetics*, 34: 73-103.
- Muthusamy, Yeshwant K., Cole, Ronald A., Oshika, Beatrice T. (1992). 'The OGI multi-language telephone speech corpus', In the Proceedings of the International Congress of Phonetic Sciences 1992, pp. 895-898.

- Nakamura, Masanobu, Iwano, Koji, and Furui, Sadaoki (2007). 'Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance'. *Computer Speech and Language*, 22: 171-184
- Niebuhr, Oliver (2008). 'Identification of highly reduced words by differential segmental lengthening'. Presentation at the First Nijmegen Speech Reduction Workshop, MPI, Nijmegen, the Netherlands.
- Pierrehumbert, Janet (2001). '[Exemplar dynamics: Word frequency, lenition, and contrast](#)'. In Bybee, J. & Hopper, P. (eds.), *Frequency Effects and the Emergence of Lexical Structure*. Amsterdam: John Benjamins, pp. 137-157.
- Pierrehumbert, Janet (2002). 'Word-specific phonetics'. In Gussenhoven, C. and Warner, N. (eds.), *Laboratory Phonology VII*. Berlin: Mouton de Gruyter, pp. 101-139.
- Pitt, Mark, Johnson, Keith, Hume, Elizabeth, Kiesling, Scott, and Raymond, William (2005). 'The Buckeye Corpus of Conversational Speech: Labeling Conventions and a Test of Transcriber Reliability'. *Speech Communication*, 45: 90-95.
- Pluymaekers, Mark, Ernestus, Mirjam, Baayen, R. Harald, & Booij, Geert (2006). 'The role of morphology in fine phonetic detail: The case of Dutch –igheid', in Fougeron, C., et al. (eds.), *Variation, detail and representation: 10th Conference on Laboratory Phonology*. Berlin: Mouton, pp. 53-54.
- Ranbom, Larissa J. and Connine, Cynthia M. (2007). 'Lexical representation of phonological variation in spoken word recognition'. *Journal of Memory and Language*, 57: 273-298.

- Schilling-Estes, Natalie (2002). 'Investigating stylistic variation', in Chambers, J.K. et al. (eds.), *The Handbook of Language Variation and Change*. Malden, MA: Blackwell, pp. 375-401.
- Scobbie, James M., and Stuart-Smith, Jane (Submitted). 'Sociolinguistic sampling in laboratory-based phonological experimentation'. This volume.
- Shattuck-Hufnagel, Stephanie, and Veilleux, Nanette M. (2007). 'Robustness of acoustic landmarks in spontaneously-spoken American English', in the Proceedings of the International Congress of Phonetic Sciences, Saarbrücken, Germany, August 2007. (<http://www.icphs2007.de/>)
- Shockey, Linda (2008). 'Understanding casual English pronunciation: Poles apart'. Presentation at the First Nijmegen Speech Reduction Workshop, MPI, Nijmegen, the Netherlands.
- Smiljanić, Rajka. and Bradlow, Ann R. (2009) 'Speaking and hearing clearly: Talker and listener factors in speaking style changes'. *Linguistics and Language Compass*, 3: 236–264.
- Torreira, Francisco, Adda-Decker, Martine, and Ernestus, Mirjam. Submitted. 'The Nijmegen Corpus of Casual French.'
- Tucker, Benjamin. V. (2007). *Spoken Word Recognition of the Reduced American English Flap*. Ph.D. dissertation, University of Arizona.
- Warner, Natasha (submitted). 'Reduction', in van Oostendorp, M., Ewen, C., Hume, E., and Rice, K. (eds.), *The Blackwell Companion to Phonology*.

- Warner, Natasha, and Arai, Takayuki (2001). The role of the mora in the timing of spontaneous Japanese speech. *Journal of the Acoustical Society of America* 109: 1144-1156.
- Warner, Natasha, Brenner, Dan, Woods, Anna, Tucker, Benjamin V., and Ernestus, Mirjam (2009). 'Were we or are we? Perception of reduced function words in spontaneous conversations'. *Journal of the Acoustical Society of America*, 125: 2655 (abstract).
- Warner, Natasha, Fountain, Amy, and Tucker, Benjamin V. (In press). 'Cues to Perception of Reduced Flaps'. *Journal of the Acoustical Society of America*.
- Warren, Paul, and Hay, Jen (Submitted). 'Experimental design and data collection'. This volume.

Figure Captions

Fig. 1. Waveform and spectrogram from conversational speech, "...what weekend were you guys..." Symbols in parentheses are segments for which there is little acoustic evidence. Heard in isolation, the portion corresponding to "-end were" does not consist of any identifiable segments, but the entire utterance sounds quite natural and clear.

Fig. 2. Schematic representation of carefulness, speech rate, and acoustic reduction as three distinct dimensions along which speech can fall. The carefulness dimension indicates specific examples of speech settings on the right, and ranges covered by the terms non-read, connected, spontaneous, and conversational on the left. The ordering of specific examples on the right is approximate. For example, target words in frame sentences might be more or less careful than non-read responses to prompts in a particular experimental task.

Fig. 3. Waveforms of two steps on a flap duration continuum for the word "needle" (Warner et al. in press), resynthesized to simulate reduction vs. careful speech.

Fig. 1.

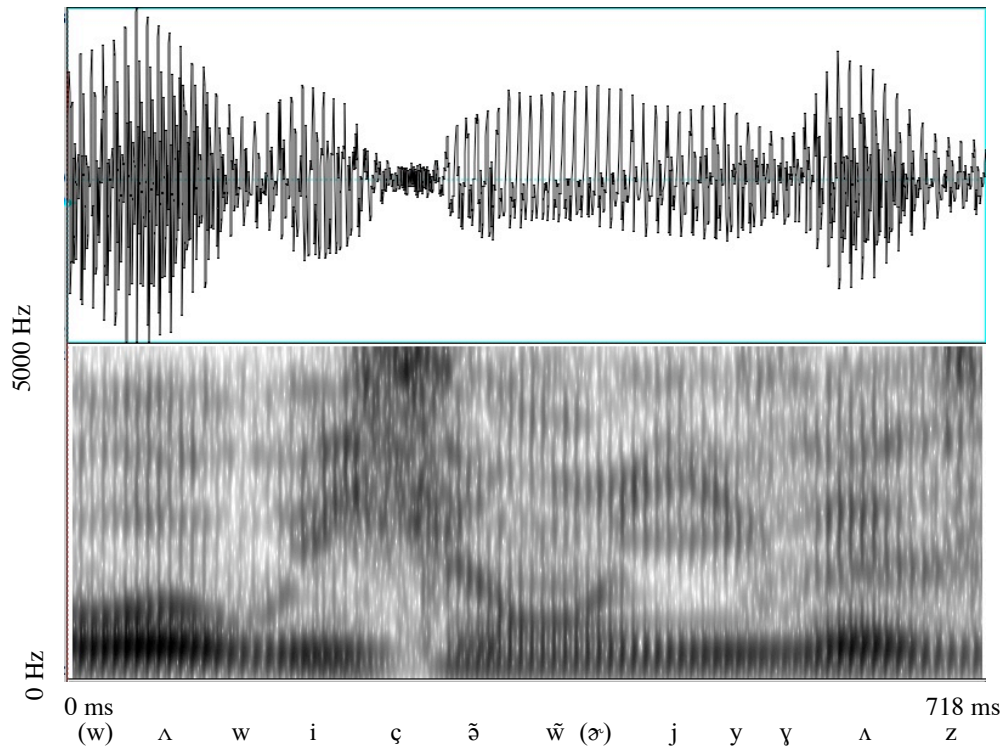


Fig 2.

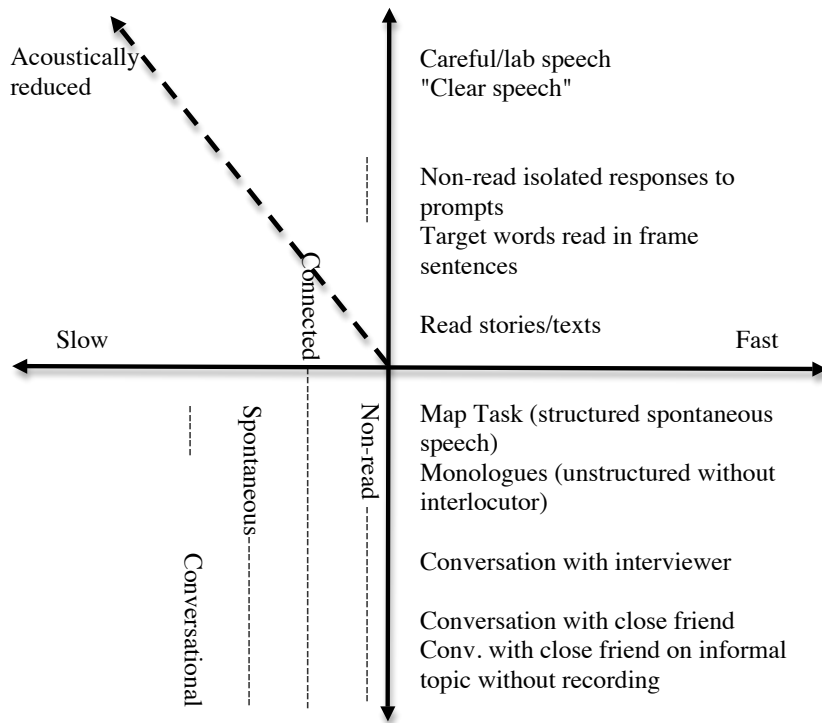
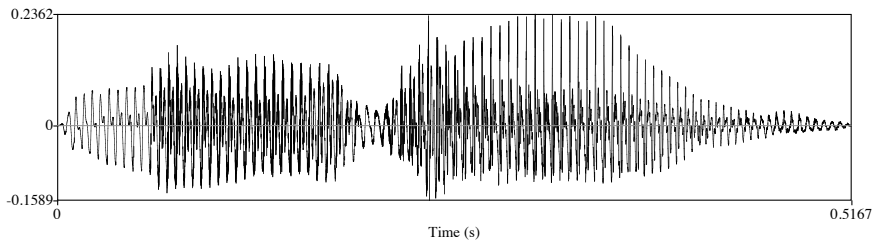
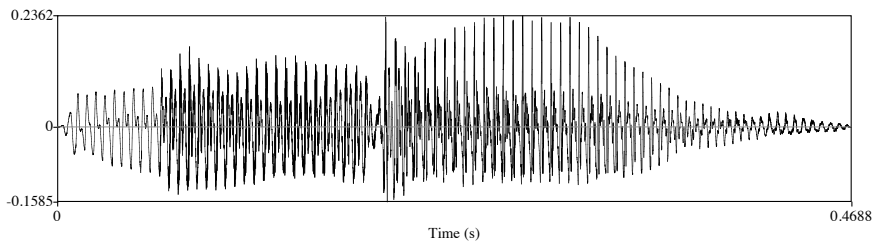


Fig. 3.



n i r l

Additional requested material:

Index terms:

reduction, spontaneous, conversation, corpus, perception

Bio:

Natasha Warner is an Associated Professor at the University of Arizona, Department of Linguistics. Her research interests are in speech reduction, speech acoustics and perception, and the three-way interface of phonetics, phonology, and psycholinguistics, along with a separate interest in language revitalization. Her language interests are primarily in English, Dutch, Japanese, and Mutsun.

Abstract:

This chapter reviews methods for studying spontaneous, conversational, or reduced speech. This is the speech we use in our daily lives, rather than the more careful varieties usually studied. The chapter examines methods for recording spontaneous speech for acoustic analysis, and also methods for obtaining spontaneous speech perception stimuli. Two overall approaches are possible: 1) using speech that has characteristics of spontaneous speech, while maintaining control over the material, or 2) using speech that is as natural as possible, with a resulting lack of control. No method is perfect, and a great many methods are being tried, because the topic is developing rapidly. The author advocates for the combined use of more controlled (less natural) methods with more

naturalistic (less controlled) methods. We hope such combinations will lead to convergent evidence, as well as to further methodological innovation.

Key words:

Index terms:

reduction, spontaneous speech, conversation, corpora, speech perception, acoustic phonetics