LANGUAGE, COGNITION AND NEUROSCIENCE, 2017
VOL. 32, NO. 10, 1344–1356
https://doi.org/10.1080/23273798.2017.1354129

Routledge
Taylor & Francis Group

REGULAR ARTICLE

Word segmentation from noise-band vocoded speech

Tina M. Grieco-Calub[a], Katherine M. Simeon[a], Hillary E. Snyder[a] and Casey Lew-Williams[b]

[a]The Roxelyn & Richard Pepper Department of Communication Sciences and Disorders, Northwestern University, Evanston, IL, USA;
[b]Department of Psychology, Princeton University, Princeton, NJ, USA

ABSTRACT
Spectral degradation reduces access to the acoustics of spoken language and compromises how learners break into its structure. We hypothesised that spectral degradation disrupts word segmentation, but that listeners can exploit other cues to restore detection of words. Normal-hearing adults were familiarised to artificial speech that was unprocessed or spectrally degraded by noise-band vocoding into 16 or 8 spectral channels. The monotonic speech stream was pause-free (Experiment 1), interspersed with isolated words (Experiment 2), or slowed by 33% (Experiment 3). Participants were tested on segmentation of familiar vs. novel syllable sequences and on recognition of individual syllables. As expected, vocoding hindered both word segmentation and syllable recognition. The addition of isolated words, but not slowed speech, improved segmentation. We conclude that syllable recognition is necessary but not sufficient for successful word segmentation, and that isolated words can facilitate listeners' access to the structure of acoustically degraded speech.

ARTICLE HISTORY
Received 8 July 2016
Accepted 2 July 2017

KEYWORDS
Word segmentation; noise-band vocoding; spectral degradation; speech rate; isolated words

Introduction

Language is replete with structure, and normal-hearing listeners are equipped to detect it. For decades, researchers have been drawn to understanding how learners discover words in continuous speech, an inherently challenging task given that connected speech has no reliable pause-defined cues to word boundaries (Romberg & Saffran, 2010; Saffran, Aslin, & Newport, 1996). One cue that listeners can use to segment words is the co-occurrence relation between sounds and syllables, often referred to as transitional probability (TP). For example, given syllables X and Y, learners are sensitive to the probability with which X will transition to Y (and vice versa), and this domain-general sensitivity to TPs has been demonstrated in several perceptual domains (Aslin, Saffran, & Newport, 1998; Fiser & Aslin, 2002; Kirkham, Slemmer, & Johnson, 2002; Lew-Williams, Pelucchi, & Saffran, 2011). This mechanism is posited to not only enable language learning in infants but also to facilitate word segmentation in adults.

An underlying assumption in previous research is that successful word segmentation from contiguous speech hinges both on accurate recognition of individual speech units, such as phonemes and syllables, as well as on tracking of syllable sequences over time. Insufficient spectral fidelity, however, compromises successful discrimination of speech. Environmental factors (i.e. background noise), biological differences (i.e. hearing loss), and the use of hearing devices (i.e. cochlear implants) restrict access to spectral cues that are important for speech unit recognition (Donaldson & Kreft, 2006; Gordon-Salant, Yeni-Komshian, Fitzgibbons, & Cohen, 2015; Munson, Donaldson, Allen, Collison, & Nelson, 2003; Xu & Pfingst, 2008; Zhou, Xu, & Lee, 2010). When recognition is impaired, there are consequences for processing both within and beyond the domain of language. A range of studies using behavioural, physiological, and neuroimaging methods provide robust evidence that encoding degraded speech is also cognitively demanding (Davis & Johnsrude, 2007; Mattys, Davis, Bradlow, & Scott, 2012; Rönnberg et al., 2013). For example, the use of dual-task paradigms has documented robust declines in secondary task performance as speech becomes more degraded in a primary task (e.g. Broadbent, 1958; Downs & Crum, 1978; Grieco-Calub, Ward, & Brehm, 2017; Pals, Sarampalis, & Başkent, 2013; Pichora-Fuller, Schneider, & Daneman, 1995; Rabbitt, 1966; Rakerd, Seitz, & Whearty, 1996; Sarampalis, Kalluri, Edwards, & Hafter, 2009; Ward, Shen, Souza, & Grieco-Calub, 2017). Pupillometry studies have also shown increased cognitive effort associated with processing of degraded speech input (e.g. Winn, Edwards, & Litovsky, 2015). These findings are supported by neuroimaging work showing more

CONTACT Tina M. Grieco-Calub ✉ tinagc@northwestern.edu

distributed neural activation during tasks that involve recognition of degraded vs. unprocessed speech (Hervais-Adelman, Carlyon, Johnsrude, & Davis, 2012; Obleser, Wise, Dresner, & Scott, 2007; Wild et al., 2012). Given the cognitive demands of listening to degraded speech, and given that listeners have different levels of access to the acoustic subtleties in speech, there may be important individual differences in the ability to track relations between speech units (such as syllables) across time.

Following the prediction that spectrally degraded speech will interfere with word segmentation, the question arises as to whether clarity is a prerequisite for successful detection of co-occurrence, or if listeners can rely on other cues in the input. In natural speech, TPs are just one cue to structure, and other cues are readily available (Johnson & Jusczyk, 2001). For example, people often use isolated words, which affect the prosody and time course of incoming information by providing silent pauses and reducing the rate of incoming speech. Inserting pauses in continuous speech isolates certain sequences in time, providing an overt word boundary that could potentially prevent difficulty with segmentation. Previous studies have shown that isolated words are a common feature of child-directed speech and facilitate language learning (Aslin, Woodward, LaMendola, & Bever, 1996; Brent & Siskind, 2001; Church, Bernhardt, Shi, & Pichora-Fuller, 2005; Jusczyk, 1999; Jusczyk & Aslin, 1995; Lew-Williams et al., 2011). A study by Brent and Siskind (2001) showed that 9% of mothers' child-direct utterances contained isolated words, and that the frequency of hearing a word in isolation was a significant and unique predictor of later knowledge of that word. While isolated words are not required for speech segmentation, they may be able to serve as a temporal and/or prosodic cue that enhances a listener's ability to track sequential statistics across time (Bortfeld, Morgan, Golinkoff, & Rathbun, 2005; Cunillera, Càmara, Laine, & Rodríguez-Fornells, 2010b; Cunillera, Laine, & Rodríguez-Fornells, 2016; Lew-Williams et al., 2011). We predicted that isolated words would highlight sequences in an otherwise continuous stream of spectrally degraded speech and, thereby, support segmentation.

Another temporal cue that may benefit listeners in degraded listening conditions is speech rate, or the number of morphemes or words produced per unit of time. Recent work by Palmer and Mattys (2016) demonstrated that slower syllable rates improved performance on a segmentation task in adults, even when they controlled for the total duration of the speech stream. They showed that adults who were familiarised to an artificial language at a slow rate (i.e. 2.27 syllables per second) correctly segmented a greater proportion of sequences than adults who were familiarised to the language at a normal or fast rate (4.17 or 7.45 syllables per second, respectively). In a follow-up experiment, adults performed either a phonological or visual two-back task while being familiarised to the artificial language. They found that the inclusion of either task eliminated the benefit of the slower speech rate, suggesting that increased cognitive load impaired segmentation. Given these findings, we predicted that reducing the speech rate would support listeners' abilities to represent individual units in degraded listening conditions and, in turn, support the tracking of units over time.

The present study was designed to test the hypotheses that successful word segmentation is contingent on full access to acoustical speech cues, and that temporal cues – such as isolated words and reduced speech rate – aid learning from spectrally degraded speech. In Experiment 1, normal-hearing adults participated in word segmentation and syllable recognition tasks using speech that was either unprocessed or spectrally degraded by a 16-channel (16-ch) or 8-channel (8-ch) noise-band vocoder. Adults listened to an artificial language consisting of four trisyllabic nonsense words (Lew-Williams & Saffran, 2012), and were then asked to distinguish between previously heard trisyllabic sequences vs. trisyllabic sequences that never occurred in the speech stream. We predicted that (1) successful word segmentation from the artificial speech stream will be dependent on spectral fidelity; (2) successful segmentation will rely on accurate recognition of individual speech units; and (3) the addition of temporal cues to the speech stream will facilitate segmentation. Together, these experiments provide insight into the effects of degraded speech on the detection of patterns, the cues that support word segmentation, and the scalability of word segmentation tasks to a previously untested dimension of natural learning conditions.

Experiment 1

Methods

Participants: Participants were 60 native English-speaking adults (mean = 22.0 years, range = 18–33 years). All participants reported normal hearing and no significant medical or otologic history. Participants completed an informed consent process prior to participation and were compensated for their time. Five additional participants were tested but excluded from analyses due to equipment malfunction (2), the presence of tinnitus (1), previous exposure to the stimuli (1), and performing more than 3 standard deviations below the mean on the word segmentation task (1). All procedures were approved by the Institutional Review Board of Northwestern University.

Stimuli: A native English-speaking female recorded 24 CV syllables. Twelve syllables (/bi/, /bu/, /da/, /do/, /go/, /ku/, /la/, /pa/, /pi/, /ro/, /ti/, /tu/) were concatenated to generate a pause-free, monotone, artificial speech stream. The same speech stimuli were used in Lew-Williams and Saffran (2012). To maintain natural coarticulation, each syllable was recorded in the middle of a three-syllable sequence, in every possible coarticulation context. Middle syllables were spliced using Praat (Boersma & Weenink, 2009) to generate four trisyllabic nonsense words (Table 1) that were repeated in quasi-random order to create a continuous speech stream with consistent rate (3.1 syllables/second) and pitch (F0 = 196 Hz). Successive syllables within the four trisyllabic words had TPs of 1.0. Two of the words were high-frequency words, appearing twice as often as two low-frequency words (70 vs. 35, respectively). The duration of the concatenated speech stream consisted of 210 words, was 3 minutes 15 seconds, and contained no acoustic cues to word boundaries. The test stimuli were the two low-frequency *words* from the familiarisation phase (TP = 1.0); two frequency-matched *part-words*, consisting of the last syllable of one high-frequency word and the first two syllables of the other high-frequency word (TP = 0.5); and two trisyllabic *non-words*, consisting of syllables that were used in the familiarisation language but never co-occurred (*tirodo*, *robaku*, *lagupi*, *dolati*; TP = 0). There were two counterbalanced artificial languages, such that each test item was a word for half of participants and a part-word for the other half (Table 1). The remaining 12 syllables (/bo/, /du/, /ga/, /gu/, /ka/, /ki/, /li/, /lo/, /po/, /ri/, /ru/, /ta/) were created to ensure that each consonant and vowel occurred an equal number of times during the syllable recognition task, which contained all 24 syllables.

Noise-band vocoding: Spectral degradation of speech stimuli was accomplished by noise-band vocoding using TigerCIS software (publicly available). Noise-band vocoding provides a way to systematically vary the amount of spectral information by indicating the number of independent frequency channels while maintaining the slow-changing temporal and amplitude features of the speech waveform. To create noise-band vocoded stimuli for the present study, the auditory stimuli were pre-filtered to include frequencies between 200 and 7000 Hz and then subsequently divided into 16 or 8 independent frequency channels using the Greenwood function, which approximates the frequency distribution of the basilar membrane of the inner ear (Greenwood, 1990). The selection of 16-ch and 8-ch conditions was based on both prior work as well as extensive pilot testing. Prior work suggests that normal-hearing adults encode speech with as little as four spectral channels (Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995), but these studies often require extended training and previously known speech. Pilot testing confirmed this: listeners were unable to segment sequences from an artificial language with spectral fidelity of four channels. Better performance was observed with eight channels, which also approximates the average spectral resolution of most cochlear implant users. A 16-ch condition was included to provide more spectral fidelity, short of the fine structure that improves speech perception. Within each frequency channel (16 or 8), the temporal envelope was extracted using half-wave rectification and low-pass filtering at 400 Hz (24 dB/octave slope), which is the process by which the temporal fine structure is stripped from the signal. The extracted envelope from each channel was then multiplied by bandpass noise with the same frequency bandwidth as the frequency channel. Finally, channel-specific output was summed and converted to a digital signal.

Table 1. Word segmentation task.

Familiarisation		Test	
Language 1	Language 2	Language 1	Language 2
pabiku[a]	tudaro[a]	pabiku[c]	pabiku[d]
tibudo[a]	pigola[a]	tibudo[c]	tibudo[d]
golatu[b]	bikuti[b]	tudaro[d]	tudaro[c]
daropi[b]	budopa[b]	pigola[d]	pigola[c]

Note: Pronunciation: /a/="ah"; /i/="ee"; /o/="oh"; /u/="oo".

[a] Low-frequency.

[b] High-frequency.

[c] Word.

[d] Part-word.

Procedure: Participants were randomly assigned to one of three listening conditions: unprocessed speech, 16-ch noise-band vocoded speech, or 8-ch noise-band vocoded speech. In each listening condition, participants performed a word segmentation task and a syllable recognition task. Participants were seated at a computer, in front of Genelec 8030A loudspeakers. *Word segmentation:* Before the task, participants were told to listen to a "set of sounds" presented through the loudspeakers. During familiarisation to the artificial language, speech was presented at 65 dB SPL, and the text, "Listen to the sound clip", was presented on the computer monitor. The artificial speech stream was repeated twice for a total of 6 minutes 30 seconds. At test, participants heard a pair of isolated trisyllabic words, presented sequentially with 500 milliseconds of silence between each word, and had 3 seconds to press a button indicating the word that they perceived to be from the artificial language. In a two-alternative forced-choice task, participants were tested on *words* vs. *non-*

words (12 trials) and *part-words* vs. *non-words* (12 trials), for a total of 24 trials presented in two blocks of 12 trials, with trial order counterbalanced. In several previous investigations of word segmentation using artificial languages, participants were tested directly on words vs. part-words. However, pilot testing in both the unprocessed and degraded listening conditions revealed that participants were consistently at chance performance when these two word types were directly compared. Thus, in our forced-choice test phase, words were tested against non-words, and part-words were tested against to non-words, thereby determining whether or not participants showed success in discriminating sequences that had vs. had not been heard previously. Trials were quasi-randomly presented, with a 1-min break between each block of 12 test trials. This design is consistent with published methods that revealed successful word segmentation from clear speech in normal-hearing adults (Saffran, Newport, & Aslin, 1996). *Syllable recognition task*: The computer screen displayed a custom graphical user interface with an 8 × 3 grid of pushbuttons designed in Matlab. Each pushbutton was assigned a label corresponding to one of the 24 syllables (see Stimuli). On each trial, participants heard a syllable and were asked to select the pushbutton that corresponded to the syllable. Presentation order was randomised without replacement. Each syllable was presented twice, resulting in 48 trials.

Statistical analysis: Accuracy on each trial in the word segmentation and syllable recognition tasks was binary – correct and incorrect responses were coded as 1 and 0, respectively – and is reported as percent correct in the Results. Logistic mixed effects modelling using the *lme4* package in R (R Core Team, 2012) was used to statistically evaluate accuracy, the dependent variable, in each task. For the word segmentation task, the fixed effects included condition (categorical variable: unprocessed, 16-ch, 8-ch) and word type (categorical variable: part-words, words). The random effects structure was designed to account for variability associated with the participants across each condition and test items. Specifically, we included intercepts of participants and test items as well as slopes of condition and word type for test items ($N = 8$; 4 test sequences × 2 non-word competitors). For the syllable recognition task, condition was the only fixed effect and the random effects structure included intercepts of participants and test items as well as the slopes of condition for test items ($N = 24$). In each model, condition was coded to test successive differences using the contr.sdif function: the first contrast compared the difference between the unprocessed and 16-ch condition; the second contrast compared the difference between the 16-ch and 8-ch conditions.

Results and discussion

Participants were tested on their word segmentation ability by selecting familiar vs. novel trisyllabic words. On average, participants who were exposed to the speech stream with full spectral representation (unprocessed condition) distinguished familiar from novel words at accuracies that were significantly above chance performance (unprocessed: 71.4% ± 4.9%, mean ± SE; chance = 50%, $t[19] = 4.4$, $p < .001$; Figure 1 (A)). Mean accuracy was not statistically different from chance performance in the 16-ch condition (59% ± 5%; $t[19] = 1.8$, $p = .09$) or 8-ch condition (55.2% ± 3.7%; $t[19] = 1.4$, $p = .18$). The results of logistic mixed effects modelling showed that participants in the unprocessed condition segmented words statistically better than participants in the 16-ch condition ($\beta = -0.68$, $z = -2.05$, $p < .05$). Participants in the 16-ch and 8-ch conditions, respectively, did not differ in their word segmentation ability ($\beta = -0.18$, $z = -0.54$, $p = .59$). Across the conditions, participants segmented the two word types, words and part-words, equivalently ($\beta = 0.10$, $z = 0.68$, $p = .49$), suggesting that higher TPs within the trisyllabic sequences (i.e. TP = 1 vs. TP = 0.5) did not confer an additional benefit for segmentation (Figure 2(A)).

Consistent with prior work (e.g. Shannon et al., 1995), syllable recognition varied systematically with spectral fidelity (Figure 3(A)). The results of the logistic mixed effects model showed that participants in the unprocessed condition identified significantly more syllables than participants in the 16-ch condition ($\beta = -1.94$, $z = -3.1$, $p = .002$). Additionally, participants in the 16-ch condition identified a significantly greater number of syllables than participants in the 8-ch condition ($\beta = -2.59$, $z = -6.18$, $p < .001$).

Central to our investigation is whether syllable recognition supports word segmentation. To statistically test the relation between participants' syllable recognition and word segmentation, pairwise comparisons were implemented. Results showed that participants' ability to recognise individual syllables did not statistically relate to their ability to segment words from the artificial language (unprocessed: Pearson's $R = -.124$, $p = .60$; 16-ch: Pearson's $R = .12$, $p = .62$; Pearson's $R = -.16$, $p = .5$). In the unprocessed condition, participants were at ceiling performance on the syllable recognition task, and all but 3 (17/20) participants segmented words from the artificial language at percentages that were statistically greater than chance performance. Participants in the 8-ch condition showed poor accuracy both in syllable recognition and word segmentation. In contrast, performance on the two tasks diverged in the 16-ch condition: although the majority of participants

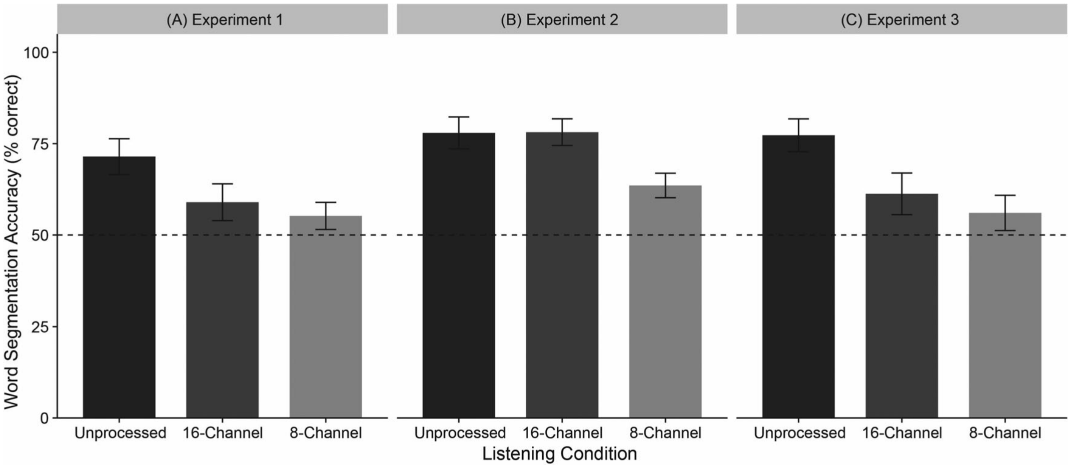

Figure 1. Mean (±SE) accuracy of word segmentation. Accuracy is defined as the percent of correctly selected syllable sequences from the artificial language on a two-alternative-forced-choice (2-AFC) task. 16-channel: 16-channel noise-band vocoded; 8-channel: 8-channel noise-band vocoded. The dotted line represents chance performance.

recognised syllables (e.g. 16/20 participants had accuracies of >87.5%), they were unsuccessful at segmenting words from the speech stream. These results highlight the fact that although individual syllable recognition is necessary for word segmentation (as evidenced in the 8-ch condition), it is not sufficient. Additionally, the results raise the possibility that word segmentation under degraded conditions may not be related to recognition of individual syllables, but rather to listeners' inability to track the statistics of the artificial language due to the cognitive load imposed by listening through degraded auditory input (e.g. Mattys et al., 2012; Rönnberg et al., 2013; Wild et al., 2012).

The objective of Experiment 1 was to determine if word segmentation is dependent on spectral fidelity. The results from the noise-band vocoded conditions suggest that spectral degradation disrupts adults' abilities to segment words from contiguous speech. This result provides the first evidence that the ability to track syllable sequences is impaired in degraded spectral conditions, even in the presence of intact syllable recognition. The dissociation between "low-level" syllable recognition and "higher-level" word segmentation in the 16-ch condition suggests that the locus of difficulty when listening to degraded speech is tracking units over time, not in recognising individual syllables. The results also raise the possibility that the cognitive processes involved in resolving degraded speech overlap with those involved in segmenting recurring syllable sequences. This finding is relevant to young children with hearing loss, whose performance on clinical tests of speech perception may underestimate their ability to track novel syllable sequences in natural discourse. Ultimately, this may be a source of individual variability in spoken language outcomes in this population because individual differences in word segmentation contribute to variability in processing of higher-order structures in language (Misyak & Christiansen, 2012).

There are two primary ways of interpreting the findings from Experiment 1. One possibility is that adults are unable to segment words from degraded speech. This is unlikely, as adults have been shown to adapt to vocoded speech over time and demonstrate successful learning (Hervais-Adelman, Davis, Johnsrude, Taylor, & Carlyon, 2011). Alternatively, adults may find it taxing to segment words from degraded speech, but be able to rely on other features of natural speech in the presence of degraded input to support the detection of structure in language. One such feature is the presence of isolated words, which provide a salient cue to word boundaries in natural speech. In Experiment 2, we inserted silent pauses before and after a subset of tokens of low-frequency words in the speech stream, thus providing a prosodic/temporal cue that could facilitate successful segmentation.

Experiment 2

Methods

Participants: Participants were 60 native English-speaking adults (mean = 22.4 years, range = 18–34 years). Inclusion criteria and consent procedures were consistent with those described in Experiment 1.

Stimuli: Experiment 2 utilised the same artificial speech stream as Experiment 1. After generating the pause-free speech stream, 20% of the low-frequency sequences (i.e.

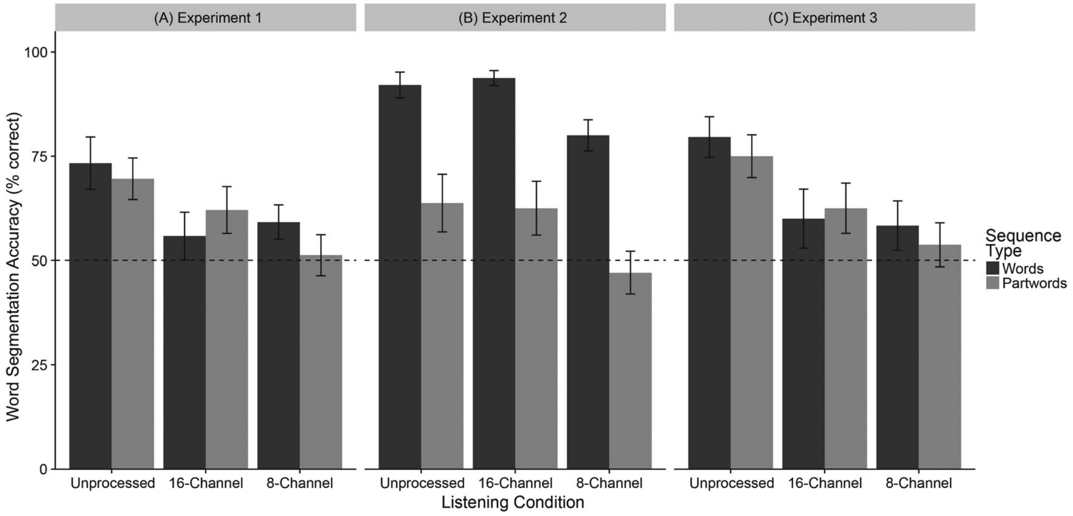

Figure 2. Mean (±SE) accuracy for segmenting words (*dark grey*) and part-words (*light grey*). Accuracy is defined as the percent of correctly selected syllable sequences from the artificial language on a two-alternative-forced choice (2-AFC) task. The dotted line represents chance performance.

TP = 1.0; sequences that served as "words" in the test phase) were preceded and followed by 500-millisecond silent pauses. This provided overt word boundaries (i.e. isolated words) for a subset of sequences (*N* = 14/70 *words*) in the stream. The targeted low-frequency words were quasi-randomly selected to ensure that they were distributed throughout the entire language, thus providing no consistent rhythmic pattern. The rationale for isolating low-frequency words (vs. other sequences in the speech stream) was to test if listeners could use their successful segmentation of the low-frequency words to segment other sequences from the speech stream. Alternatively, segmentation of low-frequency words alone would suggest that any benefits of this temporal cue were specific to the targeted sequences.

A range of pause lengths has been used in prior research on word segmentation (e.g. Lew-Williams et al., 2011; Peña, Bonatti, Nespor, & Mehler, 2002; Saffran & Thiessen, 2003), and in creating our stimuli, we selected 500 milliseconds as the optimal length for acoustically demarcating the low-frequency words. With the inclusion of pauses, the duration of the concatenated speech stream was 3 minutes 31 seconds, which is 16 seconds longer than the speech stream in Experiment 1. The speech stream was repeated twice for a total duration of 7 minutes 2 seconds.

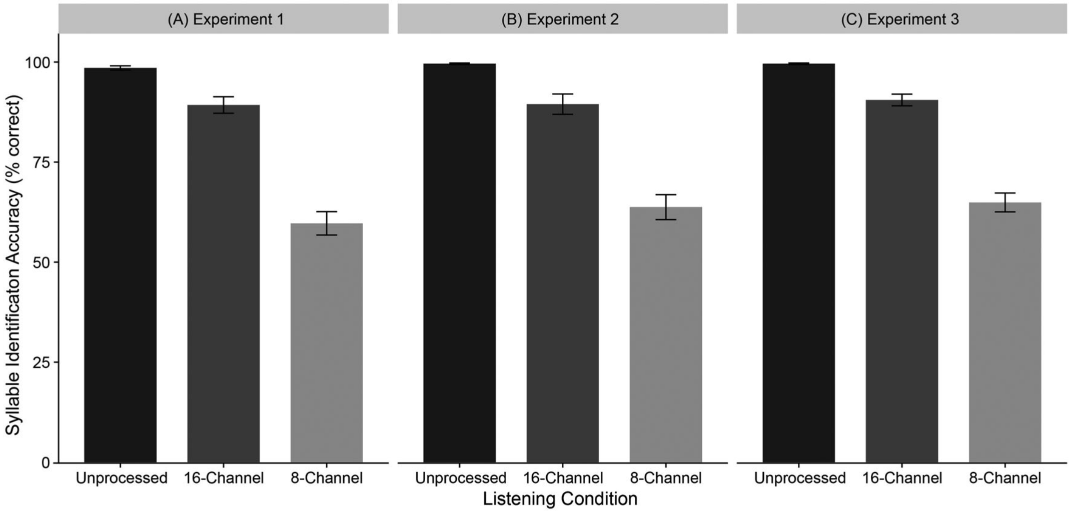

Figure 3. Mean (±SE) accuracy of syllable recognition. Accuracy is defined as the percent of correctly selected syllables on a 24-AFC task.

Procedure: Participants were randomly assigned to one of the three conditions: unprocessed, 16-ch, and 8-ch. The word segmentation and syllable recognition tasks were identical to those described in Experiment 1.

Statistical analysis: Analyses were similar to those used in Experiment 1.

Results and discussion

As in Experiment 1, participants were tested on their ability to recognise familiar vs. novel trisyllabic words. Participants segmented words from the artificial language at levels that were statistically above chance performance in the unprocessed condition (77.9% ± 4.4%, $t[19] = 6.4$, $p < .001$). Unlike Experiment 1, however, participants also segmented words at levels that were statistically above chance in both the 16-ch condition (78.1% ± 3.6%, $t[19] = 7.7$, $p < .001$) and 8-ch condition (63.5% ± 3.4%, $t[19] = 4.0$, $p < .001$; Figure 1(B)). The results of the logistic mixed effects modelling showed that participants in the 16-ch condition segmented the artificial language equally as well as participants in the unprocessed condition ($\beta = -0.09$, $z = -0.22$, $p = .82$), but better than participants in the 8-ch condition ($\beta = -1.1$, $z = -2.8$, $p < .01$). Finally, across conditions, participants had statistically better segmentation of words than part-words ($\beta = -2.2$, $z = -5.5$, $p < .001$; Figure 2(B)). Taken together, the results suggest that pauses, resulting in a small number of isolated words, supported word segmentation from spectrally degraded speech, but were mainly beneficial for segmenting the specific trisyllabic sequences that they bordered in the speech stream (i.e. words), as opposed to the other sequences. Consistent with this idea, participants in the 16-ch and 8-ch conditions segmented words, but not part-words, at levels that were statistically above chance (16-ch: words = 93.8% ± 1.8%, $t(19) = 24.3$, $p < .001$ and part-words = 62.5% ± 6.5%, $t(19) = 1.9$, $p = .07$; 8-ch: words = 80% ± 3.7%, $t(19) = 8.0$, $p < .001$ and part-words = 47.1% ± 5.1%, $t(19) = -0.57$, $p = .57$; Figure 2(B)).

Next, we tested whether participants in Experiment 2 performed differently than those in Experiment 1 by subjecting participants' accuracy on the word segmentation task to logistic mixed effects modelling. We were specifically interested in understanding the nuanced differences in performance across the three conditions of each experiment. Therefore, the model included the interactions between the fixed effects of experiment (categorical variable: Experiment 1, Experiment 2), condition, and word type. The random effects structure included intercepts of participants and test items as well as slopes of condition and word type for test items. The fixed effects of condition and word type were contrasted as described previously. The results revealed three significant interactions. First, there was an experiment by word type interaction ($\beta = 2.1$, $z = 10.1$, $p < .001$), suggesting that the difference in participants' accuracy for words and part-words was larger in Experiment 2 than in Experiment 1. Second, there was a word type by condition contrast (16-ch vs. 8-ch) interaction ($\beta = 0.72$, $z = 2.5$, $p < .05$), suggesting that the difference in accuracy for words and part-words was larger for the 16-ch condition than for the 8-ch condition. Finally, there was a three-way interaction between experiment, word type, and condition contrast (16-ch vs. 8-ch; $\beta = -1.5$, $z = -3.0$, $p < .01$), suggesting that the interaction between word-type and the 16-ch vs. 8-ch contrast was smaller in Experiment 2. Taken together, the logistic mixed effects model supports our prediction: silent pauses facilitated word segmentation from vocoded speech. This effect was larger for words vs. part-words, and larger in the 16-ch vs. 8-ch conditions.

The improved recognition of words in the 16-ch and 8-ch conditions of Experiment 2 was observed, however, in the absence of improved recognition of syllables. Specifically, participants in the unprocessed condition identified a greater proportion of syllables than participants in the 16-ch condition ($\beta = -3.3$, $z = -2.2$, $p < .05$), and participants in the 16-ch condition identified a greater proportion of syllables than participants in the 8-ch condition ($\beta = -2.84$, $z = -5.2$, $p < .001$; Figure 3(B)). Finally, syllable recognition did not predict word segmentation (unprocessed: Pearson's $R = -0.14$, $p = .55$; 16-ch: Pearson's $R = .22$, $p = .34$; 8-ch: Pearson's $R = -.19$, $p = .43$), consistent with the idea that decoding individual speech sounds is not the sole determinant of word segmentation.

Results from Experiment 2 suggest that intermittent inclusion of silent pauses, which effectively served as overt word boundaries, improved participants' abilities to segment the speech stream. We observed differential improvement in the segmentation of words over part-words, in both the 16-ch and 8-ch noise-band vocoded conditions. In both cases, participants successfully recognised the units that had been bookended with pauses during familiarisation, that is, low-frequency words, but they did not recognise part-words at above-chance levels. The results generally support the main interpretation of Experiment 1, namely that noise-band vocoding impairs the processes involved in tracking the structure of the input, even in the presence of intact syllable recognition. But the results of Experiment 2 also indicate that the inclusion of brief pauses in an otherwise continuous artificial speech stream supported segmentation. This may have resulted from the break in speech provided by isolated words themselves. However, it could also have resulted from the fact that the speech stream was slightly slower in Experiment 2 than in Experiment 1, which provided the participants with more time to track syllables patterns. To investigate these ideas further, we evaluated segmentation using a more pervasive temporal modification to the speech stream. In Experiment 3, the artificial language was slowed down by 33%. By reducing the rate of syllable presentation in the artificial speech stream, we investigated whether slower speed would support adults' tracking of its inherent structure.

Experiment 3

Methods

Participants: Participants were 60 native English-speaking adults (mean = 22.0 years; range = 19–34 years). Inclusion criteria and consent procedures were consistent with those described in Experiments 1 and 2.

Stimuli: Experiment 3 used the same artificial speech stream as the previous experiments. The pause-free speech stream was slowed by 33%, with no change in pitch, using Adobe Audition® (2.6 syllables/second; Liu & Zheng, 2006). To assess whether this manipulation had equivalent effects on different syllables, we compared the durations of 10 samples of each syllable type in Language 1 across the original and slowed speech streams. The mean percent change was 33.0% ± 0.21% (mean ± SD, range: 32.7% to 33.5%). The duration of the speech stream was 4 minutes 19 seconds, which is 1 minute 4 seconds longer than the speech stream in Experiment 1. The speech stream was repeated twice for a total duration of 8 minutes 28 seconds.

Procedure: Participants were randomly assigned to one of the three conditions: unprocessed, 16-ch, and 8-ch. The word segmentation and syllable recognition tasks were identical to those described in Experiment 1.

Results and discussion

As in Experiments 1 and 2, participants were tested on their ability to recognise familiar vs. novel words. Participants segmented words from the artificial language at levels that were statistically greater than chance performance in the unprocessed condition (77% ± 1.9%, $t[19] = 6.1$, $p < .001$) but not in the 16-ch or 8-ch conditions (16-ch: 61.2% ± 5.7%, $t[19] = 2.0$, $p = .06$; 8-ch: 56.0% ± 4.8%, $t[19] = 1.2$, $p = .22$; Figure 1(C)). The results of the logistic mixed effects modelling showed that participants in the unprocessed condition segmented sequences from the artificial language statistically better than participants in the 16-ch condition ($\beta = -1.07$, $z = -2.1$, $p < .05$), and participants in the 16-ch and 8-ch conditions did not differ in their word segmentation abilities ($\beta = -0.32$, $z = -0.71$, $p = .48$; Figure 1(C)). Participants also segmented words and part-words equally well ($\beta = 0.16$, $z = 0.63$, $p = .53$; Figure 2(C)).

The pattern of results in Experiment 3 does not reflect those in Experiment 2, suggesting that the more pervasive type of temporal modification of slowing the speech rate is a weaker cue than silent pauses for segmenting words. To evaluate this idea statistically, participants' accuracy across Experiments 2 and 3 was evaluated with a logistic mixed effects model that included the interactions between the fixed effects of experiment (categorical variable: Experiment 2, Experiment 3), condition, and word type. The random effects structure included intercepts of participants and test items as well as slopes of condition and word type for test items. The fixed effects of condition and word type were contrasted as described previously. Two statistically significant interactions were identified. First, there was an experiment by word type interaction ($\beta = -2.1$, $z = -9.8$, $p < .001$), suggesting that the difference in accuracy for words and part-words was smaller in Experiment 3 than in Experiment 2. Second, there was a three-way interaction between experiment, word type, and condition contrast (16-ch vs. 8-ch; $\beta = 1.2$, $z = 2.4$, $p < .05$), suggesting that an interaction between word-type and the 16-ch vs. 8-ch contrast was larger in Experiment 3. Thus, this model is consistent with our interpretation that participants' word segmentation from vocoded speech was facilitated only for the trisyllabic sequences that were flanked by silent pauses in Experiment 2, and not by a more universal slowing of the speech stream.

Finally, the results of the logistic mixed effect modelling of the syllable recognition task in Experiment 3 was consistent with what was observed in Experiments 1 and 2. Specifically, participants in the unprocessed condition identified a greater proportion of syllables than participants in the 16-ch condition ($\beta = -3.6$, $z = -2.6$, $p < .01$), and participants in the 16-ch condition identified a greater proportion of syllables than participants in the 8-ch condition ($\beta = -2.14$, $z = -5.1$, $p < .001$; Figure 3(C)). Syllable recognition did not predict word segmentation (unprocessed: Pearson's $R = -.42$, $p = .06$; 16-ch: Pearson's $R = -.03$, $p = .91$; 8-ch: Pearson's $R = .05$, $p = .82$).

General discussion

The present studies tested the hypotheses that spectral degradation interferes with word segmentation and that salient temporal cues can restore successful segmentation. Results from three experiments affirm these hypotheses. In Experiment 1, participants segmented words from unprocessed, acoustically rich speech in an artificial language, but they did not segment words with above-chance performance when speech was filtered into either 16 or 8 spectral channels. In the 16-ch condition, they failed to segment despite robust skill in recognising individual syllables, suggesting that resolution of speech units did not account for impaired word segmentation. In Experiment 2, providing intermittent silent pauses within the speech stream (akin to overt word boundaries) aided segmentation in the two vocoded conditions. This contrasts with results from Experiment 3, which showed that slowing the rate of speech – a pervasive temporal modification – did not improve segmentation of vocoded speech. Alongside participants' consistent accuracy in recognising individual syllables across the three 16-ch conditions, the findings of Experiments 1 and 3 indicate that vocoding disrupts the ability to track relations between syllables. The findings of Experiment 2 suggest that salient supportive cues – in this case, isolated words and/or the pauses that flank them – can partially restore successful segmentation of degraded speech.

This investigation provides new insight into our understanding of how degraded input disrupts listeners' processing of language. Previous research has documented that acoustically degraded speech affects adults' use of semantic cues to predict subsequent words (Sohoglu, Peelle, Carlyon, & Davis, 2012; Winn, 2016), as well as adults' word recognition and later memory for successfully heard words (McCoy et al., 2005). Recognition of known words is also slower and less accurate in two-year-old children who use cochlear implants (Grieco-Calub, Saffran, & Litovsky, 2009). Here, we begin to uncover the effects of degraded input on a different – and foundational – process in language learning: the tracking of patterns between sounds and syllables across time. Even when adults' recognition of individual syllables was intact, their segmentation of word-like units was less successful when listening to vocoded speech relative to unprocessed speech. Given that word segmentation feeds into other aspects of language learning, such as word learning (Estes, Evans, Alibali, & Saffran, 2007), word recognition (Lany, Shoaib, Thompson, & Graf Estes, 2016), and reading (Arciuli & Simpson, 2012; Spencer, Kaschak, Jones, & Lonigan, 2015), difficulty at the level of detecting recurring syllable sequences in degraded speech could give rise to the disrupted language processes documented in previous research.

Relatedly, our investigation – particularly Experiment 2 – contributes to what is currently known about the component processes involved in breaking into the structure of speech. A range of cues can be used by adult and/or infant listeners to help solve the segmentation problem, including transition statistics (Aslin et al., 1998; Graf Estes & Lew-Williams, 2015), isolated words (Lew-Williams et al., 2011), phrasal or sentence-level prosody (Johnson, Seidl, Tyler, & Berwick, 2014; Shukla, Nespor, & Mehler, 2007; Soderstrom, Seidl, Nelson, & Jusczyk, 2003), sentence position (Seidl & Johnson, 2006), language-specific patterns of word formation (Brent & Cartwright, 1996; Toro, Pons, Bion, & Sebastián-Gallés, 2011), lexical stress (Houston, Jusczyk, Kuijpers, Coolen, & Cutler, 2000; Johnson & Jusczyk, 2001; Thiessen & Saffran, 2003), and redundant visual cues (Cunillera, Càmara, Laine, & Rodríguez-Fornells, 2010a). Researchers are only beginning to understand how these cues to structure, separately or in combination, optimally support word segmentation across the lifespan. Here, we make progress in understanding how isolated words factor into adults' abilities to segment words when acoustical cues are less reliable.

Previous data have revealed that interspersed isolated words improve the ability to discover structure in otherwise fluent speech (Lew-Williams et al., 2011), and that the presence of isolated words uniquely predicts toddlers' later production of those words (Brent & Siskind, 2001). The data presented in Experiment 2 reveal that isolated words help listeners segment spectrally degraded speech in an artificial language, evidenced primarily by the overall improvement in accuracy. There are two primary explanations for this general enhancement in segmentation. First, the insertion of pauses in the speech stream could act as a temporal cue that gives listeners incrementally more time to process incoming speech. However, the pervasive temporal modification used in Experiment 3 renders this possibility unlikely, because, as in Experiment 1, accuracy in segmentation was not above chance in the vocoded conditions. Therefore, a second and more likely possibility is that isolated words – and the pauses that surrounded them – served as salient "anchors" that provided clear exemplars of recurring units in the speech stream (Cunillera et al., 2010b, 2016). Various investigations have proposed that the perceptual salience of familiar units, such as a high-frequency word or an isolated word, can help listeners segment sequences from the speech stream (Bortfeld et al., 2005; Dahan & Brent, 1999; Lew-Williams et al., 2011; Mersad & Nazzi, 2012). In Experiment 2, hearing occasional pauses before and after low-frequency words during familiarisation (which then served as "word" items during test) may have provided listeners with clear exemplars of the relevant units to track over time. Notably, in both of the vocoded conditions, listeners showed above-chance performance for word items but not for part-word items. This suggests that isolated words supported adults' segmentation of the particular units in the speech stream that were bookended by silent pauses, but not their neighbours. By studying vocoded speech, we show for the first time that pauses act as a compensatory cue for segmenting words in adverse listening conditions, especially in speech that maintains a minimum level of spectral fidelity.

The idea that isolated words supported learning by providing the listener with clear exemplars of the target sequences, rather than providing a pure temporal cue, is supported by the main result of Experiment 3: that slowing the presentation rate of the familiarisation language was insufficient to restore word segmentation from vocoded speech. This finding contrasts with results from Palmer and Mattys (2016), which showed improvements in segmentation when the speech stream was slowed. A plausible explanation for this inconsistency is that listeners may only be able to take advantage of slower speech under clear, unencumbered listening conditions. The results from Experiment 3 suggest that the burden of encoding spectrally degraded input may outweigh any benefits afforded by the presence of slower speech. These findings can be used as a springboard for future work investigating how real-time processing of incoming perceptual input interacts with a range of adverse listening conditions.

By studying word segmentation from noise-band vocoded speech, our three experiments provide insight into the difficulties facing hearing-impaired individuals with cochlear implants. There has been growing interest in understanding the scalability of segmentation tasks to natural learning conditions (e.g. Frank, Tenenbaum, Gibson, & Snyder, 2013; Graf Estes & Lew-Williams, 2015; Lew-Williams & Saffran, 2012; Pelucchi, Hay, & Saffran, 2009), including investigations of how learners process speech with the presence of background noise (Fu & Nogaki, 2005; Mattys, 2004; McMillan & Saffran, 2016). While listeners in quiet conditions can exploit transition statistics and many of the cues listed above, listeners in noisy conditions and listeners with limited access to acoustic hearing are less able to recruit this suite of cues. Mattys (2004) showed that listeners' ability to rely on stress and co-articulation cues when segmenting speech units from a real language varies depending on the presence or absence of background noise. Analogous to our simulations of noise-band vocoded speech, hearing loss and hearing devices obscure many natural acoustic cues because they render them inaudible or, in the case of cochlear implants, remove fast spectral transitions that support phoneme recognition. By embracing the variability in signal quality that is inherent in various naturalistic speech contexts, our investigation begins to unravel how diverse groups of listeners break into structure over time and become proficient listeners.

In order to understand the scalability of our experiments to individuals with hearing loss, who have varying levels of experience processing degraded speech, it is important to consider two nuances of the experimental design. First, the laboratory simulation of word segmentation used in our experiments was modelled on previous studies of statistical language learning, which used approximately 3 minutes of exposure to artificial speech (Aslin et al., 1998; Lew-Williams & Saffran, 2012). Here, exposure to the speech stream (in the absence of temporal modifications) was somewhat longer: 6 minutes. While this was a sufficient duration of exposure for adults to identify recurring syllable sequences in the unprocessed condition, adults in the vocoded conditions may have needed additional exposure time to successfully segment the syllable sequences. Consistent with this idea, perceptual adaptation to noise-band vocoded speech occurs with increased exposure (e.g. Davis, Johnsrude, Hervais-Adelman, Taylor, & McGettigan, 2005; Hervais-Adelman, Davis, Johnsrude, & Carlyon, 2008). Thus, future investigations will need to manipulate exposure time and stimulus complexity in order to understand the scalability of our findings to natural listening conditions. Perceptual adaptation to degraded speech over the course of minutes – or even years – may decrease the processing burden on hearing-impaired listeners, enabling increasingly more successful tracking of sounds and syllables in tandem with increases in language exposure.

A second limitation of the experimental design is that we did not assess how "passive" vs. "active" listening may have contributed to adults' success or failure in word segmentation. While the forced-choice test was unquestionably an active task, we do not know if participants listened to the speech stream in a passive or active manner during familiarisation. Key to this uncertainty is that the cognitive demands of listening to 16-ch or 8-ch vocoded speech may automatically require more "active" listening relative to unprocessed speech, because listeners have reduced access to the sounds comprising the speech stream. And moreover, participants' anticipation of a test following the familiarisation phase may have led them to engage actively with the incoming speech stream. Regardless, the laboratory context does not reflect language use in real-time communication, which engages a complex combination of active and passive processes when using speech (Batterink & Neville, 2013; Morgan-Short, Steinhauer, Sanz, & Ullman, 2012; Norris & Ortega, 2000). We do know that passive exposure in infancy is sufficient for successful word segmentation in lab tasks (Aslin et al., 1998; Graf Estes & Lew-Williams, 2015), but it remains unclear whether passive exposure is sufficient for word segmentation from degraded speech.

In summary, our experiments uncover how adults fail to segment sequences from spectrally degraded speech, and suggest possible means through which they recover the ability to do so, that is, when provided with salient cues, such as isolated words. We conclude that listeners' ability to discover patterns in speech is influenced by the spectral fidelity of incoming input and the availability of supportive speech cues to support segmentation. Exploring how listeners find structure in non-optimal listening conditions is central to understanding the nature of language learning, both in typical populations and in listeners with impaired access to the acoustics subtleties of speech.

Acknowledgments

We would like to thank Paul Reinhart, Valerie Tiscareno, and Klinton Bicknell for invaluable help with experiment design, data collection, and data analysis.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the American Hearing Research Foundation and the National Institute of Child Health and Human Development (R03HD079779).

References

Arciuli, J., & Simpson, I. C. (2012). Statistical learning is related to reading ability in children and adults. *Cognitive Science*, *36*, 286–304. doi:10.1111/j.1551-6709.2011.01200.x

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*, 321–324. doi:10.1111/1467-9280.00063

Aslin, R. N., Woodward, J. Z., LaMendola, N. P., & Bever, T. G. (1996). Models of word segmentation in fluent maternal speech to infants. *Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition*, 117–134.

Batterink, L., & Neville, H. (2013). Implicit and explicit second language training recruit common neural mechanisms for syntactic processing. *Journal of Cognitive Neuroscience*, *25* (6), 936–951. doi:10.1162/jocn_a_00354

Boersma, P., & Weenink, D. (2009). Praat: Doing phonetics by computer (Version 5.1. 10)[Computer program]. Retrieved July 8, 2009.

Bortfeld, H., Morgan, J. L., Golinkoff, R. M., & Rathbun, K. (2005). Mommy and me familiar names help launch babies into speech-stream segmentation. *Psychological Science*, *16*(4), 298–304. doi:10.1111/j.0956-7976.2005.01531.x

Brent, M. R., & Cartwright, T. A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, *61*(1), 93–125. doi:10.1016/S0010-0277(96)00719-6

Brent, M. R., & Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, *81* (2), B33–B44. doi:10.1016/S0010-0277(01)00122-6

Broadbent, D. E. (1958). *Perception and communication*. Elmsford, NY: Pergamon Press. doi:10.1037/10037-000

Church, R., Bernhardt, B., Shi, R., & Pichora-Fuller, K. (2005). Infant-directed speech: Final syllable lengthening and rate of speech. *The Journal of the Acoustical Society of America*, *117*(4.2), 2429–2430. doi:10.1121/1.4786663

Cunillera, T., Càmara, E., Laine, M., & Rodríguez-Fornells, A. (2010a). Speech segmentation is facilitated by visual cues. *The Quarterly Journal of Experimental Psychology*, *63*(2), 260–274. doi:10.1080/17470210902888809

Cunillera, T., Càmara, E., Laine, M., & Rodríguez-Fornells, A. (2010b). Words as anchors. *Experimental Psychology*, *57*(2), 134–141. doi:10.1027/1618-3169/a000017

Cunillera, T., Laine, M., & Rodríguez-Fornells, A. (2016). Headstart for speech segmentation: A neural signature for the anchor word effect. *Neuropsychologia*, *82*, 189–199. doi:10.1016/j.neuropsychologia.2016.01.011

Dahan, D., & Brent, M. R. (1999). On the discovery of novel word like units from utterances: An artificial-language study with implications for native-language acquisition. *Journal of Experimental Psychology: General*, *128*(2), 165–185. doi:10.1037/0096-3445.128.2.165

Davis, M. H., & Johnsrude, I. S. (2007). Hearing speech sounds: Top-down influences on the interface between audition and speech perception. *Hearing Research*, *229*(1–2), 132–147. doi:10.1016/j.heares.2007.01.014

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A. G., Taylor, K., & McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noisevocoded sentences. *Journal of Experimental Psychology: General*, *134*(2), 222–241. doi:10.1037/0096-3445.134.2.222

Donaldson, G. S., & Kreft, H. A. (2006). Effects of vowel context on the recognition of initial and medial consonants by cochlear implant users. *Ear and Hearing*, *27*(6), 658–677. doi:10.1097/01.aud.0000240543.31567.54

Downs, D. W., & Crum, M. A. (1978). Processing demands during auditory learning under degraded listening conditions. *Journal of Speech, Language, and Hearing Research*, *21*(4), 702–714. doi:10.1044/jshr.2104.702

Estes, K. G., Evans, J. L., Alibali, M. W., & Saffran, J. R. (2007). Can infants map meaning to newly segmented words? Statistical segmentation and word learning. *Psychological Science*, *18* (3), 254–260. doi:10.1111/j.1467-9280.2007.01885.x

Fiser, J., & Aslin, R. N. (2002). Statistical learning of higher-order temporal structure from visual shape sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(3), 458–467. doi:10.1037/0278-7393.28.3.458

Frank, M. C., Tenenbaum, J. B., Gibson, E., & Snyder J. (2013). Learning and long-term retention of large-scale artificial languages. *PLoS One*, *8*(1), e52500. doi:10.1371/journal.pone.0052500

Fu, Q.-J., & Nogaki, G. (2005). Noise susceptibility of cochlear implant users: The role of spectral resolution and smearing. *Journal of the Association for Research in Otolaryngology*, *6* (1), 19–27. doi:10.1007/s10162-004-5024-3

Gordon-Salant, S., Yeni-Komshian, G. H., Fitzgibbons, P. J., & Cohen, J. I. (2015). Effects of age and hearing loss on recognition of unaccented and accented multisyllabic words. *The Journal of the Acoustical Society of America*, *137*(2), 884–897. doi:10.1121/1.4906270

Graf Estes, K., & Lew-Williams, C. (2015). Listening through voices: Infant statistical word segmentation across multiple speakers. *Developmental Psychology*, *51*(11), 1517–1528. doi:10.1037/a0039725

Greenwood, D. D. (1990). A cochlear frequency-position function for several species – 29 years later. *The Journal of the Acoustical Society of America*, *87*(6), 2592–2605. doi:10.1121/1.399052

Grieco-Calub, T. M., Saffran, J. R., & Litovsky, R. Y. (2009). Spoken word recognition in toddlers who use cochlear implants. *Journal of Speech Language and Hearing Research*, *52*(6), 1390–1400. doi:0.1044/1092-4388(2009/08-0154)

Grieco-Calub, T. M., Ward, K. M., & Brehm, L. (2017). Multitasking during degraded speech recognition in school-Age children. *Trends Hear*, *21*, 1–14. doi:10.1177/2331216516686786

Hervais-Adelman, A., Davis, M. H., Johnsrude, I. S., & Carlyon, R. P. (2008). Perceptual learning of noise vocoded words: Effects of feedback and lexicality. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(2), 460–474. doi:10.1037/0096-1523.34.2.460

Hervais-Adelman, A. G., Carlyon, R. P., Johnsrude, I. S., & Davis, M. H. (2012). Brain regions recruited for the effortful comprehension of noise-vocoded words. *Language and Cognitive Processes*, *27*(7-8), 1145–1166. doi:10.1080/01690965.2012.662280

Hervais-Adelman, A. G., Davis, M. H., Johnsrude, I. S., Taylor, K. J., & Carlyon, R. P. (2011). Generalization of perceptual learning of vocoded speech. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(1), 283–295. doi:10.1037/a0020772

Houston, D. M., Jusczyk, P. W., Kuijpers, C., Coolen, R., & Cutler, A. (2000). Cross-language word segmentation by 9-month-olds. *Psychonomic Bulletin & Review*, *7*(3), 504–509. doi:10.3758/BF03214363

Johnson, E. K., & Jusczyk, P. W. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language*, *44*(4), 548–567. doi:10.1006/jmla.2000.2755

Johnson, E. K., Seidl, A., Tyler, M. D., & Berwick R. C. (2014). The edge factor in early word segmentation: Utterance-level prosody enables word form extraction by 6-month-olds. *PloS one*, *9*(1), e83546. doi:10.1371/journal.pone.0083546

Jusczyk, P. W. (1999). How infants begin to extract words from speech. *Trends in Cognitive Sciences*, *3*(9), 323–328. doi:10.1016/S1364-6613(99)01363-7

Jusczyk, P. W., & Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, *29*(1), 1–23. doi:10.1006/cogp.1995.1010

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition*, *83*(2), B35–B42. doi:10.1016/S0010-0277(02)00004-5

Lany, J., Shoaib, A., Thompson, A., & Graf Estes, K. (2016). Is statistical-learning ability related to real-time language processing? In *Proceedings of the 40th Annual Boston University Conference on Language Development* (pp. 203–215). Somerville, MA: Cascadilla Press.

Lew-Williams, C., Pelucchi, B., & Saffran, J. R. (2011). Isolated words enhance statistical language learning in infancy. *Developmental Science*, *14*(6), 1323–1329. doi:10.1111/j.1467-7687.2011.01079.x

Lew-Williams, C., & Saffran, J. R. (2012). All words are not created equal: Expectations about word length guide infant statistical learning. *Cognition*, *122*(2), 241–246. doi:10.1016/j.cognition.2011.10.007

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

Mattys, S. L. (2004). Stress versus coarticulation: Toward an integrated approach to explicit speech segmentation. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(2), 397–408. doi:10.1037/0096-1523.30.2.397

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes*, *27*(7–8), 953–978. doi:10.1080/01690965.2012.705006

McCoy, S. L., Tun, P. A., Cox, L. C., Colangelo, M., Stewart, R. A., & Wingfield, A. (2005). Hearing loss and perceptual effort: Downstream effects on older adults' memory for speech. *The Quarterly Journal of Experimental Psychology Section A*, *58*, 22–33. doi:10.1080/02724980443000151

McMillan, B. T. M., & Saffran, J. R. (2016). Learning in complex environments: The effects of background speech on early word learning. *Child Development*, *87*(6), 1841–1855. doi:10.1111/cdev.12559

Mersad, K., & Nazzi, T. (2012). When mommy comes to the rescue of statistics: Infants combine top-down and bottom-up cues to segment speech. *Language Learning and Development*, *8*, 303–315. doi:10.1080/15475441.2011.609106

Misyak, J. B., & Christiansen, M. H. (2012). Statistical learning and language: An individual differences study. *Language Learning*, *62*(1), 302–331. doi:10.1111/j.1467-9922.2010.00626.x

Morgan-Short, K., Steinhauer, K., Sanz, C., & Ullman, M. T. (2012). Explicit and implicit second language training differentially affect the achievement of native-like brain activation patterns. *Journal of Cognitive Neuroscience*, *24*(4), 933–947. doi:10.1162/jocn_a_00119

Munson, B., Donaldson, G. S., Allen, S. L., Collison, E. A., & Nelson, D. A. (2003). Patterns of phoneme perception errors by listeners with cochlear implants as a function of overall speech perception ability. *The Journal of the Acoustical Society of America*, *113*(2), 925–935. doi:10.1121/1.1536630

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, *50*(3), 417–528. doi:10.1111/0023-8333.00136

Obleser, J., Wise, R. J. S., Dresner, M. A., & Scott, S. K. (2007). Functional integration across brain regions improves speech perception under adverse listening conditions. *Journal of Neuroscience*, *27*(9), 2283–2289. doi:10.1523/jneurosci.4663-06.2007

Palmer, S. D., & Mattys, S. L. (2016). Speech segmentation by statistical learning is supported by domain-general processes within working memory. *The Quarterly Journal of Experimental Psychology*, 1–12. doi:10.1080/17470218.2015.1112825

Pals, C., Sarampalis, A., & Başkent, D. (2013). Listening effort with cochlear implant simulations. *Journal of Speech Language and Hearing Research*, *56*(4), 1075–1084. doi:10.1044/1092-4388(2012/12-0074)

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009). Statistical learning in a natural language by 8-month-old infants. *Child Development*, *80*, 674–685. doi:10.1111/j.1467-8624.2009.01290.x

Peña, M., Bonatti, L. L., Nespor, M., & Mehler, J. (2002). Signal-driven computations in speech processing. *Science*, *298* (5593), 604–607. doi:10.1126/science.1072901

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

Rabbitt, P. M. (1966). Recognition: Memory for words correctly heard in noise. *Psychonomic Science*, *6*(8), 383–384. doi:10.3758/bf03330948

Rakerd, B., Seitz, P., & Whearty, M. (1996). Assessing the cognitive demands of speech listening for people with hearing losses. *Ear and Hearing*, *17*(2), 97–106. doi:10.1097/00003446-199604000-00002

R Core Team. (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(6), 906–914. doi:10.1002/wcs.78

Rönnberg, J., Lunner, T., Zekveld, A., et al. (2013). The ease of language understanding (ELU) model: Theory, data, and clinical implications. *Frontiers in Systems Neuroscience*, *7*, 1–17. doi:10.3389/fnsys.2013.00031

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by eight-month old infants. *Science*, *274*, 1926–1928. doi:10.1126/science.274.5294.1926

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*(4), 606–621. doi:10.1006/jmla.1996.0032

Saffran, J. R., & Thiessen, E. D. (2003). Pattern induction by infant language learners. *Developmental Psychology*, *39*(3), 484–494. doi:10.1037/0012-1649.39.3.484

Sarampalis, A., Kalluri, S., Edwards, B., & Hafter, E. (2009). Objective measures of listening effort: Effects of background noise and noise reduction. *Journal of Speech Language and Hearing Research*, *52*(5), 1230–1240. doi:10.1044/1092-4388(2009/08-0111)

Seidl, A., & Johnson, E. K. (2006). Infant word segmentation revisited: Edge alignment facilitates target extraction. *Developmental Science*, *9*(6), 565–573. doi:10.1111/j.1467-7687.2006.00534.x

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, *270*(5234), 303–304. doi:10.1126/science.270.5234.303

Shukla, M., Nespor, M., & Mehler, J. (2007). An interaction between prosody and statistics in the segmentation of fluent speech. *Cognitive Psychology*, *54*(1), 1–32. doi:10.1016/j.cogpsych.2006.04.002

Soderstrom, M., Seidl, A., Nelson, D. G. K., & Jusczyk, P. W. (2003). The prosodic bootstrapping of phrases: Evidence from prelinguistic infants. *Journal of Memory and Language*, *49*(2), 249–267. doi:10.1016/S0749-596X(03)00024-X

Sohoglu, E., Peelle, J. E., Carlyon, R. P., & Davis, M. H. (2012). Predictive top-down integration of prior knowledge during speech perception. *Journal of Neuroscience*, *32*(25), 8443–8453. doi:10.1523/JNEUROSCI.5069-11.2012

Spencer, M., Kaschak, M. P., Jones, J. L., & Lonigan, C. J. (2015). Statistical learning is related to early literacy-related skills. *Reading and Writing*, *28*(4), 467–490. doi:10.1007/s11145-014-9533-0

Thiessen, E. D., & Saffran, J. R. (2003). When cues collide: Use of stress and statistical cues to word boundaries by 7-to 9-month-old infants. *Developmental Psychology*, *39*(4), 706–716. doi:10.1037/0012-1649.39.4.706

Toro, J. M., Pons, F., Bion, R. A., & Sebastián-Gallés, N. (2011). The contribution of language-specific knowledge in the selection of statistically-coherent word candidates. *Journal of Memory and Language*, *64*(2), 171–180. doi:10.1016/j.jml.2010.11.005

Ward, K. M., Shen, J., Souza, P. E., & Grieco-Calub, T. M. (2017). Age-related differences in listening effort during degraded speech recognition. *Ear and Hearing*, *38*(1), 74–84. 10.1097/AUD.0000000000000355

Wild, C. J., Yusuf, A., Wilson, D. E., Peelle, J. E., Davis, M. H., & Johnsrude, I. S. (2012). Effortful listening: The processing of degraded speech depends critically on attention. *Journal of Neuroscience*, *32*(40), 14010–14021. doi:10.1523/jneurosci.1528-12.2012

Winn, M. (2016). Rapid release from listening effort resulting from semantic context, and effects of spectral degradation and cochlear implants. *Trends in Hearing*, *20*, 1–17. doi:10.1177/2331216516669723

Winn, M. B., Edwards, J. R., & Litovsky, R. Y. (2015). The impact of auditory spectral resolution on listening effort revealed by pupil dilation. *Ear and Hearing*, 36(4), e153-e165. doi:10.1097/AUD.0000000000000145

Xu, L., & Pfingst, B. E. (2008). Spectral and temporal cues for speech recognition: Implications for auditory prostheses. *Hearing Research*, *242*(1), 132–140. doi:10.1016/j.heares.2007.12.010

Zhou, N., Xu, L., & Lee, C.-Y. (2010). The effects of frequency-place shift on consonant confusion in cochlear implant simulations. *The Journal of the Acoustical Society of America*, *128* (1), 401–409. doi:10.1121/1.3436558