Prosody, intelligibility and familiarity in speech perception

Daniel Robert McCloy

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2013

Reading Committee:

Richard A. Wright, Chair

Frederick J. Gallun

Sharon L. Hargus

Gina-Anne Levow

Program Authorized to Offer Degree:

Linguistics

University of Washington

Abstract

Prosody, intelligibility and familiarity in speech perception

Daniel Robert McCloy

Chair of the Supervisory Committee:
Professor Richard A. Wright
Department of Linguistics

This thesis concerns the relationship between speech intelligibility and speech prosody, and the role that speech prosody plays in the perceptual advantage that occurs when listening to a familiar talker. A parallel corpus of 90 sentences (each spoken by three talkers of varying intelligibility) was used to create resynthesized stimuli in which fundamental frequency (f_0), intensity, and patterns of syllable duration were swapped between all possible pairs of talkers. An additional 90 sentences were reserved for use as training stimuli.

Findings from speech-in-noise tasks suggest that the contribution of prosody to intelligibility varies considerably across talkers, evidenced by differences in sentence intelligibility after prosodic replacement from different prosodic donors. In particular, high-intelligibility talkers need not have particularly "good" or "helpful" prosody if their intelligibility rests on articulatory strategies that emphasize robust segmental cues, while talkers with relatively good prosody may nonetheless have low intelligibility (due to non-prosodic factors).

Post hoc acoustic analyses of the stimuli (interpreted in light of the behavioral results) suggest that many acoustic measures index both prosodic and non-prosodic aspects of the speech signal, whereas acoustic measures that reflect only the prosodic component of intelligibility are harder to find. Mean f_0 range and mean f_0 dynamicity appear to be the most promising measures in this regard, but utterance-final creaky voicing presents a challenge to interpretation due to its exaggeration of f_0-related measures.

Results of the familiarization experiments were inconclusive; listeners trained on different talkers showed different degrees of task adaptation during familiarization/training, but were unable to generalize to a testing phase involving multiple talkers presented in random order. Thus the contribution of prosody to the familiar talker advantage remains unclear.

Table of Contents

List of Figures

List of Tables

Acknowledgments

First and foremost I thank my advisor, mentor, and dissertation chair, Richard Wright. For the last three years he has (among other things) challenged, inspired, critiqued, funded, encouraged, frustrated, and befriended me. One of the main reasons I finished graduate school was that I truly enjoyed going to our meetings. Sharon Hargus has also been a generous and understanding presence, and I am grateful for the guidance she has provided and the example of uncompromising integrity that she sets.

I owe a large debt to the other members of my dissertation committee, Erick Gallun and Gina-Anne Levow, who raised keen questions and offered seasoned advice along the way, and struggled through my inelegant draft prose to help refine this work. I would be remiss if I failed to also mention Pam Souza, who has been a great collaborator, and a model example of science in the service of humanity.

Jennifer Haywood, Gus McGrath, and Heather Morrison get thanks here as well for their work on the PN/NC corpus. Gus and Heather in particular deserve special mention for having so quickly transitioned from students to trainees to collaborators, and for doing the dirty work of hand-correcting the automated alignments to make our spoken corpus fit for research use.

My colleagues in the UW Linguistics Department have been fine companions, and I have profited from many stimulating discussions over the years. Special thanks go to Josh Crowgey, Valerie Freeman, Justin Goodenkauf, Michael Goodman, Prescott Klassen, Bill McNeill, Julia Miller, Meghan Oxley, Sarala Puthuval, John Riebold, Lisa Tittle, and most especially Darren

Tanner and Steve Moran. Most of my conversations with these folks had little to do with my dissertation research, but rather were a source of enrichment and distraction when it was sorely needed. To the extent that I seem like a well-rounded scholar, it is often because of the little facets of knowledge that I gleaned from each of them.

I must also acknowledge my dear friends outside of academia, who for several years have tolerated my absence from parties, dinners, concerts, openings, performances and festivals, and yet still embraced me fondly when I did manage to show up once in a while. My parents have been especially tolerant in this regard, and deserve thanks for so much more than I can describe here.

Along my winding path through education, I have had a number of gifted teachers who quietly inspired me to remain curious and to someday be as good a teacher as they were. I have always wanted to acknowledge them publicly; this seems as good a place as any. Thanks to (in chronological order): Sherry Brown (chemistry), David Adams (calculus), Dennis Lamb (Greek and Roman literature), Bill Moody (cellular neurobiology), Larry BonJour (epistemology), Bill Talbott (moral and social philosophy), Andrea Woody (philosophy of science), David Knechteges (Chinese literature), Bi Nyan-Ping (Chinese language), Zev Handel (Chinese historical linguistics), and Chris Stecker (psychoacoustics).

Dedication

To Sarala, *sine qua non*.

Chapter 1. Introduction

This thesis is concerned with two main questions. The first question — *what is the relationship between prosody and intelligibility?* — is investigated using careful resynthesis of read sentences, which are then presented in a listening task. By "prosody" I mean the patterns of pitch, duration and loudness that combine to form patterns of prominence and comprise the rhythms of speech. The second question — *what is the relationship between prosody and the familiar talker advantage?* — is investigated in a training task utilizing the same stimuli created to address the first question.

1.1 Overview of the thesis

The first part of the thesis comprises a review of relevant literature (Chapter 2), followed by a formal statement of research questions and description of experimental designs (Chapter 3). A detailed description of methods (including stimulus creation, data analysis, and *post hoc* measurements) follows in Chapter 4. Results of the two experiments are presented in Chapter 5, followed by a general discussion in Chapter 6.

1.2 Abbreviations and acronyms used in the thesis

Following is a table of acronyms and abbreviations used in this thesis. I have endeavored to spell out each abbreviation or acronym at the point it is first used in the document, but all of them are included here as well for convenient reference.

Table 1.1: Abbreviations and acronyms used in the thesis

Abbreviation	Explanation
CRM	Coordinate response measure (a corpus of parallel sentences of the form "Ready ⟨CALLSIGN⟩ go to ⟨COLOR⟩ ⟨DIGIT⟩ now"). In a typical competing speech usage, the listener is told to attend to the speech stream containing a particular callsign, and to repeat the color and digit from that stream. See Bolia *et al.* (2000) for details.
dB	Decibels (a dimensionless base-10 logarithmic unit commonly used to express the intensity or power of a sound, relative to a known reference value).
dB HL	Intensity in decibels relative to frequency-specific threshold averages representative of normal-hearing humans (ANSI, 2004). "HL" stands for "hearing level".
dB SPL	Intensity in decibels relative to 20 micropascals (μPa) (ANSI, 1994). "SPL" stands for "sound pressure level".
DTW	Dynamic Time Warping (a method of non-linear time-scaling of a sequence that relies on inserting or removing elements to align key landmarks in the sequences being equated). See Kruskal & Liberman (1983).
f_0	Fundamental frequency (of speech). Often represented as F_0 or F0, the form f_0 is preferred because it emphasizes that the fundamental frequency is not a formant (which are typically represented as F1, F2, *etc*). The form f0 is an acceptable substitute when glyph choice is restricted to ASCII characters.
F2×F1	The acoustic space defined by the center frequencies of the first and second formants of sonorant speech sounds (usually vowels).
HL	See dB HL.
L1, L2, *etc.*	First (native) language, second language, *etc.*
LTAS	Long-term average spectrum (a representation of the mean sound power present across various frequencies in a signal).
MCMC	Markov-chain Monte Carlo (an algorithm for sampling from probability distributions; useful for simulating the parameters of mixed models for, *e.g.*, obtaining estimated p-values).

continued on next page

Table 1.1: Abbreviations and acronyms *(continued from previous page)*

Abbreviation	Explanation
PSOLA™	Pitch synchronous overlap and add (a method for time- or frequency-domain resynthesis of speech; see Charpentier & Moulines 1988; Moulines & Charpentier 1990). PSOLA™ is a trademark of France Télécom.
RMS	Root mean square, a measure of sound magnitude, defined as: $\sqrt{\frac{1}{n}(x_1^2 + x_2^2 + \cdots + x_n^2)}$ where $x_1 \ldots x_n$ are the samples of a digital signal.
RMSE	Root mean squared error, a measure of the difference between two signals or time series, defined as: $\sqrt{\frac{1}{n}((x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2)}$ where $x_1 \ldots x_n$ and $y_1 \ldots y_n$ are values of two time series at time points n.
SNR	Signal-to-noise ratio, usually expressed in decibels: $\text{SNR}_{\text{dB}} = 20 \log_{10} \frac{A_{\text{signal}}}{A_{\text{noise}}}$ (where A_{signal} and A_{noise} are RMS amplitudes).
SPL	See dB SPL.
SRT	Speech reception threshold (the intensity level or SNR at which a listener achieves 50% correct in a speech recognition task).

Chapter 2. Background

This chapter presents an overview of research on speech intelligibility, prosody, and talker familiarity. Because the experiments described in this thesis involve speech obscured by background noise, a review of auditory masking research is also given. The discussion of talker familiarity touches on both short-term familiarity (*i.e.*, training and exposure studies) and long-term familiarity. Finally, the discussion of prosody focuses on pitch, loudness, and duration as they relate to speech perception.

2.1 Auditory masking

Auditory masking is the phenomenon whereby the perception of a (target) sound is altered or impaired by the presence of other (masker) sounds. It is well known that auditory masking is dependent on both the spectrotemporal characteristics of the masking sound as well as its relationship to the target sound. For example, even the early experiments of Wegel & Lane (1924) — which involve only pure-tone targets and maskers — reveal the importance of the relationship between target and masker sounds. The experiments of Wegel and Lane demonstrate that, regardless of the target tone frequency, masker tones *close in frequency to the target tone* are the best maskers (except when the frequencies are so similar as to cause beating).

In the nearly ninety years since, many studies have reinforced the view that what makes something a good masker is at least partly dependent on properties of the signal one is trying to

mask. This relationship may involve spectrotemporal properties of target and masker, or properties of the signal that are not strictly spectrotemporal in nature (*e.g.*, when the target signal is speech: phonotactics, transitional probabilities based on usage frequency, contextual probabilities based on semantics, *etc*). Indeed, sometimes it seems that every study of the target-masker relationship turns up new factors that have a role to play in our understanding of auditory masking.

In this section, discussion of those factors is organized in terms of the division between "energetic" and "informational" masking. It is acknowledged that the terms "energetic masking" and "informational masking" have historically been somewhat ill-defined (cf. discussions in Durlach *et al.* 2003a and Watson 2005), though for present purposes they will suffice as an organizing principle for review of the literature. The discussion of informational masking is further subdivided into two broad categories (following Kidd Jr. *et al.* 2002 *et seq*): masking due to target-masker similarity within a stimulus or trial, and masking due to unpredictability (and, consequently, listener uncertainty) across trials.

2.1.1 Energetic and informational masking

Early masking experiments with speech targets were often performed with noise maskers (*e.g.*, Hawkins & Stevens, 1950; Tolhurst, 1957b; Pollack & Pickett, 1958). The noise used was typically a random (Gaussian) signal, usually without any amplitude modulation ("stationary maskers") or frequency shaping ("white noise", as opposed to "pink noise", "speech-shaped noise", *etc*). Such random signals do not contain any meaningful information (at least from the perspective of a human listening to such sounds) and it seems generally agreed upon that stationary noise is not sufficiently speech-like to be processed by the parts of the brain that are

specialized for speech understanding. As such, elevation of speech reception thresholds (SRT) due to masker sounds that are both random (Gaussian) and stationary (temporally unmodulated in both amplitude and frequency domains) is typically termed "energetic masking", reflecting the idea that the masking is a consequence of elevated signal detection thresholds in the auditory filters of the cochlea (cf. the discussion in Moore 2008, 96–97, and the anatomically motivated notion of "peripheral masking" proposed by Durlach *et al.* 2003a).

In contrast to stationary noise masking, most everyday situations involve listening to speech in the presence of identifiable environmental noises and/or competing speech streams — a situation poorly modeled by stationary white noise maskers, since most environmental sounds are spectrotemporally dynamic, and speech is both amplitude modulated and frequency shaped (indeed, the frequency shaping of speech is itself dynamic). Thus when speech is used *as a masker*, the amplitude modulations in the masker speech create temporal variations in signal-to-noise ratio (SNR) that allow "glimpses" of the target stream (Miller & Licklider, 1950; Festen & Plomp, 1990, *inter alia*). The existence of such glimpses make speech a (potentially) worse masker than stationary noise, since listeners can use information gleaned during the glimpses to reconstruct neighboring spans of the target stream that are less clearly heard.[1]

Of course, glimpses can also occur with any amplitude-modulated masker (including random noise maskers), and if noise maskers are both amplitude-modulated and frequency-shaped so as to be maximally comparable to speech maskers in their global spectrotemporal properties, what we find is that target perception accuracy is still lower when masked by speech than when

[1]The reconstruction of masked portions of signal from adjacent glimpses may leverage lexical knowledge, phonotactics, transitional probabilities between words based on usage frequency, contextual probabilities based on semantic relationships between words, *etc.* See Section 2.2.6 for discussion.

masked by noise (Carhart *et al.*, 1969; Lewis *et al.*, 1988; Simpson & Cooke, 2005, *inter alia*). That additional masking is (by definition) one form of "informational masking" (less commonly: "perceptual masking" or "central masking"), and is usually attributed to the (linguistic) information contained in the masker signal competing for (language) processing resources at higher levels of the auditory processing streams in the brain (Durlach *et al.*, 2003a). Teasing apart the energetic and informational components of masking phenomena has been (and continues to be) an important question motivating speech perception research.

Returning to the notion of glimpsing, Cooke (2006) points out that glimpsing is not necessarily a strictly temporal phenomenon: in any given temporal span of a mixed target-and-masker sound, there can be spectral regions of the target signal that are relatively more or less obscured by the masker signal. Moreover, glimpsing need not involve an amplitude-modulated masker, so long as there is amplitude modulation present in the target signal. For example, the intensity of the formants of a relatively loud vowel might exceed the intensity of a stationary masker noise — even at relatively low SNRs — while the quieter frequency components of the vowel's spectrum are effectively obscured (cf. Figure 2.1). In such cases the glimpsed formants may convey sufficient information for the vowel to be recognized by a listener.

The main lesson to draw from glimpsing research is that energetic masking can vary along both temporal and spectral dimensions, just as the information contained in speech cues varies in its spectrotemporal distribution (compare brief broadband transients as cues for release bursts, to slower-changing formant frequency patterns as cues to vowel quality, for example). The fact that a single speech sound may have multiple non-simultaneous cues, combined with variation in the "robustness" and "precision" of those cues adds another layer of complexity to

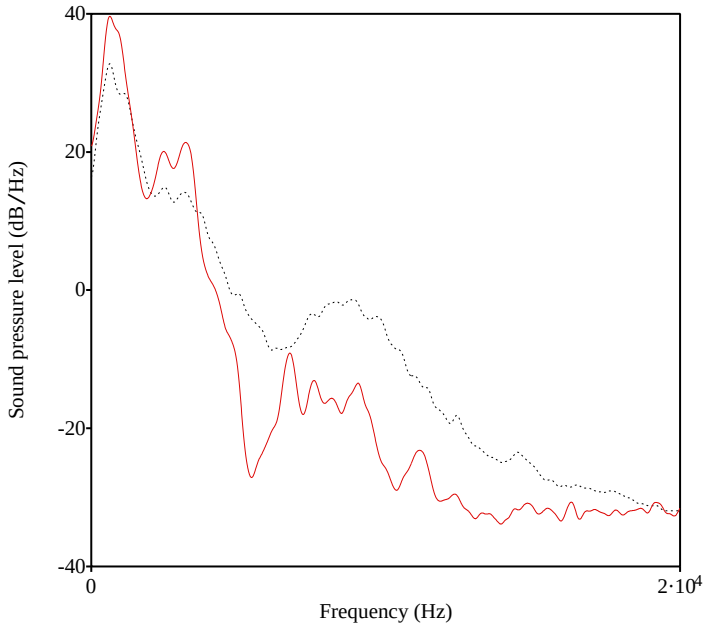

Figure 2.1: Partial glimpsing across frequencies at a particular time point, illustrated by power spectrum of an [ɑ] vowel (solid red line) poking above the long-term average spectrum of the sentence from which the [ɑ] was extracted (dotted black line), at a signal to noise ratio of 0 dB. Both spectra have undergone cepstral smoothing with a 500 Hz bandwidth.

the picture.[2] Thus for a particular instance of speech masking it is difficult to quantify how much masking has occurred in terms of the *information* recoverable. Rather, measures of masking are often expressed in terms of the change in masker intensity necessary to offset the difference in listener performance between two experimental conditions in a speech perception task.

The difficulty of pinning down degree of energetic masking is exacerbated by the way that energetic and informational masking covary in certain masker types. For example, in competing speech tasks one can reduce glimpses (both the likelihood of their occurrence and their spectrotemporal extent) — and thereby increase energetic masking — by increasing the

[2]Here I use "robust" to describe a cue that is maximally perceptible under a range of listening conditions (*i.e.*, resistant to environmental masking), and "precise" to mean a cue that is maximally perceptually distinct from cues for other speech sounds in a given language (cf. definitions in Wright, 2001, 2004b, Benkí, 2003, and Henke *et al.*, 2012, 72 ff.).

total number of background speech streams (*i.e.*, by using a "multitalker babble masker"). However, this creates a situation where the background speech streams "mask" one another in addition to masking the target stream, such that the background speech itself becomes harder to understand. This has the effect of decreasing informational masking, presumably by decreasing competition for lexical processing resources at higher levels of the brain's speech processing stream. In other words, although most research using babble maskers shows that the degree of total masking increases as the number of competing speech streams increases (*e.g.*, Miller, 1947; Brungart *et al.*, 2001), this result seems to be a combination of energetic masking going up even as informational masking goes down, making it difficult to separate the relative contribution of each to the total masking that is observed. Recent research supports this view: Simpson & Cooke (2005) report that total masking seems to plateau around 6–8 background talkers, and further increasing the number of talkers causes the masking to recede back to a lower level seen with stationary, frequency-shaped random noise. This can be understood as a decrease in spectrotemporal variation in the masker stream due to an increase in number of background talkers (as background talkers fill in the gaps between one another's words), such that the babble masker more and more closely approximates a random noise masker with spectral shape matching the long-term average spectrum of speech (Simpson & Cooke, 2005).

To summarize, research on glimpsing shows us that not all forms of energetic masking are equal, and in a typical speech-in-noise task much depends on the frequency and timing of spectrotemporal glimpses with regard to the spectrotemporal location of target speech cues. This is due to the fact that listeners use whatever information is available to make their best guess as to the content of the target speech, drawing on their perceptual experience at many

levels of representation to "fill in the details" as the signals travel up through the language-processing pathways of their brains. The next two sections discuss how masker signals can interfere at those higher levels to cause informational masking, either by sharing some crucial features with the target signal (target-masker similarity), or by virtue of frequent changes in some relevant aspect of the target-masker relationship (yielding low predictability from trial to trial, *i.e.*, listener uncertainty).

2.1.2 Target-masker similarity

Target-masker similarity can occur along a variety of dimensions, from fine-grained comparisons of spectral similarity to more abstract properties of the signals, such as the lexical token frequency of words. This section reviews various dimensions of target-masker similarity that have been shown to affect masking.

One well-studied property of target-masker similarity is the perceived spatial origin of the competing streams. Fundamentally, the localization of sound sources is perceived on the basis of interaural differences in timing and level of incoming sounds (*e.g.*, Hirsh, 1950; Kock, 1950; see Darwin, 2008 for review). Numerous studies have shown a release from masking due to (actual or simulated) spatial separation between the target and masker streams (*e.g.*, Carhart *et al.*, 1968; Freyman *et al.*, 1999; Brungart & Simpson, 2002; Freyman *et al.*, 2004; Gallun *et al.*, 2005; Kidd Jr. *et al.*, 2005; Johnstone & Litovsky, 2006); the neural circuits underlying this ability are thought to reside in the lateral superior olive in the auditory brainstem (Park *et al.*, 2004; Tollin & Yin, 2005). Strictly speaking, interaural timing and level differences are not properties of the target and masker signals *per se*, but of their relationship to the auditory system of a binaural listener. Nonetheless, spatial origin of the signals is often treated as one of the

dimensions along which target-masker similarity increases masking.

Similarity of the spectral content of targets and maskers has also been shown to decrease perceptual accuracy for speech sounds. In a series of binaural listening experiments, Gallun *et al.* (2007) show that in demanding listening tasks, noise maskers presented contralaterally to the target signal can nonetheless mask the target in cases where the target and masker are spectrally similar. In other words, ignoring a background stream is more difficult when it is spectrally similar to the target (even when target and masker are maximally separated spatially). Similarity between the voices in the target and masker streams is also predictive of masking: perception of the target stream is most difficult when the voices in the target and masker streams belong to the same person, moderately difficult when the voices belong to same-gendered talkers, and least difficult when the target and masker voices belong to a gender-mismatched pair (Brungart, 2001). It is unknown what properties of the talkers' voices underlie this phenomenon, though it has been speculated that both low-level spectral discrimination and auditory attention may be at play (Helfer & Freyman, 2008).

Another signal property known to affect masking is the dialect, accent, or language of the target and masker streams. For example, native English listeners perform better on English target speech when the masker speech is Spanish (García Lecumberri & Cooke, 2006), Dutch (Brouwer *et al.*, 2012), or Modern Standard Chinese (Van Engen & Bradlow, 2007) rather than English. Analogous experiments by Rhebergen *et al.* (2005) showed similar effects for Dutch listeners exposed to Dutch target speech with either Dutch or Swedish maskers. The release from masking seen with mismatched target and masker languages is smaller or absent in cases where the masker language is comprehensible to the listener, even if it's not their native

language (García Lecumberri & Cooke, 2006; Van Engen, 2010; Brouwer *et al.*, 2012). Likewise, for English target speech and native English-speaking listeners, masking is higher for native English competing speech than for foreign-accented English competing speech (Calandruccio *et al.*, 2010).[3]

Although findings regarding dialect, accent, or language mismatch between target and masker are sometimes attributed to target-masker similarity, they are perhaps better understood in terms of the availability, accessibility or perceptibility of the information in the masker stream (Van Engen & Bradlow, 2007; Calandruccio *et al.*, 2010; Brouwer *et al.*, 2012). In other words, it is not necessarily the similarity of the target and masker languages or dialects that matters, but rather the degree to which elements of the masker stream are recognizable by the listener, on the assumption that such recognition causes competition between masker and target at later stages of auditory processing. Support for this view comes from the finding that informational masking (indexed by reaction time) correlates with the lexical token frequency of the words in the masker speech (Boulenger *et al.*, 2010). Relatedly, Brouwer *et al.* (2012) showed that semantically anomalous English sentences provide less masking than semantically coherent ones, but no similar effect was found for English target speech masked by semantically well-formed *vs.* anomalous Dutch speech.

Taken together, these findings support the view that some cases of informational masking can be explained by "lexical pop-out" of specific words in the masker speech. According to such an explanation, comprehensible background speech can compete with target perception relatively late in the auditory processing stream (*i.e.*, at the point at which lexemes are

[3]Corollary experiments confirmed that the difference was due to variation in the intelligibility of the background talker, not spectrotemporal differences between the masker streams (Calandruccio *et al.*, 2010).

recognized or accessed). In contrast (so the explanation goes), incomprehensible background speech (such as foreign-language babble) is unlikely to trigger lexical competition, except by dint of accidental similarity to native words (a very low probability occurrence). Experiments comparing normal to time-reversed masker speech also provide support for the lexical pop-out explanation of informational masking, since SRTs are generally higher with forward (comprehensible) speech than with reversed (incomprehensible) speech.[4] These studies also suggest that a listener's past experience (in the form of word frequency) and expectations (in the form of contextual probabilities of words) both play an important role in the formation of the speech percept.

This reasoning can be extended to the findings of Calandruccio *et al.* 2010 (discussed above), which showed that high intelligibility (native) speech masks better than low intelligibility (foreign-accented) speech, on the assumption that words are high-intelligibility precisely because they more readily trigger lexical activation (due to, *e.g.*, greater cue redundancy). The findings of Brouwer *et al.* 2012 (showing weaker masking from babble comprising semantically anomalous sentences) might also be explained by appeal to lexical pop-out, though the explanation is somewhat more convoluted: a lack of coherent semantic relationships between early and late words leads to a reduction of semantic priming effects, thus decreasing the likelihood of lexical activations due to words in the masker stream, thereby reducing lexical competition from the masker stream.

In sum, there seems to be value in reclassifying most reported cases of target-masker linguistic "similarity" as cases of lexical interference from the masker speech stream.[5] Of

[4]See, *e.g.*, Hoen *et al.* (2007), though cf. Rhebergen *et al.* 2005 for a review of issues related to energetic forward masking in time-reversed speech.

course, languages may be similar along many dimensions, both above and below the lexical level (*e.g.*, typical grammatical or intonational patterns, the presence or absence of particular speech sounds or sequences, *etc*). The effect of such similarities in speech-on-speech masking situations is thusfar not well studied. Nevertheless, target-masker similarity does seem an important consideration with regard to talker voices and spatial location. The next section introduces the effect of listener uncertainty on speech perception, which offers additional insights into many of the findings discussed above.

2.1.3 Masking and listener uncertainty

In laboratory settings, another apparent source of informational masking is the trial-to-trial regularity of the target and masker stimuli. Experiments with pure-tone targets and pure-tone complexes as maskers have shown significant release from masking when the masker is held constant for the two intervals of each trial, even when the masker still varies spectrotemporally between trials (Neff & Green, 1987; Neff & Callahan, 1988). Similar results (again using tone complexes) have been obtained for masker spatial location (Fan *et al.*, 2008) and temporal location (Bonino & Leibold, 2008). Uncertainty about the target is also relevant: the presence of notched-noise maskers or off-frequency tone complex maskers induces a considerable cost (quantified via target detection threshold) when listeners must attend to two different frequency regions in a tone-perception experiment (Kidd Jr. *et al.*, 2008). Additionally, stimulus uncertainty has been shown to interact with target-masker similarity, such that masking under conditions of stimulus uncertainty is stronger when target and masker are more similar (Durlach *et al.*, 2003b). Such results suggest that listener uncertainty (or conversely, stimulus

[5] Cf. the distinction in Watson (2005) between target-masker "structural similarity" and target-masker "representational similarity".

predictability) is important to a full account of informational masking.

This ability to "tune out" predictable maskers and "tune in" to predictable targets extends also to speech targets and speech maskers. For example, early studies of listener uncertainty using speech stimuli showed higher word-recognition scores for blockwise presentation of a single talker compared to mixed-talker blocks (Sommers *et al.*, 1994), suggesting that listeners can accommodate to a particular voice when target talker identity is predictable (a finding replicated in Brungart & Simpson 2004 and Ericson *et al.* 2004; cf. also the discussion of training studies in Section 2.3.1). Similar results are seen when holding fixed other dimensions of the stimuli besides talker voice: holding fixed the lexico-semantic content of the competing speech leads to a release from masking even when the voice speaking the masker sentence changes from trial to trial (Brungart & Simpson, 2004). Other studies have shown that holding fixed the spatial location of the target speech (in the presence of competing, spatially distinct speech streams) results in a release from masking compared to cases where the target's spatial location varies from trial to trial (Ericson *et al.*, 2004; Kidd Jr. *et al.*, 2005).

However, not all experimental manipulations show the same effect of tuning out maskers or tuning in to targets. In a study using nonsense sentences as targets and maskers, Freyman, Helfer, and Balakrishnan found little to no effect of stimulus variability (and hence listener uncertainty) along the dimensions SNR, talker voice in the masking stream, and masker lexical content (Freyman *et al.*, 2007). The fact that Freyman, Helfer, and Balakrishnan report no effect of masker content predictability seems to contradict the findings of Brungart & Simpson (2004); one possible explanation is that Brungart and Simpson's study used small-vocabulary response sets (the CRM corpus), whereas Freyman, Helfer, and Balakrishnan used semantically anomalous

sentences, which may have made the listening task more challenging.

In general, experiments exploring stimulus uncertainty using speech maskers have been less common than studies using non-speech stimuli, presumably due to the difficulty of varying nonspectral dimensions of speech (like semantic content or talker identity) while controlling and quantifying the spectrotemporal similarity of the speech (as can be done fairly readily with tone complexes). Despite this relative lack of research, stimulus uncertainty may turn out to play a role in explaining previous findings thought to be related to target-masker similarity. For example, increased masking from native- *vs.* foreign-language babble maskers (*e.g.*, Rhebergen *et al.*, 2005; García Lecumberri & Cooke, 2006; Van Engen & Bradlow, 2007; Brouwer *et al.*, 2012) might be explained as a difference in stimulus uncertainty: in the foreign-language babble, the accessible semantic content is constant and predictable (*i.e.*, null) whereas the native-language babble has varying semantic content with regard to lexical activation (in that the identity and timing of words that pop out of the babble can vary from trial to trial). Further research is needed to disentangle the roles of similarity and uncertainty under such circumstances.

2.1.4 Summary of auditory masking research

The studies reviewed above show various ways in which speech perception degrades under adverse conditions. The picture that emerges is one in which listeners use whatever information is available to aid in the perceptual task, relying on information at many levels in the speech processing stream. One final study bears mentioning here, which shows that listeners adapt their listening strategy to the particular type of adversity they are confronted with. Mattys *et al.* (2009) show that severe signal degradation (*i.e.*, energetic masking) tends to promote reliance on acoustic speech cues that escape the noise, and reduce reliance on higher level lexico-semantic

cues. In contrast, situations involving depleted cognitive resources or divided attention (*i.e.*, informational masking) tend to increase reliance on lexico-semantic cues. Thus for any investigation of speech perception under adverse conditions, one must always be cognizant of the different types of information available to the listener (either in the signal or as part of her linguistic experience and expectations), and consider which types of information are likely to be compromised by a given masker type.

2.2 Intelligibility

The previous section discussed a variety of factors that impact speech perception under adverse conditions. However, it is well known that the intelligibility of speech can vary from talker to talker — or from utterance to utterance by the same talker — even when listening conditions are unchanged. The sources of such variation can in principle lie in intrinsic properties of a talker's vocal tract, or in her habits of speech pronunciation, or indeed in properties of the particular words chosen to express an idea.

Most studies of speech intelligibility focus on measurements of the acoustic signal, and look for low-level acoustic properties of speech whose presence or magnitude correlate with differences in intelligibility. Another common approach is to observe ways in which a particular talker's speech varies with different types of communicative goals. Examples include speech directed at listeners believed to have hearing impairments (often called "intentionally clear speech"), child-directed speech (AKA, "motherese" or "parentese"), speech directed at non-native speakers of the talker's language ("foreigner-directed speech"), and speech produced with intent to overcome environmental background noise ("Lombard speech"). Such studies often still rely

on measurements of the acoustic signal to quantify differences between speech styles; thus this section begins by reviewing past findings on intelligibility with regard to four acoustic dimensions of speech: vowel formant frequencies, duration, intensity, and fundamental frequency. A discussion of speech styles follows in Section 2.2.5, with special attention to intentionally clear speech and Lombard effects. Finally, listener contributions to speech perception are reviewed, focusing on language proficiency, hearing loss, cognitive factors, and attention.

Throughout this section, it bears remembering that talkers are generally unable to control the various acoustic dimensions of speech independently, and thus with regard to intelligibility it is often difficult to definitively establish the primacy of one particular dimension over and above others it covaries with. Moreover, because of the redundancy of information that is built into speech, the degradation or absence of one particular type of speech cue may or may not manifest as a reduction in intelligibility, depending on the presence or absence of other cues sufficient to transmit the information successfully. For these reasons, it should be remembered that a correlation between intelligibility and some aspect of the acoustic signal does not establish that acoustic dimension as a necessary, sufficient, or causal prerequisite for high-intelligibility speech.

2.2.1 Intelligibility and vowel formant frequencies

Perhaps the best-studied aspect of speech acoustics is the property known as "formant frequencies", which are frequency regions of the glottal source spectrum that are most strongly amplified by the resonant properties of the vocal tract. The lowest two formant frequencies during vocalic portions of speech (abbreviated F1 and F2) are the primary cue to vowel category

(Syrdal & Gopal, 1986); formants also carry information related to the identity of flanking consonantal sounds, in the form of the direction and velocity of changes in formant frequencies at the edges of vowel sounds (Delattre *et al.*, 1955).

Much research supports the belief that speech intelligibility is linked to properties of the F2×F1 vowel space (*e.g.*, Bond & Moore, 1994; Bradlow *et al.*, 1996; Hazan & Markham, 2004; Neel, 2008), though precisely which properties of the vowel space best correlate with intelligibility is still an unresolved question. From both word identification and sentence comprehension experiments, there seems to be consistent evidence for the overall size of the vowel space as a reliable predictor of intelligibility, though different studies have used different metrics of vowel space size. Size-related measures reported to positively correlate with intelligibility include F1 range (Bradlow *et al.*, 1996), F2 range (Hazan & Markham, 2004), mean distance of vowel tokens from the center of the vowel space (Bradlow *et al.*, 1996), and area of the polygon formed by vowel means (Neel, 2008; McCloy *et al.*, 2012).[6]

The diversity of methods for quantifying vowel space size reflects the lack of a well-established consensus in the field. In addition to the methods mentioned above, the area of the convex polygonal hull encompassing all vowel tokens has intuitive appeal, as it captures the total extent of vowel space better than metrics based on vowel means, and more accurately reflects the shape of the vowel space than linear metrics like F1 or F2 range. Moreover, measures like F1 and F2 range may be more sensitive to biases due to dialectal variation. For example, Hazan and Markham measured F2 range based only on tokens of /i/ and /u/, which

[6]Note that each of these studies used a different set of vowel means to construct the polygons: Bradlow *et al.* (1996) (who did not find polygonal area to significantly correlate with intelligibility) used /i ɔ a/; Neel (2008) used /i æ ɑ u/; and McCloy *et al.* (2012) used /i ɪ e ɛ æ ɑ o ʊ u/.

could be influenced by dialectal or gender differences in /u/-fronting (a known feature of many dialects of American English, including Pacific Northwestern English; see Reed 1952, Ward 2003, ch. 4). Values for mean distance from the center of the vowel space may also be influenced by dialectal variation, and furthermore are apt to vary in their utility depending on which vowels are included in the analysis. For example, inclusion of English mid-vowels /e ɛ o ɔ/ and non-peripheral vowels /ɪ ɛ ɔ ʊ/ may decrease the correlation between mean distance from center and other measures of vowel space size, effectively adding noise to the metric.[7]

Clearly more research is needed to determine which of these metrics is most stable across languages and talker groups, and which most reliably indexes differences in intelligibility. For now, we must be content with a somewhat vague acceptance that vowel space size (however it is quantified) is likely to correlate with speech intelligibility, presumably because an expansive vowel space offers more "room" to differentiate each vowel category from its neighbors, providing more easily discriminable perceptual targets for the listener. Note, however, that the relevance of listener discriminability highlights the fact that size isn't everything: two talkers' vowel spaces may be quite similar in overall extent, but vary considerably in the relative positions and extents of the vowel categories comprising them. For this reason, other metrics related to vowel space size are of interest: a measure of the extent or spread of individual vowel categories, or a measure of the degree of vowel phoneme overlap, crowding, or encroachment. To date such measures have not been widely adopted; McCloy *et al.* (submitted) measures category spread via mean vowel cluster size and finds a correlation with intelligibility, and it has been proposed that phonemic crowding can be indexed by the total repulsive force of the vowel system (see Wright, 2004a; McCloy *et al.*, submitted).

[7] See McCloy *et al.* (submitted) for discussion.

2.2.2 Intelligibility and duration: Speech rate and speech rhythm

One of the earliest studies of the relation between intelligibility and speech rate was performed by Tolhurst (1957a), who found a significant effect for duration among three speech styles (prolonged, normal, and staccato) in an analysis of variance test. In more recent studies, evidence has been mixed as to whether speech rate is predictive of intelligibility. For example, Sommers *et al.* (1994) found a significant effect of variation in speech rate on intelligibility, suggesting that speech rate is one of the factors relevant to masking due to listener uncertainty (cf. Section 2.1.3). In contrast, Bradlow *et al.* (1996) found no correlation between mean sentence duration and intelligibility, and Krause & Braida (2002) found that the intelligibility difference between (high-intelligibility) clear speech and (lower-intelligibility) conversational speech is apparently independent of speaking rate.[8] Examining word durations, Bond & Moore (1994) found differences in isolated word durations between two talkers of varying intelligibility, but no difference in words in sentential context. Finally, in a modern study echoing Tolhurst (1957a), Mayo *et al.* (2012) examined the perception of plain, infant-directed, computer-directed, foreigner-directed, and shouted speech, finding a correlation between sentence duration and word error rate across speech styles, as well as significant segment length differences across speech styles for stops, fricatives, nasals, vowels, and diphthongs (but not affricates, liquids or glides).

Mayo, Aubanel, and Cooke interpret their findings to mean that "much of the prosody-related intelligibility gain comes from durational increases", and certainly when speech is masked by an amplitude-modulated signal — be it environmental sounds, modulated noise or

[8] See Section 2.2.5 for a discussion of clear speech.

competing speech — we would expect increased duration to increase the likelihood of glimpsing, and thereby increasing the likelihood of accurate perception of the target speech. However, in natural speech, duration co-varies with other dimensions of linguistic variability (*e.g.*, segmental reduction), making it somewhat unclear whether correlations seen in the literature reflect the effects of duration alone, or whether duration is merely a convenient surrogate for other, harder-to-measure properties of speech.

There is also some evidence that patterns of duration (*i.e.*, rhythm) can contribute to speech intelligibility. The proper description and classification of linguistic rhythm has been a long-standing debate in the field, and a thorough rehashing of those issues is beyond our purposes here.[9] Nonetheless, two studies on speech rhythm bear reporting here for their relation to intelligibility. First, with regard to duration in foreign-accented speech, Quené & van Delft (2010) found that the speech reception thresholds (SRT) of high-intelligibility non-native Dutch sentences were improved when the sentences were resynthesized to match the durational patterns of a native Dutch speaker, while the SRT of native Dutch speech was degraded when the native sentence was resynthesized to have the durational patterns of non-native speech. Second, in a 2×2 design investigating native duration patterns and syllable-isochronous duration patterns where all stimuli had been monotonized, it was found that target speech perception was above chance only when the target had native-like duration patterns and the competing speech had syllable-isochronous duration patterns (Cushing & Dellwo, 2010). In other words, subjects performed near chance when either the masker had native-like durations, or the target had syllable-isochronous durations.

[9] A succinct summary of the issues and evidence regarding the classification of linguistic rhythm can be found in Ramus (2002).

Both of those studies suggest an intelligibility benefit for stimuli with native-like durational patterns. The relationship between rhythmic variations (*e.g.*, between native-like and non-native-like rhythms), and summary measures of duration like mean sentence length or mean syllable duration, is not well understood.[10] One recent study looking at differences in temporal structure between conversational and clear speech found differences in occurrences of vowel and consonant reduction and deletion, leading to differences in the number of prosodic phrases, but an overall consistency across speech styles in the ratio of consonantal to vocalic stretches and the variability of each (Smiljanić & Bradlow, 2008). In other words, consonants and vowels were both subject to durational reduction in conversational style (in roughly equal measure), and the variability in consonant and vowel duration remained constant across speaking styles (relative to the overall speech rate).

Of course, there is more to rhythm than simply duration patterns, and none of the studies mentioned above accounted for differences in intensity patterns between the native and non-native conditions (though they did equate the mean intensities of all stimuli). A discussion of the relation between intensity and intelligibility follows in Section 2.2.3; unfortunately there will be little to say about rhythmic patterns of intensity as related to speech intelligibility, as such research is still in its infancy (though see Tilsen & Johnson 2008 for a recent inroad into quantifying speech intensity rhythms through Fourier analysis).

2.2.3 Intelligibility and intensity: Audibility, rollover and SNR

When considering the relationship between the intensity of speech and its intelligibility, the most obvious consideration is whether the speech is loud enough to be above the listener's

[10] Cf. discussion in Ramus (2002).

auditory thresholds in the relevant frequency regions. In other words, for speech to be intelligible it must be audible. However, there are also aspects of suprathreshold hearing and speech perception that are level-dependent. For example, Dubno *et al.* (2012) report that speech envelope cues (*e.g.*, manner cues that help distinguish fricatives from stops) are most readily perceived at around 60 dB SPL, with performance declining at both lower and higher intensities. The decline in speech reception at high intensities is usually termed "rollover", and has been shown to be somewhat frequency-dependent: rollover is stronger for high-frequency and broadband sounds than for low-frequency sounds, in both normal-hearing and hearing-impaired listeners (Summers & Cord, 2007). However, the level at which rollover begins varies between normal-hearing and hearing-impaired populations (Summers & Cord, 2007), and in general listeners with hearing impairment gain less advantage from supra-threshold changes in signal level than do normal hearing listeners (Summers & Molis, 2004).

Returning to considerations of audibility, in real-world settings audibility is affected not only by the source power of the talker's speech, but also by the distance between talker and listener, their relative head orientations, the presence of background noise, and any irregularities in auditory thresholds of the listener (*i.e.*, hearing loss or congenital deficit). In contrast, laboratory settings offer controlled signal levels, minimal background noise, and minimal or no effect of head orientation due to the use of headphones (or fixed head position with free-field sound sources). Irregularities in auditory thresholds can also be accounted for by audiometric testing of listeners, and to some degree compensated for by selective amplification of frequencies or other signal processing techniques. Thus in the laboratory audibility is for the most part well within the experimenter's control, and the primary audibility consideration

becomes the variations in intensity of the target signal itself, especially in relation to the intensity of different masking signals.

In all speech stimuli there is considerable variation in intensity across the temporal span of the stimulus. Some of this variation is due to inherent differences in intensity of different speech sounds. For example, in general the vowel [ɑ] is louder than [i] (Ladefoged, 1967), [f] is quieter than [s] (Ladefoged & Maddieson, 1996), and vowels are usually much louder than consonants (Horii *et al.*, 1971). In sentence-length or longer stimuli, there is also variation in the mean intensity of each syllable or word. This variation is not random, but rather is determined by linguistic properties of the speech, in particular lexical stress and phrasal accent (Fry, 1955; Sluijter & van Heuven, 1993, 1996; Plag *et al.*, 2011). There is also a cross-linguistic tendency for words at the ends of utterances to be quieter than words near the start of the utterance (Strik & Boves, 1995; Trouvain *et al.*, 1998).

For speech presented in stationary background noise, the inherent intensity variations in the target speech can lead to differences in SNR across the span of the utterance, and increase or decrease the probability of glimpsing a particular word depending on its position or role in the sentence. One approach to mitigate this effect is to use an amplitude-modulated masker that tracks the amplitude envelope of the target speech, thereby keeping SNR effectively constant at a fairly short temporal scale (instead of calculating SNR at the whole-stimulus level, as is more typical). However, some cues to the identity of speech sounds are preserved in the amplitude envelope, (*e.g.*, intensity rise time as a cue to fricative/stop contrasts; Shinn & Blumstein 1984). This makes envelope-tracking maskers less effective than stationary maskers in some cases (Horii *et al.* 1971; Van Tasell *et al.* 1987; Bashford *et al.* 1996; see Wright 2004b for review). For

this reason, the choice between stationary and amplitude-modulated maskers necessarily involves a trade-off with regard to which cue(s) are most likely to be masked. Moreover, given the complex interaction between information redundancy on one hand, and cue precision and cue robustness on the other, it is difficult to separate exogenous and endogenous sources of information in a speech perception task. That is, it is difficult to estimate how much information in the signal has escaped masking and actually been heard, and how much has merely been reconstructed by the listener based on sparse, robust cues that escaped masking combined with past experience or lexico-semantic cues.

In sum, aside from considerations of audibility, there seems to be little relationship between intensity and intelligibility at longer time scales (*i.e.*, at the sentential level and above) as long as intensity variations are small enough to avoid rollover effects. Rather, the variations in intensity that are most relevant to intelligibility in typical situations are the variations that are intrinsic to the speech signal, either as part of the segmental content of speech or superimposed prosodic patterns.

2.2.4 Intelligibility and pitch: Gender and voice quality

Some studies have found that speech produced by female talkers is more intelligible than speech produced by male talkers (Bradlow *et al.*, 1996; Hazan & Markham, 2004), while others show no correlation between intelligibility and talker gender for normal-hearing listeners (Kiliç & Ögüt, 2004; Neel, 2008). Due to human sexual dimorphism of the vocal tract organs, the fundamental frequency (f_0) of speech is typically much lower for adult males than females, so a natural question to ask is whether gender differences in intelligibility are due (in part, at least) to differences in f_0. However, what is typically found is that static measures of pitch such as

mean f_0 do not correlate well with speech intelligibility (*e.g.*, Picheny *et al.*, 1986; Bradlow *et al.*, 1996; Hazan & Markham, 2004; Lu & Cooke, 2009), suggesting that gender differences in intelligibility are merely *indexed* by differences in mean pitch, rather than *caused* by them (cf. discussion in Bradlow *et al.*, 1996).

In contrast, measures that reflect a talker's dynamic use of pitch do seem to be relevant to intelligibility. For example, both Bradlow *et al.* (1996) and McCloy *et al.* (submitted) report a positive correlation between a talker's f_0 range and intelligibility, regardless of talker gender. Further evidence that dynamic properties of f_0 matter to intelligibility come from studies of speech stimuli with manipulated f_0 contours. Binns & Culling (2007) report a significant increase in SRT in an English competing speech task when the f_0 contour of target speech is flattened or inverted, but no effect when the masker speech is similarly manipulated. Watson & Schlauch (2008) report similar results for speech with flattened f_0 in white noise. Predictably, flattening pitch contours also impacts intelligibility in lexical tone languages like Modern Standard Chinese (Patel *et al.*, 2010).

Changes in f_0 may also have corollary effects on the signal that impact intelligibility. Laryngealization or "creaky voicing" is common as a prosodic marker (Lehiste, 1979; Kreiman, 1982; Dilley *et al.*, 1996), and is associated with a drop in both f_0 and intensity (Gordon & Ladefoged, 2001). The drop in intensity makes creaky-voiced words particularly susceptible to masking due to elevated word-level SNR, particularly with stationary noise maskers. Creaky voicing is especially likely in utterance-final position due to pitch declination in common utterance types (neutral declarative prosody in most languages involves pitch declination).

Pitch and voice quality can also be exploited for sociolinguistic purposes (*e.g.*, to index gender or sexual orientation; see McConnell-Ginet 1978; Gaudio 1994; Podesva 2011, *inter alia*), and the linguistic uses of pitch are not necessarily consistent across languages (Majewski *et al.*, 1972; Todaka, 1993; Yuasa, 2008; Keating & Kuo, 2012), dialects (Grabe & Post, 2002; Clopper & Smiljanić, 2011), or sociolects (McLemore, 1991; Britain, 2008). Moreover, style-based changes in pitch are often accompanied by changes in other dimensions relevant to intelligibility (*viz.* intensity, duration, and vowel formant frequencies, cf. Section 2.2.5). For example, infant-directed speech may show several simultaneous changes, including increased pitch range, exaggerated short/long vowel duration contrasts, and increased intensity (Beckford-Wassink *et al.*, 2007); other researchers have shown expanded vowel space size in infant-directed speech (Kuhl *et al.*, 1997). Such facts make correlations among intelligibility, gender, and pitch difficult to disentangle from other acoustic dimensions of speech, and any individual correlation should be interpreted cautiously. It also underscores the need for studies of intelligibility that vary these acoustic dimensions independently (a feat only possible with resynthesized or fully synthetic speech).

Returning to the question of gender differences in intelligibility, results regarding the dynamic use of f_0 (*i.e.*, the studies using flattened f_0 contours mentioned above) might explain the findings of Bradlow *et al.* (1996), who report a gender difference in intelligibility using sentential stimuli. Hazan & Markham (2004) also report a gender difference in intelligibility using word list stimuli, though they admit that the gender difference seen in their data was much less clear than that reported by Bradlow, Torretta, and Pisoni. Given the lack of findings regarding gender differences in other studies, it is possible that idiosyncracies of the talker

samples are an issue. In other words, there are many correlates of speech intelligibility, and any of them may lend the impression of a gender difference in intelligibility if left unmeasured, or improperly modeled statistically. On the other hand, studies of gender differences in the sociolinguistic use of f_0 make gender differences in intelligibility seem plausible or even likely, though such studies suggest shifting the locus of difference from anatomical to social factors. More research is required to clarify these issues.

2.2.5 Intelligibility and speech styles: Clear, reduced, and Lombard speech

It is well known that talkers have some voluntary control over the intelligibility of their speech, in that a single talker can adopt different speaking styles that vary in intelligibility. Early work by Tolhurst on intelligibility and speech styles was discussed in Section 2.2.2 with regard to speaking rate (Tolhurst, 1957a); related studies examined the effect of instructions to talkers to speak clearly (Tolhurst, 1954, 1955). More recently, research into the intelligibility of speech styles has mostly focused on "clear speech" — speech directed toward listeners who are hard of hearing and produced with intent to aid comprehension — a line of research pioneered by Durlach, Braida and colleagues (*e.g.*, Picheny *et al.*, 1985, 1986, 1989; Uchanski *et al.*, 1996; Krause & Braida, 2004). These and related studies show intelligibility improvements in the range of 10–20% over conversational speech.

The acoustic differences between clear and conversational speech span a range of acoustic parameters discussed in Sections 2.2.1–2.2.4, including changes in vowel formant frequencies, speech rate (due to both longer words and more pauses in clear speech), intonation patterns, and segmental reduction or deletion (see Picheny *et al.*, 1986; Li & Loizou, 2008; Smiljanić & Bradlow, 2008; Hazan & Baker, 2011, *inter alia*). However, precise characterization of each

parameter's contribution to intelligibility is elusive, due at least in part to processing artifacts that arise in speech resynthesis, which is necessary to manipulate these parameters independently (cf. discussions in Picheny *et al.*, 1989; Uchanski *et al.*, 1996; Liu & Zeng, 2006; Krause & Braida, 2009).

A related line of research involves the changes to speech produced in the presence of background noise (often called "Lombard speech" after Etienne Lombard, who first characterized the phenomenon). Lombard speech has been shown to involve changes in intensity, pitch, formant frequencies, duration, and spectral tilt (Lane & Tranel, 1971; Summers *et al.*, 1988; Lu & Cooke, 2008), though the changes are less pronounced than with read clear speech (Hazan & Baker, 2011). Interestingly, talkers' adjustments seem to be sensitive to the content of their speech: at high noise levels, the f_0 and duration of words bearing a high informational load will be preferentially modified (Patel & Schell, 2008). Talkers' adjustments also seem to be sensitive to the content of the background noise: when speaking over amplitude-modulated noise, talkers appear to predict pauses in the background noise and time their speech to take advantage of the gaps in the masker (Cooke & Lu, 2010).

In general, speech alterations due to the Lombard effect seem to be automatic, subconscious, and difficult to suppress (Pick Jr. *et al.*, 1989), and quite variable from talker to talker (Junqua, 1993). Nonetheless, Lombard speech has been shown to be more intelligible than speech produced in quiet (Dreher & O'Neill, 1957; Summers *et al.*, 1988), though again, the acoustic properties of responsible for the difference in intelligibility is not entirely clear. One contributing factor seems to be the flatter spectral tilt of Lombard speech, which may increase spectrotemporal glimpses compared to RMS-matched normal speech (Lu & Cooke, 2009).

2.2.6 Intelligibility and linguistic content

To a first approximation, Sections 2.2.1–2.2.5 deal with correlates of speech intelligibility at the level of the acoustic signal (intensity, duration, f_0) and at the phonetic/phonological level (vowel space characteristics). This section discusses relationships between speech intelligibility and lexico-semantic aspects of speech, which are not properties of the speech signal *per se*, but rather properties of listeners' (shared) linguistic knowledge that is used to interpret the incoming signal.

A fundamental property of auditory word perception is that words may be perceived even when not all speech sounds that comprise the word are heard. An extreme example of this is the perception of words where a portion of its speech sounds have been completely excised and replaced by environmental sounds of equivalent duration (Warren, 1970). The ability of listeners to do this successfully depends on the "neighborhood density" of the word; neighborhood density is usually estimated by counting the number of other words in the lexicon that differ by only one phoneme by substitution, addition, or deletion (Luce & Pisoni, 1998). Words from sparse lexical neighborhoods are typically more readily perceived than words from dense lexical neighborhoods (Vitevitch & Luce, 1998; Ziegler *et al.*, 2003), although this does not seem to be universally true of all languages (Vitevitch & Rodríguez, 2005).

Another aspect of words that can influence their intelligibility in noise is how frequent or common they are. Numerous studies have shown a perceptual advantage for high-frequency words in lexical decision tasks (word/non-word judgments) in both reaction time and accuracy, as well as higher intelligibility in speech-in-noise tasks (*e.g.*, Howes, 1957; Savin, 1963; Vitevitch

& Luce, 1998; Dirks *et al.*, 2001; Takayanagi *et al.*, 2002; Vitevitch & Rodríguez, 2005). In competing speech tasks with sentence-length stimuli, high-frequency words are also more likely than low-frequency words to pop out of the background and cause distraction (*i.e.*, retard reaction time or cause a misperception of a target word for a background competitor; Boulenger *et al.* 2010). A third property of words relevant to speech perception is "neighborhood frequency", or the mean lexical frequency of all the phonological neighbors of a word; a high neighborhood frequency generally impedes word perception (Luce & Pisoni, 1998).

Finally, the semantic context of words in continuous speech contributes to their intelligibility. For example, Lewis *et al.* (1988) showed that words that are predictable from context are harder to mask, in that they are recoverable (or at least guessable) at lower SNRs than the same words in low-context sentences. At the same time, words that are predictable from context are typically articulated less distinctively, as though the talker were balancing her own articulatory effort with expectations that the listener is paying attention and can recover some words more easily than others (Lindblom, 1990; Wright, 2004a).

Overall, the effect of linguistic content on intelligibility presents a difficult challenge for speech scientists. Listener contributions to speech perception in the form of prior linguistic experience can only be roughly approximated, by reference to community-level aggregates like word frequency in corpora or estimates of lexical neighborhood based on standard pronunciations and expected levels of word knowledge. This is in sharp contrast to the spectrotemporal content of stimuli, which is under complete experimental control in the case of synthetic stimuli and is at least measurable in the case of recorded speech. In the next section, other forms of listener contribution to intelligibility are also considered.

2.2.7 Listener contributions to intelligibility

There are at least three aspects of the listener that are relevant to considerations of speech intelligibility: hearing impairment, cognitive capacity, and language proficiency. A fourth factor, familiarity with the voice of the talker, will be discussed in detail in Section 2.3.

A thorough characterization of the types and causes of hearing impairment is beyond the scope of this work.[11] However, for our purposes it suffices to state that most types of age-related hearing loss typically involve elevated detection thresholds (especially in frequencies above 4 kHz), and decreased frequency selectivity (*i.e.*, broader auditory filter bandwidths) across all frequencies. This increases the difficulty of the signal detection task that underlies auditory perception generally (making it hard to separate target from background), and also impairs perception of the spectral composition of sounds (which can make certain speech sounds more difficult to discriminate).

Whether listeners have normal or impaired hearing, speech perception (like most perceptual tasks) becomes more difficult under increased cognitive load. For example, Francis & Nusbaum (1996) show that listeners performing a syllable recognition task with a simultaneous digit span recall task have slower reaction time under high cognitive load when there is variability in speech rate of the stimuli. Francis and Nusbaum interpret this finding as evidence that speech perception tasks like talker- or rate-normalization are active (as opposed to passive) perceptual processes and thus induce a cognitive load. Recent research supports this view of speech perception as cognitive load: Mattys *et al.* (2009) report that competing talker tasks increase

[11] See Pickles (2008, chap. 10) or Moore (2008, 117–119) for overviews, or Sataloff & Sataloff (2005) and Gordon-Salant *et al.* (2010) for detailed treatments.

reliance on lexico-semantic cues, whereas noise masking increases reliance on low-level acoustic cues. The link between reliance on lexico-semantic cues and cognitive load comes from findings that older listeners (with diminished cognitive resources) benefit more from contextual (lexico-semantic) cues than do younger listeners (Pichora-Fuller *et al.*, 1995; Sommers & Danielson, 1999).

Further evidence for the importance of cognitive capacity comes from the study of auditory attention and the distribution of attentional resources. In a divided listening task using the CRM corpus, Best, Gallun, Ihlefeld, and Shinn-Cunningham showed that listeners perform substantially worse when asked to report keywords from both of two simultaneous speech streams, especially when the two speech streams are widely separated spatially (Best *et al.*, 2006). This contrasts with results showing that spatial separation of target and masker streams *aids* the perceptual task when only one stream is the focus of attention, as does prior knowledge of the target's spatial location (cf. Kidd Jr. *et al.* 2005, and discussion of spatial separation in Section 2.1.2). In a related study, Kitterick, Bailey, and Summerfield show that prior knowledge of how to direct attention along any of several stimulus dimensions can help listeners overcome high cognitive load due to multiple competing speech streams. Knowledge of talker identity, target spatial location, or target onset time improved SRT of target speech in a 26-talker listening task in which pairs of talkers began synchronously (with the 13 pairs staggered at regular intervals), but such knowledge had much less benefit in an easier 13-talker task in which single talkers (rather than pairs of talkers) were presented (Kitterick *et al.*, 2010).

As discussed in Section 2.1.2, one of the factors that influences the degree of masking in a speech-on-speech task is the relationship between the native language of the listener and the

language of the masker speech; masker speech that is comprehensible to the listener masks more strongly than speech that the listener can't understand. This relationship applies to the target speech as well: a monolingual English speaker will obviously find Tamil target speech completely unintelligible, and as she begins to learn Tamil there will almost certainly be a period of first-language (AKA, L1) interference in her perception of Tamil speech sounds. In one study demonstrating this, Hazan & Simpson (2000) showed that in an English consonant identification task, the errors made by L1 Spanish speakers and L1 Japanese speakers differed in a way that was partly predictable from differences between English and the listeners' native phonologies. The pattern of errors was preserved even for stimuli with cue enhancement (selective amplification of formant transitions and release bursts), showing the robustness of L1 phonological interference.

In addition to interference from L1 segmental phonology, there are also subtle biases in cue weighting built into listener perceptual processes. For example, Zielinski reports that native English speakers listening to non-native (L2) English speech preferentially rely on syllable stress patterns, and segmental cues within prominent syllables. In other words, native listeners "[apply] native speech processing strategies to a non-native speech signal," in a way that makes errors of stress placement particularly destructive to intelligibility (Zielinski, 2008, 80).

The degree and type of listener exposure to the language of the target speech is also an important consideration. For example, Pinet *et al.* (2011) showed that L1 French listeners with low-proficiency L2 English performed better when target speech was French-accented English than when target speech was native (Southern England) English, but the pattern was reversed for L1 French listeners with high English proficiency. Pinet, Iverson, and Huckvale attribute this

finding to listener experience and exposure: the listeners with low-proficiency English had mostly been exposed to French-accented English in classroom settings, whereas the listeners with high-proficiency English had more broad exposure to the language. Similar results have been reported for Spanish learners of English (Imai *et al.*, 2005), Chinese and Korean learners of English (Bent & Bradlow, 2003), and speakers of many other languages.

2.2.8 Summary of intelligibility research

Much of the research on intelligibility described in preceding sections could be classified as one of two types, which could be called the "modeling approach" and the "controlled manipulation approach". Studies using the modeling approach examine two or more speakers (or speaker populations) and attempt to correlate differences in intelligibility with differences in other attributes of the speakers (*e.g.*, speech rate, pitch range, gender, *etc*).[12]

A major shortcoming of the modeling approach is that naturalistic speech varies among many dimensions simultaneously, so it is difficult to segregate the contributions of the various dimensions being measured. This makes it difficult to know whether variation on a given dimension (*e.g.*, gender) is in fact genuinely contributing to differences in intelligibility, or merely indexing some other unmeasured dimension (*e.g.*, vowel space expansion) that is in fact what listeners are relying on in cases where the perceptual task is successful. High-quality spoken corpora are also labor-intensive to create and to analyse acoustically, and automated measurements still introduce sufficient measurement noise to be unsuitable for many (if not most) research questions. Finally, though there may be general agreement about which aspects of speech are likely to be relevant to intelligibility broadly speaking, as yet there is no

[12]Studies of this type include Bond & Moore (1994); Bradlow *et al.* (1996); Hazan & Markham (2004); Neel (2008).

consensus among researchers regarding how those dimensions should be quantified (cf. the various ways of quantifying vowel space size discussed in Section 2.2.1).

In contrast, studies using the controlled manipulation approach take a dimension believed to be relevant to the speech perception task, and create stimuli with a known manipulation along that dimension.[13] Such studies allow researchers to segregate the contribution of different dimensions of speech (a good example of this is the separation of talker characteristics "who, where, and when" in Kitterick *et al.* 2010). However, such studies can be criticized for being too far removed from real-world scenarios of speech perception, either because the speech content is too restricted (*i.e.*, heavily stereotyped sentences such as those in the CRM corpus), or because the experimental conditions are too contrived (*e.g.*, knowing that the target speech will occur in the fourth of the thirteen pairs of talkers, which start at 800 ms intervals, as in Kitterick *et al.* 2010).

The ideal middle ground between the modeling and manipulation approaches involves fully synthetic speech, such that all aspects of the signal are under experimenter control. Until such an approach becomes feasible, a complete understanding of speech intelligibility is likely outside our grasp. Regardless of which approach has been taken, few studies have investigated the contribution of speech prosody to intelligibility, so there is still little understanding of what constitutes "good" prosody from an intelligibility perspective.

[13] Examples of this approach include Kidd Jr. *et al.* (2005); Kitterick *et al.* (2010); Dubno *et al.* (2012).

2.3 Talker familiarity

Section 2.2.7 mentioned several studies in which familiarity with the language or dialect of the target speech conferred a perceptual advantage (*e.g.*, Bent & Bradlow, 2003; Imai *et al.*, 2005; Pinet *et al.*, 2011). A variety of studies have shown that familiarity with a talker's voice is also advantageous. Many such studies involve training listeners to reliably identify previously unknown talkers by voice alone. Typical methodology involves a period of exposure to voices paired with names, followed by a training phase where listeners must supply the names but are given immediate feedback, and a final testing phase where listeners supply the names of talkers and no feedback is given. The training studies discussed in Section 2.3.1 all use some variant of this procedure, but vary in the number of talkers to be identified, the type of training and/or testing stimuli (words *vs.* sentences; native *vs.* foreign speech), the presence and type of masking noise in the training and/or testing phase, and various other subtle differences (*e.g.*, associating voices with pictures instead of names). Section 2.3.2 discusses two studies in which longer-term familiarity with a talker's voice is investigated.

2.3.1 Training studies

A foundational study establishing a familiarity advantage in speech perception was Nygaard *et al.* (1994), in which listeners trained for nine days to associate ten voices with ten common first names (five male, five female) based on single-word utterances. On the tenth day, listeners correctly identified talkers based on new words not in the training set, showing that successful talker identification was not limited to the actual words used in the training phase. More importantly, listeners were presented words in various levels of background noise, and were

significantly better at word identification in noise when the word was spoken by a familiar talker.

Another important early finding is that the familiarity advantage does not readily generalize across stimulus types. Listeners trained to identify talkers based on single words do not benefit from talker familiarity when tested on sentence-length stimuli in noise; likewise, listeners trained to identify talkers based on sentential stimuli do not benefit from talker familiarity when tested on single-word utterances (Nygaard & Pisoni, 1998; Yonan & Sommers, 2000). A corollary finding is that talker identification training progresses much faster when based on sentential stimuli than when based on single words (three training sessions *vs.* nine to achieve comparable levels of talker identification success; Nygaard & Pisoni 1998).

The finding that talker identification training does not generalize to different stimulus types has been repeated in a variety of ways. Levi *et al.* (2011) show that training monolingual English listeners to discriminate (bilingual) talkers based on German words confers no advantage when the listeners are tested on English words.[14] Van Engen (2012) reports similar results when the training is carried out in different kinds of background noise: if listeners are tested on speech with a babble masker, they perform better when they were trained with a babble masker (rather than speech-spectrum noise) and when the language of the babble masker matches the language of the masker they were trained with.

Finally, there are age-related differences in the ability to identify talkers and the ability to benefit from familiarity. In one study, young listeners (age 18–24) were near ceiling (98%) on a

[14]In fact, just learning to correctly discriminate talkers is difficult if the training speech is not readily comprehensible to the listener, either because it is a foreign language (Perrachione & Wong, 2007) or because of a phonological deficit such as dyslexia (Perrachione *et al.*, 2011).

four-talker identification task on both days of training, while older listeners (age 66–87) hovered around 75%. However, older listeners paradoxically showed a greater benefit for talker familiarity in a speech-in-noise perception test at the end of the training, suggesting that older listeners were benefiting from implicit learning or exposure even if they could not make use of it in a talker identification task (Yonan & Sommers, 2000). A second experiment confirmed this: old and young listeners who were merely exposed to talkers in an orthogonal task showed the same pattern in a speech-in-noise task as the listeners who were expressly learning talker identification (namely, familiar voices were more intelligible, and older listeners benefited more from exposure than young listeners).

2.3.2 Long-term familiarity studies

The training studies discussed in Section 2.3.1 involved talker familiarization periods ranging from two to ten days. Relatively few studies have investigated the benefit of longer-term familiarity, on the order of weeks, months or years. One such study involved college students shadowing a target voice belonging to the professor of one of their current courses in the presence of a competing talker. Students who were explicitly told that the voice belonged to their professor performed better on the shadowing task than both classmates who were unaware of the talker's identity and other students who had no prior experience with the talker (Newman & Evers, 2007).

In another study, Souza *et al.* (in press) recruited older participants in pairs (spouses and long-term friends) and recorded sentential stimuli from one member of each pair; the other member listened to sentences from all of the talkers in quiet, speech spectrum noise, and babble. Duration of relationship ranged from 7 to 58 years, and all participants attested to conversing

with their partner a minimum of 3 hours per week. Results showed that, for all listeners, speech perception was best for the talker whom they were familiar with (regardless of whether the listener consciously realized that their partner was among the talkers presented), and the familiarity advantage was greatest in the heaviest levels of noise.

2.3.3 Summary of talker familiarity research

Despite the above-mentioned progress in our understanding of talker familiarity, it is still not well understood what underlies the advantage gained from familiarity with a talker's voice. This question is the motivation for the familiarity experiment described in this thesis. The hypothesis — that the familiar talker advantage lies at least partially in prosodic patterns — is based on the results showing that voice familiarity is more readily acquired with sentential stimuli than with isolated words, and that advantages are greatest in the most challenging conditions.

2.4 Summary

This chapter reviewed numerous studies relating to auditory masking, intelligibility, and talker familiarity in speech perception. From these studies, a complex picture emerges regarding how speech perception occurs under adverse conditions. Given the redundancy of information present in the speech signal, listeners can rely on robust segmental cues when more precise cues are missing, distorted, or masked. When acoustic cues are insufficient to resolve the speech signal, listeners can supplement impoverished information from the speech signal by drawing on a vast store of past experience with language, including transitional probabilities between speech sounds, knowledge of word frequencies and the contextual probabilities of words given surrounding context, past experience of a particular talker's speech idiosyncracies,

etc. Because listeners make use of information at several levels, misinformation at any of those levels can contribute to a perceptual error.

Turning to studies of intelligibility, numerous acoustic dimensions of speech have been identified that correlate with talker intelligibility, given a particular sample of talkers and a particular type of listening task. Unfortunately most of those dimensions cannot be independently controlled by talkers, nor can they be easily manipulated with current signal processing techniques. Thus our understanding of intelligibility is still in its infancy, and new methods are needed to advance our understanding beyond studies of specific talkers and tasks, to a model of intelligibility that encompasses variation across languages and usage domains.

Chapter 3. Research questions & experimental designs

In Section 2.2.8, a broad distinction was drawn between experiments that *model* intelligibility variation between groups based on measurements of the speech signal, and experiments that *manipulate* some dimension of the stimulus and observe how intelligibility changes as a result. An ideal compromise would be a series of perception experiments using fully synthetic speech, with complete experimental control over all aspects of the signal. At present such an approach is prohibitively labor-intensive: parametric synthesis of a single sentence can take hours to days, and naturalistic results are difficult to achieve. Moreover, even if naturalistic synthesis were easy to achieve, listener experience can in no case be controlled so completely, and thus a fairly large corpus of synthetic stimuli would be needed to allow subtle differences in stimulus intelligibility to emerge above the noise inherent in a between-subjects experimental design.

The experiments described in this thesis take a hybrid approach between modeling, manipulation, and full-blown synthesis. Stimuli are recorded speech that have been processed to neutralize variation along three dimensions (intensity, f_0, and duration) while preserving variation in others. In this way, the experiments are most similar to the modeling approach, but with a "top-down" approach to predictor selection. In other words, rather than starting with low-level measurables (*e.g.*, mean f_0, vowel space size, *etc*), these experiments begin with a broad distinction between segmental and suprasegmental aspects of speech, and attempt to determine the contribution of each to speech intelligibility.

This approach aligns well with the distinction between "envelope" and "fine structure" cues (Rosen, 1992), although the emphasis here is less on temporal scale and more on the linguistic concept of prosody (patterns of duration, intensity, and f_0) *vs.* segmental content.[1] The approach also trades on particular facts about English compared to other languages: namely, that the dimensions of duration, loudness, and pitch are not primary cues to lexical contrasts in English (at least not in English as commonly spoken in the Pacific Northwest region of the United States). In other words, English is not a lexical tone language, has no phonemic length contrasts, and there are few lexemes that are distinguished on the basis of syllabic stress location (for which pitch and duration *are* important cues, along with vowel quality). To the extent that these experiments concern the relative importance to intelligibility of prosodic *vs.* segmental aspects of the signal, the experiments described here also bear some similarity to cue weighting studies, as they constitute an implicit comparison of prosodic cues (taken as a group) *vs.* segmental cues (taken as a group).

3.1 Research questions

The first research question addressed in this thesis is: *how does prosody relate to intelligibility?* In other words, are differences in intelligibility between talkers attributable (at least in part) to their prosody? A related question is whether an unintelligible talker can be made more intelligible through prosodic changes alone (without changes to segmental phonetic features like formant transitions, lenited consonants, missing release bursts, *etc*).

The second research question concerns the intersection of prosody, intelligibility, and talker

[1] Note that segmental content includes both "fine structure" and "periodicity" cues, collapsing the three-way distinction in (Rosen, 1992).

familiarity: *to what extent is the familiarity advantage relying on prosodic aspects of the talker's speech?* In other words, when a listener is sufficiently attuned to a talker's voice such that they receive a perceptual advantage from that familiarity, what is it about the talker's voice that they are "tuning in" to? More concretely, is a novel talker more intelligible if his prosody mimics the prosody of a familiar talker? Or alternately, will the familiarity advantage persist even when the familiar talker's prosody changes to mimic the prosody of a novel talker?

3.2 Experimental designs

Both of the research questions above are investigated through experiments using sentential stimuli resynthesized via PSOLA™ (Moulines & Charpentier, 1990), which allows manipulation of the fundamental frequency and duration of speech (intensity is trivially manipulated by scaling sample magnitudes). Recordings of a low-intelligibility talker can be resynthesized with the prosodic characteristics of a high-intelligibility talker (and *vice versa*), allowing comparisons between stimuli that differ in segmental content only or in prosodic content only. Experiment 1 tests the effect of prosodic replacement on intelligibility by comparing unmodified recordings to stimuli created with prosodic replacement, using three talkers known to vary in their intelligibility in noise. The processing artifacts that arise during resynthesis (mentioned in Section 2.2.5) are minimized by several methodological means (described in Chapter 4) and modeled statistically.

Experiment 2 investigates the relationship between talker familiarity and prosody, by comparing groups of listeners whose "training talker" did or did not match one of the three talkers used in the test sentences. Experiment 2 includes test sentences resynthesized via the

Table 3.1: Schematic table of stimulus types and comparisons for the three experiments described in this thesis. Unmodified recordings of the three original talkers are represented by letters A, B, and C (Talker A being the most intelligible, and Talker C being the least). Resynthesized "talkers" are represented by combinations of the letters A, B, and C, with the first letter indicating the "segmental donor" and the second letter indicating the "prosodic donor" of the resynthesized stimuli. For experiments involving familiarity, the talker used for training is indicated in (parentheses) preceding the test talker.

	Question	**Hypothesis**	**Intelligibility prediction**
1	Can we shift the intelligibility of a talker by replacing his prosody?	yes	if A > B > C, then AB > AC; BA > BC; CA > CB
2	Can listeners gain a familiarity advantage based only on prosody?	yes	if A = B = C, then (A)BA > (A)BC
3	Does the talker familiarity advantage persist after prosodic replacement?	yes	if A = B = C, then (A)AC > (A)BC

same methods used in Experiment 1. Listeners in Experiment 2 were grouped based on whether the training talker they heard was or was not among the three talkers in the set of test sentences.

Table 3.1 schematizes the comparisons of interest in Experiments 1 and 2. In describing these experiments, the convention adopted is to represent resynthesized "talkers" as two-letter combinations of the talkers used to achieve the resynthesis. For example, Talker AB indicates a stimulus made from a recording of Talker A, resynthesized to have the prosody from Talker B's recording of the same sentence. Question 1 in Table 3.1 is addressed by Experiment 1; Questions 2 and 3 are addressed by Experiment 2.

One difficulty with this experimental design is that the predictions, as formulated in Table 3.1, are conditional statements, whose antecedents cannot all be true given a fixed choice

of three talkers. In other words, the contribution of prosody to intelligibility is most easily seen when Talkers A, B, and C vary in their base intelligibility, whereas the effect of familiarity is most easily measured when the base intelligibility of the three talkers is known to be equal. A further complexity is that even the most careful resynthesis introduces some distortion as an artifact of the signal processing, such that listener performance on resynthesized stimuli is expected to be somewhat less than performance on unresynthesized speech. In analyzing the data reported here, both of these problems are addressed by statistical means using mixed-effects regression, such that the effects of prosody, familiarity, and resynthesis are estimated simultaneously, along with estimates of group variance for listener and sentence (see Section 4.6 for details).

Chapter 4. Methods

This chapter describes the methods used for stimulus creation, data collection, and data analysis. Scripts used for stimulus creation are referenced in the text, and are collected in Appendix B.

4.1 The PN/NC corpus

Stimuli for the experiments described here were recordings of the IEEE "Harvard" sentences (Rothauser *et al.*, 1969) drawn from the PN/NC corpus (McCloy *et al.*, 2013). The 180 sentences in the PN/NC corpus were selected from among the 720 IEEE sentences based on absence of alliteration or rhyming, avoidance of focus/contrast readings, and lack of marked locutions. The full list of sentences used is given in Appendix A.

Based on within-dialect intelligibility scores from McCloy *et al.* (submitted), three talkers were chosen for the present experiments, herein referred to as Talkers A, B, and C. These talkers were selected from among the Pacific Northwest male talkers because, as a group, the Pacific Northwest male talkers exhibited the largest spread in intelligibility in the corpus, and those three talkers formed the endpoints and midpoint of that group, with Talker A being the most intelligible and Talker C the least intelligible (cf. Figure 4.1).

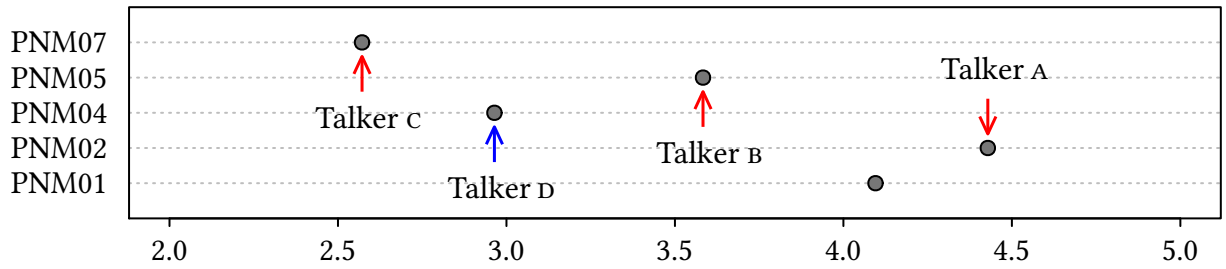

Figure 4.1: By-talker mean keywords correct (across 15 dialect-matched listeners) in speech-shaped noise at +2 dB SNR (data from McCloy *et al.* submitted). The three talkers selected for resynthesis are indicated by red arrows; the talker used as control (Talker D) in the familiarity experiment is indicated by a blue arrow.

4.2 Stimulus creation

Resynthesized stimuli underwent prosodic replacement, using the PSOLA™ algorithm as implemented in Praat (Boersma & Weenink, 2013). In all cases, the sentential content of the target file and prosodic donor file were identical (*i.e.*, the contributing recordings were of two different talkers reading the same sentence). The general approach builds on the methodology in Yoon (2007), with several modifications to improve the accuracy of the measurements on which the algorithm depends, and thereby improve the quality of the resynthesized waveforms. Required measurements include intensity and duration for both the target file and the prosodic donor file, as well as information about glottal pulse timing of the target file and pitch contour information from the prosodic donor file. Methods used to obtain each of these measures are described below.

4.2.1 Duration

Duration measurements for target and donor files were made at the syllable level. Syllable boundaries were based primarily on local minima in the intensity contour of the sentence. In

cases where intensity contour minima disagreed with phonological syllable affinity, the phonological considerations were given priority.[1] Where intensity contour minima were absent from the syllable transition region, boundaries were marked based on inflection points in the intensity contour or waveform envelope, or in the absence of inflection points, on aspects of waveform or spectrogram morphology. The segmentation process was aided by a custom Praat script (see Script B.1).

In order to maximize inter-speaker agreement in number of durational units for a given sentence, the intensity-based syllable-wise method of segmentation described above was chosen over the more common spectrogram-based phonemic/allophonic segmentation method (*e.g.* Peterson & Lehiste, 1960; Turk *et al.*, 2006). For the same reason, the method of marking only periodic/aperiodic transitions was also avoided, though otherwise it is particularly well-suited to PSOLA™ resynthesis. Whereas there were often inter-speaker disagreements on the number of segments (*e.g.*, eight segments in [pʰɑɹkˀtʰɹʌk] *vs.* nine in [pʰɑɹkt.tʰɹʌk] for "parked truck") or the number of periodic/aperiodic transitions (*e.g.*, five in [əpʰɑɾətʰi] *vs.* seven in [əpʰɑtəvtʰi] for "a pot of tea"), virtually all sentences in the corpus showed inter-speaker agreement on number of syllables. The rare disagreements were cases of extreme reduction, *e.g.*, sentence 22–07 "It is hard to erase blue or red ink", in which one talker contracted the initial syllables to "It's". In such extreme cases the sentence was excluded from resynthesis, but in most cases it was still possible to separate the reduced form into two syllable-like units using the same acoustic landmarks mentioned above. Such sentences were therefore retained to avoid introducing a

[1] This arose primarily in cases of voiceless fricative-stop onset clusters, where the intensity minimum occurred during the stop closure, effectively grouping the fricative into the coda of the preceding syllable. In such cases, the lack of aspiration on the stop is taken as evidence that it is not in syllable-initial position, and that the syllable boundary should therefore fall before the fricative.

systematic bias by eliminating the most heavily reduced sentences.

4.2.2 Fundamental frequency

Fundamental frequency (f_0) information was measured with Praat using a semi-automated process (see Script B.2). For each recording, f_0 tracks generated by Praat's pitch tracking algorithm were displayed over a narrowband spectrogram, and algorithm parameters were adjusted as necessary to eliminate spurious pulse detections, pitch halving errors, and other irregularities. The corrected parameters were then used to create Praat manipulation objects, which allow manual editing of the detected pulses. After hand-correction of the pulses, sparse f_0 tracks were re-generated with a single value at the midpoint between each pair of pulses in contiguous voicing regions.

Hand-correction of the pulse points was often necessary in recordings involving creaky voicing, either because the frequency dropped below the algorithm's pitch floor, or because the amount of jitter/shimmer present in the creaky voicing exceeded the level at which the algorithm was willing to mark periodicity (see Figure 4.2a). In a few such cases, better resynthesis results (as judged auditorily) were achieved by marking each cycle — despite cycle-to-cycle irregularities — rather than omitting alternating pulses in pursuit of a smoother, lower-frequency pitch track (see Figure 4.2 for illustration). The hand-correction process was aided by Script B.2.

Prior to being mapped onto the target signal, the locations of donor f_0 points were shifted via dynamic time warping, with the temporal ratios determined by the relative durations of corresponding syllables in the target and donor files (see Script B.3). Pitch points were also

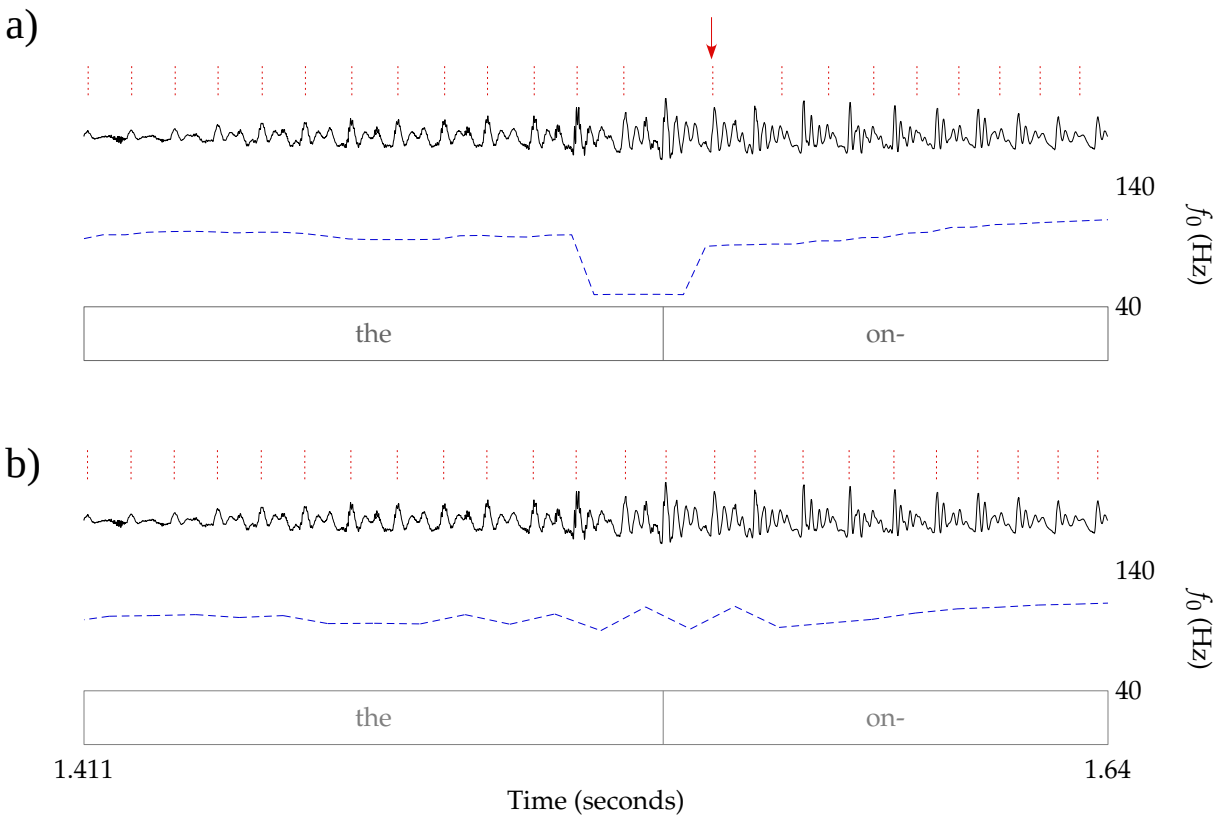

Figure 4.2: Illustration of method for handling creaky-voiced portions of speech. (a) Excerpt of Talker B's recording of sentence 33–05, in which vowel hiatus is resolved with light creaky voicing. The waveform (black) is overlaid with Praat's auto-detected pulse marks (dotted red lines) and pitch track (dashed blue line). Note the missed cycles on either side of the red arrow, and the phase-shift of all pulses right of the arrow. (b) The same span of speech after manual correction of pulses, and a (jittery, but continuous) pitch track generated from the corrected pulses.

Table 4.1: Mean and standard deviation of root mean squared error (RMSE) across sentences for each pair of talkers.

	RMSE (Hz)	
Talkers	**Mean**	**St. dev.**
Talkers A & B	12.4	3.9
Talkers A & C	15.4	6.1
Talkers B & C	16.6	5.9

shifted in frequency so as to equate the mean pitch of the stimulus before and after resynthesis. This was done to minimize absolute magnitudes of pitch shifts (in hopes of minimizing distortion due to pitch manipulations), on the assumption that mean f_0 is irrelevant to the intelligibility of speech (on this point see Section 2.2.4).

To compare the magnitude of f_0 contour differences among the talkers, the root mean squared error (RMSE) was calculated between the original f_0 values and the time-warped f_0 values of the prosodic donor (after mean-shifting) for each stimulus. The mean and standard deviation of these RMSE values across sentences is shown in Table 4.1. As expected, the RMSE values were symmetrical (*i.e.*, the RMSE values comparing Talker A unmodified with Talker B time-warped were nearly identical to the RMSE values comparing Talker B unmodified with Talker A time-warped). Consequently, only one value is reported for each pair of talkers.

4.2.3 Intensity

Intensity was altered by first multiplying the signal by the difference of the maximum intensity and the inverted intensity contour (see Figure 4.3), then multiplying the resulting signal by the intensity contour of the replacement prosody and scaling as needed to achieve the desired RMS amplitude. To do this, the donor intensity contour first underwent dynamic time warping to

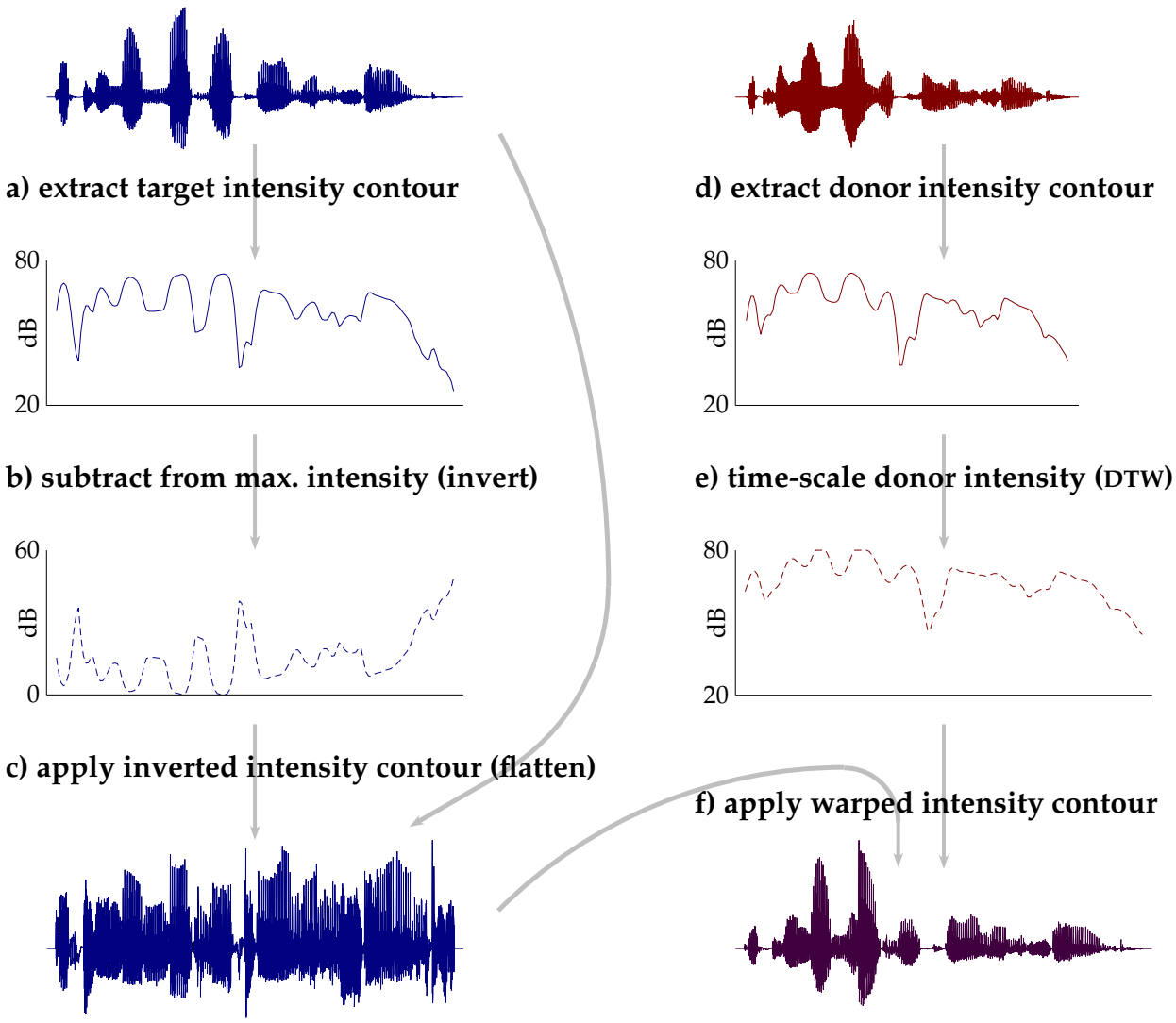

Figure 4.3: Illustration of the method used to scale intensity. (a) Intensity contour is extracted from target waveform (blue). (b) The intensity contour of the target signal is inverted, by subtracting each intensity point from the maximum intensity. (c) Target intensity is "neutralized" or "flattened" by multiplying the target signal by the inverted intensity contour. (d) Intensity contour is extracted from the prosodic donor signal (red). (e) The prosodic donor intensity contour is time-scaled syllable-by-syllable via dynamic time warping (DTW) to match the temporal pattern of the target signal. (f) The intensity-neutralized target signal is multiplied by the time-warped donor intensity contour (purple). The signal is now ready for PSOLA™ resynthesis of duration and pitch.

match the durational patterns of the target signal (see Script B.3).

After resynthesis, all stimuli were assessed auditorily for excessive distortion. In some cases, problems could be remedied by readjusting the pulses and pitch tracks and re-running the resynthesis script; in other cases the problems were irremediable, and the unacceptable donor/target/sentence combination was excluded from the experimental stimuli. Most cases of irremediable stimuli arose from one of two sources: intra-syllable segment duration and intensity mismatches (see Figure 4.4), or complete devoicing of a syllable by the target talker (see Figure 4.5).

4.3 Experiment sessions

Except where otherwise noted, stimuli were presented with a stationary Gaussian masker noise, frequency shaped to match the long term spectral average of the corpus of stimuli, at 0 dB SNR. This SNR was chosen to avoid ceiling and floor effects, based on a pilot study testing five SNRs ranging from −1 to +3 dB. To ensure target audibility, the level of the speech was held constant at 67 dB SPL (dB RMS in a 6 cc coupler) and the masker noise was digitally added to the speech to achieve the desired SNR, yielding a final presentation level of approximately 70 dB SPL. The noise extended past the beginning and end of the speech by 50 ms in each direction, and linear onset and offset ramps were applied to this excess noise to prevent clicks during stimulus playback.

The combined speech-and-noise signal was presented in a sound-insulated booth over closed-back supra-aural headphones (Sennheiser HD 25–1 II). Listeners were instructed to repeat each sentence they heard, to give partial answers when they only heard some words, and

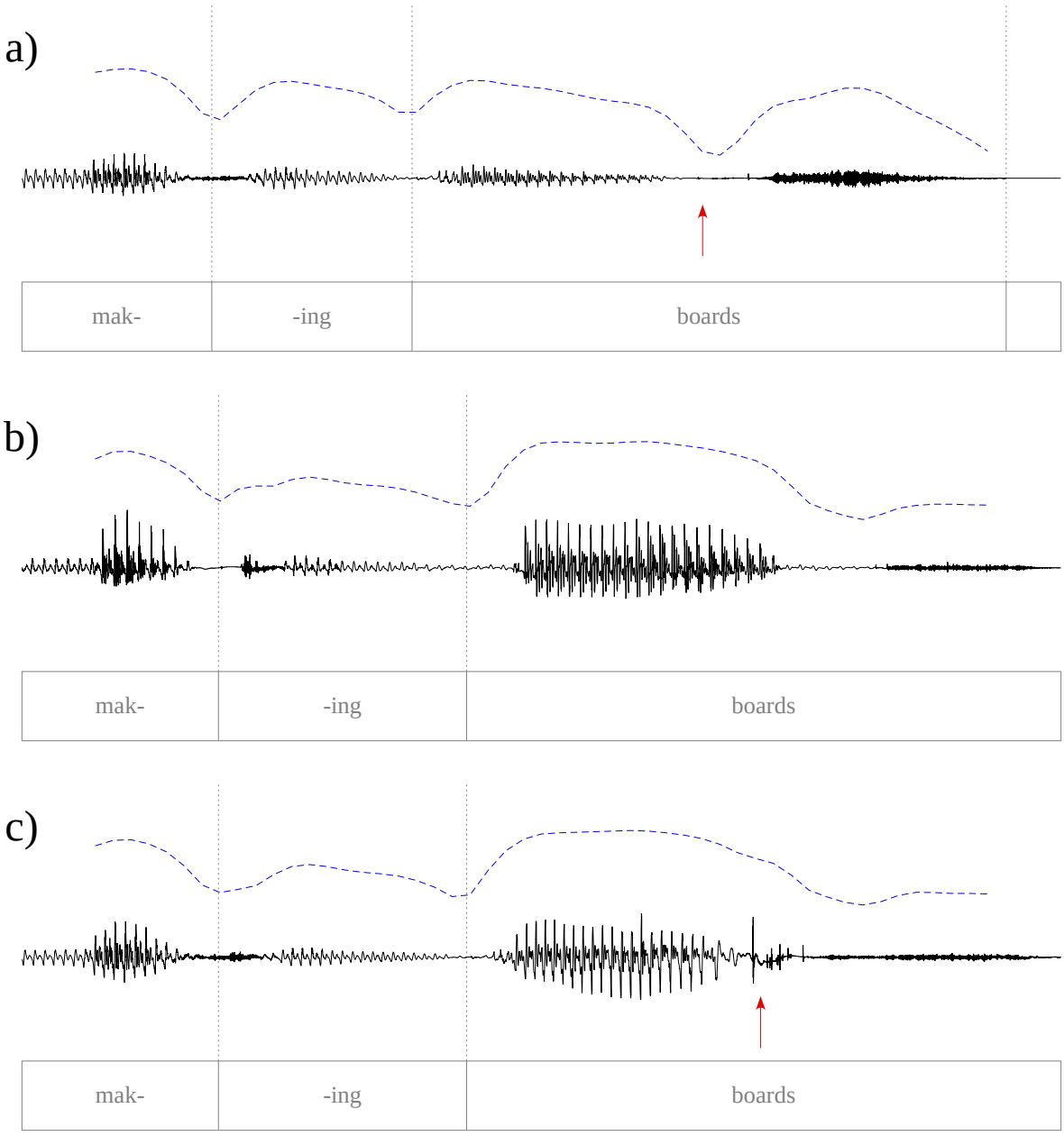

Figure 4.4: Illustration of intra-syllable segment duration and intensity mismatch, which led to unnatural patterns of intensity after resynthesis. (a) Waveform (black line) of Talker C's pronunciation of "boards" in sentence 06–05, overlaid with intensity contour (blue dashed line). Note that the syllable duration is split roughly 50/50 between the periodic nucleus and the aperiodic coda. (b) Talker A's recording of the same sentence. Note the relatively longer vocalic nucleus and relatively shorter coda. (c) Talker C's waveform after resynthesis to match Talker A's intensity, pitch, and duration. The red arrow marks a nearly silent portion of Talker C's speech that was amplified to vowel-like intensity levels.

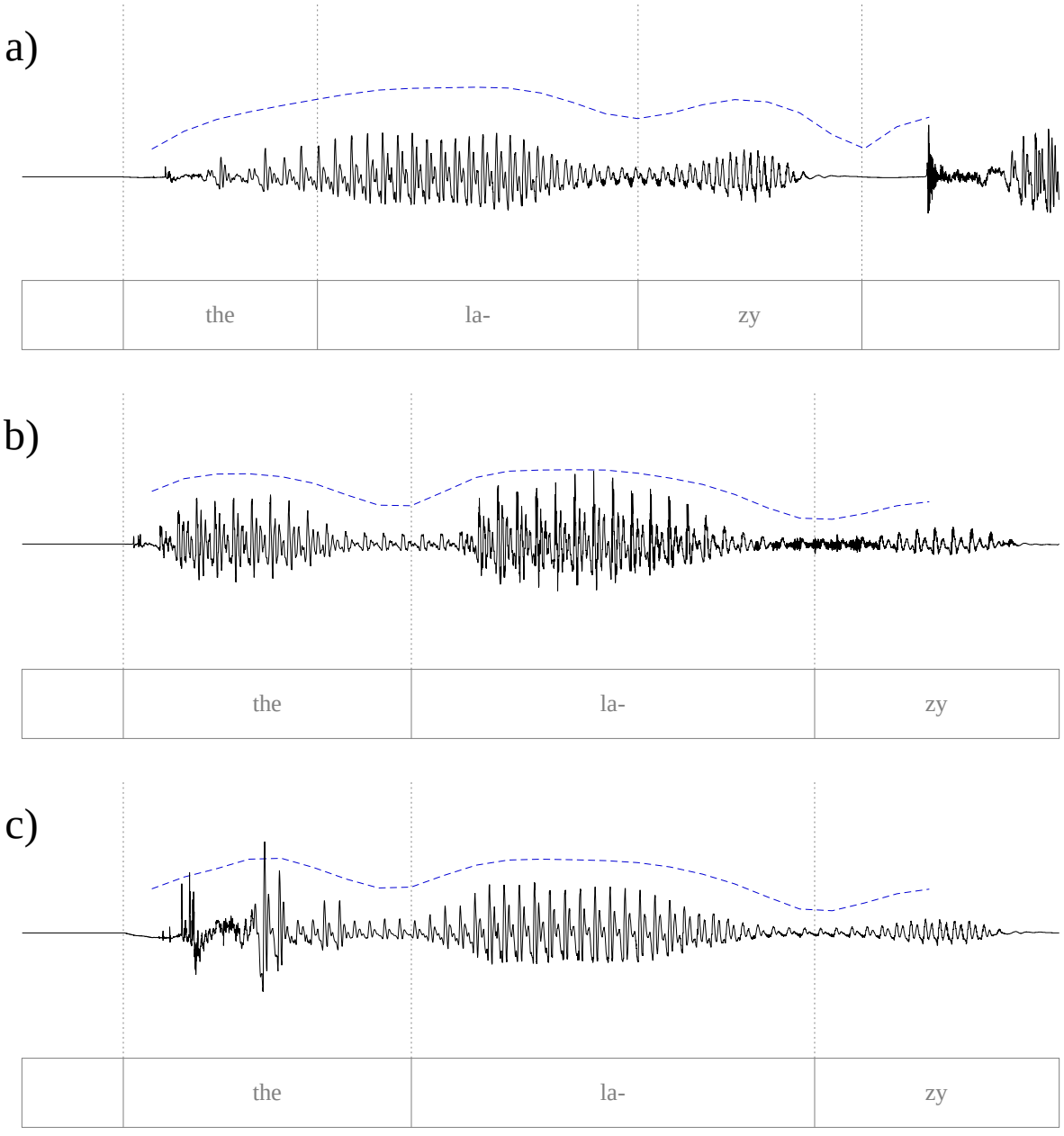

Figure 4.5: Illustration of how devoicing can lead to unacceptable levels of distortion during resynthesis. (a) Devoicing of the word "the" in Talker B's recording of sentence 05–05 (first ≈500 ms shown). The intensity contour (blue dashed line) is overlaid on the waveform (black). (b) Talker A's recording of the same sentence. (c) Talker B's waveform after resynthesis to match Talker A's intensity, pitch, and duration. Note the excessively high amplitude in the first syllable.

to guess when they were unsure. Trials were scored 0–5 on keywords correct during the task. An audio recording was made of listener responses, and scoring uncertainties were resolved offline.

Of the 180 sentences in the corpus, half were set aside for use as training/exposure sentences in Experiment 2; the remaining 90 sentences were designated as test sentences and resynthesized versions of those sentences were created. Experiment 1 presented the 90 test sentences to each listener, with equal numbers of stimuli from each "talker" (*i.e.*, ten from each of the three unmodified talkers A, B, and C, plus ten from each of the six resynthesized "talkers" AB, AC, BA, BC, CA, and CB). Talker-sentence combinations were random and unique for each listener, subject to the above-mentioned constraints (*i.e.*, each listener heard each talker an equal number of times, and each listener heard each sentence only once). Stimuli were advanced by the researcher only after listeners had finished responding, to eliminate any effects of time pressure that might vary across listeners.

In Experiment 2, the 90 training sentences were presented in speech-shaped noise at 0 dB SNR. After each listener had finished responding, the sentence was played again without background noise, and the text of the sentence was simultaneously presented on a computer screen for the listener to read. For a given listener, all 90 training sentences were recordings of the same talker, but the test sentences were again drawn in equal numbers from among the three unmodified talkers (A, B, and C) and the six resynthesized "talkers" (AB, AC, BA, BC, CA, and CB).

This design of the training phase — to include stimuli both with and without noise maskers — was chosen for several reasons. First, stimuli with noise were included to maximize similarity

between the training and testing phases, and also to first expose listeners to only the speech cues that are sufficiently robust to escape from the masker. At the same time, auditory feedback presented without a noise masker was included so that listeners had access to all possible speech cues during training, to provide both reinforcement of the robust speech cues and additional complementary information that would (hopefully) facilitate the process of talker familiarization.

4.4 Scoring

Each trial was scored 0–5 based on number of keywords correct. Keywords were all content words, though a few sentences included pronouns among the keywords. This likely added some variability in the difficulty of sentences, since pronominal forms are more likely to undergo reduction in speech, and are often highly confusable (*e.g.*, "him" *vs.* "them"). However, differences in difficulty from sentence to sentence were unavoidable in any case, since across sentences the keywords varied in their lexical frequency and their predictability from sentential context. Both of these shortcomings of the stimuli are overcome by explicitly including variability in sentence difficulty as part of the statistical model (see Section 4.6).

Another potential problem with scoring based on keywords is that discrete scores from 0 to 5 are not necessarily well-modeled as a continuous outcome in statistical models. One solution to this problem is to model the outcome as an ordered multinomial variable, though such models are complicated and difficult to implement computationally. An alternative solution is to score sentences in an "all-or-nothing" fashion, and model the outcome as a binomial distribution (for which well-tested computational methods exist). The drawback of binary scoring is that

information about partially correct responses is lost, and models are less intuitive to interpret (model coefficients are "logits" instead of keywords correct).

In analyzing these experiments, a hybrid approach is taken: both continuous and binomial models are constructed, and as long as the continuous model agrees with the corresponding binomial model in the direction and significance of the predictors, the results of the continuous model are reported to facilitate the interpretation of model results. This approach comes with the caveat that the absolute magnitudes of the coefficients should be interpreted somewhat cautiously.

4.5 Participants

Listeners were all native English speakers who lived in the Pacific Northwest (Washington, Oregon, or Idaho) throughout their period of primary and secondary education (*i.e.*, ages 6–18). All reported English as the primary home language, and all had learned or studied at least one other language as an adolescent or adult (though this was not a criterion for inclusion). All listeners had bilaterally normal hearing, defined as pure-tone thresholds of 20 dB HL or better at octave intervals from 250 Hz to 8 kHz (re: ANSI, 2004). Listeners were recruited from the UW campus community and were compensated for their participation; there were 17 listeners in Experiment 1, and 20 listeners in Experiment 2. One participant was excluded from Experiment 1 due to a mild monaural threshold elevation at 8 kHz. Listener demographics are summarized in Table 4.2.

Table 4.2: Listener demographics for Experiments 1 and 2. Experiment 2a represents the experimental group trained on Talker C; Experiment 2b represents the control group trained on Talker D (who was not among the test talkers).

		Experiment		
		1	**2a**	**2b**
Gender	Female	8	9	8
	Male	8	1	2
Age	mean	22.3	20.2	24.7
	st. dev.	6.9	1.4	4.7
	min.	18	19	18
	max.	44	22	32

4.6 Data analysis

As mentioned in Section 3.2, the contribution of prosody to intelligibility is most easily seen when talkers vary in their base intelligibility, whereas the effect of familiarity is most easily seen when the base intelligibility of the talkers is known to be equal. Analysis using mixed-effects regression addresses this problem by simultaneously estimating the influence of multiple predictors or "fixed effects", as well as estimating residual within-group variability due to uncontrolled factors or "random effects" (*e.g.*, variability in item difficulty or listener ability). In other words, mixed-effects models can estimate the effect of familiarity *a*s if the base intelligibility of the talkers were equal.

In these experiments, the fixed effects are all binary variables or multi-level factors dummy-coded as binary variables, so the estimated coefficients of the linear model represent differences between groups of stimuli. Data were analyzed using R (R Development Core Team, 2013), with the packages `lme4` for mixed-effects regression (Bates *et al.*, 2012) and `languageR` for Markov-

chain Monte Carlo simulation of model parameters (Baayen, 2011).

For Experiment 1 (testing the effect of prosodic replacement on intelligibility), the fixed-effect predictors used in the model are `resynth`, `segDonor` and `proDonor`. `Resynth` is a Boolean variable that is true for Talkers AB, AC, BA, BC, CA and CB, and is included as a control for the distortion introduced by the resynthesis process. `SegDonor` represents the talker used as the base waveform for the resynthesis, and `proDonor` represents the talker whose prosody was used in the resynthesis. Both `segDonor` and `proDonor` are three-level factors, each recoded into a pair of dummy-coded binary variables for modeling purposes. For stimuli that were not resynthesized, the value of `segDonor` and `proDonor` are the same. Random effects are included for both sentence and listener, to account for the fact that not all sentences are equally difficult (cf. Section 4.4), and the fact that listeners are not necessarily equally skilled at speech-in-noise tasks. A manual "drop-1" procedure using likelihood-ratio tests was used to validate the inclusion of each of the fixed-effects predictors; in all cases, a significantly better model fit was realized with each predictor than without it.

For Experiment 2 (testing the role of prosody in the familiar talker advantage), the fixed-effect predictors include all those used for Experiment 1, plus two additional predictors: `segTrain` is a Boolean variable indicating match between the training talker and the segmental donor of the test stimulus, and `proTrain` likewise indicates a match between the training talker and the prosodic donor of the test stimulus. As in Experiment 1, random effects are included for both sentence and listener, and a manual "drop-1" procedure confirmed the validity of including each fixed-effect predictor.

4.7 *Post hoc* acoustic analyses

To better understand the experimental results, a variety of acoustic measurements were performed on the stimuli, in hopes of identifying the acoustic dimensions that underlie the effects seen in the statistical models. The measures are broadly divided into segmental measures (presence of stop release bursts, properties of the vowel space) and prosodic measures (mean pitch range, pitch velocity, intensity velocity, and speech rate).

4.7.1 Segmental measures

The number of reduced stop consonants and affricates was measured in the unmodified recordings of Talkers A, B and C, as one estimate of the degree of segmental reduction (cf. Picheny *et al.*, 1986; Li & Loizou, 2008). The presence of an unreduced stop consonant was quantified as either (a) greater than 20 ms silence in the waveform, or (b) the presence of a release transient in the waveform (for utterance-final stops and affricates, only the second criterion was used). Separate counts were made for stop consonants in onset and coda position of syllables; intervocalic stops were counted as onsets. A random sample of half of the test sentences was selected for this analysis (the same 45 sentences from each talker were analyzed). The total number of unreduced stops counted across the 45 sentences for each talker was then divided by the total number of phonemic stops, affricates and flaps present in phonemic transcriptions of the sentence transcripts, assuming standard English citation form for all words. Flaps were considered to be underlyingly stops and were therefore included in the total counts; this was done not for any theoretical reason, but because in some sentences there was variation among the talkers regarding whether flaps or short-duration stops were produced in

intervocalic position.

In order to remain agnostic about which measurements best characterize overall vowel space size, several different metrics were calculated, based on hand-measurements of 5 tokens per vowel per talker for the ten vowels /i ɪ e ɛ æ ɑ o ʊ u ʌ/. Measured vowels were drawn from keywords in positions throughout the sentence, with a preference for vowels with obstruent flanking consonants to avoid coloring by adjacent nasals, rhotics, or laterals. These hand-measured formant values were converted using the bark transform (Traunmüller, 1990) prior to further analysis.

Five measures of vowel space size were calculated from the formant data: F1 and F2 ranges, mean Euclidean distance from the center of the vowel space (cf. Bradlow *et al.*, 1996), area of the vowel polygon based on phonemic means (cf. Bradlow *et al.*, 1996; Neel, 2008), and area of the convex hull encompassing all vowel tokens. In addition, two measures related to the composition of the vowel space were also calculated: total repulsive force of the vowel system (cf. Liljencrants & Lindblom, 1972; Wright, 2004a) and mean vowel cluster size. Repulsive force (sometimes called "total energy") was calculated as the sum of inverse squared distances between all pairs of vowel tokens not belonging to the same phoneme, as in Equation 4.1 (where /i/ and /j/ represent the phonemic categories of the vowel tokens being compared, and r is the Euclidean distance formula).

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{1}{r_{ij}^2}, \text{/i/} \neq \text{/j/} \tag{4.1}$$

This measures the degree to which neighboring vowel phonemes in a system encroach on one another, with higher values of repulsive force corresponding to greater degrees of phoneme overlap or encroachment. The calculation seen here differs from both Liljencrants & Lindblom

1972 and Wright 2004a in calculating force based on individual vowel tokens rather than mean values for each vowel. Vowel cluster size was calculated as the area of the ellipse encompassing 68.27% of the data points along a bivariate normal density contour (equivalent to ±1 standard deviation from the bivariate mean for each vowel), and the mean of all vowel cluster sizes was calculated for each talker.

4.7.2 Prosodic measures

To quantify variation in talkers' use of duration, mean and variance of syllable duration was calculated. Because stimuli were RMS normalized, mean intensity across stimuli is identical, but for each stimulus the mean rate of change of intensity ("mean intensity velocity") was calculated as in Equation 4.2.

$$\sum_{k=1}^{n-1} \frac{I_{k+1} - I_k}{(t_{k+1} - t_k) \times (n-1)} \tag{4.2}$$

Additionally, the mean *absolute value* of the rate of change in intensity was calculated ("mean intensity dynamicity"), as in Equation 4.3. In both Equations 4.2 and 4.3, n indicates the number of intensity samples, I_k is the intensity value at sample k, and t_k is the time (in seconds) at sample k.

$$\sum_{k=1}^{n-1} \frac{|I_{k+1} - I_k|}{(t_{k+1} - t_k) \times (n-1)} \tag{4.3}$$

Stimulus-level means for intensity velocity were then averaged within talkers, in hopes of capturing a talker's tendency to "trail off" at the ends of utterances, or conversely to maintain a more consistent level across all the keywords in the sentence. Likewise, stimulus-level means for intensity dynamicity were averaged within talkers in hopes of capturing differences in the magnitude of envelope modulations between talkers.

To quantify pitch, the hand-corrected pitch tracks for the 90 test sentences were used to calculate average f_0 range magnitude, as well as the mean rate of change of f_0 ("pitch velocity") and the mean absolute value of rate of change in f_0 ("pitch dynamicity"). Calculations for these values were identical to those used for intensity velocity and intensity dynamicity, respectively. As with the intensity values, mean velocity is meant to capture overall trend in f_0, while mean dynamicity is meant to capture differences in the magnitude of f_0 modulations.

Chapter 5. Results

This chapter describes the results of the two experiments conducted for this thesis. Statistical models and post-hoc acoustic analyses that help to clarify and explain the results are also presented.

5.1 Experiment 1

Experiment 1 tests the role of prosody in intelligibility, by comparing mean sentence scores for unmodified talkers against resynthesized stimuli with prosodic replacement. A quartile analysis indicated a significant improvement in performance between the first and last quartiles (t=−3.016 on 709.3 degrees of freedom, p<0.01; see Figure 5.1). The magnitude of the difference between the first and fourth quartiles is approximately 0.4 keywords.

The mean sentence score for each talker across listeners is shown in Figure 5.2. Darker bars indicate unmodified stimuli, while light bars represent resynthesized stimuli. It is noteworthy that not all resynthesized stimuli have lower mean scores compared to their unmodified counterparts, suggesting that distortion due to the resynthesis process was relatively minimal. This is likely due to several factors; one is probably the painstaking hand-correction of the pulse marks during stimulus preparation, which ensured consistent phase of the f_0 epochs throughout voiced spans of speech. Another possible explanation for the low levels of distortion is the choice to maintain each talker's natural mean pitch on each sentence, and map only the *s*hape of the pitch contour of the prosodic donor during resynthesis.

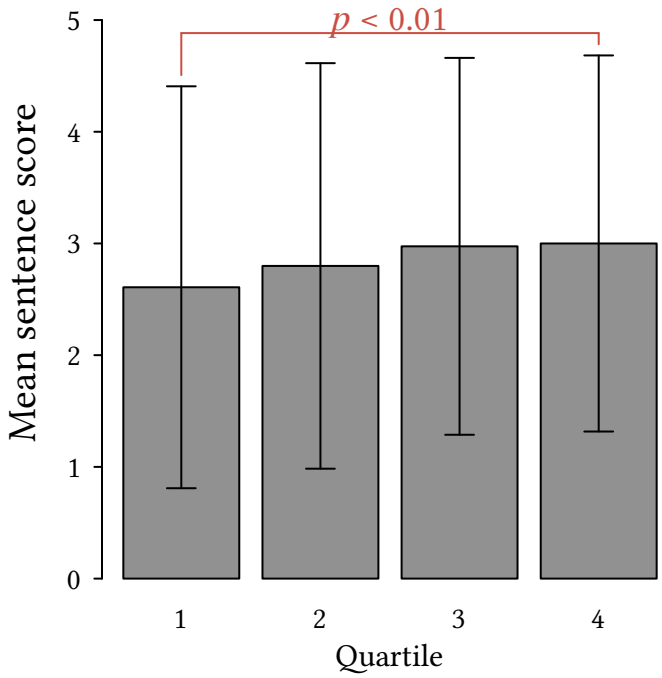

Figure 5.1: Mean sentence scores by quartile for Experiment 1. Error bars are ±1 standard error.

With regard to Research Question 1 — *how does prosody relate to intelligibility?* — a complex picture emerges. Talker B suffers dramatically when his prosody is replaced, whereas Talker C is unchanged or slightly improved, and Talker A is unchanged or slightly worsened. One possible explanation for these results is to postulate that Talker A, despite being highly intelligible, does not have especially good prosody, evidenced in particular by the fact that scores for Talker CA are lower than the scores for Talker CB. On the same grounds, and on the additional observation that scores for Talker AB are greater than for Talker AC, we might conclude that Talker B has more intelligible prosody than the other two talkers, and his middling base intelligibility scores are due to segmental factors.

With regard to the question of whether low-intelligibility talkers can be made more intelligible through prosody alone, it would appear that the answer is "yes": Talker CB appears

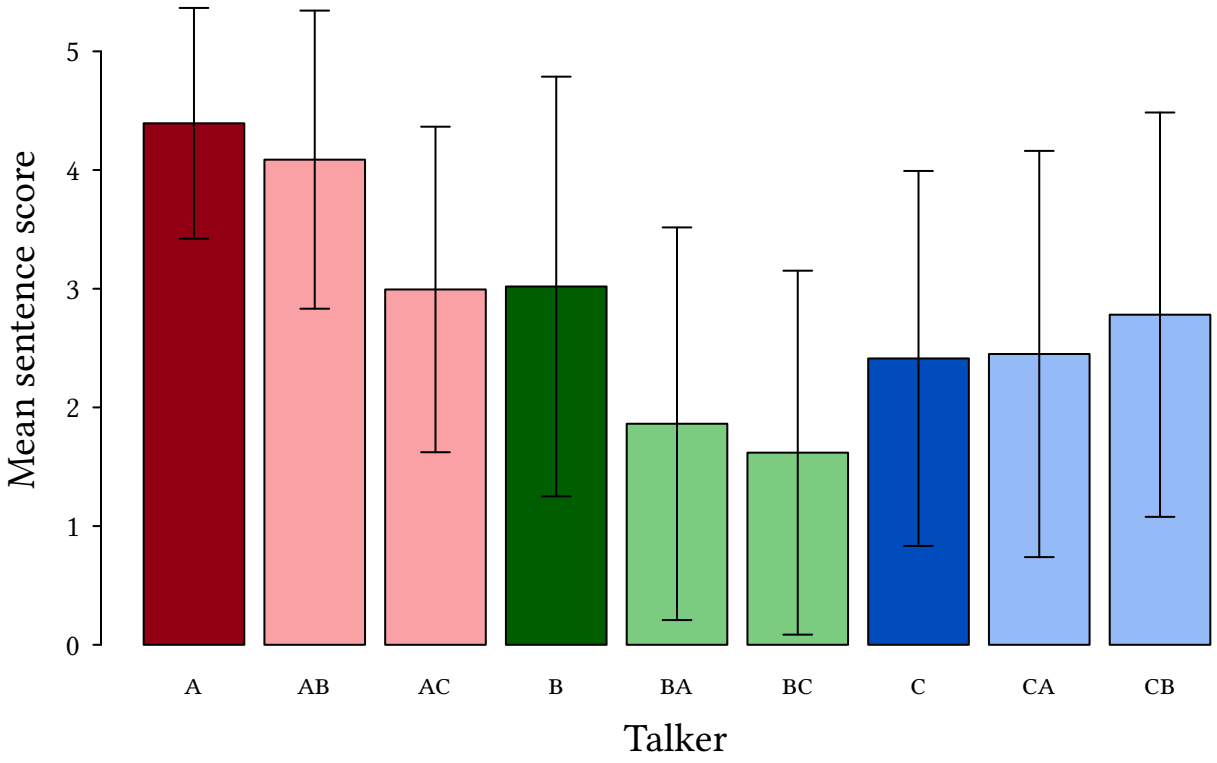

Figure 5.2: Barplot of mean sentence scores for Experiment 1. Error bars are ±1 standard error; lighter colors indicate resynthesized talkers, with the second letter indicating the prosodic donor (see Section 3.2 for full explanation of talker codes). Similar hues indicate shared segmental donors.

to have higher scores than Talker C, despite the likelihood that Talker CB's recordings suffer some amount of distortion from the resynthesis process, however small. In other words, any degradation due to resynthesis seems to have been more than overcome by the benefit of having Talker B's prosody mapped onto Talker C's signal. However, it is noteworthy that the greatest improvement to Talker C's intelligibility does not come from Talker A's prosody, even though Talker A has the highest overall intelligibility. This suggests that high-intelligibility talkers do not necessarily use prosody that enhances their intelligibility.

5.1.1 Experiment 1 statistical model

To further probe these results, the scores were submitted to a mixed-effects linear regression model, shown here:

```
lmer(sentScore~resynth+segDonor+proDonor+trial+(1|listener)+(1|sentence))
```

In this model, the dependent variable `sentScore` is the number of keywords correct (0–5), and is being predicted by several independent variables: `resynth` is a Boolean variable that is true for Talkers AB, AC, BA, BC, CA and CB, and `segDonor` and `proDonor` are three-level factors indicating the talker in the target signal and the talker from whom the prosodic information was drawn, respectively. For the unmodified original recordings, `segDonor` and `proDonor` are defined as the talker himself, even though those recordings were not resynthesized. The effect of task familiarization seen in Figure 5.1 is accounted for by the fixed-effect predictor `trial` (a numeric value ranging from 1–90). Two random-effects predictors (`1|listener` and `1|sentence`) are also included to model variability in listener performance and sentence difficulty, respectively.

A summary of fixed-effects predictors for this model is given in Table 5.1. All fixed-effects predictors were significantly different from zero, and there was no evidence of correlation of fixed effects (correlation coefficients all less than 0.1; not shown). The baseline condition is Talker A, with a value at the intercept of about 4.1 words correct. The coefficients reveal similar patterns to those seen in Figure 5.2: firstly, the model supports the interpretation that Talker B has the most intelligible prosody. Having the prosody of Talker B represents a net gain of 0.3 words correct over the prosody of Talker A, whereas having the prosody of Talker C represents a net loss of more than 0.6 words correct compared to Talker A. The model also supports the idea

Table 5.1: Summary of fixed effect predictors in the statistical model of Experiment 1. *s*: standard error of the coefficient estimate; *t*: *t*-value of coefficient estimate; *p*: *p*-value of coefficient estimate (calculated via MCMC).

Summary of fixed effects (N=1440; log-likelihood=−2551)				
Predictor	**Coefficient**	***s***	***t***	***p***
Intercept	4.132	(0.145)	28.44	$<10^{-16}$
resynth = TRUE	−0.662	(0.077)	−8.63	$<10^{-16}$
segDonor = B	−1.673	(0.089)	−18.86	$<10^{-16}$
segDonor = C	−1.278	(0.089)	−14.31	$<10^{-16}$
proDonor = B	0.307	(0.088)	3.48	$<10^{-3}$
proDonor = C	−0.646	(0.088)	−7.36	$<10^{-12}$
trial	0.006	(0.001)	4.02	$<10^{-4}$

that Talker A's intelligibility stems in large part from non-prosodic factors, given that the other levels of `segDonor` both have strongly negative coefficients. The estimated degradation due to resynthesis is about −0.7 keywords correct. Because trial was a continuous variable ranging from 1–90, the coefficient for the effect of trial is misleadingly small; the predicted difference between the first and the last trial due to listener adaptation to the task is actually 90×0.005549, or 0.5 keywords.

A summary of the random effects in the model for Experiment 1 are shown in Table 5.2. The results here are not unexpected: the variance in intercepts due to particular listeners doing systematically better or worse on the task is only about 2% of the total residual variance. This suggests that, by and large, all listeners were performing equally well on the task. The variance in intercepts due to varying difficulty of particular sentences is somewhat larger (about 22% of the total residual variance, or a standard deviation of 0.7 keywords correct),[1] suggesting that there was indeed some value in modeling the sentences as varying in their difficulty (cf. the

Table 5.2: Summary of random effects in the statistical model of Experiment 1. s^2: estimated variance; s: standard error; HPD: highest posterior density interval.

Summary of random effects			MCMC (nsim=10 000)	
Group	s^2	s	**mean**	**95% HPD**
Sentence (intercept)	0.506	0.711	0.600	(0.498 0.700)
Listener (intercept)	0.043	0.207	0.220	(0.100 0.346)
Residual	1.767	1.329	1.344	(1.294 1.396)

discussion of scoring in Section 4.4).

5.2 Experiment 2

Experiment 2 tests the role of prosody in the familiar talker advantage, by comparing mean sentence scores for various talkers across two groups of listeners: those trained on one of the test talkers (Talker C), and those trained on a control talker (Talker D). The first question to be addressed is whether the training phase was in fact effective for both groups of listeners.

Across all listeners, performance shows an upward trend in mean sentence score across training quartiles (t=−4.00 on 878.3 degrees of freedom, p<0.0001), but no significant changes across quartiles of the testing phase (see Figure 5.3). This suggests that training was successful in general, and that any familiarization effects were complete before the start of the testing phase.

Considering the control and experimental listener groups separately, both show an upward trend in mean sentence score across training quartiles, and t-tests performed on the first and

[1]The MCMC estimate for variability due to listener is in close agreement with the fitted model. The MCMC estimate for variability due to sentence is slightly smaller than the value in the fitted model, at 16% of the total variance (*vs.* 22%), with a standard deviation for sentence of about 0.6 keywords (*vs.* 0.7).

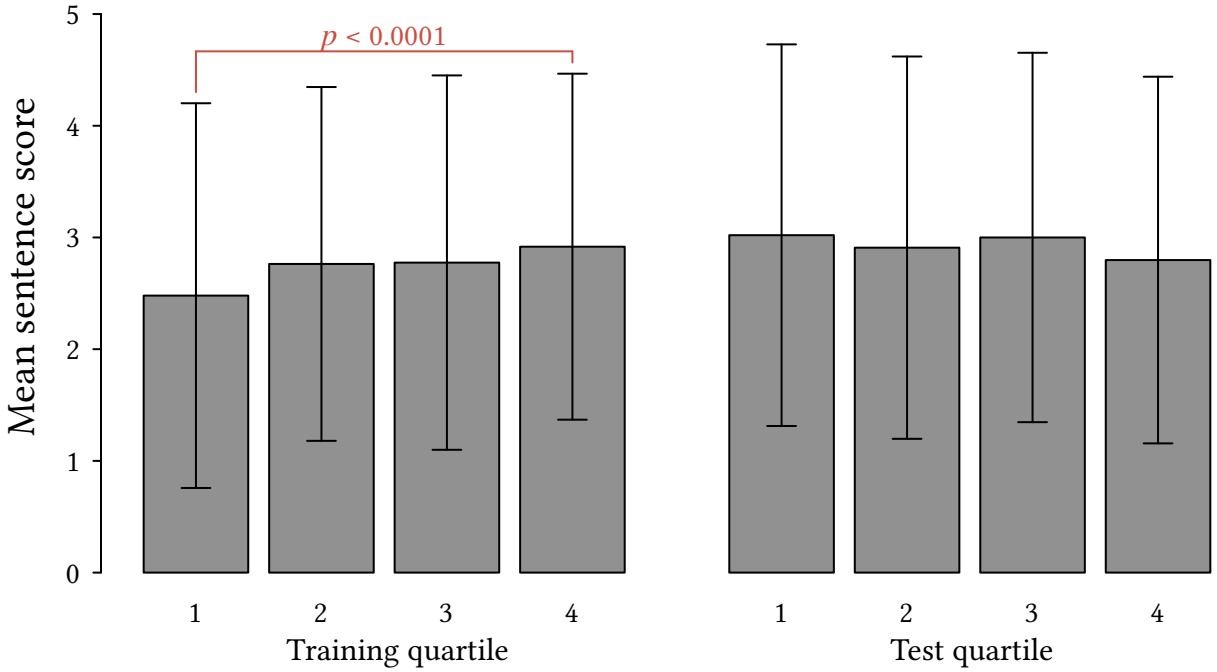

Figure 5.3: Quartile analysis of training and testing phases in Experiment 2 (all listeners combined). Significant improvement is seen during the training phase, but not during testing.

fourth quartile of each group show a statistically significant improvement in both groups of listeners (control group: t=−2.898 on 437.2 degrees of freedom, p<0.01; experimental group: t=−2.816 on 438.1 degrees of freedom, p<0.01; see Figure 5.4). The magnitude of the improvement differs slightly between the control group (0.50 keywords) and the experimental group (0.38 keywords).

For the experimental group, it appears that familiarization with Talker C during training did not confer an advantage on Talker C during testing (compare the last quartile of the training phase to the score on Talker C during the testing phase in Figure 5.4). In light of this, it is perhaps unsurprising that training did not confer a reliable perceptual advantage on test stimuli resynthesized to have the training talker's prosody; this is seen in the barplot of mean sentence

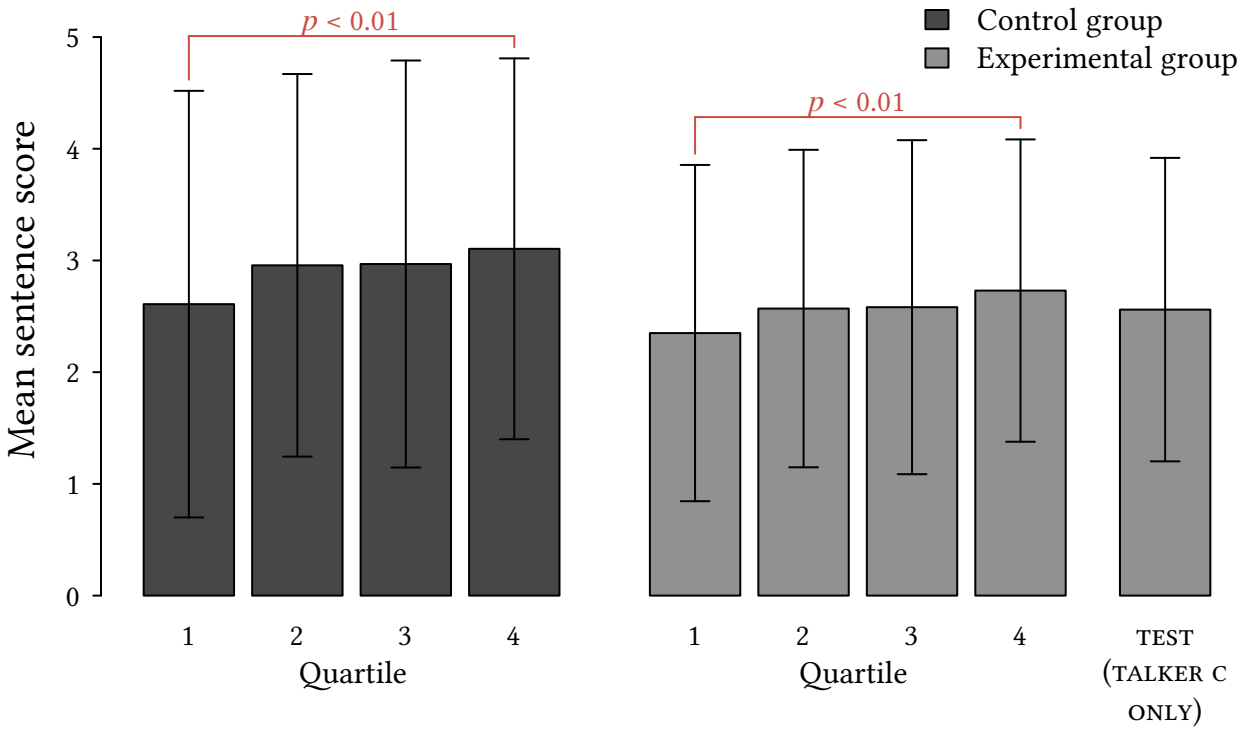

Figure 5.4: Quartile analysis of training phase in Experiment 2. Improvement is seen during training for both the control group (trained on Talker D) and the experimental group (trained on Talker C). The adaptation in the experimental group appears not to have persisted through the testing phase.

scores for the testing phase of Experiment 2 (Figure 5.5). Even without controlling for multiple comparisons, none of the *t*-tests comparing the experimental and control groups within talker are significant, suggesting that there was no familiar talker advantage enjoyed by the experimental group. If there had been a familiar talker advantage, we would have expected the colored bars to be higher than their corresponding gray bars for Talker C, and perhaps for Talkers CA, CB, AC and BC (depending on whether and how the advantage extended to resynthesized talkers).

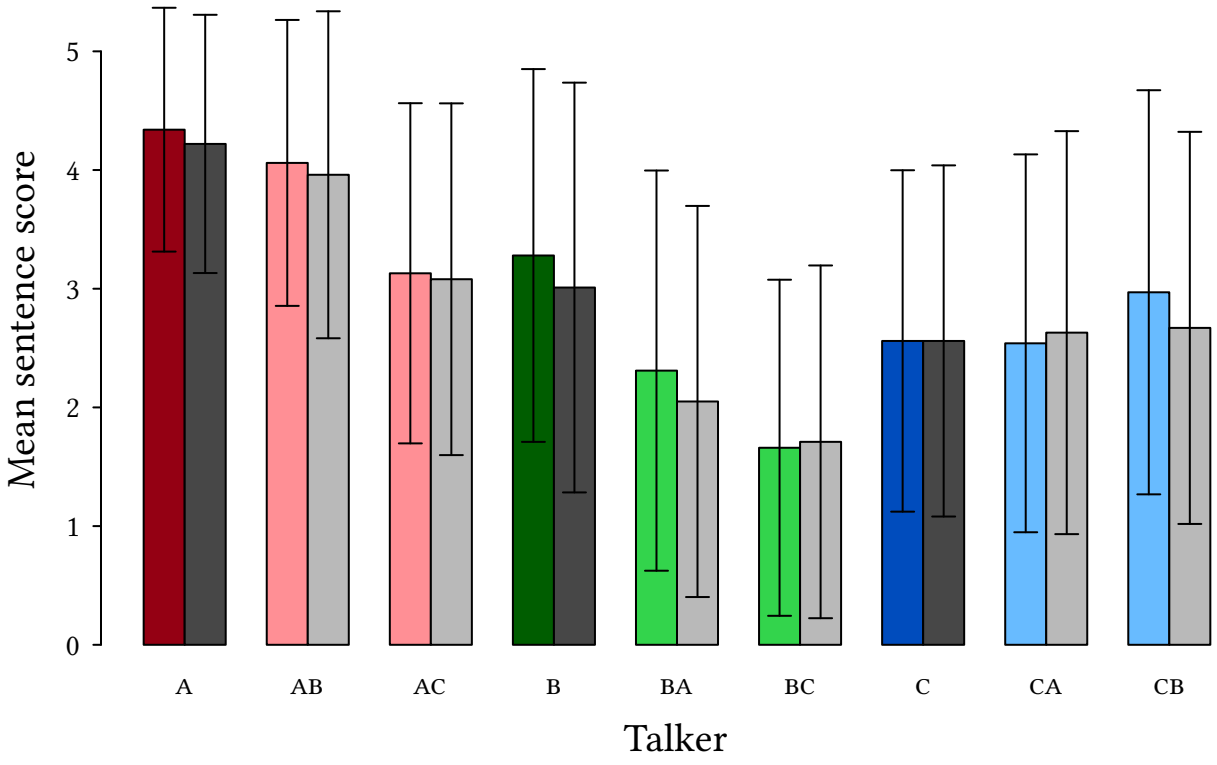

Figure 5.5: Barplot of mean sentence scores for Experiment 2. Error bars are ±1 standard error. Colored bars indicate the experimental group; grayscale bars the control group. Lighter colors indicate resynthesized talkers, and similar hues indicate shared segmental donors.

5.2.1 Experiment 2 statistical model

Full results of the statistical model for Experiment 2 are shown in Table 5.3. Predictor codes are the same as in the model for Experiment 1, with the addition of Boolean variables `segTrain` (indicating match between the training talker and the segmental donor of the test stimulus) and `proTrain` (indicating match between the training talker and the prosodic donor of the test stimulus).

Unlike the model for Experiment 1, there is no significant effect for `trial` (unsurprising given Figure 5.3), most likely because listeners had already undergone a training phase and were already familiarized to the task. Neither of the two new predictors `segTrain` and `proTrain`

Table 5.3: Summary of fixed effect predictors in the statistical model of Experiment 2. *s*: standard error of the coefficient estimate; *t*: *t*-value of coefficient estimate; *p*: *p*-value of coefficient estimate (calculated via MCMC).

Summary of fixed effects (N=1800; log-likelihood=−3103)				
Predictor	**Coefficient**	***s***	***t***	***p***
Intercept	4.366	(0.139)	31.50	$<10^{-16}$
resynth = TRUE	−0.603	(0.066)	−9.15	$<10^{-16}$
segDonor = B	−1.466	(0.075)	−19.42	$<10^{-16}$
segDonor = C	−1.159	(0.098)	−11.80	$<10^{-16}$
proDonor = B	0.270	(0.075)	3.60	$<10^{-3}$
proDonor = C	−0.584	(0.098)	−5.95	$<10^{-8}$
segTrain = TRUE	0.008	(0.125)	0.06	0.95
proTrain = TRUE	−0.117	(0.125)	−0.94	0.35
trial	−0.001	(0.001)	−0.64	0.53

were significant, suggesting that listeners were not realizing a familiarity advantage due to training. Aside from the lack of effect for `trial` and the two additional non-significant predictors `segTrain` and `proTrain`, the model is nearly identical to the model for Experiment 1; the magnitude, direction, and significance of the other fixed-effect predictors are all unchanged from Experiment 1.

A summary of the random effects in the model of Experiment 2 are shown in Table 5.4. The results are also very similar to Experiment 1, with estimates of listener accounting for about 4% of the total unexplained variability (with a standard deviation of 0.3 keywords) and sentence accounting for about 24% of total unexplained variability (with a standard deviation of about 0.7 keywords).[2]

[2]Again, MCMC estimates for the effect of sentence were slightly smaller than the fitted model (18% *vs.* 24%), and there was close agreement between the two for the effect of listener.

Table 5.4: Summary of random effects in the statistical model of Experiment 2. s^2: estimated variance; s: standard error; HPD: highest posterior density interval.

Summary of random effects			MCMC (nsim=10 000)	
Group	s^2	s	**mean**	**95% HPD**
Sentence (intercept)	0.537	0.733	0.617	(0.526 0.712)
Listener (intercept)	0.083	0.288	0.297	(0.192 0.425)
Residual	1.617	1.271	1.285	(1.242 1.329)

5.3 *Post hoc* acoustic analyses

Overall, the statistical models for Experiments 1 and 2 support the view that both prosodic and non-prosodic factors contribute to differences in the intelligibility of talkers. To better understand these results, a variety of acoustic measurements were performed on the stimuli, in hopes of identifying the acoustic dimensions that underlie the effects seen in the statistical models. The measures are broadly divided into segmental measures (presence of stop release bursts, properties of the vowel space) and prosodic measures (mean pitch range, pitch velocity, pitch dynamicity, intensity velocity, intensity dynamicity, and syllable duration).

5.3.1 Segmental measures

The results of the stop consonant reduction analysis are shown in Figure 5.6. The percentages given are calculated as the total number of stops and affricates counted (across the 45 sentences measured) divided by the total number of expected stop consonants based on citation-form phonemic transcriptions; there was little difference between this method and the method of calculating the mean of unreduced stop percentage calculated on a per-stimulus basis. The relative ordering of the talkers accords with their relative intelligibility (*i.e.*, Talker A >

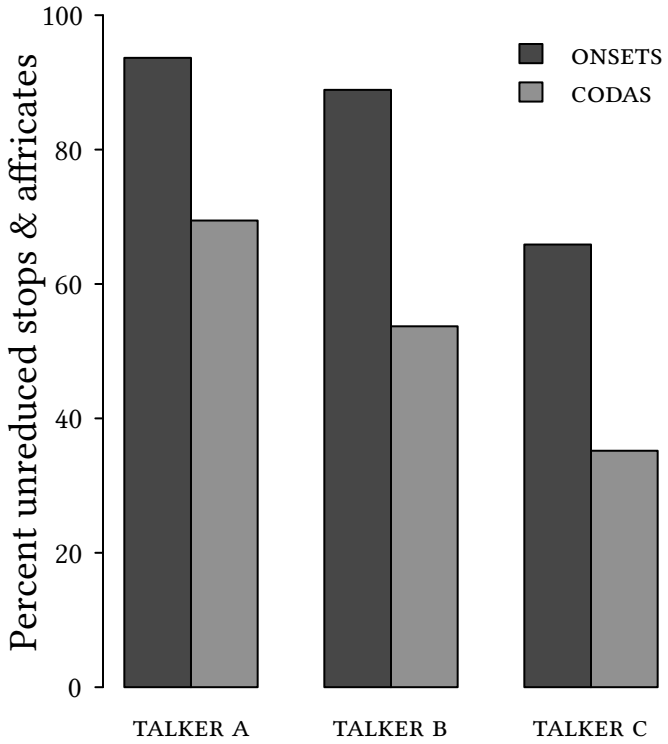

Figure 5.6: Barplot of mean proportion of unreduced stop consonants present, calculated across half of the test stimuli (45 sentences per talker).

Talker B > Talker C), but does not correspond to the expected ordering based on coefficients for segmental donor (A > C > B).

A possible explanation for the ordering of talkers is that stop reduction may index both segmental and prosodic information, for two reasons: first, stop consonants often occur at word edges, which are loci for word-level prosodic marking, and second, stop consonants are often strengthened in prosodically prominent syllables at higher levels of phrasal structure (cf., *e.g.*, de Jong 1995; Fougeron & Keating 1997; Cho *et al.* 2007; Cole *et al.* 2007; see Keating 2006 for review). In other words, because stop consonant reduction was measured across all words in the sentence, it may reflect a combination of each talker's prosodic habits (*i.e.*, tendency toward more *vs.* fewer intonational phrases) as well as purely segmental pronunciation habits (which

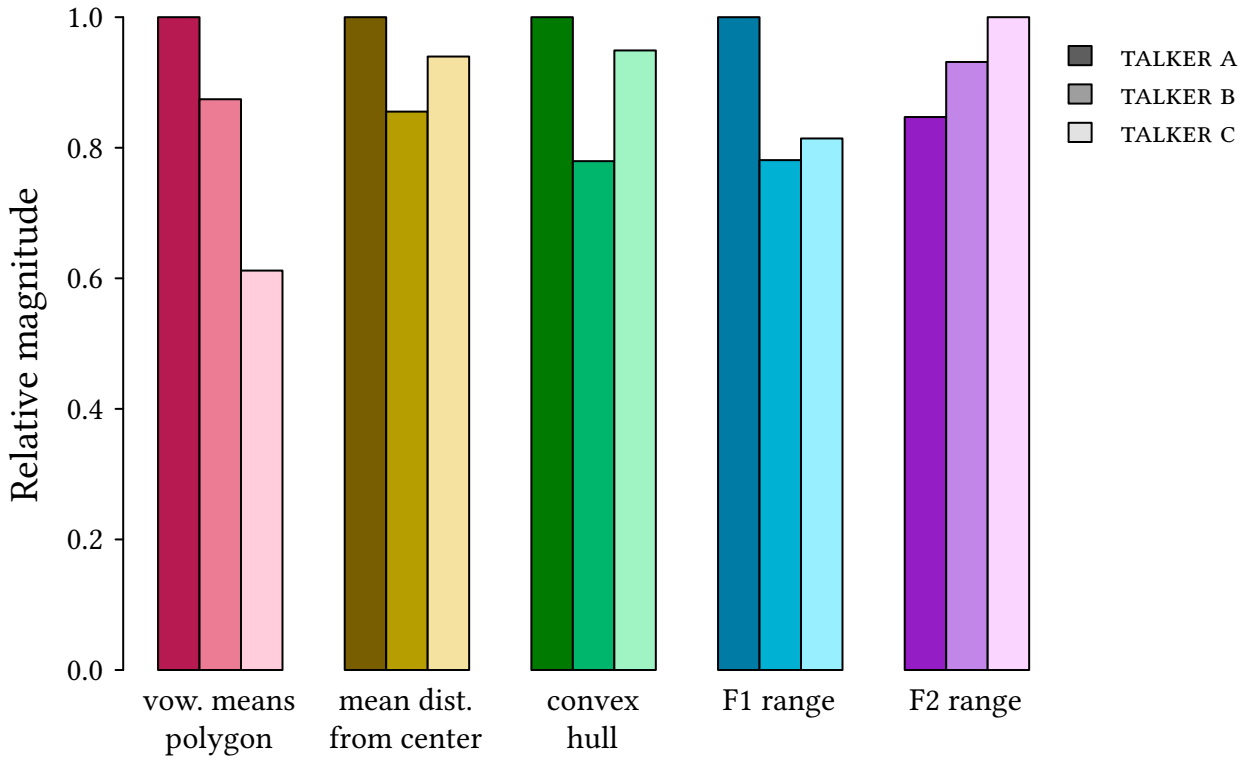

Figure 5.7: Barplot of the relative magnitudes of several vowel space size metrics. Each metric has been scaled by dividing by the maximum value among the three talkers.

might have been more easily observed from words lists). Unfortunately, the PN/NC corpus does not include recordings of word lists, so this explanation must remain speculative.

Results of the *post hoc* analysis of vowel space size is shown in Figure 5.7. Here the expected pattern of A > C > B (based on coefficients for segmental donor from Experiments 1 and 2) is seen in three of the five measures: mean distance from center, area of the convex hull, and F1 range. F2 range does not correlate with any expected intelligibility-related pattern, while area of the polygon based on vowel means seems to correlate with overall intelligibility scores for each talker (*i.e.*, A > B > C).

The difference between area of the convex hull and area of the vowel means polygon is especially interesting. Figure 5.8 plots both the convex hull and vowel means polygons on data

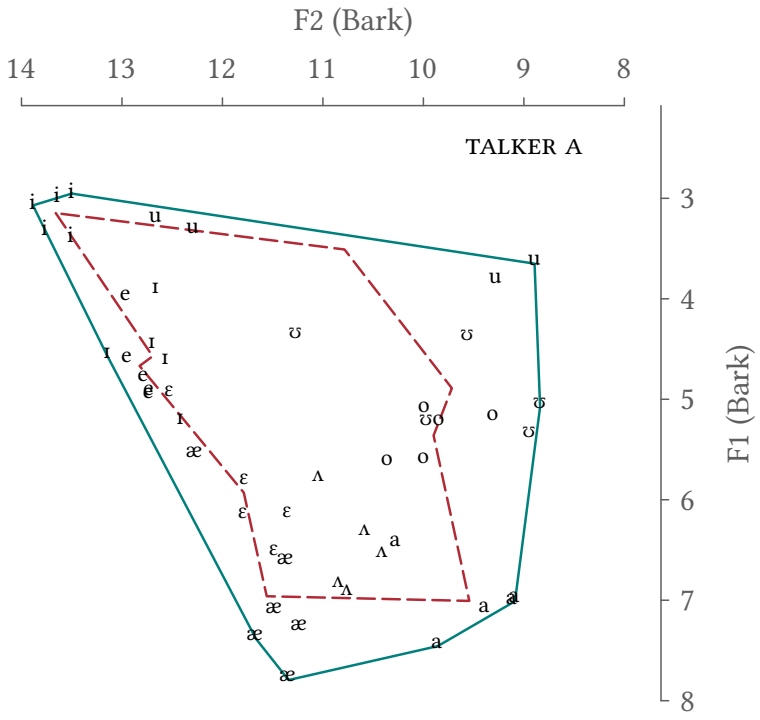

Figure 5.8: Illustration of two vowel space area metrics (data from Talker A). The area of the polygon described by vowel means (dashed red line) is contrasted with the convex polygonal hull (solid blue line). Note the difference between the polygons in the high-back and low-front regions due to /u/-fronting and /æ/-raising (respectively), as well as differences due to reduction (particularly in the /o/, /ʊ/ and /ɛ/ regions).

from Talker A. This figure illustrates one possible explanation for the difference in patterning between these two measures of vowel space size: the convex hull effectively ignores vowel tokens that are reduced (for any reason) — and thus indexes a talker's most extreme vowel productions — whereas the area of the vowel means polygon includes information from multiple vowel tokens, which may show reduction due to prosodic differences between talkers (because the tokens came from various positions within the sentence). Considering that all the vowel tokens measured came from lexically stressed syllables in content words, any vowel quality reduction present in these tokens is *not* reduction due to lack of *lexical* stress; thus it is quite likely that the reduction seen is prosodic in origin.

The distribution of vowel tokens in Figure 5.8 also shows some within-category variation that is probably not due to prosodically-driven reduction. For example, the large variation in F2 of the high back vowel /u/ is likely an example of the widespread change-in-progress known as /u/-fronting (Labov *et al.*, 2006, chap. 12). Similar variation is seen in F1 values for /æ/ (cf. reports of /æ/-raising before velars in the Pacific Northwest: Reed 1952; Beckford-Wassink *et al.* 2009). Such cases also affect measures of vowel means polygonal area, but are less likely to impact measures of convex hull area (as seen in the high-back and low-front regions of Figure 5.8). In this way, area of the vowel means polygon could be said to contain information about both segmental and prosodic dimensions of speech (much like stop consonant reduction discussed above), which may explain why it corresponds to the ordering of talkers based on general intelligibility scores, rather than the ordering of talkers based on segmental donor coefficients in the statistical models. In contrast, area of the convex hull may be a better index of purely segmental properties of speech, since it ignores within-category variation of vowels (at least some of which is prosodically-driven); this would explain why area of the convex hull correlates with the ordering of talkers based on signal donor coefficients.[3]

The preceding discussion suggests that a thorough analysis of the vowel space would segregate measures of overall vowel space expansion from measures of individual vowel category size, and would include some measure of category overlap or encroachment. To explore this need, measures of mean cluster size and repulsive force of each talker's vowel system are shown in Figure 5.9 (see Section 4.7.1 for definition of these measures).

[3]Note that although area of the convex hull appears to be less sensitive than the area of the vowel means polygon to phonological variation such as /u/-fronting and /æ/-raising, it is not completely immune to such influence. As such, care should be used when comparing across dialects or sociolects where the presence or degree of such features is not consistent across groups.

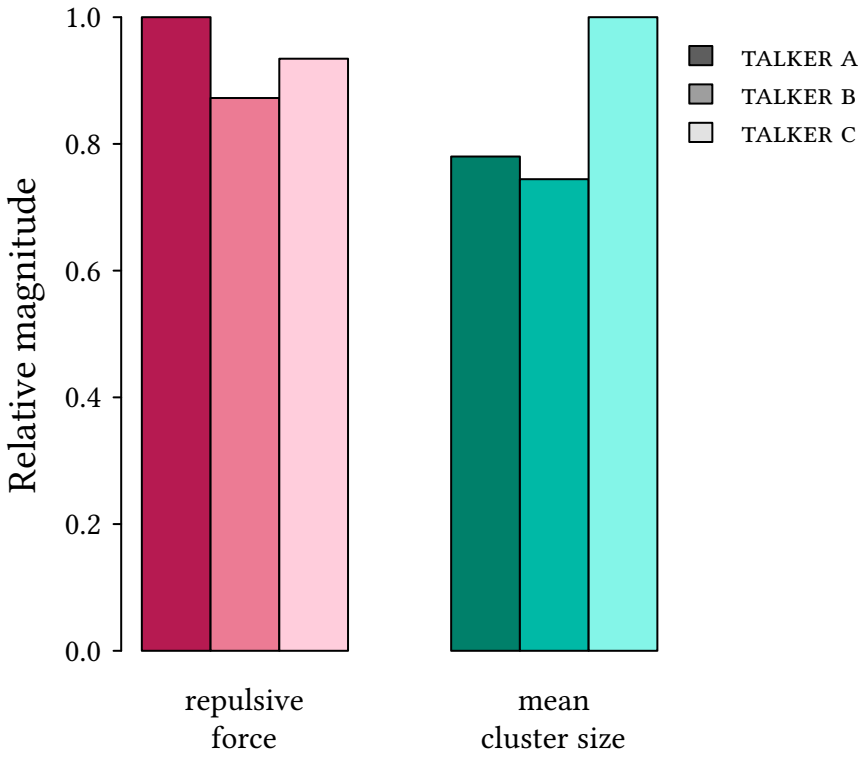

Figure 5.9: Barplot of the relative magnitudes of mean vowel cluster size and repulsive force. Each metric has been scaled by dividing by the maximum value among the three talkers.

A high degree of repulsive force indexes a high degree of phonemic overlap, which ought to correspond to *lower* intelligibility (on the assumption that words would be more confusable in talkers whose phonemic categories are not well segregated). Thus the expected pattern based on the statistical model coefficients for segmental donor ought to be B > C > A, which is precisely the opposite of the pattern seen in Figure 5.9. It is possible that fifty vowel tokens (five per vowel) was simply not a large enough number to give an accurate picture of the internal structure of the vowel space. In any case, the likelihood of lexical confusion due to vowel phoneme overlap is contingent on the existence and lexical probability of the competing form. For example, no matter how much /u/-fronting a talker exhibits, the /u/ vowel in "poodles" (sentence 15–06) is unlikely to be heard as /i/. Duration is also relevant: although "kits" and "Kate's" are potentially confusable based on formant values (see Talker A's vowel plot in

Figure 5.8), there are likely to be duration differences that listeners can rely on to disambiguate the two forms. Thus it is perhaps unsurprising that repulsive force is a poor predictor in a sentence perception task such as this one (as opposed to an isolated word perception task in which stimuli are chosen to ensure the viability of competing forms, and plausibility is not constrained by semantic context).

Mean cluster size should index within-category variability of vowel formants, and is predicted (like repulsive force) to be inversely correlated with the non-prosodic component of intelligibility. However, like repulsive force, the magnitudes of mean cluster size also fail to track the expected pattern of B > C > A. Again, a possible explanation is that an insufficient number of vowel tokens were available to give accurate estimates of the spectral extent of within-category variation. Another explanation is that because mean cluster size collapses information across vowels, it does not distinguish differences in crowded parts of the vowel space (likely to cause confusions) from differences in sparse regions (unlikely to cause confusions). This suggests the need for a hybrid measure somehow encompassing both within-category variation and category overlap or encroachment. More research is needed to determine how such a measure could be derived.

5.3.2 Prosodic measures

Post hoc analyses of the prosodic measures related to f_0 are shown in Figure 5.10. The expected pattern based on statistical model coefficients for prosodic donor is Talker B > Talker A > Talker C. None of the three measures match this pattern exactly; the closest is mean f_0 dynamicity, which shows a pattern of B ≈ C > A. Mean f_0 range shows a pattern of C > B > A, as does mean f_0 velocity (in the case of velocity, ">" meaning "more negative").[4] One explanation

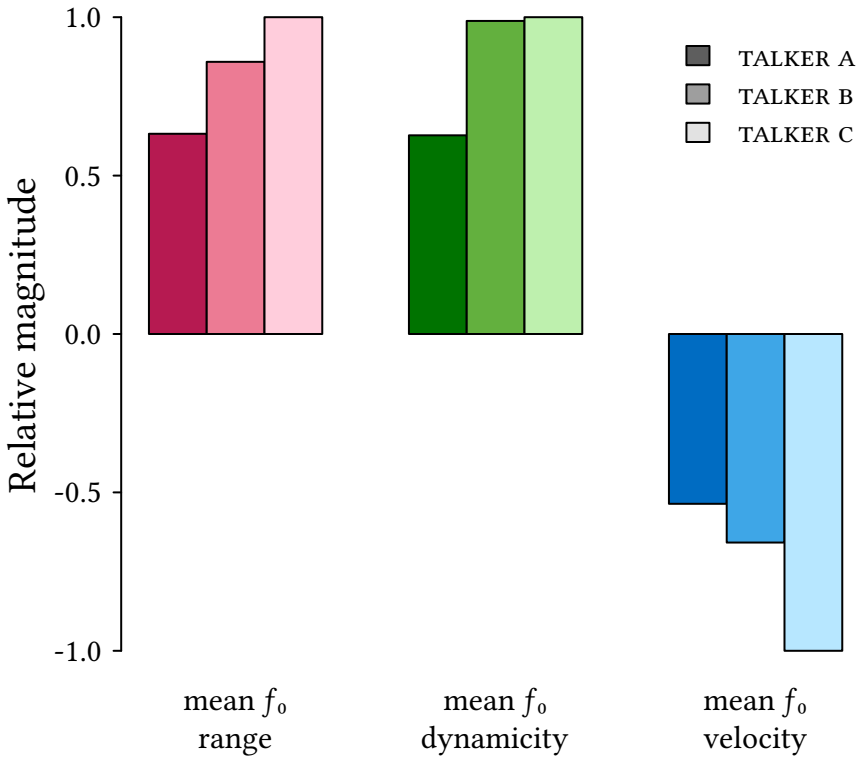

Figure 5.10: Barplot of the relative magnitudes of three f_0-related metrics. Each metric has been scaled by dividing by the maximum absolute value among the three talkers.

that ties these three measures together is the fact that Talker C exhibits a lot of creaky voicing in his speech, especially utterance-finally, which accounts for his highly negative value of f_0 velocity (nearly twice the magnitude of Talker A). The relatively frequent occurrence of creaky voicing in Talker C's stimulus sentences can be seen in Figure 5.11, by the high number of "tails" protruding downward from the main mass of pitch tracks at the ends of several sentences.

The consistent use of creaky voicing also inflates Talker C's values for mean f_0 range and mean f_0 dynamicity. Re-examining Figure 5.10 in this light, we see that both mean f_0 range and mean f_0 dynamicity show the expected relationship between Talkers A and B. Figure 5.11 also

[4]Recall that dynamicity is a measure of the mean of the *absolute value* in rate of change of f_0, whereas velocity is the mean of the signed rate of change. As such, velocity reflects overall trends in pitch across the utterance, whereas dynamicity better indexes the magnitude of pitch movements throughout the utterance.

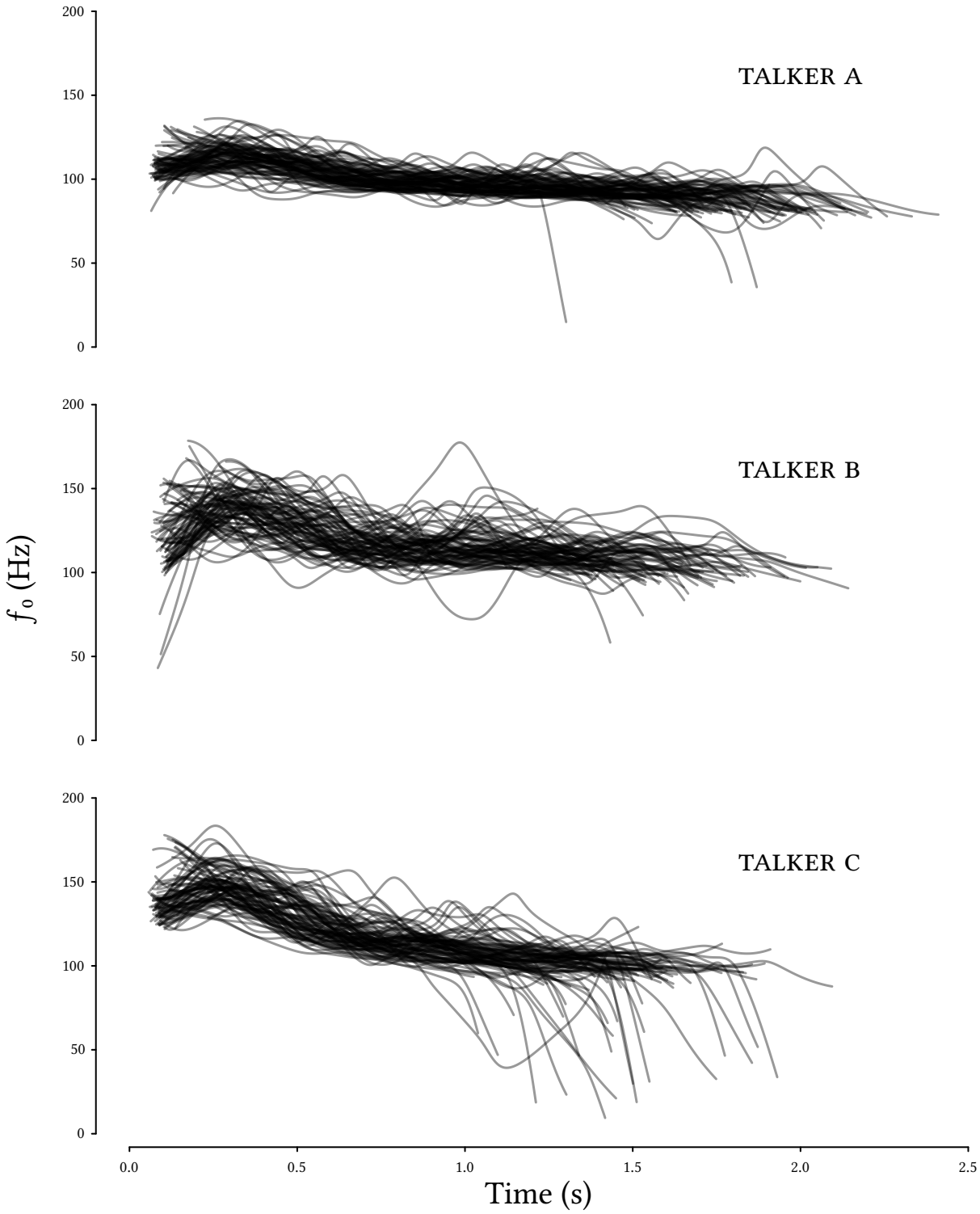

Figure 5.11: Overlaid pitch tracks for Talkers A, B and C for the 90 test sentences. Pitch tracks have been smoothed by fitting a local polynomial weighted by a Gaussian kernel with a bandwidth of 75 ms. Note the frequent occurrence of utterance-final creaky voicing (indicated by extreme drops in f_0) for Talker C.

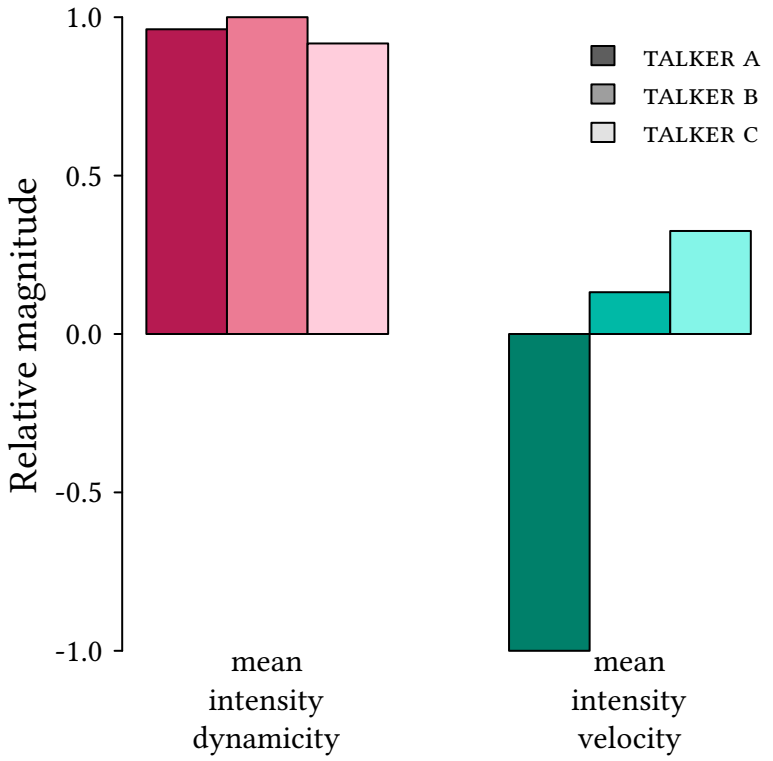

Figure 5.12: Barplot of the relative magnitudes of two intensity-related metrics. Each metric has been scaled by dividing by the maximum absolute value among the three talkers.

visually illustrates Talker A's low f_0 range (the tighter clustering of his pitch tracks compared to Talker B) and dynamicity (the straightness of the lines in his pitch tracks compared to Talker B).

Post hoc analyses of the prosodic measures related to intensity are seen in Figure 5.12. The values for mean intensity dynamicity follow the predicted pattern of B > A > C, but only the difference between Talkers B and C is significant. The small differences between talkers is probably due to the fact that intensity dynamicity mostly tracks something like obstruent/sonorant alternation rate, since the largest-magnitude intensity modulations in speech (at all but the shortest time scales) are the inherent differences between segment types — voiceless stop closures at one extreme, and open vowels at the other (cf. the well-known

linguistic notion of "sonority"). Because all talkers read the same set of sentences, the patterns of obstruent-sonorant alternation were by and large identical across talkers (modulo variation due to segmental reduction) and therefore any variation due to inter-talker differences in prosody are likely to be relatively small in comparison (and thus hard to detect statistically).

Values for mean intensity velocity show an unusual pattern in which Talker A shows an extreme negative value, with Talkers B and C showing small positive values (see Figure 5.12). Some insight into this pattern is available if we examine the overlaid intensity tracks for each talker, as seen in Figure 5.13. At the expense of showing detail at smaller time scales, the wide (150 ms) smoothing bandwidth highlights the overall trends in intensity in each recording, and reveals the strong intensity drop-off at the end of most of Talker A's sentences, the smaller drop-off in Talker B's speech, and the tendency for Talker C's speech to decline more gradually in intensity across the sentence. The strong drop-off is responsible for Talker A's large negative value for mean intensity velocity seen in Figure 5.12; in fact, given that mean intensity velocity is anticorrelated with talker intelligibility, we can infer that the strong drop-off is probably less damaging to the intelligibility of speech in noise than a gradual decline. This makes sense if Talker A's tendency for utterance-final drop-off likely only affects the last word, whereas Talker C's tendency for more gradual intensity decline gradually reduces SNR for words occurring mid-sentence.

Post hoc analyses of syllable duration are seen in Figure 5.14. Mean syllable duration for Talkers B and C is nearly identical, and thus cannot explain the difference in intelligibility between them. This is consistent with findings in the literature suggesting that within-talker

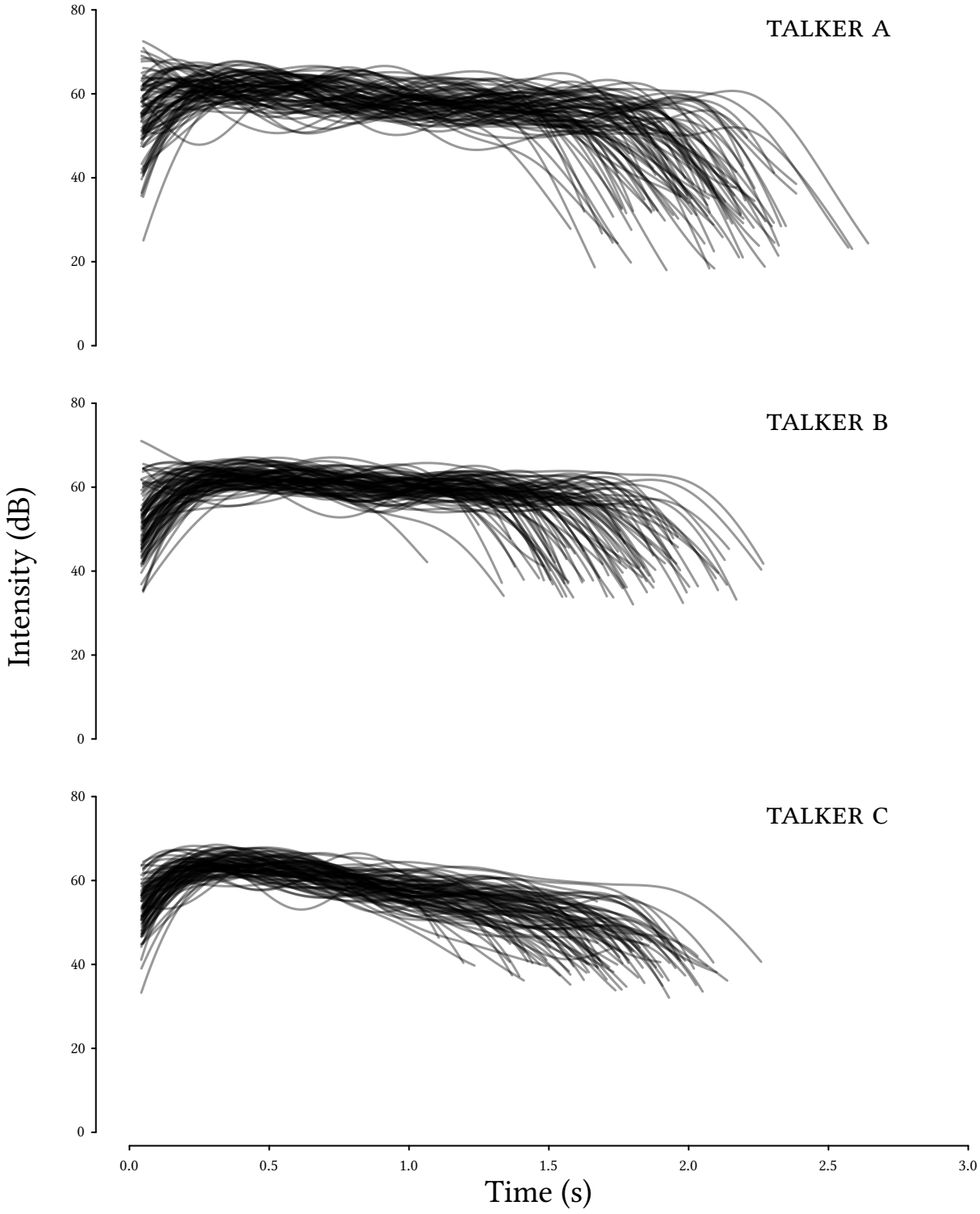

Figure 5.13: Overlaid pitch tracks for Talkers A, B and C for the 90 test sentences. Pitch tracks have been heavily smoothed by fitting a local polynomial weighted by a Gaussian kernel with a bandwidth of 150 ms.

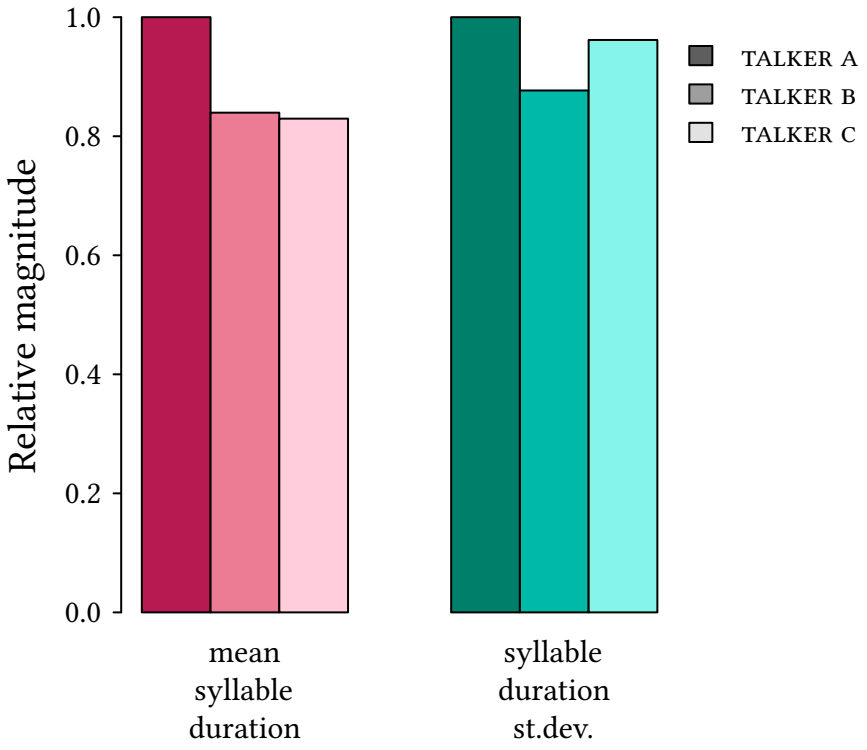

Figure 5.14: Barplot of the relative magnitudes of two duration-related metrics. Each metric has been scaled by dividing by the maximum absolute value among the three talkers.

changes in intelligibility are not necessarily accompanied by a change in speech rate (Krause & Braida, 2002); further interpretation of mean speech rate values is limited by the small number of talkers investigated here. Interestingly, *variability* in syllable duration appears to follow the pattern of non-prosodic coefficients (*i.e.*, A > C > B). This is unexpected, since large variability in syllable duration is expected to index more extreme duration contrasts between prominent and reduced syllables (more a prosodic phenomenon than a segmental or lexical one). This result also merits further investigation with a larger sample of talkers.

Chapter 6. Discussion

It is well-known that speakers of a language may use different articulatory means to achieve the same goals. Examples of such free variation abound, from the various articulations of English /ɹ/ (Hagiwara, 1995; Campbell *et al.*, 2010) to the gestural combinations used in conveying stress or prominence (de Jong, 1995). In that sense, the finding that talkers vary in their strategies toward intelligibility is unsurprising, and has a clear theoretical forerunner in the H&H theory of Lindblom (1990).

Nonetheless, the individual differences in intelligibility strategies seen here are interesting when viewed in light of the information present in the speech signal. That is, some talkers (such as Talker A in these experiments) seem to maintain a higher information density at the segmental level of their speech, making them relatively intelligible even when the low-frequency modulations of their speech do little to enhance the successful transmission of information. In contrast, other talkers (such as Talker B) compensate for lower information density at the segmental level by maintaining prosodic patterns that are especially conducive to successful transmission of that information. Put simply, once we set aside the most common causes of misunderstood speech in everyday life (unpredictable background noises, unfamiliar dialects or foreign accents, unexpected content given context, *etc*), we still find that not all (un)intelligible talkers are (un)intelligible for the same reasons.

6.1 Talker familiarity

With regard to talker familiarity, it was surprising that listeners failed to realize any advantage due to talker familiarity in Experiment 2. One possible explanation is that the training was too brief, and the apparent learning seen in the quartile analysis is merely a task familiarization effect that was not tied to any particular talker. However, previous studies involving speech-in-noise tests with non-native English speakers showed that listeners adapt to even the lowest-intelligibility talkers over the course of a single 64-sentence test session (Bradlow & Bent, 2008); thus it was predicted that a 90-sentence training phase that included both noise-masked and clear presentations of each sentence would be sufficient to realize a familiarity advantage that would persist through the testing phase. Moreover, the fact that the experimental and control groups differed in the magnitude of improvement during training suggests that the listener adaptation that took place was talker-specific, at least to some extent.

Further support for the belief that listener adaptation was talker-specific comes if we compare the training phase of Experiment 2 to the testing phase of Experiment 1. Since Experiment 1 did not involve a training phase, listener adaptation during Experiment 1 occurred during exposure to all nine test "talkers" (three unmodified and six resynthesized), presented in random order. Thus the improvement seen across quartiles in Experiment 1 (an improvement of 0.39 keywords) cannot possibly be due to talker-specific adaptations on the part of the listeners. Compare this to the magnitude of improvement in the testing phase of Experiment 2, in which listeners in the experimental group (training on Talker C) improved by 0.38 keywords, while listeners in the control group (trained on Talker D) improved by 0.50 keywords. The fact that

listeners in the control group showed greater adaptation during training suggests that there was something about their training talker that enabled that greater adaptation to take place.

This difference also raises the question of what attribute(s) of Talker D's speech facilitated the greater adaptation seen in the control group listeners. It is known from the pilot study that Talker D's intelligibility is slightly better than Talker C (mean 3.0 *vs.* 2.6 keywords correct in a similar speech-in-noise task, cf. Section 4.1). Unfortunately, because Talker D was a control talker and did not appear in the testing phase, it is unknown whether that adaptation would have persisted for the listeners in the control group. Also unfortunate is the fact that Talker D was not used in the creation of resynthesized stimuli, so it is not known whether his intelligibility rests more on segmental or prosodic aspects of his speech.

6.2 Predicting intrinsic intelligibility

Regardless of the differences in listener adaptation between listener groups, the fact remains that a listener's perceptual gain from rapid adaptation to a talker's speaking style is relatively small in magnitude, compared to the differences in intrinsic intelligibility seen across talkers. In other words, familiarity with a talker's speaking style is not enough to transform a mumbler into a clear talker in the ears of the trained listener. Nonetheless, the finding that individuals vary in their intelligibility strategies also helps clarify previous literature, in particular the diversity of acoustic predictors shown to correlate with intelligibility in various studies. Given that dramatic individual differences were seen in these experiments even among just three talkers, it is unsurprising that prior studies using different samples of talkers diverge in which acoustic measurements best predict the intelligibility of the talkers in their sample. This raises

questions regarding the generalizability of findings for studies where only one or a few talkers are taken to represent the population at large, even in cases where more prominent sources of variability (like dialect or foreign accent) are controlled.

In light of this problem, it would be unwise to interpret *any* of the *post hoc* analyses in this thesis as solid facts about English in general or even Pacific Northwestern English in particular. Nonetheless, some of the measures seemed more promising than others, and deserve to be highlighted here. First, using area of the convex hull to index vowel space size seems superior to previously described methods (*e.g.*, area of the vowel means polygon), since it is less sensitive to potentially confounding factors such as prosodically-driven reduction and phonemic within-category variation. On the other hand, both of those "confounds" may also be relevant to intelligibility, and ways of quantifying them cleanly are also needed; it is unclear whether the proposed measures tested here (mean cluster size and repulsive force) are fit for the job.

A second set of measures worth highlighting are f_0 range and f_0 dynamicity, which show promise for indexing prosodic contributions to intelligibility (on the assumption that the confounding effect of creaky voicing is addressed in some way). If these measures turn out to be reliable in indexing prosodic contributions to intelligibility, it raises the question of whether the increase in intelligibility *is actually due to* differences in f_0 range and dynamicity (or perhaps equivalently, pitch accent magnitude), or whether here again we have found an acoustic measure that merely covaries with whatever signal property is actually *responsible* for the intelligibility boost. Once again, more research is needed.

A final theoretical question concerns the generalizability of the patterns described here. As mentioned in Chapter 3, the operationalization of prosody as suprasegmental patterns of f_0,

intensity, and duration trades on particular facts about English (*i.e.*, its lack of phonemic length distinctions and lexical tones). Obviously, such a division between segmental and suprasegmental is not so clear-cut in a lexical tone or accent language. The small number of talkers involved also raises the question of whether the variation seen is typical, and whether patterns of segmental dominance *vs.* prosodic dominance vary across dialects, languages, or social groups.

6.3 Methodological lessons

The methodology employed seems to have been relatively successful at minimizing distortion due to resynthesis, as evidenced by the low coefficient for `resynth=TRUE` in the statistical models. Nonetheless, certain changes to the methodology might improve future results. First, the scaling of intensity based on intensity contours with high temporal fidelity led to the problem of intra-syllable segment duration and intensity mismatches, seen in Figure 4.4 and discussed in Section 4.2. In retrospect, a method of scaling the intensity of whole syllables by a fixed factor (presumably a ratio of intensity maxima or means between talkers), followed by RMS normalizing the resulting files might have introduced less distortion in the form of misalignment between intensity contour inflection points and segmental transitions.

Another possible methodological change would be to eliminate prosodic variability by mapping all test talkers to the prosody of a carefully selected non-test talker. Another way of neutralizing prosody would be to create a "mean prosody" for each sentence, by combining the mean syllable-wise durations across talkers with the mean pitch and intensity contours (the pitch and intensity contours would have to be dynamically time-warped to the time scale of the

mean duration pattern before averaging).

6.4 Future directions

Perhaps the most interesting application of these findings relates to the fact that some talkers rely on low-frequency modulations (*i.e.*, prosody) as a strategy for maintaining or enhancing intelligibility (indeed, a full ⅓ of the variability in intelligibility appears to be due to prosodic differences). It has been shown that listeners with hearing impairment rely more strongly on speech envelope cues (Lorenzi *et al.*, 2006); thus listeners with hearing impairment ought to perform differently on talkers whose intelligibility relies more on prosody *vs.* talkers whose intelligibility relies more on segmental factors (even in cases where the two talkers are equally intelligible to unimpaired listeners). That is, for a group of talkers whose unmodified speech is equally intelligible to a group of normal-hearing listeners, the talkers could nonetheless be discriminated based on the intelligibility of their prosody when mapped onto some other talker's speech. The expected result is that talkers whose prosody degraded intelligibility through resynthesis would be less intelligible to listeners with hearing loss. Moreover, from a signal-processing perspective, low-frequency modulations are easier to modify than segmental attributes of the signal (such as consonant release bursts), especially in real-time processing applications such as found in hearing aids.

Another question raised by this research concerns the relative contributions of duration, f_0, and intensity to intelligibility. These questions could be addressed with methods similar to those used in this thesis (*i.e.*, prosodic replacement via resynthesis), but they also introduce new complexities. In particular, the manipulation of one of the dimensions of prosody while holding

others constant introduces the possibility of conflicting cues to which syllables are prominent in the utterance. To do justice to such questions, a preliminary study of how such conflicts are resolved by listeners (and hence how prosodic cues to prominence are weighted) is required.

A final set of questions involves talker familiarity and listener adaptation. Unfortunately, the experiments in this thesis gave no answer to the question of what underlies the familiar talker advantage. More research is needed to determine whether listener adaptation to a particular talker rests primarily on segmental or prosodic aspects of speech, or whether some other talker attribute (such as voice quality) is at play. Given the individual differences seen in these experiments, there are also still questions about individual differences in listener adaptive ability, and whether listeners vary in which dimensions of speech they "tune in" to when adapting to a new talker's speech. Ultimately, as is often the case in scientific research, we are left with more questions than answers.

Bibliography

American National Standards Institute (1994). ANSI S1.1-1994 (R2004): Acoustical terminology.

American National Standards Institute (2004). ANSI/ASA S3.21-2004 (R2009): Methods for manual pure-tone threshold audiometry. Acoustical Society of America.

Baayen, R. H. (2011). languageR: Data sets and functions with "Analyzing Linguistic Data: A practical introduction to statistics", version 1.4.
URL: http://cran.R-project.org/package=languageR

Bashford, J. A., Warren, R. M., & Brown, C. A. (1996). Use of speech-modulated noise adds strong "bottom-up" cues for phonemic restoration. *Perception & Psychophysics*, *58*(3), 342–350. DOI: 10.3758/BF03206810.

Bates, D., Maechler, M., & Bolker, B. (2012). lme4: Linear mixed-effects models using S4 classes, version 0.999999-0. URL: http://cran.R-project.org/package=lme4

Beckford-Wassink, A., Squizzero, R., Schirra, R., & Conn, J. (2009). Effects of gender and style on fronting and raising of /e/, /e:/ and /æ/ before /g/ in Seattle English.
URL: http://www.sociolinguistics.uottawa.ca/nwav38/abstracts/Wassink(2009)Effects_of_Gender.pdf

Beckford-Wassink, A., Wright, R. A., & Franklin, A. D. (2007). Intraspeaker variability in vowel production: An investigation of motherese, hyperspeech, and Lombard speech in Jamaican speakers. *Journal of Phonetics*, *35*(3), 363–379. DOI: 10.1016/j.wocn.2006.07.002.

Benkí, J. R. (2003). Analysis of English nonsense syllable recognition in noise. *Phonetica*, *60*(2), 129–157. DOI: 10.1159/000071450.

Bent, T., & Bradlow, A. R. (2003). The interlanguage speech intelligibility benefit. *The Journal of the Acoustical Society of America*, *114*(3), 1600–1610. DOI: 10.1121/1.1603234.

Best, V., Gallun, F. J., Ihlefeld, A., & Shinn-Cunningham, B. G. (2006). The influence of spatial separation on divided listening. *The Journal of the Acoustical Society of America*, *120*(3), 1506–1516. DOI: 10.1121/1.2234849.

Binns, C., & Culling, J. F. (2007). The role of fundamental frequency contours in the perception of speech against interfering speech. *The Journal of the Acoustical Society of America*, *122*(3), 1765–1776. DOI: 10.1121/1.2751394.

Boersma, P., & Weenink, D. (2013). Praat: Doing phonetics by computer, version 5.3.41.
URL: http://www.praat.org/

Bolia, R. S., Nelson, W. T., Ericson, M. A., & Simpson, B. D. (2000). A speech corpus for multitalker communications research. *The Journal of the Acoustical Society of America*, *107*(2), 1065–1066. DOI: 10.1121/1.428288.

Bond, Z. S., & Moore, T. J. (1994). A note on the acoustic-phonetic characteristics of inadvertently clear speech. *Speech Communication*, *14*(4), 325–337. DOI: 10.1016/0167-6393(94)90026-4.

Bonino, A. Y., & Leibold, L. J. (2008). The effect of signal-temporal uncertainty on detection in bursts of noise or a random-frequency complex. *The Journal of the Acoustical Society of America*, *124*(5), EL321–EL327. DOI: 10.1121/1.2993745.

Boulenger, V., Hoen, M., Ferragne, E., Pellegrino, F., & Meunier, F. (2010). Real-time lexical competitions during speech-in-speech comprehension. *Speech Communication*, *52*(3), 246–253. DOI: 10.1016/j.specom.2009.11.002.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*(2), 707–729. DOI: 10.1016/j.cognition.2007.04.005.

Bradlow, A. R., Torretta, G. M., & Pisoni, D. B. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication*, *20*(3-4), 255–272. DOI: 10.1016/S0167-6393(96)00063-5.

Britain, D. (2008). Linguistic change in intonation: The use of high rising terminals in New Zealand English. *Language Variation and Change*, *4*(01), 77. DOI: 10.1017/S0954394500000661.

Brouwer, S., Van Engen, K. J., Calandruccio, L., & Bradlow, A. R. (2012). Linguistic contributions to speech-on-speech masking for native and non-native listeners: Language familiarity and semantic content. *The Journal of the Acoustical Society of America*, *131*(2), 1449–1464. DOI: 10.1121/1.3675943.

Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America*, *109*(3), 1101–1109. DOI: 10.1121/1.1345696.

Brungart, D. S., & Simpson, B. D. (2002). Within-ear and across-ear interference in a cocktail-party listening task. *The Journal of the Acoustical Society of America*, *112*(6), 2985–2995. DOI: 10.1121/1.1512703.

Brungart, D. S., & Simpson, B. D. (2004). Within-ear and across-ear interference in a dichotic cocktail party listening task: Effects of masker uncertainty. *The Journal of the Acoustical Society of America*, *115*(1), 301–310. DOI: 10.1121/1.1628683.

Brungart, D. S., Simpson, B. D., Ericson, M. A., & Scott, K. R. (2001). Informational and energetic masking effects in the perception of multiple simultaneous talkers. *The Journal of the Acoustical Society of America*, *110*(5), 2527–2538. DOI: 10.1121/1.1408946.

Calandruccio, L., Dhar, S., & Bradlow, A. R. (2010). Speech-on-speech masking with variable access to the linguistic content of the masker speech. *The Journal of the Acoustical Society of America*, *128*(2), 860–869. DOI: 10.1121/1.3458857.

Campbell, F., Gick, B., Wilson, I., & Vatikiotis-Bateson, E. (2010). Spatial and temporal properties of gestures in North American English /r/. *Language and Speech*, *53*(1), 49–69.

Carhart, R., Tillman, T. W., & Greetis, E. S. (1969). Perceptual masking in multiple sound backgrounds. *The Journal of the Acoustical Society of America*, *45*(3), 694–703. DOI: 10.1121/1.1911445.

Carhart, R., Tillman, T. W., & Johnson, K. R. (1968). Effects of interaural time delays on masking by two competing signals. *The Journal of the Acoustical Society of America*, *43*(6), 1223–1230. DOI: 10.1121/1.1910971.

Charpentier, F. J., & Moulines, É. (1988). Text-to-speech algorithms based on FFT synthesis. In *Proceedings of ICASSP-88*, vol. 1, 667–670. DOI: 10.1109/ICASSP.1988.196674.

Cho, T., McQueen, J. M., & Cox, E. A. (2007). Prosodically driven phonetic detail in speech processing: The case of domain-initial strengthening in English. *Journal of Phonetics*, *35*(2), 210–243. DOI: 10.1016/j.wocn.2006.03.003.

Clopper, C. G., & Smiljanić, R. (2011). Effects of gender and regional dialect on prosodic patterns in American English. *Journal of Phonetics*, *39*(2), 237–245. DOI: 10.1016/j.wocn.2011.02.006.

Cole, J., Kim, H., Choi, H., & Hasegawa-Johnson, M. (2007). Prosodic effects on acoustic cues to stop voicing and place of articulation: Evidence from Radio News speech. *Journal of Phonetics*, *35*(2), 180–209. DOI: 10.1016/j.wocn.2006.03.004.

Cooke, M. (2006). A glimpsing model of speech perception in noise. *The Journal of the Acoustical Society of America*, *119*(3), 1562–1573. DOI: 10.1121/1.2166600.

Cooke, M., & Lu, Y. (2010). Spectral and temporal changes to speech produced in the presence of energetic and informational maskers. *The Journal of the Acoustical Society of America*, *128*(4), 2059. DOI: 10.1121/1.3478775.

Cushing, I. R., & Dellwo, V. (2010). The role of speech rhythm in attending to one of two simultaneous speakers. Paper presented at the 5th International Conference on Speech Prosody. In *SP-2010*, paper 039.
URL: http://www.isca-speech.org/archive/sp2010/sp10_039.html

Darwin, C. J. (2008). Spatial hearing and perceiving sources. In W. Yost, A. N. Popper, & R. R. Fay (Eds.) *Auditory perception of sound sources*, 215–232. Boston: Springer.

de Jong, K. J. (1995). The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *The Journal of the Acoustical Society of America*, *97*(1), 491–504. DOI: 10.1121/1.412275.

Delattre, P. C., Liberman, A. M., & Cooper, F. S. (1955). Acoustic loci and transitional cues for consonants. *The Journal of the Acoustical Society of America*, *27*(4), 769–773. DOI: 10.1121/1.1908024.

Dilley, L. C., Shattuck-Hufnagel, S., & Ostendorf, M. (1996). Glottalization of word-initial vowels as a function of prosodic structure. *Journal of Phonetics*, *24*(4), 423–444. DOI: 10.1006/jpho.1996.0023.

Dirks, D. D., Takayanagi, S., & Moshfegh, A. (2001). Effects of lexical factors on word recognition among normal-hearing and hearing-impaired listeners. *Journal of the American Academy of Audiology*, *12*(5), 233–244. URL: http://www.audiology.org/resources/journal/Documents/JAAA12/JAAA_12_05_02.pdf

Dreher, J. J., & O'Neill, J. (1957). Effects of ambient noise on speaker intelligibility for words and phrases. *The Journal of the Acoustical Society of America*, *29*(12), 1320–1323. DOI: 10.1121/1.1908780.

Dubno, J. R., Ahlstrom, J. B., Wang, X., & Horwitz, A. R. (2012). Level-dependent changes in perception of speech envelope cues. *Journal of the Association for Research in Otolaryngology*, *13*(6), 835–852. DOI: 10.1007/s10162-012-0343-2.

Durlach, N. I., Mason, C. R., Kidd Jr., G., Arbogast, T. L., Colburn, H. S., & Shinn-Cunningham, B. G. (2003a). Note on informational masking. *The Journal of the Acoustical Society of America*, *113*(6), 2984–2987. DOI: 10.1121/1.1570435.

Durlach, N. I., Mason, C. R., Shinn-Cunningham, B. G., Arbogast, T. L., Colburn, H. S., & Kidd Jr., G. (2003b). Informational masking: Counteracting the effects of stimulus uncertainty by decreasing target-masker similarity. *The Journal of the Acoustical Society of America*, *114*(1), 368–379. DOI: 10.1121/1.1577562.

Ericson, M. A., Brungart, D. S., & Simpson, B. D. (2004). Factors that influence intelligibility in multitalker speech displays. *The International Journal of Aviation Psychology*, *14*(3), 313–334. DOI: 10.1207/s15327108ijap1403_6.

Fan, W. L., Streeter, T. M., & Durlach, N. I. (2008). Effect of spatial uncertainty of masker on masked detection for nonspeech stimuli. *The Journal of the Acoustical Society of America*, *124*(1), 36–39. DOI: 10.1121/1.2932257.

Festen, J. M., & Plomp, R. (1990). Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *The Journal of the Acoustical Society of America*, *88*(4), 1725–1736. DOI: 10.1121/1.400247.

Fougeron, C., & Keating, P. A. (1997). Articulatory strengthening at edges of prosodic domains. *The Journal of the Acoustical Society of America*, *101*(6), 3728–3740. DOI: 10.1121/1.418332.

Francis, A. L., & Nusbaum, H. C. (1996). Paying attention to speaking rate. Paper presented at the 4th International Conference on Spoken Language Processing. In *ICSLP-1996*, vol. 3, 1537–1540. DOI: 10.1109/ICSLP.1996.607911.

Freyman, R. L., Balakrishnan, U., & Helfer, K. S. (2004). Effect of number of masking talkers and auditory priming on informational masking in speech recognition. *The Journal of the Acoustical Society of America*, *115*(5), 2246–2256. DOI: 10.1121/1.1689343.

Freyman, R. L., Helfer, K. S., & Balakrishnan, U. (2007). Variability and uncertainty in masking by competing speech. *The Journal of the Acoustical Society of America*, *121*(2), 1040–1046. DOI: 10.1121/1.2427117.

Freyman, R. L., Helfer, K. S., McCall, D. D., & Clifton, R. K. (1999). The role of perceived spatial separation in the unmasking of speech. *The Journal of the Acoustical Society of America*, *106*(6), 3578–3588. DOI: 10.1121/1.428211.

Fry, D. B. (1955). Duration and intensity as physical correlates of linguistic stress. *The Journal of the Acoustical Society of America*, *27*(4), 765–768. DOI: 10.1121/1.1908022.

Gallun, F. J., Mason, C. R., & Kidd Jr., G. (2005). Binaural release from informational masking in a speech identification task. *The Journal of the Acoustical Society of America*, *118*(3), 1614–1625. DOI: 10.1121/1.1984876.

Gallun, F. J., Mason, C. R., & Kidd Jr., G. (2007). The ability to listen with independent ears. *The Journal of the Acoustical Society of America*, *122*(5), 2814–2825. DOI: 10.1121/1.2780143.

García Lecumberri, M. L., & Cooke, M. (2006). Effect of masker type on native and non-native consonant perception in noise. *The Journal of the Acoustical Society of America*, *119*(4), 2445–2454. DOI: 10.1121/1.2180210.

Gaudio, R. P. (1994). Sounding gay: Pitch properties in the speech of gay and straight men. *American Speech*, *69*(1), 30–57. URL: http://www.jstor.org/stable/455948

Gordon, M., & Ladefoged, P. (2001). Phonation types: A cross-linguistic overview. *Journal of Phonetics*, *29*(4), 383–406. DOI: 10.1006/jpho.2001.0147.

Gordon-Salant, S., Frisina, R. D., Popper, A. N., & Fay, R. R. (Eds.) (2010). *The aging auditory system*. New York: Springer.
URL: http://www.springer.com/medicine/otorhinolaryngology/book/978-1-4419-0992-3

Grabe, E., & Post, B. (2002). Intonational variation in the British Isles. Paper presented at the International Conference on Speech Prosody. In *SP-2002*, 343–346.
URL: http://www.isca-speech.org/archive_open/sp2002/sp02_343.html

Hagiwara, R. (1995). Acoustic realizations of American /r/ as produced by women and men. *UCLA Working Papers in Phonetics*, *90*. URL: http://escholarship.org/uc/item/8779b7gq

Hawkins, J. E., & Stevens, S. S. (1950). The masking of pure tones and of speech by white noise. *The Journal of the Acoustical Society of America*, *22*(1), 6–13. DOI: 10.1121/1.1906581.

Hazan, V., & Baker, R. (2011). Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions. *The Journal of the Acoustical Society of America*, *130*(4), 2139. DOI: 10.1121/1.3623753.

Hazan, V., & Markham, D. (2004). Acoustic-phonetic correlates of talker intelligibility for adults and children. *The Journal of the Acoustical Society of America*, *116*(5), 3108–3118.
DOI: 10.1121/1.1806826.

Hazan, V., & Simpson, A. (2000). The effect of cue-enhancement on consonant intelligibility in noise: Speaker and listener effects. *Language and Speech*, *43*(3), 273–294.
DOI: 10.1177/00238309000430030301.

Helfer, K. S., & Freyman, R. L. (2008). Aging and speech-on-speech masking. *Ear and Hearing*, *29*(1), 87–98. DOI: 10.1097/AUD.0b013e31815d638b.

Henke, E., Kaisse, E. M., & Wright, R. A. (2012). Is the Sonority Sequencing Principle an epiphenomenon? In S. G. Parker (Ed.) *The sonority controversy*, 65–100. Berlin: de Gruyter Mouton. DOI: 10.1515/9783110261523.65.

Hirsh, I. J. (1950). The relation between localization and intelligibility. *The Journal of the Acoustical Society of America*, *22*(2), 196–200. DOI: 10.1121/1.1906588.

Hoen, M., Meunier, F., Grataloup, C.-L., Pellegrino, F., Grimault, N., Perrin, F., Perrot, X., & Collet, L. (2007). Phonetic and lexical interferences in informational masking during speech-in-speech comprehension. *Speech Communication*, *49*(12), 905–916. DOI: 10.1016/j.specom.2007.05.008.

Horii, Y., House, A. S., & Hughes, G. W. (1971). A masking noise with speech-envelope characteristics for studying intelligibility. *The Journal of the Acoustical Society of America*, *49*(6B), 1849–1856. DOI: 10.1121/1.1912590.

Howes, D. (1957). On the relation between the intelligibility and frequency of occurrence of English words. *The Journal of the Acoustical Society of America*, *29*(2), 296–305. DOI: 10.1121/1.1908862.

Imai, S., Walley, A. C., & Flege, J. E. (2005). Lexical frequency and neighborhood density effects on the recognition of native and Spanish-accented words by native English and Spanish listeners. *The Journal of the Acoustical Society of America*, *117*(2), 896–907. DOI: 10.1121/1.1823291.

Johnstone, P. M., & Litovsky, R. Y. (2006). Effect of masker type and age on speech intelligibility and spatial release from masking in children and adults. *The Journal of the Acoustical Society of America*, *120*(4), 2177–2189. DOI: 10.1121/1.2225416.

Junqua, J.-C. (1993). The Lombard reflex and its role on human listeners and automatic speech recognizers. *The Journal of the Acoustical Society of America*, *93*(1), 510–524. DOI: 10.1121/1.405631.

Keating, P., & Kuo, G. (2012). Comparison of speaking fundamental frequency in English and Mandarin. *The Journal of the Acoustical Society of America*, *132*(2), 1050–1060. DOI: 10.1121/1.4730893.

Keating, P. A. (2006). Phonetic encoding of prosodic structure. In J. Harrington, & M. Tabain (Eds.) *Speech production: Models, phonetic processes, and techniques*, 167–186. New York: Psychology Press.

Kidd Jr., G., Arbogast, T. L., Mason, C. R., & Gallun, F. J. (2005). The advantage of knowing where to listen. *The Journal of the Acoustical Society of America*, *118*(6), 3804–3815. DOI: 10.1121/1.2109187.

Kidd Jr., G., Mason, C. R., & Arbogast, T. L. (2002). Similarity, uncertainty, and masking in the identification of nonspeech auditory patterns. *The Journal of the Acoustical Society of America*, *111*(3), 1367–1376. DOI: 10.1121/1.1448342.

Kidd Jr., G., Richards, V. M., Mason, C. R., Gallun, F. J., & Huang, R. (2008). Informational masking increases the costs of monitoring multiple channels. *The Journal of the Acoustical Society of America*, *124*(4), EL223–EL229. DOI: 10.1121/1.2968302.

Kiliç, M. A., & Ögüt, F. (2004). The effect of the speaker gender on speech intelligibility in normal-hearing subjects with simulated high frequency hearing loss. *Revue de laryngologie - otologie - rhinologie*, *125*(1), 35–38.

Kitterick, P. T., Bailey, P. J., & Summerfield, A. Q. (2010). Benefits of knowing who, where, and when in multi-talker listening. *The Journal of the Acoustical Society of America*, *127*(4), 2498–2508. DOI: 10.1121/1.3327507.

Kock, W. E. (1950). Binaural localization and masking. *The Journal of the Acoustical Society of America*, *22*(6), 801–804. DOI: 10.1121/1.1906692.

Krause, J. C., & Braida, L. D. (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. DOI: 10.1121/1.1509432.

Krause, J. C., & Braida, L. D. (2004). Acoustic properties of naturally produced clear speech at normal speaking rates. *The Journal of the Acoustical Society of America*, *115*(1), 362–378. DOI: 10.1121/1.1635842.

Krause, J. C., & Braida, L. D. (2009). Evaluating the role of spectral and envelope characteristics in the intelligibility advantage of clear speech. *The Journal of the Acoustical Society of America*, *125*(5), 3346–3357. DOI: 10.1121/1.3097491.

Kreiman, J. (1982). Perception of sentence and paragraph boundaries in natural conversation. *Journal of Phonetics*, *10*(2), 163–175.

Kruskal, J. B., & Liberman, M. Y. (1983). The symmetric time-warping problem: From continuous to discrete. In D. Sankoff, & J. B. Kruskal (Eds.) *Time warps, string edits, and macromolecules: The theory and practice of sequence comparison*, 125–161. Cambridge: Cambridge University Press.

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., Stolyarova, E. I., Sundberg, U., & Lacerda, F. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science*, *277*(5326), 684–686. DOI: 10.1126/science.277.5326.684.

Labov, W., Ash, S., & Boberg, C. (2006). *The Atlas of North American English: Phonetics, Phonology and Sound Change*. New York: Mouton de Gruyter. DOI: 10.1515/9783110167467.

Ladefoged, P. (1967). Stress and respiratory activity. In *Three areas of experimental phonetics*, 1–49. London: Oxford University Press.

Ladefoged, P., & Maddieson, I. (1996). *The sounds of the world's languages*. Oxford: Blackwell.

Lane, H., & Tranel, B. (1971). The Lombard sign and the role of hearing in speech. *Journal of Speech and Hearing Research*, *14*(4), 677–709. URL: http://jslhr.asha.org/cgi/content/abstract/14/4/677

Lehiste, I. (1979). Perception of sentence and paragraph boundaries. In B. Lindblom, S. E. G. Öhman, & G. Fant (Eds.) *Frontiers of speech communication research*, 191–201. London: Academic Press.

Levi, S. V., Winters, S. J., & Pisoni, D. B. (2011). Effects of cross-language voice training on speech perception: Whose familiar voices are more intelligible? *The Journal of the Acoustical Society of America*, *130*(6), 4053–4062. DOI: 10.1121/1.3651816.

Lewis, H. D., Benignus, V. A., Muller, K. E., Malott, C. M., & Barton, C. N. (1988). Babble and random-noise masking of speech in high and low context cue conditions. *Journal of Speech and Hearing Research*, *31*(1), 108–114. URL: http://jslhr.asha.org/cgi/content/abstract/31/1/108

Li, N., & Loizou, P. C. (2008). The contribution of obstruent consonants and acoustic landmarks to speech recognition in noise. *The Journal of the Acoustical Society of America*, *124*(6), 3947–3958. DOI: 10.1121/1.2997435.

Liljencrants, J., & Lindblom, B. (1972). Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language*, *48*(4), 839–862.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In W. J. Hardcastle, & A. Marchal (Eds.) *Speech production and speech modeling*, 403–439. Dordrecht: Kluwer. DOI: 10.1007/978-94-009-2037-8_16.

Liu, S., & Zeng, F.-G. (2006). Temporal properties in clear speech perception. *The Journal of the Acoustical Society of America*, *120*(1), 424–432. DOI: 10.1121/1.2208427.

Lorenzi, C., Gilbert, G., Carn, H., Garnier, S., & Moore, B. C. J. (2006). Speech perception problems of the hearing impaired reflect inability to use temporal fine structure. *Proceedings of the National Academy of Sciences*, *103*(49), 18866–18869. DOI: 10.1073/pnas.0607364103.

Lu, Y., & Cooke, M. (2008). Speech production modifications produced by competing talkers, babble, and stationary noise. *The Journal of the Acoustical Society of America*, *124*(5), 3261. DOI: 10.1121/1.2990705.

Lu, Y., & Cooke, M. (2009). The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise. *Speech Communication*, *51*(12), 1253–1262. DOI: 10.1016/j.specom.2009.07.002.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, *19*(1), 1–36. URL: http://journals.lww.com/ear-hearing/Abstract/1998/02000/Recognizing_Spoken_Words__The_Neighborhood.1.aspx

Majewski, W., Hollien, H., & Zalewski, J. (1972). Speaking fundamental frequency of Polish adult males. *Phonetica*, *25*(2), 119–125. DOI: 10.1159/000259375.

Mattys, S. L., Brooks, J., & Cooke, M. (2009). Recognizing speech under a processing load: Dissociating energetic from informational factors. *Cognitive Psychology*, *59*(3), 203–243. DOI: 10.1016/j.cogpsych.2009.04.001.

Mayo, C., Aubanel, V., & Cooke, M. (2012). Effect of prosodic changes on speech intelligibility. Paper presented at the 13th Annual Conference of the International Speech Communication Association. In *INTERSPEECH-2012*.
URL: http://interspeech2012.org/accepted-abstract.html?id=661

McCloy, D. R., Souza, P. E., Wright, R. A., Haywood, J., Gehani, N., & Rudolph, S. (2013). The PN/NC corpus, version 1.0. URL: http://depts.washington.edu/phonlab/resources/pnnc/

McCloy, D. R., Wright, R. A., & McGrath, A. T. D. (2012). Modelling talker intelligibility variation in a dialect-controlled corpus. Poster presented at the 164th Meeting of the Acoustical Society of America.
URL: http://students.washington.edu/drmccloy/pubs/McCloyEtAl2012_asaIntelPoster.pdf

McCloy, D. R., Wright, R. A., & Souza, P. E. (submitted). Models of intelligibility variation: Prosodic and vowel-space predictors. *The Journal of the Acoustical Society of America*.
URL: http://students.washington.edu/drmccloy/pubs/McCloyEtAl_IntelligibilityModeling.pdf

McConnell-Ginet, S. (1978). Intonation in a man's world. *Signs*, *3*(3), 541–559.
URL: http://www.jstor.org/stable/3173170

McLemore, C. A. (1991). *The pragmatic interpretation of English intonation: Sorority speech*. Doctoral dissertation, University of Texas at Austin.

Miller, G. A. (1947). The masking of speech. *Psychological Bulletin*, *44*(2), 105–129.
DOI: 10.1037/h0055960.

Miller, G. A., & Licklider, J. C. R. (1950). The intelligibility of interrupted speech. *The Journal of the Acoustical Society of America*, *22*(2), 167–173. DOI: 10.1121/1.1906584.

Moore, B. C. J. (2008). *An introduction to the psychology of hearing*. Bingley: Emerald, 5th ed.

Moulines, É., & Charpentier, F. J. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, *9*(5-6), 453–467.
DOI: 10.1016/0167-6393(90)90021-Z.

Neel, A. T. (2008). Vowel space characteristics and vowel identification accuracy. *Journal of Speech, Language, and Hearing Research*, *51*(3), 574–585. DOI: 10.1044/1092-4388(2008/041).

Neff, D. L., & Callahan, B. P. (1988). Effective properties of multicomponent simultaneous maskers under conditions of uncertainty. *The Journal of the Acoustical Society of America*, *83*(5), 1833–1838. DOI: 10.1121/1.396518.

Neff, D. L., & Green, D. M. (1987). Masking produced by spectral uncertainty with multicomponent maskers. *Perception & Psychophysics*, *41*(5), 409–415.
DOI: 10.3758/BF03203033.

Newman, R. S., & Evers, S. (2007). The effect of talker familiarity on stream segregation. *Journal of Phonetics*, *35*(1), 85–103. DOI: 10.1016/j.wocn.2005.10.004.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics*, *60*(3), 355–376. DOI: 10.3758/BF03206860.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, *5*(1), 42–46. URL: http://www.jstor.org/stable/40062340

Park, T. J., Klug, A., Holinstat, M., & Grothe, B. (2004). Interaural level difference processing in the lateral superior olive and the inferior colliculus. *Journal of Neurophysiology*, *92*(1), 289–301. DOI: 10.1152/jn.00961.2003.

Patel, A. D., Xu, Y., & Wang, B. (2010). The role of F0 variation in the intelligibility of Mandarin sentences. Paper presented at the 5th International Conference on Speech Prosody. In *SP-2010*, paper 890. URL: http://www.isca-speech.org/archive/sp2010/sp10_890.html

Patel, R., & Schell, K. W. (2008). The influence of linguistic content on the Lombard effect. *Journal of Speech, Language, and Hearing Research*, *51*(1), 209–220. DOI: 10.1044/1092-4388(2008/016).

Perrachione, T. K., del Tufo, S. N., & Gabrieli, J. D. E. (2011). Human voice recognition depends on language ability. *Science*, *333*(6042), 595–595. DOI: 10.1126/science.1207327.

Perrachione, T. K., & Wong, P. C. M. (2007). Learning to recognize speakers of a non-native language: Implications for the functional organization of human auditory cortex. *Neuropsychologia*, *45*(8), 1899–1910. DOI: 10.1016/j.neuropsychologia.2006.11.015.

Peterson, G. E., & Lehiste, I. (1960). Duration of syllable nuclei in English. *The Journal of the Acoustical Society of America*, *32*(6), 693–703. DOI: 10.1121/1.1908183.

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*(1), 96–103. URL: http://jslhr.asha.org/cgi/content/abstract/28/1/96

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*(4), 434–446. URL: http://jslhr.asha.org/cgi/content/abstract/29/4/434

Picheny, M. A., Durlach, N. I., & Braida, L. D. (1989). Speaking clearly for the hard of hearing III: An attempt to determine the contribution of speaking rate to differences in intelligibility between clear and conversational speech. *Journal of Speech and Hearing Research*, *32*(3), 600–603. URL: http://jslhr.asha.org/cgi/content/abstract/32/3/600

Pichora-Fuller, M. K., Schneider, B. A., & Daneman, M. (1995). How young and old adults listen to and remember speech in noise. *The Journal of the Acoustical Society of America*, *97*(1), 593–608. DOI: 10.1121/1.412282.

Pick Jr., H. L., Siegel, G. M., Fox, P. W., Garber, S. R., & Kearney, J. K. (1989). Inhibiting the Lombard effect. *The Journal of the Acoustical Society of America*, *85*(2), 894–900. DOI: 10.1121/1.397561.

Pickles, J. O. (2008). *An introduction to the physiology of hearing*. Bingley: Emerald, 3rd ed.

Pinet, M., Iverson, P., & Huckvale, M. (2011). Second-language experience and speech-in-noise recognition: Effects of talker–listener accent similarity. *The Journal of the Acoustical Society of America*, *130*, 1653–1662. DOI: 10.1121/1.3613698.

Plag, I., Kunter, G., & Schramm, M. (2011). Acoustic correlates of primary and secondary stress in North American English. *Journal of Phonetics*, *39*(3), 362–374. DOI: 10.1016/j.wocn.2011.03.004.

Podesva, R. J. (2011). Salience and the social meaning of declarative contours: Three case studies of gay professionals. *Journal of English Linguistics*, *39*(3), 233–264. DOI: 10.1177/0075424211405161.

Pollack, I., & Pickett, J. M. (1958). Masking of speech by noise at high sound levels. *The Journal of the Acoustical Society of America*, *30*(2), 127–130. DOI: 10.1121/1.1909503.

Quené, H., & van Delft, L. E. (2010). Non-native durational patterns decrease speech intelligibility. *Speech Communication*, *52*(11-12), 911–918. DOI: 10.1016/j.specom.2010.03.005.

R Development Core Team (2013). R: A language and environment for statistical computing, version 2.15.3. URL: http://www.R-project.org/

Ramus, F. (2002). Acoustic correlates of linguistic rhythm: Perspectives. Paper presented at the International Conference on Speech Prosody. In *SP-2002*, 115–120. URL: http://www.isca-speech.org/archive_open/sp2002/sp02_115.html

Reed, C. E. (1952). The pronunciation of English in the state of Washington. *American Speech*, *27*(3), 186–189.

Rhebergen, K. S., Versfeld, N. J., & Dreschler, W. A. (2005). Release from informational masking by time reversal of native and non-native interfering speech. *The Journal of the Acoustical Society of America*, *118*(3 Pt 1), 1274–1277. DOI: 10.1121/1.2000751.

Rosen, S. (1992). Temporal information in speech: Acoustic, auditory and linguistic aspects. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *336*(1278), 367–373. URL: http://www.jstor.org/stable/55906

Rothauser, E. H., Chapman, W. D., Guttman, N., Hecker, M. H. L., Nordby, K. S., Silbiger, H. R., Urbanek, G. E., & Weinstock, M. (1969). IEEE recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics*, *17*, 225–246. DOI: 10.1109/TAU.1969.1162058.

Sataloff, R. T., & Sataloff, J. (2005). *Hearing loss*. New York: Taylor & Francis, 4th ed.

Savin, H. B. (1963). Word-frequency effect and errors in the perception of speech. *The Journal of the Acoustical Society of America*, *35*(2), 200–206. DOI: 10.1121/1.1918432.

Shinn, P., & Blumstein, S. E. (1984). On the role of the amplitude envelope for the perception of [b] and [w]. *The Journal of the Acoustical Society of America*, *75*(4), 1243–1252. DOI: 10.1121/1.390677.

Simpson, S. A., & Cooke, M. (2005). Consonant identification in N-talker babble is a nonmonotonic function of N. *The Journal of the Acoustical Society of America*, *118*(5), 2775–2778. DOI: 10.1121/1.2062650.

Sluijter, A. M. C., & van Heuven, V. J. (1993). Perceptual cues of linguistic stress: Intensity revisited. In *PROSODY-1993*, 246–249.
URL: http://www.isca-speech.org/archive_open/prosody_93/pro3_246.html

Sluijter, A. M. C., & van Heuven, V. J. (1996). Acoustic correlates of linguistic stress and accent in Dutch and American English. Paper presented at the 4th International Conference on Spoken Language Processing. In *ICSLP-1996*, vol. 2, 630–633. DOI: 10.1109/ICSLP.1996.607440.

Smiljanić, R., & Bradlow, A. R. (2008). Temporal organization of English clear and conversational speech. *The Journal of the Acoustical Society of America*, *124*(5), 3171–3182. DOI: 10.1121/1.2990712.

Sommers, M. S., & Danielson, S. M. (1999). Inhibitory processes and spoken word recognition in young and older adults: The interaction of lexical competition and semantic context. *Psychology and Aging*, *14*(3), 458–472. DOI: 10.1037/0882-7974.14.3.458.

Sommers, M. S., Nygaard, L. C., & Pisoni, D. B. (1994). Stimulus variability and spoken word recognition I: Effects of variability in speaking rate and overall amplitude. *The Journal of the Acoustical Society of America*, *96*(3), 1314–1324. DOI: 10.1121/1.411453.

Souza, P. E., Gehani, N., Wright, R. A., & McCloy, D. R. (in press). The advantage of knowing the talker. *Journal of the American Academy of Audiology*.
URL: http://students.washington.edu/drmccloy/pubs/SouzaEtAl_FamiliarTalkerPrepub.pdf

Strik, H., & Boves, L. (1995). Downtrend in F0 and Psb. *Journal of Phonetics*, *23*(1-2), 203–220. DOI: 10.1016/S0095-4470(95)80043-3.

Summers, W. V., & Cord, M. T. (2007). Intelligibility of speech in noise at high presentation levels: Effects of hearing loss and frequency region. *The Journal of the Acoustical Society of America*, *122*(2), 1130–1137. DOI: 10.1121/1.2751251.

Summers, W. V., & Molis, M. R. (2004). Speech recognition in fluctuating and continuous maskers: Effects of hearing loss and presentation level. *Journal of Speech, Language, and Hearing Research*, *47*(2), 245–256. DOI: 10.1044/1092-4388(2004/020).

Summers, W. V., 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. DOI: 10.1121/1.396660.

Syrdal, A. K., & Gopal, H. S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *The Journal of the Acoustical Society of America*, *79*(4), 1086–1100. DOI: 10.1121/1.393381.

Takayanagi, S., Dirks, D. D., & Moshfegh, A. (2002). Lexical and talker effects on word recognition among native and non-native listeners with normal and impaired hearing. *Journal of Speech, Language, and Hearing Research*, *45*(3), 585–597.
URL: http://jslhr.asha.org/cgi/content/abstract/45/3/585

Tilsen, S., & Johnson, K. (2008). Low-frequency Fourier analysis of speech rhythm. *The Journal of the Acoustical Society of America*, *124*(2), EL34–EL39. DOI: 10.1121/1.2947626.

Todaka, Y. (1993). *A cross-language study of voice quality: Bilingual Japanese and American English speakers*. Doctoral dissertation, UCLA, Los Angeles.

Tolhurst, G. C. (1954). The effect on intelligibility scores of specific instructions regarding talking. Tech. Rep. AD0051810, Ohio State University Research Foundation, Columbus, OH.

Tolhurst, G. C. (1955). The effects of an instruction to be intelligible upon a speakers intelligibility, sound pressure level, and message duration. Tech. Rep. AD0094574, Ohio State University Research Foundation, Columbus, OH.

Tolhurst, G. C. (1957a). Effects of duration and articulation changes on intelligibility, word reception and listener preference. *Journal of Speech and Hearing Disorders*, *22*(3), 328–334.

Tolhurst, G. C. (1957b). The relationship of speaker intelligibility to the sound pressure level of continuous noise environments of various spectra and octave-band widths. Tech. Rep. AD0135195, Ohio State University Research Foundation, Columbus, OH.

Tollin, D. J., & Yin, T. C. T. (2005). Interaural phase and level difference sensitivity in low-frequency neurons in the lateral superior olive. *Journal of Neuroscience*, *25*(46), 10648–10657. DOI: 10.1523/JNEUROSCI.1609-05.2005.

Traunmüller, H. (1990). Analytical expressions for the tonotopic sensory scale. *The Journal of the Acoustical Society of America*, *88*(1), 97–100. DOI: 10.1121/1.399849.

Trouvain, J., Barry, W. J., Nielsen, C., & Andersen, O. (1998). Implications of energy declination for speech synthesis. Paper presented at the 3rd ESCA/COCOSDA Workshop on Speech Synthesis. In *SSW3-1998*, 47–52.
URL: http://www.isca-speech.org/archive_open/ssw3/ssw3_047.html

Turk, A., Nakai, S., & Sugahara, M. (2006). Acoustic segment durations in prosodic research: A practical guide. In S. Sudhoff, D. Lenertova, R. Meyer, S. Pappert, P. Augurzky, I. Mleinek, N. Richter, & J. Schließer (Eds.) *Methods in empirical prosody research*, 1–28. Berlin: Walter de Gruyter. DOI: 10.1515/9783110914641.1.

Uchanski, R. M., Choi, S. S., Braida, L. D., Reed, C. M., & Durlach, N. I. (1996). Speaking clearly for the hard of hearing IV: Further studies of the role of speaking rate. *Journal of Speech and Hearing Research*, *39*(3), 494–509. URL: http://jslhr.asha.org/cgi/content/abstract/39/3/494

Van Engen, K. J. (2010). Similarity and familiarity: Second language sentence recognition in first- and second-language multi-talker babble. *Speech Communication*, *52*(11-12), 943–953. DOI: 10.1016/j.specom.2010.05.002.

Van Engen, K. J. (2012). Speech-in-speech recognition: A training study. *Language and Cognitive Processes*, *27*(7–8), 1089–1107. DOI: 10.1080/01690965.2012.654644.

Van Engen, K. J., & Bradlow, A. R. (2007). Sentence recognition in native- and foreign-language multi-talker background noise. *The Journal of the Acoustical Society of America*, *121*(1), 519–526. DOI: 10.1121/1.2400666.

Van Tasell, D. J., Soli, S. D., Kirby, V. M., & Widin, G. P. (1987). Speech waveform envelope cues for consonant recognition. *The Journal of the Acoustical Society of America*, *82*(4), 1152–1161. DOI: 10.1121/1.395251.

Vitevitch, M. S., & Luce, P. A. (1998). When words compete: Levels of processing in perception of spoken words. *Psychological Science*, *9*(4), 325–329. URL: http://www.jstor.org/stable/40063346

Vitevitch, M. S., & Rodríguez, E. (2005). Neighborhood density effects in spoken word recognition in Spanish. *Journal of Multilingual Communication Disorders*, *3*(1), 64–73. DOI: 10.1080/14769670400027332.

Ward, M. (2003). *Portland dialect study: The fronting of /ow, u, uw/ in Portland, Oregon*. Masters thesis, Portland State University, Portland, OR.

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science*, *167*(3917), 392–393. DOI: 10.1126/science.167.3917.392.

Watson, C. S. (2005). Some comments on informational masking. *Acta Acustica united with Acustica*, *91*(3), 502–512. URL: http://www.ingentaconnect.com/content/dav/aaua/2005/00000091/00000003/art00012

Watson, P. J., & Schlauch, R. S. (2008). The effect of fundamental frequency on the intelligibility of speech with flattened intonation contours. *American Journal of Speech-Language Pathology*, *17*(4), 348–355. DOI: 10.1044/1058-0360(2008/07-0048).

Wegel, R. L., & Lane, C. E. (1924). The auditory masking of one pure tone by another and its probable relation to the dynamics of the inner ear. *Physical Review*, *23*(2), 266–285.

Wright, R. A. (2001). Perceptual cues in contrast maintenance. In E. V. Hume, & K. Johnson (Eds.) *The role of speech perception in phonology*, 251–277. San Diego: Academic Press.

Wright, R. A. (2004a). Factors of lexical competition in vowel articulation. In J. Local, R. Ogden, & R. Temple (Eds.) *Phonetic interpretation: Papers in laboratory phonology IV*, 75–87. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511486425.005.

Wright, R. A. (2004b). A review of perceptual cues and cue robustness. In B. Hayes, R. Kirchner, & D. Steriade (Eds.) *Phonetically based phonology*, 34–57. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511486401.002.

Yonan, C. A., & Sommers, M. S. (2000). The effects of talker familiarity on spoken word identification in younger and older listeners. *Psychology and Aging*, *15*(1), 88–99. DOI: 10.1037/0882-7974.15.1.88.

Yoon, K. (2007). Imposing native speakers' prosody on non-native speakers' utterances: The technique of cloning prosody. 현대영미어문학회 *(The Journal of Modern British & American Language & Literature)*, *25*(4), 197–215.

Yuasa, I. P. (2008). *Culture and gender of voice pitch: A sociophonetic comparison of the Japanese and Americans*. London: Equinox.

Ziegler, J. C., Muneaux, M., & Grainger, J. (2003). Neighborhood effects in auditory word recognition: Phonological competition and orthographic facilitation. *Journal of Memory and Language*, *48*(4), 779–793. DOI: 10.1016/S0749-596X(03)00006-8.

Zielinski, B. W. (2008). The listener: No longer the silent partner in reduced intelligibility. *System*, *36*(1), 69–84. DOI: 10.1016/j.system.2007.11.004.

Appendix A. Stimulus sentences

Table A.1: IEEE "Harvard" sentences used as stimuli. Here "number" combines the "list" and "sentence" numbers from the original numeration given in Rothauser *et al.* (1969)

Number	Text
01-07	The box was thrown beside the parked truck.
02-01	The boy was there when the sun rose.
02-06	A pot of tea helps to pass the evening.
03-10	Read verse out loud for pleasure.
04-02	Take the winding path to reach the lake.
04-04	Wipe the grease off his dirty face.
04-05	Mend the coat before you go out.
04-08	The young girl gave no clear response.
05-02	The ship was torn apart on the sharp reef.
05-05	The lazy cow lay in the cool grass.
05-07	The rope will bind the seven books at once.
06-01	The frosty air passed through the coat.
06-04	The show was a flop from the very start.
06-05	A saw is a tool used for making boards.
06-09	Place a rosebush near the porch steps.
06-10	Both lost their lives in the raging storm.
07-08	This is a grand season for hikes on the road.
07-09	The dune rose from the edge of the water.
07-10	Those words were the cue for the actor to leave.

continued on next page

Table A.1: IEEE "Harvard" sentences used as stimuli *(continued from previous page)*

Number	Text
08-04	The walled town was seized without a fight.
08-07	The horn of the car woke the sleeping cop.
10-06	Bail the boat to stop it from sinking.
10-09	Ten pins were set in order.
10-10	The bill was paid every third week.
11-05	Add the sum to the product of these three.
11-07	The ripe taste of cheese improves with age.
12-03	The pennant waved when the wind blew.
12-10	We find joy in the simplest things.
13-01	Type out three lists of orders.
13-03	The boss ran the show with a watchful eye.
13-07	It caught its hind paw in a rusty trap.
14-04	Two plus seven is less than ten.
14-06	Bring your problems to the wise chief.
15-06	Pure bred poodles have curls.
15-07	The tree top waved in a graceful way.
15-09	Mud was spattered on the front of his white shirt.
16-01	The empty flask stood on the tin tray.
16-02	A speedy man can beat this track mark.
17-03	Drop the two when you add the figures.
17-05	An abrupt start does not win the prize.
17-06	Wood is best for making toys and blocks.
18-01	Steam hissed from the broken valve.

continued on next page

Table A.1: IEEE "Harvard" sentences used as stimuli *(continued from previous page)*

Number	Text
18-02	The child almost hurt the small dog.
18-04	The sky that morning was clear and bright blue.
18-05	Torn scraps littered the stone floor.
18-06	Sunday is the best part of the week.
19-02	Fairy tales should be fun to write.
19-06	Add the column and put the sum here.
19-07	We admire and love a good cook.
19-10	She has a smart way of wearing clothes.
20-04	The paper box is full of thumb tacks.
22-01	The cement had dried when he moved it.
22-02	The loss of the second ship was hard to take.
22-03	The fly made its way along the wall.
22-04	Do that with a wooden stick.
22-06	The large house had hot water taps.
22-07	It is hard to erase blue or red ink.
23-01	A pencil with black lead writes best.
23-07	The shaky barn fell with a loud crash.
23-08	Jazz and swing fans like fast music.
23-09	Rake the rubbish up and then burn it.
24-02	They are pushed back each time they attack.
24-07	Some ads serve to cheat buyers.
24-09	A waxed floor makes us lose balance.
25-01	On the islands the sea breeze is soft and mild.

continued on next page

Table A.1: IEEE "Harvard" sentences used as stimuli *(continued from previous page)*

Number	Text
25-02	The play began as soon as we sat down.
25-04	Add salt before you fry the egg.
26-09	These pills do less good than the others.
27-06	The rude laugh filled the empty room.
27-07	High seats are best for football fans.
27-09	A dash of pepper spoils beef stew.
28-01	The horse trotted around the field at a brisk pace.
28-08	The junk yard had a moldy smell.
30-02	The gold ring fits only a pierced ear.
30-04	Watch the log float in the wide river.
30-06	The heap of fallen leaves was set on fire.
31-03	The beam dropped down on the workman's head.
31-04	Pink clouds floated with the breeze.
31-10	The fight will end in just six minutes.
32-09	The purple tie was ten years old.
33-01	Fill the ink jar with sticky glue.
33-05	The crunch of feet in the snow was the only sound.
33-08	The plush chair leaned against the wall.
34-01	Nine rows of soldiers stood in line.
34-02	The beach is dry and shallow at low tide.
34-05	Pages bound in cloth make a book.
34-06	Try to trace the fine lines of the painting.
34-08	The zones merge in the central part of town.

continued on next page

Table A.1: IEEE "Harvard" sentences used as stimuli *(continued from previous page)*

Number	Text
34-10	Code is used when secrets are sent.
36-01	Pour the stew from the pot into the plate.
36-08	Our plans right now are hazy.
36-10	It takes a good trap to capture a bear.
37-01	Feed the white mouse some flower seeds.
37-05	Plead to the council to free the poor thief.
37-09	He crawled with care along the ledge.
08-02	The two men met while playing on the sand.
16-10	The sofa cushion is red and of light weight.
22-08	Write at once or you may forget it.
31-01	Slide the box into that empty space.
38-02	Mark the spot with a sign painted red.
47-01	The music played on while they talked.
56-07	A brown leather bag hung from its strap.
65-03	They sang the same tunes at each party.
39-01	He wrote down a long list of items.
39-04	Roads are paved with sticky tar.
39-10	The sun came up to light the eastern sky.
40-06	The desk was firm on the shaky floor.
40-07	It takes heat to bring out the odor.
40-09	Raise the sail and steer the ship northward.
40-10	A cone costs five cents on Mondays.
41-02	Jerk the dart from the cork target.

continued on next page

Table A.1: IEEE "Harvard" sentences used as stimuli *(continued from previous page)*

Number	Text
41-08	The sense of smell is better than that of touch.
41-09	No hardship seemed to keep him sad.
42-09	The point of the steel pen was bent and twisted.
44-10	The marsh will freeze when cold enough.
45-01	They slice the sausage thin with a knife.
45-02	The bloom of the rose lasts a few days.
45-05	Bottles hold four kinds of rum.
45-08	Drop the ashes on the worn old rug.
45-10	Throw out the used paper cup and plate.
46-02	The couch cover and hall drapes were blue.
46-05	The clothes dried on a thin wooden rack.
47-04	He sent the figs, but kept the ripe cherries.
48-01	The kite flew wildly in the high wind.
48-03	The tin box held priceless stones.
49-05	He offered proof in the form of a large chart.
49-08	They told wild tales to frighten him.
50-01	A man in a blue sweater sat at the desk.
50-06	Tuck the sheet under the edge of the mat.
51-04	The dusty bench stood by the stone wall.
51-06	Smile when you say nasty words.
52-07	The room was crowded with a wild mob.
52-10	The beetle droned in the hot June sun.
53-01	Press the pedal with your left foot.

continued on next page

Table A.1: IEEE "Harvard" sentences used as stimuli *(continued from previous page)*

Number	Text
53-03	The black trunk fell from the landing.
53-09	His wide grin earned many friends.
55-01	Those last words were a strong statement.
55-02	He wrote his name boldly at the top of the sheet.
55-04	Down that road is the way to the grain farmer.
55-07	If you mumble your speech will be lost.
56-02	Clams are small, round, soft, and tasty.
56-08	A toad and a frog are hard to tell apart.
56-09	A white silk jacket goes with any shoes.
56-10	A break in the dam almost caused a flood.
57-02	The child crawled into the dense grass.
57-06	A round hole was drilled through the thin board.
57-09	A vent near the edge brought in fresh air.
58-03	It was hidden from sight by a mass of leaves and shrubs.
58-08	The lobes of her ears were pierced to hold rings.
59-04	Drive the screw straight into the wood.
59-05	Keep the hatch tight and the watch constant.
60-03	Slide the tray across the glass top.
60-07	Dull stories make her laugh.
60-09	Get the trust fund to the bank early.
60-10	Choose between the high road and the low.
61-09	A six comes up more often than a ten.
62-06	The early phase of life moves fast.

continued on next page

Table A.1: IEEE "Harvard" sentences used as stimuli *(continued from previous page)*

Number	Text
63-05	She flaps her cape as she parades the street.
64-05	Crouch before you jump or miss the mark.
64-06	Pack the kits and don't forget the salt.
65-08	Pile the coal high in the shed corner.
65-09	A gold vase is both rare and costly.
66-01	The rarest spice comes from the far East.
66-05	The aim of the contest is to raise a great fund.
67-01	Hang tinsel from both branches.
67-05	Pick a card and slip it under the pack.
67-06	A round mat will cover the dull spot.
67-09	The mail comes in three batches per day.
67-10	You cannot brew tea in a cold pot.
68-02	Put the chart on the mantel and tack it down.
68-08	We don't like to admit our small faults.
70-02	It was a bad error on the part of the new judge.
70-04	Take the match and strike it against your shoe.
70-06	The baby puts his right foot in his mouth.
70-09	The streets are narrow and full of sharp turns.
71-05	The big red apple fell to the ground.
71-06	The curtain rose and the show was on.
71-08	He sent the boy on a short errand.
72-05	Small children came to see him.
72-06	The grass and bushes were wet with dew.

Appendix B. Praat scripts

This appendix contains the scripts used in stimulus creation for this thesis, and together represent a more-or-less complete workflow for high-quality PSOLA™ resynthesis using Praat (with the possible exception of initial RMS normalization of the original recordings, which functionality is built into Praat and for which numerous scripts are available online).

B.1 Syllabic segmentation by intensity

This script takes a directory of sound files and, for each file, creates a new TextGrid and prepopulates an interval tier with boundaries at each local minimum of the sound file's intensity contour. It then presents the user with a TextGrid editor for the opportunity to adjust boundaries, add new ones, delete spurious ones, and add notes if desired. The file name, notes (if any) and a sequential number are written to a log file. Users can stop the script at any time, and resume work on the same directory of sound files by entering a "starting file number" when re-initiating the script.

```
# COLLECT ALL THE USER INPUT
form Create syllable tier using intensity
  sentence Sound_directory ~/Desktop/SoundFiles/
  sentence Textgrid_output_directory ~/Desktop/SyllableTextgrids/
  sentence Sound_extension .wav
  sentence logFile ~/Desktop/CreateSyllableTierFromIntensity.log
  integer textgrid_tier 1
  comment Set window length:
  real Zoom_duration 0
  boolean prepopulateMinima 1
  boolean prepopulateMaxima 0
  comment You can pick up where you left off if you like:
  integer startingFileNum 1
endform

# INITIATE THE OUTPUT FILE
if fileReadable (logFile$)
  beginPause ("The log file already exists!")
    comment ("The log file already exists!")
    comment ("You can overwrite the existing file, or append new data to the end of it.")
  overwrite_setting = endPause ("Append", "Overwrite", 1)
  if overwrite_setting = 2
    filedelete 'logFile$'
    headerline$ = "number'tab$'filename'tab$'notes'newline$'"
    fileappend "'logFile$'" 'headerline$'
  endif
else
  # THERE IS NOTHING TO OVERWRITE, SO CREATE THE HEADER ROW FOR THE NEW OUTPUT FILE
  headerline$ = "number'tab$'filename'tab$'notes'newline$'"
  fileappend "'logFile$'" 'headerline$'
```

```
endif

# MAKE A LIST OF ALL SOUND FILES IN THE DIRECTORY
Create Strings as file list... soundFiles 'sound_directory$'*'sound_extension$'
fileList = selected("Strings")
fileCount = Get number of strings

# THE NEXT LINE IS A BOOLEAN FOR HANDLING EDITOR WINDOW SETTINGS
firstFile = 1
curSent$ = ""
curName$ = ""
persistentName$ = ""

# LOOP THROUGH THE LIST OF FILES...
for curFile from startingFileNum to fileCount

 # GET THE NEXT FILE
 notes$ = ""
 select Strings soundFiles
 soundfile$ = Get string... curFile
 prvSent$ = curSent$
 curSent$ = left$(soundfile$,5)

 # IF WE'VE STARTED A NEW SENTENCE, CLEAR OLD ONE
 if prvSent$ <> curSent$ and prvSent$ <> ""
  select Sound 'persistentName$'
  plus TextGrid 'persistentName$'
  plus Intensity 'persistentName$'
  Remove
 endif

 # READ IN THE SOUND
 Read from file... 'sound_directory$''soundfile$'
 totalDur = Get total duration
 prvName$ = curName$
 curName$ = selected$ ("Sound", 1)

 # CLEAR INTERMEDIATE FILES (KEEPING CURRENT & PERSISTENT ONLY)
 if prvSent$ <> curSent$
  persistentName$ = curName$
 endif
 if prvName$ <> persistentName$ and prvName$ <> ""
  select Sound 'prvName$'
  plus TextGrid 'prvName$'
  plus Intensity 'prvName$'
  Remove
 endif

 # CREATE TEXTGRID
 select Sound 'curName$'
 To Intensity... 80 0 yes
 timeStep = Get time step
 numFrames = Get number of frames
 select Sound 'curName$'
```

```
To TextGrid... intensyl
Insert boundary... 1 0.05
Insert boundary... 1 totalDur-0.05

# PREPOPULATE TEXTGRID WITH INTENSITY MINIMA / MAXIMA
if prepopulateMinima = 1 or prepopulateMaxima = 1
 for fr from 2 to numFrames-1
  select Intensity 'curName$'
  a = Get value in frame... fr-1
  b = Get value in frame... fr
  c = Get value in frame... fr+1
  localExtremum = -1
  if prepopulateMaxima = 1 and b > a and b > c
   localExtremum = Get time from frame number... fr
  elif prepopulateMinima = 1 and b < a and b < c
   localExtremum = Get time from frame number... fr
  endif
  if localExtremum > 0
   select Sound 'curName$'
   localZero = Get nearest zero crossing... 1 'localExtremum'
   if abs(localExtremum - localZero) < 0.01
    localExtremum = localZero
   endif
   select TextGrid 'curName$'
   Insert boundary... 1 'localExtremum'
  endif
 endfor
endif

# OPEN THE FILES IN THE EDITOR AND ZOOM IN
echo 'curFile'
select Sound 'curName$'
plus TextGrid 'curName$'
View & Edit
editor TextGrid 'curName$'
 if firstFile = 1
  # SHOW ONLY SOUND FILE AND INTENSITY
  Show analyses... no no yes no no 5
  firstFile = 0
 endif
 if zoom_duration = 0
  Show all
 else
  Zoom... 0 zoom_duration
 endif

 # START US OFF IN INTERVAL #2 (THE FIRST ONE AFTER THE INITIAL 50ms SILENCE PADDING)
 Move cursor to... 0.04
 Select next interval

 # SHOW A U.I. FOR FINDING LOCAL INTENSITY MAXIMA AND MINIMA
 repeat
  beginPause ("Correct boundaries")
   comment ("Add/del boundaries. Put cursor near extremum before")
```

```
    comment ("using FindMin/FindMax. When done, click Save.")
    sentence ("Notes", notes$)
   clicked = endPause ("FindMin", "FindMax", "Save", 3)
   pt = Get cursor
   if clicked < 3
    endeditor
    select Intensity 'curName$'
    if clicked = 1
     localExtremum = Get time of minimum... 'pt'-0.05 'pt'+0.05 Parabolic
    elif clicked = 2
     localExtremum = Get time of maximum... 'pt'-0.05 'pt'+0.05 Parabolic
    endif
    editor TextGrid 'curName$'
     Move cursor to... 'localExtremum'
     Move cursor to nearest zero crossing
     localZero = Get cursor
     if abs(localExtremum-localZero) < 0.01
      localExtremum = localZero
     endif
    endeditor
    select TextGrid 'curName$'
    Insert boundary... 1 'localExtremum'
    editor TextGrid 'curName$'
   endif
  until clicked = 3

  if clicked = 3
   Save TextGrid as text file... 'textgrid_output_directory$''curName$'.TextGrid
  endif
 endeditor

 # WRITE TO FILE
 resultline$ = "'curFile''tab$''soundfile$''tab$''notes$''newline$'"
 fileappend "'logFile$'" 'resultline$'

 # GO ON TO NEXT FILE...
 select Strings soundFiles

endfor

# REMOVE THE STRINGS LIST
select Strings soundFiles
# plus Table 'talkList$'
Remove
clearinfo
files_read = fileCount - startingFileNum + 1
printline Done! 'files_read' files read.'newline$'
```

Script B.1: Praat script for semi-automated syllable-level segmentation by intensity

B.2 Semi-auto pulse correction

This script facilitates semi-automatic creation of manipulation objects from .wav files. It takes a directory of sound files and, for each file, displays the pitch contour over a narrowband spectrogram, and prompts the user to either: (a) accept the pitch settings, (b) adjust the pitch floor/ceiling and redraw, or (c) mark the file as unmeasurable, before continuing on to the next file. An advancedInterface option is available for users who want full control over all pitch parameters during the process. Filename, duration, and pitch settings are saved to a tab-delimited log file. The option outputType allows users to (a) continue to next file after finalizing pitch settings, (b) silently create and save manipulation objects using final pitch settings before continuing, or (c) create manipulation objects and open them for hand-correction before continuing to the next file.

```
# COLLECT ALL THE USER INPUT
form Pitch settings tool: Select directories & starting parameters
  sentence Sound_directory ~/Desktop/SoundFiles/
  sentence Pitch_directory ~/Desktop/PitchObjects/
  sentence Manip_directory ~/Desktop/ManipObjects/
  sentence logFile ~/Desktop/SoundToManipulation.log
  optionmenu outputType: 3
    option only log pitch settings
    option save manipulation objects
    option edit manip. objs. before save
  integer startingFileNum 1
  real defaultZoomDuration 0
  integer viewMin 1
  integer viewMax 300
  integer defMinPitch 50 (=PSOLA minimum)
  integer defMaxPitch 300
  boolean advancedInterface 0
  boolean carryover 0
  optionmenu autoplay: 1
    option off
    option on first view
    option on each redraw
  real defTimeStep 0 (=auto)
  integer defMaxCand 15
  boolean defHighAcc 1
  real defSilThresh 0.03
  real defVoiThresh 0.45
  real defOctCost 0.01
  real defJumpCost 0.5
  real defVoiCost 0.14
endform

# INITIATE THE OUTPUT FILE
procedure initializeOutfile
  headerline$ = "Number'tab$'Filename'tab$'Duration'tab$'PitchFloor'tab$'PitchCeiling'tab$'TimeStep'tab$'
  ...MaxCandidates'tab$'HighAccuracy'tab$'SilenceThreshold'tab$'VoicingThreshold'tab$'OctaveCost'tab$'
  ...JumpCost'tab$'VoicingCost'tab$'Notes'newline$'"
  fileappend "'logFile$'" 'headerline$'
endproc
```

```
if fileReadable (logFile$)
 beginPause ("The log file already exists!")
  comment ("The log file already exists!")
  comment ("You can overwrite the existing file, or append new data to the end of it.")
 overwrite_setting = endPause ("Append", "Overwrite", 1)
 if overwrite_setting = 2
  filedelete 'logFile$'
  call initializeOutfile
 endif
else
 # THERE IS NOTHING TO OVERWRITE, SO CREATE THE HEADER ROW FOR THE NEW OUTPUT FILE
 call initializeOutfile
endif

# MAKE A LIST OF ALL SOUND FILES IN THE FOLDER
Create Strings as file list... list 'sound_directory$'*.wav
fileList = selected("Strings")
fileCount = Get number of strings

# INITIALIZE SOME VARIABLES
minPitch = defMinPitch
maxPitch = defMaxPitch
timeStep = defTimeStep
maxCand = defMaxCand
highAcc = defHighAcc
silThresh = defSilThresh
voiThresh = defVoiThresh
octCost = defOctCost
jumpCost = defJumpCost
voiCost = defVoiCost

# LOOP THROUGH THE LIST OF FILES...
for curFile from startingFileNum to fileCount

 # READ IN THE SOUND...
 select Strings list
 soundname$ = Get string... curFile
 basename$ = mid$(soundname$,7,11)
 Read from file... 'sound_directory$''soundname$'
 Rename... 'basename$'
 filename$ = selected$ ("Sound", 1)
 totalDur = Get total duration

 # SHOW THE EDITOR WINDOW
 zoomStart = 0
 if defaultZoomDuration > 0
  zoomEnd = defaultZoomDuration
 else
  zoomEnd = totalDur
 endif
 select Sound 'filename$'
 View & Edit
 editor Sound 'filename$'
```

```
 # HIDE THE SPECTROGRAM & ANALYSES TO PREVENT ANNOYING FLICKERING
 Show analyses... no no no no no 10
 Zoom... zoomStart zoomEnd

 # SET ALL THE RELEVANT SETTINGS
 Spectrogram settings... 0 2500 0.025 50
 Advanced spectrogram settings... 1000 250 Fourier Gaussian yes 100 6 0
 if carryover = 0
  Pitch settings... defMinPitch defMaxPitch Hertz cross-correlation speckles
  Advanced pitch settings... viewMin viewMax defHighAcc defMaxCand defSilThresh defVoiThresh defOctCost
  ... defJumpCost defVoiCost
 else
  Pitch settings... minPitch maxPitch Hertz cross-correlation automatic
  Advanced pitch settings... viewMin viewMax highAcc maxCand silThresh voiThresh octCost jumpCost voiCost
 endif

 # DISPLAY NARROWBAND SPECTROGRAM AND PITCH (MAKING SURE "MAX ANALYSIS" IS LONG ENOUGH
 # SO THAT THE SPECTROGRAM ACTUALLY SHOWS UP)
 Show analyses... yes yes no no no zoomEnd-zoomStart+1
endeditor

# INITIALIZE SOME VARIABLES FOR THE PAUSE U.I.
clicked = 1
notes$ = ""
if autoplay > 1
 firstview = 1
else
 firstview = 0
endif
# THE maxPitch=0 CONDITION PREVENTS ERRORS WHEN carryover=1 AND THE PREV. FILE WAS SKIPPED
if carryover = 0 or maxPitch = 0
 minPitch = defMinPitch
 maxPitch = defMaxPitch
 timeStep = defTimeStep
 maxCand = defMaxCand
 highAcc = defHighAcc
 silThresh = defSilThresh
 voiThresh = defVoiThresh
 octCost = defOctCost
 voiCost = defVoiCost
 jumpCost = defJumpCost
endif

# SHOW A U.I. WITH PITCH SETTINGS.  KEEP SHOWING IT UNTIL THE USER ACCEPTS OR CANCELS
repeat
 beginPause ("Adjust pitch analysis settings")
  comment ("File 'filename$' (file number 'curFile' of 'fileCount')")
  comment ("You can change the pitch settings if the pitch track doesn't look right.")
  integer ("newMinPitch", minPitch)
  integer ("newMaxPitch", maxPitch)
  if advancedInterface = 1
   real ("newTimeStep", timeStep)
   integer ("newMaxCandidates", maxCand)
```

```
    boolean ("newHighAccuracy", highAcc)
    real ("newSilenceThreshold", silThresh)
    real ("newVoicingThreshold", voiThresh)
    real ("newOctaveCost", octCost)
    real ("newVoicingCost", voiCost)
    real ("newJumpCost", jumpCost)
  endif
  comment ("clicking RESET will reset parameters to the default values and redraw;")
  comment ("Clicking REDRAW will redraw the pitch contour with the settings above;")
  comment ("clicking SKIP will write zeros to the log file and go to next file.")
  sentence ("Notes", notes$)
  # AUTOPLAY
  if autoplay = 3 or firstview = 1
    editor Sound 'filename$'
      Play... 0 totalDur
    endeditor
  endif
  firstview = 0
clicked = endPause ("Play","Reset", "Redraw", "Accept", "Skip", 3)

# STILL NEED TO PASS ALONG THE ADVANCED SETTINGS, EVEN IF USER DIDN'T CHANGE THEM
if advancedInterface = 0
  newTimeStep = timeStep
  newMaxCandidates = maxCand
  newHighAccuracy = highAcc
  newSilenceThreshold = silThresh
  newVoicingThreshold = voiThresh
  newOctaveCost = octCost
  newVoicingCost = voiCost
  newJumpCost = jumpCost
endif

# IF THE USER CLICKS "PLAY"
if clicked = 1
  editor Sound 'filename$'
    Play... 0 totalDur
  endeditor

# IF THE USER CLICKS "RESET"
elif clicked = 2
  minPitch = defMinPitch
  maxPitch = defMaxPitch
  timeStep = defTimeStep
  maxCand = defMaxCand
  highAcc = defHighAcc
  silThresh = defSilThresh
  voiThresh = defVoiThresh
  octCost = defOctCost
  voiCost = defVoiCost
  jumpCost = defJumpCost
  # REDRAW THE PITCH CONTOUR
  editor Sound 'filename$'
    Pitch settings... minPitch maxPitch Hertz cross-correlation speckles
    Advanced pitch settings... viewMin viewMax highAcc maxCand silThresh voiThresh octCost jumpCost voiCost
```

```
    endeditor

  # IF THE USER CLICKS "REDRAW"
  elif clicked = 3
    minPitch = newMinPitch
    maxPitch = newMaxPitch
    timeStep = newTimeStep
    maxCand = newMaxCandidates
    highAcc = newHighAccuracy
    silThresh = newSilenceThreshold
    voiThresh = newVoicingThreshold
    octCost = newOctaveCost
    voiCost = newVoicingCost
    jumpCost = newJumpCost
    # REDRAW THE PITCH CONTOUR
    editor Sound 'filename$'
      Pitch settings... minPitch maxPitch Hertz cross-correlation speckles
      Advanced pitch settings... viewMin viewMax highAcc maxCand silThresh voiThresh octCost jumpCost voiCost
    endeditor
  endif
until clicked >3

# IF THE USER SKIPS, WRITE OVERRIDE VALUES
if clicked = 5
  minPitch = 0
  maxPitch = 0
  timeStep = 0
  maxCand = 0
  highAcc = 0
  silThresh = 0
  voiThresh = 0
  octCost = 0
  voiCost = 0
  jumpCost = 0
endif

select Sound 'filename$'
# IF WE'RE DOING MORE THAN JUST WRITING SETTINGS LOG
if outputType > 0
  To Pitch (cc)... timeStep minPitch 15 highAcc silThresh voiThresh octCost jumpCost voiCost maxPitch
  Save as text file... 'pitch_directory$''filename$'.Pitch
  plus Sound 'filename$'
  To Manipulation
  # CREATE BACKUP OF PULSES FOR "RESET" BUTTON
  Extract pulses
  Rename... ppBackup
  select Manipulation 'filename$'

  # IF WE'RE EDITING THE MANIPULATION OBJECT BEFORE SAVING
  if outputType = 3
    editor Sound 'filename$'
    Close
    select Manipulation 'filename$'
    View & Edit
```

```
  # UI FOR PULSE CORRECTION (OR PITCH/DURATION TIER IF DESIRED...)
  repeat
   beginPause ("Correct pulses")
    comment ("File 'filename$' (file number 'curFile' of 'fileCount')")
   clicked = endPause ("Play","Reset", "Redraw", "Finished", 3)

   # IF THE USER CLICKS "PLAY"
   if clicked = 1
    select Sound 'filename$'
    Play

   # IF THE USER CLICKS "RESET"
   elif clicked=2
    # undo all the changes to the pulses and pitch tier
    select Manipulation 'filename$'
    plus PointProcess ppBackup
    Replace pulses
    minus Manipulation 'filename$'
    To PitchTier... 1/minPitch
    select Manipulation 'filename$'
    plus PitchTier ppBackup
    Replace pitch tier
    select PitchTier ppBackup
    Remove
    select Manipulation 'filename$'

   # IF THE USER CLICKS "REDRAW"
   elif clicked=3
    # redraw the PitchTier from the new PointProcess
    select Manipulation 'filename$'
    Extract pulses
    To PitchTier... 1/minPitch
    # previous line: 20ms is the longest gap allowed by PSOLA algorithm in treating a span as voiced, so the
    # lowest pitch it will track is 50Hz. Might want to hard-code the value of 0.02 instead of 1/minPitch.
    select Manipulation 'filename$'
    plus PitchTier 'filename$'
    Replace pitch tier
    select PointProcess 'filename$'
    plus PitchTier 'filename$'
    Remove
    select Manipulation 'filename$'
   endif

  until clicked>3
  select PointProcess ppBackup
  Remove
 endif

 # SAVE FILES
 select Manipulation 'filename$'
 Save as binary file... 'manip_directory$''filename$'.Manipulation
 plus Pitch 'filename$'
 plus Sound 'filename$'
endif
```

```
 Remove

 # WRITE TO LOG FILE
 resultline$ = "'curFile''tab$''filename$''tab$''totalDur''tab$''minPitch''tab$''maxPitch''tab$''timeStep''tab$'
 ...'maxCand''tab$''highAcc''tab$''silThresh''tab$''voiThresh''tab$''octCost''tab$''jumpCost''tab$''voiCost'
 ...'tab$''notes$''newline$'"
 fileappend "'logFile$'" 'resultline$'
endfor

# REMOVE THE STRINGS LIST AND GIVE A SUCCESS MESSAGE
select Strings list
Remove
clearinfo
files_read = fileCount - startingFileNum + 1
printline Done! 'files_read' files read.'newline$'
```

Script B.2: Praat script for semi-automated correction of glottal pulses within a manipulation object.

B.3 Prosody replacement with PSOLA™

This script takes as input two manipulation objects and two TextGrids and maps the pitch, duration, and intensity patterns from one manipulation object onto the other. The manipulation objects must have the waveform embedded, but may be either text or binary.

```
# COLLECT ALL THE USER INPUT
form Neutralize Prosody: Select directories & starting parameters
 sentence Segmental_donor ~/Desktop/ManipulationObjects/NWM02_01-07.Manipulation
 sentence Seg_donor_textgrid ~/Desktop/TextGrids/NWM02_01-07.TextGrid
 integer Seg_donor_tier 1
 sentence Prosodic_donor ~/Desktop/ManipulationObjects/NWM07_01-07.Manipulation
 sentence Pro_donor_textgrid ~/Desktop/TextGrids/NWM07_01-07.TextGrid
 integer Pro_donor_tier 1
 sentence Output_directory ~/Desktop/ResynthesizedFiles/
endform

# INITIALIZE GLOBAL VARIABLE (used in the duration tier to prevent points from coinciding)
offset = 0.00000000001

# READ IN ALL THE FILES
segManip = Read from file... 'segmental_donor$'
segTextGrid = Read from file... 'Seg_donor_textgrid$'
proManip = Read from file... 'prosodic_donor$'
proTextGrid = Read from file... 'pro_donor_textgrid$'

# MAKE SURE THEY HAVE THE SAME NUMBER OF INTERVALS
select segTextGrid
segInt = Get number of intervals... seg_donor_tier
select proTextGrid
proInt = Get number of intervals... pro_donor_tier
if segInt <> proInt
 exit The two TextGrids do not have the same number of intervals.
endif

# EXTRACT PITCH TIERS AND SOME PITCH INFO TO BE USED LATER
select segManip
segPitch = Extract pitch tier
segPitchMean = Get mean (curve)... 0 0
segPitchPts = Get number of points
segPitchTable = Down to TableOfReal... Hertz
Insert column (index)... 3

select proManip
proPitch = Extract pitch tier
proPitchMean = Get mean (curve)... 0 0
proPitchPts = Get number of points
proPitchTable = Down to TableOfReal... Hertz
Insert column (index)... 3

pitchDiff = segPitchMean - proPitchMean

# EXTRACT INTENSITY TIERS AND SOME INTENSITY INFO TO BE USED LATER
```

```
select segManip
segSound = Extract original sound
segIntensityRMS = Get intensity (dB)
segIntensity = To Intensity... 60 0 yes
segIntensityMax = Get maximum... 0 0 Parabolic
segIntensityTier = Down to IntensityTier
segIntensityTable = Down to TableOfReal
Insert column (index)... 3

select proManip
proSound = Extract original sound
proIntensityRMS = Get intensity (dB)
proIntensity = To Intensity... 60 0 yes
proIntensityMax = Get maximum... 0 0 Parabolic
proIntensityTier = Down to IntensityTier
proIntensityTable = Down to TableOfReal
Insert column (index)... 3

# EXTRACT (EMPTY) DURATION TIERS
select segManip
segDurationTier = Extract duration tier
select proManip
proDurationTier = Extract duration tier

# STEP THROUGH EACH INTERVAL IN THE TEXTGRIDS
for intNum to segInt

 # GET DURATION OF INTERVALS
 select segTextGrid
 segIntStart = Get start point... seg_donor_tier intNum
 segIntEnd = Get end point... seg_donor_tier intNum
 segIntDur = segIntEnd - segIntStart

 select proTextGrid
 proIntStart = Get start point... pro_donor_tier intNum
 proIntEnd = Get end point... pro_donor_tier intNum
 proIntDur = proIntEnd - proIntStart

 proSegRatio = proIntDur / segIntDur
 segProRatio = segIntDur / proIntDur

 # CREATE DURATION TIER POINTS FOR CURRENT INTERVAL IN TARGET OBJECT
 select segDurationTier
 Add point... segIntStart+offset proSegRatio
 Add point... segIntEnd proSegRatio

 # DO THE SAME FOR THE PROSODY DONOR IN CASE SWAP=TRUE, OR IF REPLACING PITCH W/O DURATION, ETC
 select proDurationTier
 Add point... proIntStart+offset segProRatio
 Add point... proIntEnd segProRatio

 # WARP TIME DOMAIN OF PITCH AND INTENSITY VALUES AND STORE IN (PREVIOUSLY EMPTY) COLUMN 3 OF THE TABLES.
 select segPitchTable
  Formula... if col = 3 and self[row,1] > segIntStart and self[row,1] <= segIntEnd
```

```
  ... then proIntStart + (self[row,1] - segIntStart) * proSegRatio else self fi
 select proPitchTable
  Formula... if col = 3 and self[row,1] > proIntStart and self[row,1] <= proIntEnd
  ... then segIntStart + (self[row,1] - proIntStart) * segProRatio else self fi
 select segIntensityTable
  Formula... if col = 3 and self[row,1] > segIntStart and self[row,1] <= segIntEnd
  ... then proIntStart + (self[row,1] - segIntStart) * proSegRatio else self fi
 select proIntensityTable
  Formula... if col = 3 and self[row,1] > proIntStart and self[row,1] <= proIntEnd
  ... then segIntStart + (self[row,1] - proIntStart) * segProRatio else self fi

# DONE STEPPING THROUGH EACH INTERVAL OF THE TEXTGRIDS
endfor

# CREATE NEW PITCH AND INTENSITY TIERS WITH WARPED TIME DOMAINS
select segSound
segDur = Get total duration
segProPitchWarped = Create PitchTier... proPitchWarped 0 'segDur'
segProIntensityWarped = Create IntensityTier... proIntensityWarped 0 'segDur'

select proPitchTable
proPitchRows = Get number of rows
for r to proPitchRows
 select proPitchTable
 t = Get value... r 3
 v = Get value... r 2
 select segProPitchWarped
 Add point... t v
endfor
Shift frequencies... 0 'segDur' 'pitchDiff' Hertz

select proIntensityTable
proIntensityRows = Get number of rows
for r to proIntensityRows
 select proIntensityTable
 t = Get value... r 3
 v = Get value... r 2
 select segProIntensityWarped
 Add point... t v
endfor

# MULTIPLY TARGET SOUND BY ITS INTENSITY INVERSE, THEN BY THE TARGET INTENSITY
select segIntensity
Formula... 'segIntensityMax' - self
segIntensityInverse = Down to IntensityTier

select segSound
plus segIntensityInverse
segSoundInverse = Multiply... yes

select segSoundInverse
plus segProIntensityWarped
segSoundProIntensity = Multiply... yes
Scale intensity... proIntensityRMS
```

```
# ASSEMBLE FINAL MANIPULATION OBJECT
select segManip
plus segSoundProIntensity
Replace original sound

select segManip
plus segProPitchWarped
Replace pitch tier

select segManip
plus segDurationTier
Replace duration tier

select segManip
Save as binary file... 'output_directory$''segDonorFilename$'_'proDonorFilename$'.Manipulation
segProResynth = Get resynthesis (overlap-add)
Save as WAV file... 'output_directory$''segDonorFilename$'_'proDonorFilename$'.wav

# CLEAN UP
select segManip
plus segTextGrid
plus segPitch
plus segPitchTable
plus segSound
plus segIntensity
plus segIntensityTier
plus segIntensityTable
plus segDurationTier
plus segProPitchWarped
plus segProIntensityWarped
plus segIntensityInverse
plus segSoundInverse
plus segSoundProIntensity
plus segProResynth
Remove

select proManip
plus proTextGrid
plus proPitch
plus proPitchTable
plus proSound
plus proIntensity
plus proIntensityTier
plus proIntensityTable
plus proDurationTier
Remove
```

Script B.3: Praat script for prosodic replacement using PSOLA™.

B.4 Create noise spectrally shaped to the LTAS of the corpus

This script takes a directory of sound files and creates a Gaussian noise file that is spectrally shaped to match the long-term average spectrum (LTAS) of the stimuli. The noise file is created to match the duration of the longest stimulus (plus any noise padding specified in the arguments to the script), and scaled to match the average intensity of the stimuli. Two methods of spectral averaging are available: either (a) calculating the LTAS of each file and averaging them, or (b) concatenating the stimuli and breaking into equal-sized chunks and averaging the spectra of the chunks. The methods are expected to differ substantially only when the stimuli vary dramatically in length (in which case method a, by treating all file-level LTAS objects equally, effectively weights the final spectrum in favor of shorter files). The script also saves the LTAS object into the output directory (along with the noise file).

```
# COLLECT ALL THE USER INPUT
form Calculate LTAS of corpus
  sentence Input_directory ~/Desktop/SoundFiles/
  sentence Output_directory ~/Desktop/NoiseFiles/
  positive ltasBandwidth_(Hz) 100
  positive noisePad_(seconds) 0.05
  optionmenu method: 2
    option by file
    option by chunk
  positive Chunk_duration_(seconds) 30
  comment Chunk duration is ignored if method is "by file".
endform

# READ IN LIST OF FILES
Create Strings as file list... stimuli 'input_directory$'*.wav
n = Get number of strings
intensityRunningTotal = 0
longestFile = 0
echo 'n' WAV files in directory 'input_directory$'

# OPEN ALL SOUND FILES
for i from 1 to n
  select Strings stimuli
  curFile$ = Get string... 'i'
  tempSound = Read from file... 'input_directory$''curFile$'

  # KEEP TRACK OF INTENSITIES SO WE CAN SCALE NOISE APPROPRIATELY
  intens = Get intensity (dB)
  intensityRunningTotal = intensityRunningTotal + intens

  # KEEP TRACK OF DURATIONS SO THE NOISE IS LONG ENOUGH FOR THE LONGEST STIMULUS
  tempDur = Get total duration
  if longestFile < tempDur
    longestFile = tempDur
  endif

  if method = 1
    if i = 1
      printline Creating LTAS objects...
```

```
    endif
    # CREATE LTAS FOR EACH FILE AS IT'S OPENED, AND IMMEDIATELY CLOSE SOUND FILE
    ltas_'i' = To Ltas... ltasBandwidth
    select tempSound
    Remove
  else
    # RE-OPEN EACH FILE AS LONGSOUND, TO BE CONCATENATED AND CHUNKED LATER
    Remove
    snd_'i' = Open long sound file... 'input_directory$''curFile$'
  endif
endfor

if method = 1
  # SELECT FILEWISE LTAS OBJECTS AND AVERAGE
  printline Averaging LTAS objects...
  select ltas_1
  for i from 2 to n
    plus ltas_'i'
  endfor
  finalLTAS = Average
  Save as binary file... 'output_directory$'CorpusFilewise.Ltas
  select ltas_1
  for i from 2 to n
    plus ltas_'i'
  endfor
  Remove
else
  # CONCATENATE
  printline Concatenating corpus...
  select snd_1
  for i from 2 to n
    plus snd_'i'
  endfor
  Save as WAV file... 'output_directory$'ConcatenatedCorpus.wav
  Remove

  # SPLIT INTO EQUAL-LENGTH CHUNKS
  printline Chunking corpus...
  corpus = Open long sound file... 'output_directory$'ConcatenatedCorpus.wav
  corpusDur = Get total duration
  chunkCount = ceiling(corpusDur/chunk_duration)

  # CREATE LTAS FOR EACH CHUNK
  printline Creating LTAS objects...
  for i from 1 to chunkCount
    select corpus
    tempSound = Extract part... chunk_duration*(i-1) chunk_duration*i no
    ltas_'i' = To Ltas... ltasBandwidth
    select tempSound
    Remove
  endfor

  # CREATE FINAL LTAS
  printline Averaging LTAS objects...
```

```
 select ltas_1
 for i from 2 to chunkCount
  plus ltas_'i'
 endfor
 finalLTAS = Average
 Save as binary file... 'output_directory$'CorpusChunkwise.Ltas

 # CLEAN UP INTERIM FILES
 select corpus
 for i from 1 to chunkCount
  plus ltas_'i'
 endfor
 Remove
 filedelete 'output_directory$'ConcatenatedCorpus.wav
endif

# CREATE WHITE NOISE SPECTRUM
printline Creating speech-shaped noise...
noise = Create Sound from formula... noise 1 0 longestFile+2*noisePad 44100 randomGauss(0,0.1)
noiseSpect = To Spectrum... no
Formula... self*10^(Ltas_averaged(x)/20)
ltasNoise = To Sound

# SCALE TO AVERAGE INTENSITY OF INPUT FILES
meanIntensity = intensityRunningTotal / n
Scale intensity... meanIntensity
Save as WAV file... 'output_directory$'SpeechShapedNoise.wav

# CLEAN UP
select noise
plus noiseSpect
plus Strings stimuli
plus finalLTAS
plus ltasNoise
Remove
printline Done!
```

Script B.4: Praat script for creating speech-shaped noise.

B.5 Mix signal and noise

This script takes a noise file and a directory of sound files, and mixes the stimuli with the noise at a specified SNR, writing the files to the specified directory. The mixed signal-plus-noise files can optionally be scaled to match the original intensity of the stimulus files.

```
# COLLECT ALL THE USER INPUT
form Mix speech with noise
 sentence InputFolder ~/Desktop/SoundFiles/
 sentence NoiseFile ~/Desktop/NoiseFiles/SpeechShapedNoise.wav
 sentence OutputFolder ~/Desktop/StimuliWithNoise/
 real DesiredSNR_(dB) 0
 optionmenu finalIntensity: 1
  option match final intensity to stimulus intensity
  option maximize (scale peaks to plus/minus 1)
  option just add noise to signal (do not scale result)
endform

# NOISE
noise = Read from file... 'noiseFile$'
noiseDur = Get total duration
noiseRMS = Get root-mean-square... 0 0

# STIMULI
Create Strings as file list... stimuli 'inputFolder$'*.wav
n = Get number of strings
echo 'n' WAV files in folder 'inputFolder$'

for i from 1 to n
 # READ IN EACH STIMULUS
 select Strings stimuli
 curFile$ = Get string... 'i'
 curSound = Read from file... 'inputFolder$''curFile$'
 curDur = Get total duration
 curRMS = Get root-mean-square... 0 0
 curInten = Get intensity (dB)

 # MAKE SURE NOISE IS LONG ENOUGH. IF NOT, DOUBLE LENGTH UNTIL IT IS.
 while curDur > noiseDur
  select noise
  temp = Concatenate
  plus noise
  noise = Concatenate
  select temp
  Remove
 endwhile

 # CALCULATE NOISE COEFFICIENT THAT YIELDS DESIRED SNR
 # SNR = 20*log10(SignalAmpl/NoiseAmpl)
 # NoiseAmpl = SignalAmpl/(10^(SNR/20))
 noiseAdjustCoef = (curRMS / (10 ^ (desiredSNR / 20))) / noiseRMS

 # MIX SIGNAL AND NOISE AT SPECIFIED SNR
```

```
 select curSound
 Formula... self[col] + noiseAdjustCoef * object[noise,col]

 if finalIntensity = 1
  # SCALE TO MATCH STIMULUS
  Scale intensity... curInten
 else if finalIntensity = 2
  # SCALE TO +/- 1
  Scale peak... 0.99
 endif

 # WRITE OUT FINAL FILE
 select curSound
 Save as WAV file... 'outputFolder$''curFile$'
 Remove
endfor

# CLEAN UP
select noise
plus Strings stimuli
Remove
printline Done!
```

Script B.5: Praat script for mixing speech and noise at a specified SNR.

B.6 Ramp edges of stimuli

This script takes a directory of sound files and applies linear onset and offset ramps to the beginning and end of the file, respectively. The duration of the ramps is specified in the arguments to the script.

```
# COLLECT ALL THE USER INPUT
form Ramp edges of sound files
 sentence InputFolder ~/Desktop/StimuliWithNoise/
 sentence OutputFolder ~/Desktop/StimuliWithNoise/Ramped/
 real RampDuration_(seconds) 0.05
endform

# READ IN LIST OF FILES
Create Strings as file list... stimuli 'inputFolder$'*.wav
n = Get number of strings
echo 'n' WAV files in directory 'inputFolder$'

for i from 1 to n
 # READ IN EACH STIMULUS
 select Strings stimuli
 curFile$ = Get string... 'i'
 curSound = Read from file... 'inputFolder$''curFile$'
 printline Processing file 'i' of 'n'
 curDur = Get total duration
 offset = curDur - rampDuration

 # APPLY RAMPS
 Formula (part)... 0 rampDuration 1 2 self*(1-(rampDuration - x)/(rampDuration))
 Formula (part)... offset curDur 1 2 self*(curDur - x)/(curDur - offset)

 # WRITE OUT FINAL FILE
 select curSound
 Save as WAV file... 'outputFolder$''curFile$'
 Remove
endfor

# CLEAN UP
select Strings stimuli
Remove
printline Done!
```

Script B.6: Praat script for applying linear ramps to the beginning and end of sound files.

B.7 Vacuous resynthesis

This script takes a directory of manipulation objects and monotonizes the pitch using PSOLA™, then reapplies the original pitch contour (also using PSOLA™). The purpose of this is to intentionally introduce processing artifacts into what would otherwise be clean, unmanipulated recordings, to provide a more equivalent comparison to prosody-swapped stimuli. Note that this script was developed for, but not used in, the experiments described in this thesis. The reason it was not used is that the amount of distortion introduced by this method of vacuous resynthesis was negligible, most likely due to the high-quality pitch tracks derived from hand-corrected pulse epoch marks. Thus it was judged that treating the vacuously resynthesized stimuli to be as equally distorted as the stimuli resynthesized with prosodic replacement was inadvisable, and a choice was made to model the difference statistically instead.

```
form Vacuous manipulation
  sentence InputFolder ~/Desktop/ManipulationFiles/
  sentence OutputFolder ~/Desktop/VacuouslyManipulatedSoundFiles/
endform

# STIMULI
Create Strings as file list... manipFiles 'inputFolder$'*.Manipulation
n = Get number of strings
echo 'n' Manipulation files in folder 'inputFolder$'

for i from 1 to n
  # READ IN EACH STIMULUS
  select Strings manipFiles
  curFile$ = Get string... 'i'
  curManip = Read from file... 'inputFolder$''curFile$'

  # EXTRACT ORIGINAL PITCH
  curPitch = Extract pitch tier
  meanPitch = Get mean (curve)... 0 0

  # CREATE FLATTENED PITCH
  select curManip
  monoPitch = Extract pitch tier
  Formula... meanPitch
  plus curManip
  Replace pitch tier
  minus monoPitch
  monoSound = Get resynthesis (overlap-add)

  # RECREATE ORIGINAL PITCH FROM MONOTONIZED SOUND
  monoManip = To Manipulation... 0.01 50 300
  plus curPitch
  Replace pitch tier
  minus curPitch
  finalSound = Get resynthesis (overlap-add)
  newFileName$ = replace$("'curFile$'", ".Manipulation", ".wav", 0)
  Save as WAV file... 'outputFolder$''newFileName$'

  # CLEAN UP
```

```
  select curManip
  plus curPitch
  plus monoPitch
  plus monoManip
  plus monoSound
  plus finalSound
  Remove
endfor

select Strings manipFiles
Remove
printline Done!
```

Script B.7: Praat script for introducing processing artifacts without altering prosody, via monotonization and demonotonization.