Assessing the importance of several acoustic properties to the perception of spontaneous speech

Ryan G. Podlubny, Terrance M. Nearey, Grzegorz Kondrak, and Benjamin V. Tucker

Citation: The Journal of the Acoustical Society of America **143**, 2255 (2018); doi: 10.1121/1.5031123
View online: https://doi.org/10.1121/1.5031123
View Table of Contents: http://asa.scitation.org/toc/jas/143/4
Published by the Acoustical Society of America

Assessing the importance of several acoustic properties to the perception of spontaneous speech

Ryan G. Podlubny,[1,a)] Terrance M. Nearey,[2] Grzegorz Kondrak,[3] and Benjamin V. Tucker[2]
[1]*New Zealand Institute of Language, Brain and Behaviour, University of Canterbury, 201 Locke Building, Christchurch, Canterbury, 8140, New Zealand*
[2]*Department of Linguistics, University of Alberta, 4-32 Assiniboia Hall, Edmonton, Alberta, T6G 2E7, Canada*
[3]*Department of Computing Science, University of Alberta, 2-21 Athabasca Hall, Edmonton, Alberta, T6G 2E8, Canada*

(Received 17 October 2017; revised 28 March 2018; accepted 28 March 2018; published online 20 April 2018)

Spoken language manifests itself as change over time in various acoustic dimensions. While it seems clear that acoustic-phonetic information in the speech signal is key to language processing, little is currently known about which specific types of acoustic information are relatively more informative to listeners. This problem is likely compounded when considering *reduced* speech: Which specific acoustic information do listeners rely on when encountering spoken forms that are highly variable, and often include altered or elided segments? This work explores contributions of spectral shape, *f*0 contour, target duration, and time varying intensity in the perception of reduced speech. This work extends previous laboratory-speech based perception studies into the realm of casual speech, and also provides support for use of an algorithm that quantifies phonetic reduction. Data suggest the role of spectral shape is extensive, and that its removal degrades signals in a way that hinders recognition severely. Information reflecting *f*0 contour and target duration both appear to aid the listener somewhat, though their influence seems small compared to that of short term spectral shape. Finally, information about time varying intensity aids the listener more than noise filled gaps, and both aid the listener beyond presentation of acoustic context with duration-matched silence. © *2018 Acoustical Society of America*. https://doi.org/10.1121/1.5031123

[TCB] Pages: 2255–2268

I. INTRODUCTION

A. A brief review of phonetic reduction

Some of the variation encountered in casual speech exists as the product of phonetic reduction. Reduced forms are often characterized by speech sounds and segments that have been incompletely articulated or elided entirely (Ernestus and Warner, 2011), and have been attested in many forms of spoken language including even word list reading (e.g., Ernestus and Warner, 2011; Labov, 1972; Warner and Tucker, 2011). A problem raised by such phenomena is understanding how listeners deal with this variation. The present work explores the contributions of specific acoustic information—namely, spectral shape (SS), *f*0 contour, target duration, and time-varying intensity—in the perception of reduced speech. We report on two experiments which incorporate modified Cloze tasks (Taylor, 1953) to partly control for effects of wider context on the perception of modified target phrases, and we limit available acoustic information to explore the relative contributions of select acoustic dimensions.

Variation typical of connected speech is often viewed as a continuum spanning careful to highly reduced forms (e.g., Tucker, 2007). Warner (2012) offers an outline of what we will refer to as a careful speech continuum: At one end exists clear speech, often described as hyper-articulated; such speech may, for example, be directed toward non-native listeners and the hard of hearing. At the other end of the continuum is casual speech, also called spontaneous or conversational speech. Contrasting with more careful productions, casual speech has been characterized by more frequent hesitations, approximated articulatory gestures, and shorter prosodic units (Cutler, 1998: Mehta and Cutler, 1988).

Phonetic reduction has been shown to inhibit comprehension of spoken language in many circumstances (e.g., Ernestus *et al.*, 2002; Tucker, 2011; Tucker and Ernestus, 2016). Less severe cases include productions much like citation forms with relatively fewer and less extreme changes. In such instances, assimilations to context and various other forms of consonant and vowel reduction—including lenition and centralization—characterize productions that are still readily recognizable despite various deviations from the citation forms. Severe cases of reduction, however, are described by Johnson (2004) as dramatically altered, and involve sounds which may differ radically from citation form through deletions and/or major assimilations. When removed from context, such severely reduced productions may be deemed unintelligible.

Research has shown the degree of reduction accommodated by listeners to be influenced by several factors. Ernestus *et al.* (2002) describe the effects of syntactic and semantic context on the perception of reduced speech in an experiment offering varied degrees of context. In this study participants experienced three grades of reduction: Low, medium, and highly reduced target forms, which were

[a)]Electronic mail: ryan.podlubny@pg.canterbury.ac.nz

encountered under three contextual conditions: Full, limited, and targets in isolation. Specifically, the full context afforded participants the complete sentence in which a reduced form was produced; limited context involved reduced forms with only the adjacent vowels and intervening consonants on either side; and targets in isolation had been completely removed from the surrounding context. Each condition, therefore, provided different but specific restrictions on the amount of acoustic, phonological, and semantic/syntactic information framing a target. Results from this study show that listeners can recognize (or reconstruct) even massive reductions with a high degree of accuracy when supplied with the full context. In contrast, targets are misidentified much more frequently when context is limited and, unsurprisingly, recognition is impeded further when targets are presented in isolation. Ernestus *et al.* (2002) suggest that offering a listener syntactic/semantic context better enables recognition of highly reduced forms, and they draw specific attention to the usefulness of high frequency collocations (i.e., particular words known to group together) in reduced target identification.

Relative frequency of individual words has also been shown to aid the recognition of words in some circumstances. Pollack *et al.* (1959, 1960) describe words with higher frequency of occurrence as more accurately recognized and reproduced at lower speech to noise ratios when compared to words of lesser frequency in conditions that are otherwise equivalent. Pollack *et al.* explain that such word-frequency effects are observed with unknown message sets, though relative frequency effects do not persist when listeners are familiarized with a word list prior to the session. Thus, it appears that listeners expect to hear words of higher relative frequency unless they are primed to expect specific, relatively less common words.

An extreme view of effects governed by broader context might suggest that many cases of reduction are so far removed from their clear speech forms that the local acoustic information is virtually useless. In these cases, missing phonetic information would more or less be filled in through "guessing" given the context. However, there is clear experimental evidence that some phonetic information is available in the signal that aids recovery of even highly reduced forms.

Though Ernestus *et al.* (2002) note the elimination of immediate phonological context impedes target recognition, even in the absence of any syntactic/semantic information, van de Ven *et al.* (2012) go further by specifically investigating the relative importance of non-acoustic vs acoustic contexts in the prediction of reduced targets. In a series of experiments participants experienced either the preceding or full contextual frame as text or via auditory presentation, and were then asked to predict an excised, reduced form from four possible candidates. Each experiment in the series presented progressively more informative contextual frames to investigate the contributions of syntax/semantics and the acoustic/phonetic cues potentially surrounding a target. In these studies van de Ven *et al.* argue that syntactic and semantic cues alone are generally insufficient to predict reduced targets, and go on to emphasize the importance of phonetic cues (e.g., segmental duration and short-distance coarticulation) in the context over and above their syntactic/semantic properties.

In this work we attempt to explore the relative contributions of several readily modified acoustic cues corresponding to selected, reduced targets embedded in their unmodified contextual frames. In other words, we test how preserving or degrading specific types of acoustic information affects how well a listener can recognize (or reconstruct) reduced target forms. Before describing this work however, it is helpful to first review some theories regarding the general nature of reduced speech.

B. Some proposals on the production and perception of reduced speech

Lindblom's (1990) theory of Hyper- and Hypo-articulation, referred to herein as *Hyper/Hypo theory*, claims that reduction phenomena involve a kind of equilibrium of constraints on the speaker and hearer. The theory argues that speakers realize productions as somewhere along a continuum between hyper- and hypo-articulation (cf. Warner's careful speech continuum) based on a compromise depending roughly on both ease of articulation and some estimate of the listener's ability to recover the intended message from the signal produced in a given context. Constraints on the producer's side are both physiological and cognitive; they are limitations set by vocal anatomy and neural motor capacity, respectively. In short, there are limits on how quickly and accurately speech can be produced. Broadly speaking it is expected that even trying to approach such hard limits will incur costs to the speaker (for some direct evidence related to quantifying such costs see Moon and Lindblom, 2003).

There are also constraints that apply to the listener: Specifically, some productions are harder to decode than others. If a listener is unable to recognize the message via the speech signal they receive, the speaker is then forced to reconsider the *clarity* of their productions and alter their speech to more effectively convey the intended message. Lindblom (1996) summarizes the balance between the needs of both the speaker and the listener arguing that, with regard to the reductions and coarticulations that make speech less effortful, "…the speaker will allow himself only so much… as the listener will tolerate" (p. 1684).

Building on the principles of Hyper/Hypo theory, Aylett and Turk (2004, 2006) propose a Smooth Signal Redundancy Hypothesis to model acceptable levels of clarity during speech production. Their work defines two forms of redundancy:

(1) *Language redundancy* refers to a target's predictability based upon sentence structure and lexical meaning (i.e., the contributions of syntax/semantics).
(2) *Signal redundancy* describes the likelihood of a target being recognized by its acoustic properties alone.

According to the Smooth Signal Redundancy Hypothesis, language redundancy and signal redundancy tend to vary inversely in a way that is economical but effective for communication. Lowering levels of either signal or language

redundancy results in lower target recoverability; however, the effects of lowering one may be offset by raising the other. For example, a higher level of language redundancy means the listener can rely less on information encoded within the acoustics of the target itself. If we assume listeners predict what might have been said using collocation-frequencies at least some of the time then this theory is consistent with the findings of Tremblay and Tucker (2011), where collocations of higher frequency were generally more severely reduced. The inverse is also the case: A listener encountering strings of speech produced in a particularly clear manner can rely less on language redundancy, because in such cases there is sufficient information within the acoustics of the target for effective message transmission with little or no aid from the syntactic/semantic context.

C. Assessing the importance of four general signal characteristics

The fact that listeners are able to understand reduced speech in variable acoustic forms suggests that some subset of informative acoustic cues must persist at the core of an utterance for use during speech processing. A next step, then, is to determine what types of acoustic information a listener takes advantage of when processing reduced forms. Put another way, spoken language is the amalgam of change over time along a number of dimensions (including intensity, formants and other spectral information, segment duration, etc.), but the relative importance (or *informativeness*) of each dimension to listeners remains largely unexplored. Some of these acoustic components are very closely related and potentially difficult to isolate and test individually. At this stage of our knowledge it is impossible to achieve an exhaustive breakdown describing the relative importance of every possible acoustic component relevant to this problem, although certain characteristics of signals can be readily modified in ways that affect certain cue patterns believed to be relevant to speech perception generally, while leaving others largely intact. The present work focuses on four such isolable cues deemed likely useful to listeners in order to assess their relative informativeness.

D. Four general acoustic dimensions of interest

The four characteristics of speech compared in this study are duration pattern, fundamental frequency, SS, and intensity contour (IC). These characteristics can, to a large extent, be manipulated separately from each other with well-defined signal processing techniques. We leave details of the exact manipulations to Sec. II A 3, and first briefly review literature that suggests these characteristics are relevant for the perception of speech in general.

1. Duration pattern

The duration of various signal types (e.g., voiced vs voiceless intervals within a word) has long been known to affect the perception of consonants and vowels, as well as prosodic properties such as stress and prominence (Klatt, 1976). Duration is also relevant to perceptual effects related to disruptions in durational patterns in larger stretches of speech (beyond single words), which are readily detected by listeners. Furthermore, several studies (e.g., Port, 1979; Wayland *et al.*, 1994) have shown the perception of phonetic segments in target syllables can be affected by changing the relative temporal structure of context ("carrier") sentences.

2. Fundamental frequency (f0)

Modification of fundamental frequency, which affects the detection and perception of voice pitch, also plays a strong role in the perception of prosodic properties (Fry, 1958) such as stress and intonation (Vassière, 2005); segmental properties like the voicing state of consonants (Kingston and Diehl, 1994); and indexical characteristics including speaker identity and size (Johnson, 1990; Barreda and Nearey, 2012), and speaker dialect (Szakay, 2008).

3. IC

Along with duration and fundamental frequency, intensity-based variation is known to play some role in the perception of suprasegmental or prosodic properties of stress and intonation (Fry 1958; Vassière, 2005; Cutler, 2005, pp. 265–267). Particularly relevant is the work of Bashford *et al.* (1996) in a speech restoration paradigm. From stretches of recorded speech 400 ms in duration, the authors replaced either 200 or 250 ms segments with either silence or one of two noise types. Listeners' identification rates of the replaced portions were higher in both noise conditions than in the silent condition. However, they found that listeners perform markedly better when the fillers were speech-modulated noise rather than stochastic white noise. Their modulated noise was produced using the method described by Schroeder (1968), which we also use in the present work to preserve the intensity envelope of speech at every time scale while eliminating *f*0 and SS information.

4. SS

We use the term SS to refer to the relatively smooth envelope of a log- or dB-transformed power spectrum of a temporal section of moderate duration (a few tens of milliseconds), such as might be the result of typical linear predictive smoothing or cepstral smoothing techniques. Such envelopes follow the general patterns imposed by vocal tract resonances (formants) and such general properties as spectral tilt, but blur over the fine variation along the frequency axis due to the harmonic structure of voiced speech. Some amount of spectral resolution has clearly been shown as necessary for intelligible speech. For example, in a review of experiments with noise-excited vocoders, Shannon *et al.* (2004) show that 3 to 5 or more abutting equally wide log-frequency bands are consistently required to achieve 60%–90% recognition of "easy" sentence material. This result compares to about 20% recognition with two-channel noise vocoding, and to only low single-digit recognition in the one-channel case (corresponding to a kind of speech-modulated noise).

Having motivated four general acoustic dimensions of interest, we next explore the relative contributions of such information during the processing of reduced speech.

II. EXPERIMENT 1

A. Methodology

1. Stimuli

Stimuli were extracted from a single recorded conversation. The talker is a 22-yr old female, native speaker of Western Canadian English. The recording took place in a sound attenuated booth on the University of Alberta campus, where a cordless telephone was used to speak with an off-site interlocutor (the participant's mother); only the on-site speaker was recorded, following Warner's (2012) suggestion that this methodology puts speakers more at ease in a laboratory setting and results in a more accurate approximation of casual speech. Roughly 30 min of spontaneous conversation were captured using a Countryman E6 ear-mounted condenser microphone and a Korg MR 1000 high-resolution recording device. An Alesis Multimix 8 mixer was used as a preamp for the microphone and to supply phantom power. Recordings were captured at 44.1 kHz and 16 bits.

Reduced *targets* (or instantiations of reduced speech) were selected on an impressionistic basis by R.G.P.—also a native speaker of Western Canadian English—where 71 targets were noted for representing a variety in length (from 1 to 6 words) and degree of phonetic reduction after reviewing the recording multiple times. Contextual *frames* (or the speech/words adjacent to targets) were marked to later extract targets as well as the phrasal-context surrounding them. Targets could occur frame initially, medially, and frame finally, and were therefore not always bookended by additional speech. All frames were segmented and extracted as mono WAV files using Praat (Boersma and Weenink, 2012).[1]

We use two measures of reduction in the present work. (1) *The Deletion Ratio* was calculated by comparing two transcriptions of each target utterance: The first is a broad phonetic transcription of what was actually produced by the speaker, and the other represents an idealized citation form reflecting the phonemes which would be listed in a dictionary pronunciation of that target. All transcriptions represented a consensus of three linguistically trained judges. The ratio was generated by subtracting the number of phones actually realized from the number that would have been produced in the citation form; the difference was then divided by the number of intended phones (in the citation form), providing a normalized measure of deletion which accounts for differences in target length. Although this measure is not sensitive to phonemic alternations that may be observed as changes to features like place and manner, it was sufficient to confirm the selected targets did exemplify minor to substantial phonetic deviation from what are typically regarded as canonical productions. The second measure is described in Sec. II A 2.

2. The ALINE algorithm

We have also adopted a measure of phonetic similarity as our second reduction metric, which we expected would be a more sensitive measure of agreement. (2) *The ALINE Score* is a feature-level similarity measure based on the ALINE algorithm (Kondrak, 2000, 2003) which was initially designed to quantify similarity among putative cognates in historical linguistics, and has since proven useful in a number of natural-language processing applications (e.g., Downy *et al.*, 2008; Mani *et al.*, 2008). ALINE's feature-based grounding allows the algorithm to estimate the similarity of any pair of words or short phrases (that have been phonetically transcribed) by decomposing phonemes into elementary phonetic features.

The principal component of ALINE is a function that calculates the similarity of two phonemes that are expressed in terms of roughly a dozen binary or multi-valued phonetic features (Place, Manner, Voice, etc.). Feature values are encoded as real-valued numbers in the range [0,1]. For example, the feature "Manner" can take any of the following seven values: stop = 1.0, affricate = 0.9, fricative = 0.8, approximant = 0.6, high vowel = 0.4, mid vowel = 0.2, and low vowel = 0.0, where numerical values are meant to reflect the size of the closure during speech production. The phonetic features are assigned *salience* weights that express their relative importance.

The overall similarity score and optimal alignment of two words, computed by a dynamic programming algorithm (Wagner and Fischer, 1974), is the sum of individual similarity scores between pairs of phonemes. A constant insertion/deletion penalty is applied for each unaligned phoneme. Another constant penalty is set to reduce the relative importance of the vowel (as opposed to consonant) phoneme matches. The similarity-value is normalized by the length of the longer word.

ALINE's behavior is controlled by a number of parameters: the maximum phonemic score, the insertion/deletion penalty, the vowel penalty, and the feature salience weights. We used the default settings for the parameters, resulting in similarity values between 1 and 0.

In the following analyses we refer to both the ALINE (*Stimulus Reduction*) *Score* and the ALINE *Response Measure*. Though both describe information gained through the ALINE algorithm, these terms refer to very distinct ideas that exist in different contexts. The ALINE Score is an item-specific reduction metric based on phonetic similarity, which compares the phones realized in the speaker's production within a stimulus to the phones expected in its citation form. In the present work this score ranged from 0.32 to 1 with a median value of 0.704. The ALINE Response Measure, however, describes a value for each listener-response, generated by comparing orthographic transcriptions of stimuli (largely canonical in nature, though allowing for select lexicalized reductions, e.g., *gonna*, *wanna*, etc.) to the phones expected in that target's corresponding citation form (that is, how each target would be transcribed in a dictionary). Before such comparison can take place orthographic transcriptions must be converted to phonemic transcriptions via pronunciation dictionary lookup, or for out-of-dictionary items by using a custom text-to-phoneme algorithm allied to ALINE. This use of ALINE was explored as a means to shift our dependent variable from categorical (i.e., entire

responses scored as either correct or incorrect) to a more gradient measure of similarity.

3. Manipulations

Targets and frames were coded manually as TextGrids using Praat, which were used to automatically remove, manipulate, and re-insert targets into their original frames. This process resulted in four versions of each item, where corresponding conditions are referred to below as (1) *Original* (items remained unaltered), (2) *Stretched-Duration*, (3) *f*0, *and* (4) Signal Correlated Noise (SCN). Descriptions of manipulation are available below (Table I summarizes the acoustic properties modified within each condition).

To explore the contribution of high-resolution time varying intensity, targets were replaced with SCN (Schroeder, 1968; Benkí, 2003a). Targets were modified by randomizing the polarity but maintaining the amplitude of each sample of the original speech targets, thus matching the integrated IC of the original signals at every time scale at least as long as the sampling interval (1/44 100 s). This manipulation can also be viewed as a form of single-channel noise vocoding (Shannon *et al.*, 1995), which leads to complete whitening of the spectrum at any time scale, eliminating variation of SS patterns across time. In addition to preserving detailed intensity information, this manipulation also preserves total target duration. However, segment-level (e.g., vowel, consonant) duration and most other target-internal prosodic cues are at best severely weakened by the loss of spectral information.

Manipulation in the *f*0 condition retains information about fundamental frequency, periodicity, and short-time amplitude of the original signal, but removes all variation in the short-time shape of the spectral envelope. Original signals were first down-sampled to 16 kHz to facilitate a more stable, lower order linear predictive coding (LPC) model to account for its time-varying spectral envelope (note: all other conditions maintain original sampling rates of 44.1 kHz). A 17th order LPC was then applied to the down-sampled target using Praat's *Burg LPC* method with a window duration of 25 ms and time step of 5 ms. The signal was then inverse-filtered via Praat's *Filter* (*inverse*) function using the filter coefficient of the LPC analysis. The resulting LPC-residual signal is spectrally whitened, and can be viewed roughly as an estimate of the source signal when the vocal tract transfer function is largely removed. While the intensity of this residual signal varies in a way that is generally correlated with the original signal, short-time intensity is not exactly proportional to that of the original. This difference is a result of the "removed" vocal tract filter characteristics, which vary with time, also affecting the instantaneous intensity contour (IIC). To better match the short-time IC of the processed signal to that of the original, the intensity matching method from a Praat script developed by Mitterrer (2005) was adapted and applied to the residual signal. This procedure produced good matches of the original and processed signals when Praat's default ICs were compared [according to the Praat manual, this process involves convolving the squared signal with a 45 ms Kaiser window and an alpha parameter of 6.37 (alpha = 20/pi), then converting to decibels]. The resulting signals largely preserve the pitch and short-time energy contours of original signals, while short-time variation in the spectral envelope is almost entirely removed. One reviewer drew attention to silent periods of about 20 ms on either end of each target in this condition; these gaps were unintended. They resulted from an unanticipated side effect of how we used the Praat LPC inverse filtering command and how that procedure applied analysis windows to the targets.

Where the previous two treatments essentially eliminate some information from the signal entirely, our *Stretched-Duration* manipulation preserves all other acoustic information while systematically modifying its time course—thus affecting properties like segment duration. Artificial temporal manipulation (e.g., time compression) is a means to alter signal duration without influencing fundamental frequency (Adank and Janse, 2009); listeners are known to deal well with sentences compressed by up to 38% (Dupoux and Green, 1997), and Zhao (1997) argues that decreased speech rate (i.e., time-expanding signals) results in improved listening comprehension. Therefore, we have opted to use Praat's pitch-synchronous overlap and add function, *Lengthen* (*overlap add*), to effectively stretch targets and thus disrupt some temporal aspects of speaker prosody. The pitch floor and ceiling were set to 75 and 600 Hz, respectively, with a lengthening factor of 1.5 to expand the time axis by an extra 50%. We believed this change in temporal flow could either (1) hinder target recognition through disrupting the rhythm and relative timing patterns of targets with respect to the context, or (2) improve recognition of reduced targets by increasing time available to process and identify approximated gestures within the signal.

TABLE I. Description of properties maintained or modified by condition in experiment 1 (*f*0 = Fundamental Frequency; IC; SS).

Condition	Properties preserved	Properties modified
Original	All	None
Stretched-Duration	Patterns: *f*0, IC, SS	Duration (×1.5)
f0 / flattened-spectrum	Duration, *f*0, IC	SS (flattened)
SCN	Instantaneous IC	*f*0 (eliminated) SS (flattened)

4. Participants

Participants were 16 female and 7 male undergraduate students from the University of Alberta ($n = 23$), aged 17–26 yrs. With the exception of 5 who gained fluency before 6 yrs of age, all were native speakers of western Canadian English; other languages spoken by these multilingual participants include Punjabi, Cantonese, and Spanish (additional information regarding age of acquisition was not collected). No participant reported any known hearing impairment. Students received partial course credit in exchange for participation. One male subject was excluded for not completing the task.

5. Procedure

Participants were seated in a sound attenuated booth equipped with a computer monitor, headphones, and

QWERTY computer keyboard, and were instructed to watch the screen and "fill in the blank for each item with whatever might fit best, using the keyboard. This can be one word or more than one word." A session was split into five blocks and all conditions included each of the 71 items described above. Sessions always began with the visual Cloze condition, an open-response task (in this context) involving presentation of an orthographic transcription of each frame, where targets had been replaced by ten consecutive underscores (e.g., "*I will sit there and I will just watch TV* _____ *homework*" or "_____ *really like those*"). Each session concluded with the Original condition—thus the Visual Cloze (Block 1) and Original (Block 5) conditions bookended the three test conditions. Sequencing of those test conditions (*f*0, Stretched-Duration, and SCN) as blocks 2, 3, and 4 was randomized by subject. The visual Cloze condition was presented in quiet, and participants were prompted to put on the headphones for the second and subsequent blocks via the computer monitor.

B. Results

This experiment was designed to explore two measures of accuracy: (1) the binary measure of correct identification of the target, and (2) the graded ALINE Response Measure as a relatively continuous measure of phonetic similarity when comparing responses to transcriptions of their corresponding targets. These two dependent variables were analyzed separately; each analysis is described and compared in detail below. The collected data include some 7810 responses from 22 participants. All analyses were executed in R (R Core Team, 2017) using linear mixed-effects regression (Bates *et al.*, 2014). A backward stepwise modeling procedure was used to define the predictor structure of each model described within this work, including random effects, where non-significant predictors were pruned one by one and iterative models were compared using analysis of variance (ANOVA) testing (Baayen, 2008). A forward fitting procedure, where predictor variables were added one by one and models compared using ANOVA, was then used to investigate the contribution of random slopes. All predictors and random slopes that significantly improved the model fit were retained.

We performed an analysis with Condition as the sole fixed factor to assess the overall differences between the Visual Cloze condition and the other conditions. In this analysis Accuracy, or the number of items correctly transcribed (any error within a transcription resulted in its treatment as incorrect), was predicted by Condition (Visual Cloze, Original, Stretched-Duration, *f*0, SCN). Further, we included the following control variables: Trial (within a given condition), Phrase Rate, Deletion Ratio (or) ALINE Score, Location (initial, medial, final), and Target Frequency [as collected from the Corpus of Contemporary American English (Davies, 2008)].[2] Subject and Item were included as random effect predictors. No random slopes were found that allowed the model to converge while improving the model fit. The optimizer for this model was set to "bobyqa."

Figure 1(a) illustrates the average accuracy per condition, where responses to the Visual Cloze task were consistently the least accurate and statistically different from all other levels [Visual Cloze compared to Original ($\beta = 7.06$, SE = 0.18, $p < 0.001$), *f*0 ($\beta = 3.19$, SE = 0.14, $p < 0.001$), SCN ($\beta = 2.899$, SE = 0.137, $p < 0.001$), Stretched-Duration ($\beta = 6.88$, SE = 0.18, $p < 0.001$)]. Responses to the Original condition were 91.5% accurate, and a similarly high level of target recognition was observed in response to the Stretched-Duration condition (90.3%). We therefore set the Original condition as the intercept for comparison to the Stretched-Duration condition and found that this difference is not statistically significant ($\beta = -0.18$, SE = 0.14, $p > 0.05$). The comparisons to SCN ($\beta = -4.16$, SE = 0.14, $p < 0.001$) and *f*0 ($\beta = -3.87$, SE = 0.14, $p < 0.001$) do show a significant difference. We then set SCN as the intercept to test differences between it and the other conditions, and found that SCN was significantly less accurate than the *f*0 ($\beta = 0.29$, SE = 0.092, $p < 0.001$) and Stretched-Duration ($\beta = 3.98$, SE = 0.14, $p < 0.001$) conditions.

Having established the Visual Cloze condition elicited the least accurate responses, we focused next on the conditions of interest by removing the Visual Cloze and Original conditions from the dataset. Following this sub-setting the three test conditions remained—a total of 4686 observations. As a control for the influence of the contextual frame [i.e., non-acoustic contextual predictability (NACP)] we calculated

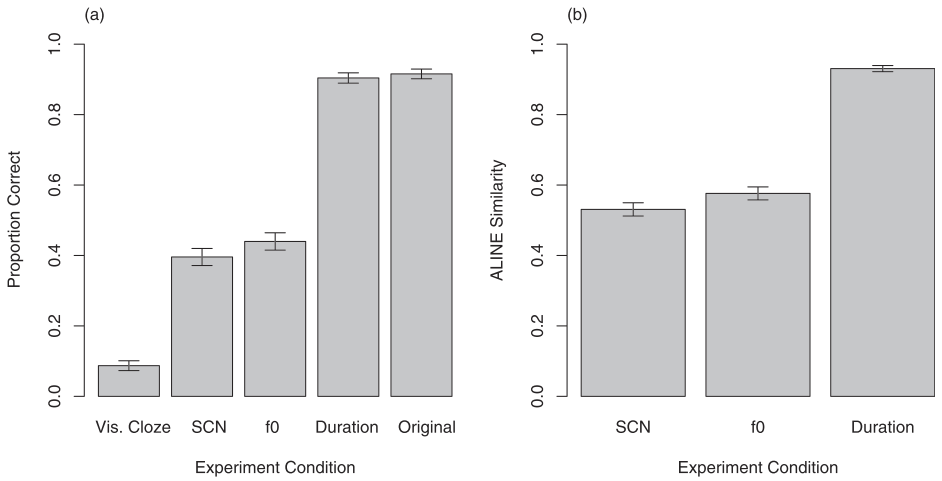

FIG. 1. (a) Average percent correct by Condition and (b) average ALINE Similarity by Condition for the three signal-based manipulations. Both plots include 95% confidence intervals.

the mean percent accuracy for each of the 71 items across all participants in the Visual Cloze condition, which ranged from 0% to 73% accuracy with over 50% of these values falling at 0. We refer to this variable herein as the *vcScore* (Visual Cloze Score) for each item, and include it in all further modeling.

Binary accuracy is a relatively coarse measure of the information transmitted during each trial. For example, a response like "I can go" would be judged wrong if the target was "I *can't* go," despite most phonemes having been correctly identified. We expected, therefore, that a graded measure of phonetic similarity, available through the more continuous ALINE Response Measure, might reveal subtle patterns left undetected through a binary metric. While the ALINE measure correlates strongly overall with binary accuracy ($r = 0.959$, across items and conditions averaged over listeners), it also contains information about the degree of mismatch in each incorrect response; these values range from 0 to 1 in our data with a median value of 0.84, indicating a prevalence of relatively highly similar responses.

Therefore, we turn to an analysis of Aline Scores to study differences among the three modified signal conditions. Mean similarity for the three modified stimulus conditions is shown in Fig. 1(b). Analysis included the following predictors: Condition (stimulus signal condition), Order (presentation block order of the conditions), Trial (trial number within a condition), Phrase Rate, Deletion Ratio (or) ALINE Score, Location (initial, medial, final), and Target Frequency. All possible two-way interactions of the predictor variables were explored as part of the modeling process, where non-significant predictors and interactions were trimmed. Numeric variables were centered and Target Frequency was log-transformed. The optimizer for this model was set to "bobyqa." The final model, shown in Table II, contains random intercepts for Subject and Item. Random slopes were included for Condition and ALINE Score by Subject and Condition by Item. Coefficients with a *t*-score greater than an absolute value of 2 were considered significant within all present model summaries (following Baayen, 2008: pp. 73–76), and have thus been bolded.[3]

Phrase Rate and log Target Frequency were the only predictors that did not significantly improve the model fit. With regard to ALINE Score, targets with a lower similarity were more difficult to recognize. We find a small effect of NACP (accuracy increasing with predictability) for the Stretched-Duration manipulation and a larger effect for both the *f*0 and SCN manipulations (Condition interacting with vcScore). The interaction of Order with vcScore indicates that the influence of NACP decreases as the experiment progresses. We also observe an interaction between Order and Condition. A study of the relevant coefficient weights suggests responses are relatively more accurate in conditions following the Stretched-Duration manipulation; though, we also find a general increase in target-similarity for the *f*0 and SCN conditions as they move from the first position toward the third position. An additional analysis using full factorial coding of all Order by Condition combinations was consistent with the outcomes reported here. Regardless of Order, we find that responses to the Stretched-Duration manipulation are significantly more accurate than those to the SCN and *f*0 manipulations (when the model was releveled we find a trend toward higher target-similarity values in the *f*0 condition than the SCN condition, though this effect does not reach significance). The interaction between vcScore and Condition indicates the slope of the vcScore effect is significantly reduced for the Stretched-Duration condition when compared to SCN and *f*0. Finally, we observe a significant interaction between condition Order and Location where participants are least accurate when targets exist utterance-initially, and there is no difference between the medial and final positions.

TABLE II. The output of a linear mixed effects model fit to data collected during test-conditions only within experiment 1, with the ALINE measure serving as the dependent variable.

	Estimate	Std. Error	*t* value
(Intercept)	0.9481	0.0140	**67.75**
Condition:SCN	−0.5928	0.0430	**−13.78**
Condition:*f*0	−0.4877	0.0356	**−13.69**
Order:2	0.0001	0.0158	0.01
Order:3	0.0165	0.0151	1.09
vcScore	0.0304	0.0110	**2.76**
Location:Beginning	−0.0570	0.0268	**−2.13**
Location:Final	−0.0109	0.0342	−0.32
ALINE Score	0.0275	0.0101	**2.73**
Condition:SCN × Order:2	0.2430	0.0448	**5.43**
Condition:*f*0 × Order:2	0.1515	0.0446	**3.4**
Condition:SCN × Order:3	0.2691	0.0476	**5.65**
Condition:*f*0 × Order:3	0.2537	0.0417	**6.08**
Condition:SCN × vcScore	0.1356	0.0225	**6.02**
Condition:*f*0 × vcScore	0.1274	0.0225	**5.66**
Order:2 × Location:Beginning	−0.0736	0.0219	**−3.36**
Order:3 × Location:Beginning	−0.1062	0.0218	**−4.87**
Order:2 × Location:Final	−0.0318	0.0282	−1.13
Order:3 × Location:Final	−0.0170	0.0280	−0.6
Order:2 × vcScore	−0.0176	0.0085	**−2.08**
Order:3 × vcScore	−0.0378	0.0083	**−4.52**

In summary, our data show that a lack of any acoustic information severely impedes target recognition/reconstruction. Moreover, we have found that different forms of acoustic information lend themselves to various—and graded—levels of improvement in this task. Surprisingly, when comparing transcriptions of time-stretched stimuli with those from the unaltered original productions, we found that listeners exhibit similar recognition-accuracy across these conditions. Finally, we have provided support for use of the ALINE algorithm in the quantification of reduced speech and in the prediction of target recognition through graded transcription accuracy, and have shown this measure is sensitive to effects that may be missed using more coarse metrics for target identification.

C. Discussion

The Visual Cloze and Original conditions represent minimum and maximum information available within target stimuli. In the former listeners can extract, by eye only, syntactic and semantic hints about a missing target. No acoustic information is provided. With the original signals, however,

listeners can extract syntactic and semantic information through an *auditory* channel as well, stemming from acoustic information in both the target and the phonological context. Figure 1 shows that the discrepancy between responses in the visual Cloze and all other conditions is extensive, thus confirming the availability of *any* acoustic information, even if highly degraded, improves target recognition substantially.

Hillenbrand (2003) describes availability of the *f*0 contour as contributing to sentence intelligibility; we therefore expected available information reflecting fundamental frequency would improve response accuracy when compared to our SCN condition. Though differences were generally found in pairwise comparisons by condition, accuracy was not significantly different when comparing performance in the *f*0 and SCN conditions (recall we observed an effect in the initial model which included only Condition as a predictor, however once sufficient controls were entered into the model the relative *f*0-advantage was no longer significant). Similar performance in these two conditions seems almost surely driven by the removal of SS, where accuracy was reduced dramatically, and similarly, when comparing these conditions to those where SS remained intact (i.e., Stretched-Duration and Original).

We were surprised that the *f*0 condition provided no substantial benefit over SCN since it might be expected to preserve pitch and at least some voicing state information that was obscured by SCN. While both manipulations follow the general IC (integrated over tens of milliseconds), SCN matches intensity of the original at any time scale. Although spectral detail, such as harmonic structure, is obliterated by SCN, some relatively weak pitch information (sometimes called "envelope pitch") may remain in amplitude-modulated noise for some range of modulation frequencies above *f*0. The frequency cutoff for SCN is effectively half the sampling rate and, hence, well above speech *f*0 range. Therefore, some cues to periodicity and intonation may be signaled through higher frequency fluctuations in the amplitude envelope (Van Tasell *et al.*, 1987). However, work in noise-vocoding often shows no increase in listener benefits for modulation rates above ~200–300 Hz (Shannon *et al.*, 1995), and voice-pitch may not be perceived reliably for pitches in the range of 220 Hz for 3 channel noise vocoders at 300 Hz modulation cutoff (Souza and Rosen, 2009). The *f*0 manipulation, on the other hand, provides a relatively faithful pitch contour while intensity variation matches that of the original only at longer time scales. Presuming the brief unintentional gaps in our *f*0 stimuli were not driving a decrease in target recognition, it is possible then that some gains achieved through the extra voice-pitch information in the *f*0 condition were offset by some aspect of very short time intensity variation in the SCN signals.

We found that disrupting temporal prosodic cues led to a small decrease in target identification compared to unmodified signals, though this difference was not statistically significant. The fact that participants responded similarly to duration-expanded speech and the unaltered signals suggests long-range relative timing effects, such as those observed by Port (1979) or Wayland *et al.* (1994), played at best a marginal role in this experiment. In short, disrupting speech rate did not appear to help *or* hinder the processing of reduced targets. Aiming to confirm minimal influence of target duration, an alternative approach to this condition is employed in experiment 2 through the Silent Gap treatment. The ALINE algorithm proved helpful in confirming substantial variation in levels of reduction in our stimuli, and also provided gradient measures of similarity of participants' responses to the intended word forms. We explored the predictability of Deletion Ratio vs ALINE Score by replacing one with the other and re-running the model; comparing alkaline information criterion values from each model indicated that including ALINE Score provided the best fit to the data. The analysis of the ALINE Response Measure, while confirming effects observed in the coarser binary accuracy score, also proved sensitive to variations in the experimental conditions that could have otherwise been missed.

One limitation of this study can be found in the stimuli selection. While focusing attention on the degree of reduction, we neglected the issue of balance in target location across the set (Initial: 14; Medial: 48; Final: 8). Location has therefore been treated as a control in modeling, and any Location-based effects should be considered with caution.

Another issue is that participants encountered the same 71 target stimuli in each of the three signal conditions—this confound seems a likely explanation for the main effect of Condition Order as well as its interaction with vcScore. If cues provided by contextual frames become less informative in subsequent blocks, it seems reasonable to assume then that participants may recognize stimuli as a session progresses. Additionally, since the Stretched-Duration condition is well recognized, its presence in an earlier block could very possibly have a strong priming effect, enhancing the likelihood of listeners recognizing the more difficult SCN and *f*0 conditions in subsequent blocks. This confound is addressed in experiment 2.

In experiment 1 we found evidence for the informativeness of certain types of acoustic information to target intelligibility, most notably through SS. In experiment 2, we aim to build on some of the findings of experiment 1. First, to explore further the issue of the importance of acoustic contextual information, we introduce a Silent Gap condition—which amounts roughly to an Auditory Cloze condition—aiming to learn more about context effects by comparing a listener's performance across visual and auditory equivalents. Second, in retrospect, because the benefits observed in our SCN condition could be the product of either the available IC or some restorative process due to *any* kind of noise filling a target gap (e.g., Warren, 1970), we have included two types of noise, including Flat Amplitude white noise (with no short-term intensity variation) as well as the previous SCN. We also explore several degrees of degradation of SS information by varying the signal-to-noise ratio (SNR) of both types of added noise. Finally, to remedy a design flaw that became apparent in the previous analysis, in experiment 2 we avoid the unintended priming effects likely resulting from repeated exposure to versions of the same signals.[4]

III. EXPERIMENT 2

A. Methodology

1. Stimuli

Stimuli in experiment 2 were based on the original 71 targets and frames used to generate stimuli for experiment 1.

2. Manipulations

Orthographic transcripts of each stimulus (less the targets) were presented as a visual Cloze test as in experiment 1. The auditory conditions were generated as follows: Using methods similar to those outlined in Sec. II A 3, targets were automatically extracted from, manipulated, and reinserted into their corresponding frames. The *Silent Gap* condition is analogous to the Visual Cloze test in that participants were auditorily presented the acoustic context with a silent gap matched for target duration (cf. Bashford *et al.*, 1996; van de Ven *et al.*, 2012) in addition to the orthographic transcription. The Silent Gap condition was designed for within-participant comparison to the visual Cloze, to partially assess the influence of target duration and the introduction of acoustic context. We also generated a series of *Noise Masked Auditory* conditions, which were much like the Silent Gap condition but further introduce auditory target signals mixed with one of two masking noise types, each at one of three SNRs. The masking noise types were SCN as in experiment 1, and a *Flat Amplitude* noise condition generated as white noise with a static amplitude contour reflecting mean intensity averaged over the target window. Our SNRs were based on the description in Benkí (2003b) which involves SCN and both nonsense- and word-syllables, and suggests SNR values ranging from −5 to −14 dB can elicit error rates spanning roughly 5% to 95%. We therefore adopt Benkí's range of values as endpoints, and add −9.5 dB as an intermediary step. Table III summarizes the acoustic properties modified within each condition.

3. Participants

Data were collected from 85 native speakers of Western Canadian English. All were undergraduate students from the University of Alberta and were enrolled in an introductory linguistics course at the time. Participants received partial class credit for their time. Eight participants were excluded for not completing the experiment, corrupted data, or for having previously undergone hearing correction.

4. Procedure

In experiment 1 we found evidence that participants could better recognize target stimuli repeated over the course of a session, likely through a kind of priming. Consequently, the presentation strategy in experiment 2 was altered to avoid any potential priming that might result from encountering relatively high-information auditory conditions before lower-information ones. Listeners instead received each target stimulus only with increasing acoustic information over the course of the three experimental blocks in a full experimental session. The Visual Cloze condition, with no-acoustic information, was always Block 1. The Silent Gap condition, with an auditory context-frame and target-duration matched silent period, was always Block 2. The Noise condition was always Block 3—participants heard each target only once in exactly one of the six noise conditions, according to one of six counterbalanced lists. Lists were generated using a Latin square design to ensure participants encounter each of the 71 targets/frames only once, but also experience all possible combinations of Noise Type + SNR. Each permutation of the experiment (i.e., with 1 of the 6 lists as Block 3) was presented to no fewer than ten listeners. A summary of experiment 2 by condition is available in Table IV.

TABLE III. Description of properties maintained or modified by condition in experiment 2 (NACP, SS, IIC). Note that, with regard to the target, both *noise* conditions introduce increasing degrees of spectral information as SNRs become more favorable.

Condition	Properties preserved	Properties modified
Visual Cloze	NACP	Phonetic Context (eliminated)
Silent Gap Auditory	NACP, Phonetic Context, Target Duration	*f*0, SS, IIC (eliminated)
Flat Amplitude Noise	NACP, Phonetic Context, Target Duration	*f*0 (decreased—subject to SNR)
		SS (decreased—subject to SNR)
		IIC (decreased—subject to SNR)
SCN	NACP, Phonetic Context, Target Duration, IIC	*f*0 (decreased—subject to SNR)
		SS (decreased—subject to SNR)

B. Results

As in the first experiment, a preliminary main-effects only analysis was performed with our binary measure of response accuracy as the dependent variable. This analysis tests for differences between the three general signal conditions (Visual close, Silent Gap and Noise—not differentiating subtypes of noise). The analysis consisted of linear mixed-effects regression with random intercepts included for Subject and Item, as well as random slopes for Location by Subject. Potential predictors included in the model were: Condition, Trial, Location, Target Frequency, and Phrase Rate. We find once again [Fig. 2(a)] that responses to all Noise conditions on average were significantly more accurate than responses to the Visual Cloze ($\beta = -4.03$, SE = 0.08, $p < 0.001$), as well as the Silent Gap condition ($\beta = -2.95$, SE = 0.07, $p < 0.001$). Releveling to allow for ready comparison of the Visual Cloze (intercept) and Silent Gap conditions reveals they too were significantly different ($\beta = 1.08$, SE = 0.08, $p < 0.001$). We also find an effect for target Location within a Phrase (Final $\beta = 1.88$, SE = 0.80, $p < 0.001$; Medial: $\beta = 2.06$, SE = 0.56, $p < 0.001$), and a modest, though significant effect for Trial ($\beta = 0.15$, SE = 0.03, $p < 0.001$). An effect of Target Frequency was also observed ($\beta = 1.16$, SE = 0.22, $p < 0.001$). Having effectively replicated the general pattern of the first analysis in experiment 1, the remainder of this section describes analyses modeling response fidelity as reflected in the ALINE Response Measure. We also now include factors allowing us to explore the influence of both noise type (i.e., Flat Amplitude vs SCN) and SNR (−14, −9.5, −5 dB). Therefore, the following analyses are restricted to data gathered during block three only.

A linear mixed-effects model was fit to explore calculated similarity using the ALINE Response Measure [see Fig. 2(b) and Table V], including main effects for Trial, Location, Noise Type, Target Frequency, Phrase Rate, and ALINE Score. Random intercepts were included for Subject and Item, as were random slopes for Trial by Subject. All

TABLE IV. Experiment 2 summarized by condition and intended statistical comparisons.

Order/Block	Condition	Stimuli	New Information Provided (additive by condition)	Intended Comparison
1	*Visual Cloze*	Visual Frame	Orthographic cues to syntactic/ semantic context	Silent Gap (within-subject)
2	*Silent Gap Auditory*	Visual Frame + Acoustic Context	Acoustic context, target duration through silent gap	Visual Cloze (within-subject)
3	*Noise*	Visual Frame + Acoustic Context + Counterbalanced mixture of the following to mask original targets: - SCN −5 dB - Flat Amplitude −5 dB - SCN −9.5 dB - Flat Amplitude −9.5 dB - SCN −14 dB - Flat Amplitude −14 dB	Some ratio of original signal + white noise replacing target shaped as: SCN—Signal shaped instantaneous intensity contour of target (or) Flat Amplitude noise—static intensity contour replacing target	Responses during SCN are compared to responses during Flat Amplitude noise (between-subject)

numeric variables were centered and Target Frequency was log-transformed.

We find an interaction between Trial and SNR such that participants' responses become more accurate as a noise condition progresses, but this increase is most substantial when the condition involved lower levels of background noise. We also find SNR interacting with Noise Type where signal information mixed with SCN is more informative under more favorable SNRs, though the usefulness of this information diminishes as relative noise-levels increase. SNR also interacts with NACP (vcScore), confirming that listeners rely more on syntactic and semantic cues as relative noise levels increase.

SNR interacts with Phrase Rate such that increasing speech rate reduces response fidelity more dramatically in the noisier (lower SNR) conditions. We also find SNR interacting with target Location, with direction varying by position: Responses are relatively more accurate at the end of a phrase in more favorable SNRs, but are more accurate phrase-medially at less favorable SNRs. (The generality of this finding should be regarded with caution due to location balance issues in the stimuli). Additionally, the interaction between Noise Type and target Location suggests that response similarity decreases for SCN targets in final position. The interaction between vcScore and target Location indicates that contextual information has a larger influence when targets occur initially than otherwise. This effect is confirmed by taking the initial location as the reference, where we find that the initial position is affected by contextual informativeness more than final ($\beta = -0.30$, SE $= 0.10$, $t = -2.96$) and medial ($\beta = -0.32$, SE $= 0.09$, $t = -3.37$) positions. The interaction between SNR and ALINE Score indicates that increasingly favorable SNRs are correlated with relatively higher Aline Scores (that is, increased target similarity); however, this interaction only reaches significance at the −5 dB ratio. Noise Type interacts with Phrase Rate indicating that increased rate results in lower fidelity, with the magnitude of this effect more pronounced for Flat Amplitude noise. Finally, the interaction between SNR and Target Frequency indicates listeners do not rely on frequency information at more favorable SNRs, but do use this information when signals are less clear.

In summary, we have found further support for graded increase in target recognition/reconstruction through an

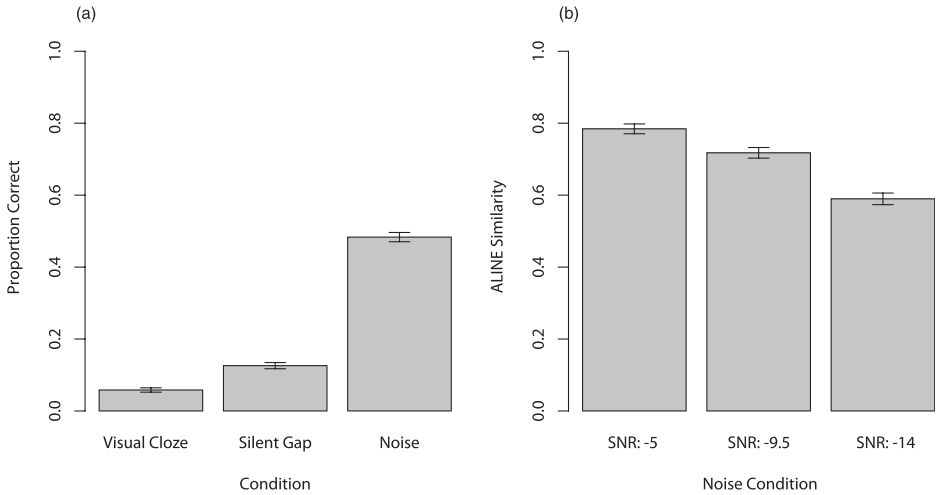

FIG. 2. (a) Proportion correct by Condition ("Noise" averaged across noise types and SNRs) and (b) ALINE Similarity by SNR (within the Noise condition, averaged across noise type). Error bars represent 95% confidence intervals.

TABLE V. A linear mixed-effects model fit to data collected during the noise condition in experiment 2, which explores gradient response accuracy using the ALINE Response Measure.

	Estimate	Std. Error	*t* value
(Intercept)	0.6474	0.0238	27.178
Trial	0.0383	0.0060	**6.427**
SNR: −5	0.1726	0.0123	**14.049**
SNR: −9.5	0.1234	0.0122	**10.121**
NoiseType:SCN	−0.0161	0.0117	−1.373
vcScore	0.0853	0.0208	**4.098**
PhraseRate	−0.0118	0.0192	−0.614
Location:Beginning	−0.1175	0.0585	**−2.007**
Location:Final	−0.0391	0.0622	−0.63
ALINEScore	0.0217	0.0193	1.122
Target Frequency	0.0501	0.0215	**2.323**
Trial × SNR: −5	−0.0210	0.0079	**−2.67**
Trial × SNR: −9.5	−0.0092	0.0078	−1.179
SNR: −5 × NoiseType:SCN	0.0359	0.0154	**2.324**
SNR: −9.5 × NoiseType:SCN	0.0118	0.0154	0.766
SNR: −5 × vcScore	−0.0400	0.0082	**−4.858**
SNR: −9.5 × vcScore	−0.0289	0.0083	**−3.497**
SNR: −5 × PhraseRate	0.0337	0.0080	**4.21**
SNR: −9.5 × PhraseRate	0.0320	0.0078	**4.095**
SNR: −5 × Location:Beginning	−0.0350	0.0215	−1.63
SNR: −9.5 × Location:Beginning	−0.0462	0.0217	**−2.128**
SNR: −5 × Location:Final	0.0571	0.0262	**2.179**
SNR: −9.5 × Location:Final	0.0014	0.0260	0.055
SNR: −5 × ALINEScore	0.0171	0.0082	**2.094**
SNR: −9.5 × ALINEScore	0.0034	0.0080	0.429
NoiseType:SCN × PhraseRate	−0.0161	0.0064	**−2.513**
NoiseType:SCN × Location:Beginning	0.0305	0.0162	1.882
NoiseType:SCN × Location:Final	−0.0311	0.0211	−1.478
vcScore × Location:Beginning	0.3150	0.0935	**3.368**
vcScore × Location:Final	0.0141	0.0558	0.252
SNR: −5 × Target Frequency	0.0104	0.0087	1.205
SNR: −9.5 × Target Frequency	0.0205	0.0087	**2.361**

increase in the available acoustic information. We have replicated within-condition improvement/adaptation-like effects observed elsewhere (e.g., Bashford *et al.*, 1996), and have shown that the influence of Noise Type varies as a function of SNR. We also find more of an effect of the ALINE reduction measure when signals are relatively less obscured, and that listeners rely less on frequency-based information when speech is encountered in relatively more favorable SNRs.

C. Discussion

The preceding analysis largely echoes the findings described in experiment 1, providing additional support for the importance of different types of acoustic information beyond the contributions of syntax and semantics; support is illustrated most clearly through the increased accuracy observed by Condition (i.e., Visual Cloze < Silent Gap < Noise). We have seen that providing the acoustic context facilitates better target recognition than an orthographic transcription of that context, and that filling the target gap with *any* form of noise improves target recognition still. When exploring data from the Noise condition alone, we find a main effect for Trial where listeners are generally better able to identify targets as a condition continues. Such improvement is not surprising as listeners may have acclimatized to specific signal conditions and adapted strategies to compensate for distorted cues as exposure increased.

We find many Location-based effects in the present work, but while these effects hint at interesting modulations involving the position of missing or obscured phonetic information within a phrase, Location has been treated as a control variable in the present work due to limitations of the stimulus set. We therefore believe Location-based effects warrant further investigation in the future using a similar paradigm—though future efforts should strive for better balance in target location.

When focusing on signal manipulations, we see a remarkable similarity to previous speech-in-noise findings. Such similarity is not necessarily unexpected, but is important to recognize because these previous works were largely based on the manipulation and perception of careful laboratory speech. That is, it was unknown how such adverse conditions would influence the processing of casual speech, some of which has undergone extensive phonetic reduction.

For example, Bashford *et al.* (1996) describe SCN as providing listeners a relatively strong advantage compared to Flat Amplitude noise. We did not find a main effect of Noise Type, though we do see Noise Type interacting with ALINE Score. We found that information made available through SCN can also benefit the listener, though our data indicate a relatively steeper slope in this regard for Flat Amplitude noise. This result suggests that access to information about time varying intensity is increasingly helpful when targets more closely resemble citation forms (i.e., less reduction). Furthermore, intensity-related information appears to become less useful as targets deviate more radically from their citation forms, which seems reasonable because articulatory gestures of reduced magnitude necessarily smooth certain dynamics within spoken language, in turn rendering a less variable IC. Thus, as prosodic cues within the speech signal diminish by way of phonetic reduction, so too may the informativeness of time varying intensity.

Other results that bear on our findings are those of Benkí (2003a,b). Benkí described how SNR influenced the recognition of laboratory speech, and found that target identification improved with an increase in available spectral information. Our results confirm similar effects for casual speech, as we found that increased spectral information (including SS and *f*0 contour) made available through more favorable SNRs improves intelligibility even for reduced targets. This finding also indirectly supports our claim that diminished short time SS information was driving a decrease in target identification in experiment 1, especially compared to disruption of durational-cues which had only a marginal effect on participant accuracy.

IV. GENERAL DISCUSSION

This work supports the notion that improved target identification arises when listeners are afforded acoustic context in addition to an orthographic transcript. In fact, we see in experiment 2 that participant-accuracy was twice as high in the

Silent Gap condition than in the Visual Cloze condition. In the former condition listeners received not only syntactic and semantic information, but also information about speech rate (among other contextual cues). It is possible this rate information interacts with target duration to inform predictions about how many words and/or syllables are likely to fill each gap. Though, while target duration may be used to some extent, listeners also received information about segments adjacent to missing speech via partial formant transitions and other forms of longer distance coarticulation (e.g., Öhman, 1966). At the very least, such information might aid the listener in discerning which phones occur at target onsets and offsets. For such reasons, the present work cannot definitively distinguish the influence of target duration from other aspects of the greater acoustic context. However, Bernhard and Tucker (2015) manipulated target duration only in auditory Cloze tasks with minimal effects on intelligibility. This finding, taken together with the similar results when comparing our Stretched-Duration and Original conditions, makes it likely that the advantage of our Silent Gap condition over the Visual Cloze was largely due to other aspects of the acoustic context.

The ALINE similarity measures have proven quite beneficial in providing more detailed information about participants' phonological recognition in various adverse conditions (i.e., phonetically reduced productions *and* acoustically limited signal manipulations) than was previously available. Participants experience lexical pressures when filling target gaps, and such pressures may be further directed by the acoustic content of that target. For example, when a listener encounters the string, *the beard*, they may misidentify portions of that signal instead reconstructing *the bard*; a binary right/wrong measure would treat this response as wholly incorrect, while the ALINE Response Measure identifies the majority of component segments as nearly identical and scores it accordingly. The analyses described above support ALINE as sensitive to many of the same effects recognized through a binary measure, as well as others that might otherwise be missed.

We found numerous interactions of stimulus conditions with aspects of the informational environment (i.e., the non-acoustic context) and with our reduction measures. However, such interactions generally modulate rather than reverse any main trends of the stimulus-based effects of primary interest, and most had no obvious interpretation. One interaction we think is worth discussing is that of SNR and Target Frequency, because it may be at odds with some previous findings. We found no statistically reliable influence of frequency at the most favorable SNR (−5 dB), but we did observe improvements driven by frequency at SNRs of −9.5 and −14 dB. The work of Pollack *et al.* (1959, p. 276) found fairly consistent increased intelligibility with increased relative-frequency over a range of white-noise SNRs spanning −10 to +15 dB. While this apparent discrepancy deserves attention in future studies, we note that this interaction reached significance only in the model predicting our relatively more sensitive ALINE Response Measure, while results in Pollack *et al.* (1959) used binary correct scores. There are many other differences in methodology between the experiments that also might be at play. However, if these apparently discrepant findings are in fact stimulus-based and not artifacts of methodological differences, then it seems the role of frequency in processing may be more complicated than previously understood. If only detectable when exploring gradient levels of target similarity, it appears any boost in accuracy supplied by way of frequency may be limited. Indeed, if frequency-based contributions when processing casual speech are restricted to only partial target reconstruction, then such contributions cannot be wholly informative on their own, nor always usable. In other words, such frequency-based boosts to target reconstruction likely contribute somewhat to language redundancy, but are only partially helpful—and, moreover, are only available when context spurs listeners to access the relevant information.

This interpretation further supports the informativeness of even fragmentary acoustic information, and may be explained by way of the Smooth Signal Redundancy Hypothesis (Aylett and Turk, 2004, 2006). It seems likely that signal redundancy was contextually maximized in our most favorable SNR, where accuracy increased as relative noise levels decreased. When speakers were less able to capitalize on acoustic information through less favorable SNRs they were forced to increasingly rely upon non-acoustic contextual information (i.e., language redundancy), which in this case would include collocation frequencies. Simply stated, when lacking sufficient acoustic information the listener's best guess is likely to be something relatively common, so long as it fits well-enough with the contextual frame. This argument raises some interesting questions about the relative weightings and contributions of signal vs language redundancies, and the specifics of the sliding scale on which they trade off [see Seyfarth (2014) for some work on this topic]. This issue deserves increased attention in future research.

In the experiments described in this paper, we focused on four acoustic dimensions in varying signal conditions. Two of these, *duration patterns* and *acoustic context*, were reviewed above; the remaining candidates are discussed briefly below. We found that access to *time varying intensity* (via SCN) at any scale aids response accuracy. Specifically, targets were better identified when mixed with SCN than they were when replaced with either duration-matched silence or Flat Amplitude noise. Our findings largely replicate those described in Bashford *et al.* (1996) with regard to the benefits of noise-filled gaps, where SCN and Flat Amplitude noise both facilitate improved recognition over a silent gap, and where there is at least *some* advantage of intensity-following noise over stochastic white noise. Thus, as a general principle, it appears that filler-stimuli more like the originals are more likely to be restored.

We also found that increases in spectral information (provided via more favorable SNRs) were related to increased response fidelity, and that removing *SS* severely inhibits target recognition. This finding builds upon the work of Benkí (2003a) which also implicates the importance of spectral information. Unlike Benkí, we have tested the effectiveness of variation in *f*0 contour in the absence of SS information, finding only a slight (and non-significant) advantage in the *f*0 condition compared to pure SCN. This finding suggests the main detriment to intelligibility at less favorable SCN levels resulted from the removal of SS, and not perceived pitch.

However, as signal processing involved in the *f*0 condition led necessarily to modification of the instantaneous intensity of the signal, it is possible that the relatively minor difference between the *f*0 and pure SCN condition was influenced by loss of some intensity-following properties in the *f*0 signals. Furthermore, we acknowledge the possibility that the unintended gaps present within our *f*0 stimuli have somehow reduced this cue's relative informativeness.

We view this research as a first step in studying the informativeness of specific acoustic properties in the processing of reduced speech, and believe it is premature to make many strong claims. However, it seems clear that spectral information plays a substantial role in this processing, and that the intensity envelope alone is not particularly effective without spectral information. This conclusion might have been anticipated from studies exploring single-channel vocoding, but has been confirmed in the present work for fluent conversational speech.

We note that we did not replicate the range of Benkí's (2003b) error rates. Even our Visual Cloze test resulted in better target recognition than Benkí's most difficult condition, where listeners heard isolated nonsense syllables at −15 dB SNR. This difference in accuracy likely stems from our inclusion of various forms of contextual information. We note also that our most favorable SNR of −5 dB resulted in accuracy of only 59%, which is much lower than the 91.5% observed in our unmasked Original condition and the ∼80% in Benkí's −5 dB condition with clear speech words. Indeed, the accuracy for our reduced signals in −5 dB of noise was roughly on par with Benkí's nonsense syllables under comparable masking (∼58%) despite supplying our participants with various forms of contextual information. This difference suggests the influence of noise masking is quite different on clear speech than for our reduced speech materials, as the reduced forms were effectively processed with the same degree of accuracy as meaningless strings of phonemes.

Perhaps messages conveyed through reduced forms can best be thought of as "fragile," where many productions may be reduced to a point where *the bare minimum* acoustic information required for effective processing has been retained. In terms of the Smooth Signal Redundancy Hypothesis, our more reduced forms likely contained relatively small amounts of signal redundancy, while Benkí's non-reduced nonsense syllables were relatively rich in this regard. It makes sense that our reduced signals were well recognized when presented without noise-masking because, along with the frame/context, they maintained a sufficient balance of language and signal redundancies. Minimal masking, therefore, seems to go a long way with reduced speech forms, as accuracy was severely reduced with the addition of relatively little noise. It also makes sense then that Benkí's signals were generally better recognized in more intense noise despite impoverished language redundancy; the increased signal redundancy likely made these signals more resistant to noise-masking. At this point it is impossible to state the degree to which different accuracies across studies were driven by differences in speech clarity (i.e., signal redundancy) and NACP (i.e., language redundancy), but we feel this is an important question for future research.

V. CONCLUSION

This work contributes to a large body of research exploring the relationship between acoustic information and various contextual factors that can inform and influence language use (e.g., Semantics: van de Ven *et al.*, 2011; Syntax: Ernestus *et al.*, 2002; Sociolinguistic: Niedzielski, 1999; Psycholinguistic: Hay *et al.*, 2017). We have extended previous laboratory-speech focused perception studies into the realm of casual speech, and have also introduced a novel method to quantify phonetic reduction in spoken language. Use of ALINE provides a more precise and informative description of reduced speech than previous metrics based primarily on deletions, and the use of this algorithm was strongly supported by multiple statistical analyses described within this work. Our data indicate that acoustic context aids language processing above and beyond information afforded through an orthographic transcription. We have also explored the relative informativeness of four acoustic dimensions within conversational speech, and their general contributions to the processing of reduced speech forms. We believe the relationships between acoustic and varied contextual cues are important and require further attention.

ACKNOWLEDGMENTS

We would like to thank the speaker whose recorded conversation provided our seed-stimuli, as well as all participants who took part in the two experiments. This work was supported in part by grants from the Social Sciences and Humanities Research Council of Canada to R.G.P. (Grant No. 752-2014-1438) and to B.V.T. (Grant No. 410-2011-0030), and from the Natural Sciences and Engineering Research Council of Canada to G.K.

[1]See supplementary material available at https://doi.org/10.1121/1.5031123 for example audio files as well as transcriptions of all stimuli.

[2]All targets, single- and multi-word, were treated as collocations; frequencies were extracted as the raw frequency of each string within the corpus.

[3]One reviewer draws attention to the fact that speech perception in adverse conditions differs between monolingual and early bilingual listeners. To test for any unknown influence in this way we have re-fit the final model to a subset of the dataset excluding the multilingual speakers. Model outputs were nearly identical (the modest interaction between vcScore and Order 2 did not maintain significance), suggesting multilingual speakers were not performing differently than the monolinguals.

[4] We did not, however, address the location balance issue in the next experiment because we thought it more important at this stage to maintain direct comparability with different signal manipulation conditions across the two studies.

Adank, P., and Janse, E. (**2009**). "Perceptual learning of time-compressed and natural fast speech," J. Acoust. Soc. Am. **126**(5), 2649–2659.

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

Aylett, M., and Turk, A. (**2006**). "Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei," J. Acoust. Soc. Am. **119**(5), 3048–3058.

Baayen, R. H. (**2008**). *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R* (Cambridge University Press, New York).

Barreda, S., and Nearey, T. M. (**2012**). "The direct and indirect roles of fundamental frequency in vowel perception," J. Acoust. Soc. Am. **131**(1), 466–477.

Bashford, J. A., Warren, R. M., and Brown, C. A. (**1996**). "Use of speech-modulated noise adds strong 'bottom-up' cues for phonemic restoration," Percept. Psychophys. **58**(3), 342–350.
Bates, D., Maechler, M., Bolker, B., and Walker, S. (**2014**). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-6. http://CRAN.R-project.org/package=lme4 (Last viewed September 2017).
Benkí, J. R. (**2003a**). "Analysis of English nonsense syllable recognition in noise," Phonetica **60**, 129–157.
Benkí, J. R. (**2003b**). "Quantitative evaluation of lexical status, word frequency, and neighborhood density as context effects in spoken word recognition," J. Acoust. Soc. Am. **113**(3), 1689–1705.
Bernhard, D., and Tucker, B. (**2015**). "The effects of duration on human processing of reduced speech," Canadian Acoust. **43**(3), 122–123.
Boersma, P., and Weenink, D. (**2012**). "Praat: Doing phonetics by computer" [Computer program], Version 5.1.43, http://www.praat.org/ (Last viewed September 2017).
Cutler, A. (**1998**). "The recognition of spoken words with variable representations," in *Sound Patterns of Spontaneous Speech* (La Baume-les-Aix, France).
Cutler, A. (**2005**). "Lexical stress," in *The Handbook of Speech Perception*, edited by D. B. Pisoni and R. E. Remez (Blackwell, Oxford).
Davies, M. (**2008**). The Corpus of Contemporary American English: 450 million words, 1990–present. Available at http://corpus.byu.edu/coca/ (Last viewed January 2014).
Downey, S. S., Hallmark, B., Cox, M. P., Norquest, P., and Lansing, J. S. (**2008**). "Computational feature-sensitive reconstruction of language relationships: Developing the ALINE distance for comparative historical linguistic reconstruction," J. Quan. Ling. **15**(4), 340–369.
Dupoux, E., and Green, K. (**1997**). "Perceptual adjustment to highly compressed speech: Effects of talker and rate changes," J. Exp. Psychol. **23**(3), 914–927.
Ernestus, M., Baayen, R. H., and Schreuder, R. (**2002**). "The recognition of reduced word forms," Brain Lang. **81**(1–3), 162–173.
Ernestus, M., and Warner, N. (**2011**). "An introduction to reduced pronunciation variants," J. Phonetics **39**(3), 253–260.
Fry, D. B. (**1958**). "Experiments in the perception of stress," Lang. Speech **1**(2), 126–152.
Hay, J., Podlubny, R., Drager, K., and McAuliffe, M. (**2017**). "Car-talk: Location-specific speech production and perception," J. Phonetics **65**, 94–109.
Hillenbrand, J. M. (**2003**). "Some effects of intonation contour on sentence intelligibility," J. Acoust. Soc. Am. **114**(4), 2338.
Johnson, K. (**1990**). "The role of perceived speaker identity in F0 normalization of vowels," J. Acoust. Soc. Am. **88**(2), 642–654.
Johnson, K. (**2004**). "Massive reduction in conversational American English," in *Proceedings of the 1st Session of the 10th International Symposium on Spontaneous Speech: Data and Analysis*, edited by K. Yoneyama and K. Maekawa (The National International Institute for Japanese Language, Tokyo, Japan), pp. 29–54.
Kingston, J., and Diehl, R. (**1994**). "Phonetic knowledge," Language **70**, 419–454.
Klatt, D. H. (**1976**). "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence," J. Acoust. Soc. Am. **59**(5), 1208–1221.
Kondrak, G. (**2000**). "A new algorithm for the alignment of phonetic sequences," in *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, pp. 288–295.
Kondrak, G. (**2003**). "Phonetic alignment and similarity," Comp. Humanities **37**(3), 273–291.
Labov, W. (**1972**). *Sociolinguistic Patterns* (University of Pennsylvania Press, Philadelphia, PA), 86 pp.
Lindblom, B. (**1990**). "Explaining phonetic variation: A sketch of the H & H theory," in *Speech Production and Speech Modeling*, edited by W. J. Hardcastle and A. Marchal (Kluwer Academic, Dordrecht, The Netherlands), pp. 403–439.
Lindblom, B. (**1996**). "Role of articulation in speech perception: Clues from production," J. Acoust. Soc. Am. **99**(3), 1683–1692.
Mani, I., Yeh, A., and Condon, S. (**2008**). "Learning to match names across languages," in *Proceedings of the Workshop Multi-source Multilingual Information Extraction and Summarization*: *Coling*, pp. 2–9. Available at http://www.aclweb.org/anthology/W08-1402 (Last viewed February 2018).
Mehta, G., and Cutler, A. (**1988**). "Detection of target phonemes in spontaneous and read speech," Lang. Speech **31**(2), 135–156.
Mitterrer, H. (**2005**). "Rotation_plus_plus.praat" (computer script). URL http://www.holgermitterer.eu/HM/rotation_plus_plus.praat (Last viewed January 2014).
Moon, S. J., and Lindblom, B. (**2003**). "Two experiments on oxygen consumption during speech production: Vocal effort and speaking tempo," in *Proceedings of the Fifteenth International Congress of Phonetic Sciences*, pp. 3129–3132.
Niedzielski, N. (**1999**). "The effect of social information on the perception of sociolinguistic variables," J. Lang. Social Psychol. **18**(1), 62–85.
Öhman, S. E. (**1966**). "Coarticulation in VCV utterances: Spectrographic measurements," J. Acoust. Soc. Am. **39**(1), 151–168.
Pollack, I., Rubenstein, H., and Decker, L. (**1959**). "Intelligibility of known and unknown message sets," J. Acoust. Soc. Am. **31**, 273–279.
Pollack, I., Rubenstein, H., and Decker, L. (**1960**). "Analysis of incorrect responses to an unknown message set," J. Acoust. Soc. Am. **32**, 454–457.
Port, R. F. (**1979**). "The influence of tempo on stop closure duration as a cue for voicing and place," J. Phonetics **7**, 45–56.
R Core Team (**2017**). "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/ (Last viewed September 2017).
Schroeder, M. R. (**1968**). "Reference signal for signal quality studies," J. Acoust. Soc. Am. **44**, 1735–1736.
Seyfarth, S. (**2014**). "Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation," Cognition **133**(1), 140–155.
Shannon, R. V., Fu, Q.-J., and Galvin, J., III (**2004**). "The number of spectral channels required for speech recognition depends on the difficulty of the listening situation," Acta Oto-Laryngologica **124**(0), 50–54.
Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**(5234), 303–304.
Souza, P., and Rosen, S. (**2009**). "Effects of envelope bandwidth on the intelligibility of sine- and noise-vocoded speech," J. Acoust. Soc. Am. **126**(2), 792–805.
Szakay, A. (**2008**). *Ethnic Dialect Identification in New Zealand: The Role of Prosodic Cues* (VDM Publishing, Saarbrücken, Germany).
Taylor, W. L. (**1953**). "'Cloze procedure': A new tool for measuring readability," Journal. Mass Commun. Quart. **30**, 415–433.
Tremblay, A., and Tucker, B. V. (**2011**). "The effects of N-gram probabilistic measures on the processing and production of four-word sequences," Mental Lexicon **6**(2), 302–324.
Tucker, B. V. (**2007**). "Spoken word recognition of the reduced American English flap," Ph.D. thesis, The University of Arizona. http://hdl.handle.net/10150/194987 (Last viewed September 2017).
Tucker, B. V. (**2011**). "The effect of reduction on the processing of flaps and /g/in isolated words," J. Phonetics **39**(3), 312–318.
Tucker, B. V., and Ernestus, M. (**2016**). "Why we need to investigate casual speech to truly understand language production, processing and the mental lexicon," Mental Lexicon **11**(3), 375–400.
Vaissière, J. (**2005**). "Perception of intonation," in *The Handbook of Speech Perception* (Blackwell, London), pp. 236–263.
van de Ven, M., Ernestus, M., and Schreuder, R. (**2012**). "Predicting acoustically reduced words in spontaneous speech: The role of semantic/syntactic and acoustic cues in context," Lab. Phonology **3**, 455–481.
Van de Ven, M., Tucker, B. V., and Ernestus, M. (**2011**). "Semantic context effects in the comprehension of reduced pronunciation variants," Mem. Cogn. **39**(7), 1301–1316.
Van Tasell, D. J., Soli, S. D., Kirby, V. M., and Widin, G. P. (**1987**). "Speech waveform envelope cues for consonant recognition," J. Acoust. Soc. Am. **82**(4), 1152–1161.
Wagner, R. A., and Fischer, M. J. (**1974**). "The string-to-string correction problem," J. Assoc. Comput. Mach. **21**(1), 168–173.
Warner, N. (**2012**). "Methods for studying spontaneous speech," in *The Oxford Handbook of Laboratory Phonology*, edited by C. Fougeron and M. Huffman (Oxford University Press, Oxford), pp. 621–633.
Warner, N., and Tucker, B. V. (**2011**). "Phonetic variability of stops and flaps in spontaneous and careful speech," J. Acoust. Soc. Am. **130**(3), 1606–1617.
Warren, R. M. (**1970**). "Perceptual restoration of missing speech sounds," Science **167**(3917), 392–393.
Wayland, S. C., Miller, J. L., and Volaitis, L. E. (**1994**). "The influence of sentential speaking rate on the internal structure of phonetic categories," J. Acoust. Soc. Am. **95**(5), 2694–2701.
Zhao, Y. (**1997**). "The effects of listeners' control of speech rate on second language comprehension," Appl. Linguist. **18**(1), 49–68.