Speech perception in simulated electric hearing exploits information-bearing acoustic change

Christian E. Stilp^{a)}

*Department of Psychological and Brain Sciences, University of Louisville, Louisville,
Kentucky 40292
christian.stilp@louisville.edu*

Matthew J. Goupell

*Department of Hearing and Speech Sciences, University of Maryland, College Park,
Maryland 20742
goupell@umd.edu*

Keith R. Kluender

*Department of Speech, Language, and Hearing Sciences, Purdue University, Lafayette,
Indiana 47907
kkluender@purdue.edu*

Abstract: Stilp and Kluender [(2010). *Proc. Natl. Acad. Sci. U.S.A.* **107**(27), 12387–12392] reported measures of sensory change over time (cochlea-scaled spectral entropy, CSE) reliably predicted sentence intelligibility for normal-hearing listeners. Here, implications for listeners with atypical hearing were explored using noise-vocoded speech. CSE was parameterized as Euclidean distances between biologically scaled spectra [measured *before* sentences were noise vocoded (CSE)] or between channel amplitude profiles in simulated cochlear-implant processing [measured *after* vocoding (CSE_{CI})]. Sentence intelligibility worsened with greater amounts of information replaced by noise; patterns of performance did not differ between CSE and CSE_{CI}. Results demonstrate the importance of information-bearing change for speech perception in simulated electric hearing.

© 2013 Acoustical Society of America

PACS numbers: 43.71.Es, 43.66.Ba [QJF]

Date Received: October 26, 2012 **Date Accepted:** January 2, 2013

1. Introduction

Much in the world is predictable from time to time and place to place, but there is little to no information in that which is known or predictable. Information is defined by uncertainty, unpredictability, or most simply, change (Shannon, 1948). The greater the unpredictability, the greater the amount of information that can potentially be transmitted. Perceptual systems respond primarily to change (e.g., Kluender *et al.*, 2003), and optimizing sensitivity to change maximizes information for perception. Information-theoretic approaches have shown this process to be fundamental to speech perception (e.g., Kluender and Alexander, 2007; Kluender *et al.*, 2013) and to auditory perception in a broader sense (e.g., Kluender *et al.*, 2003; Stilp *et al.*, 2010a).

Stilp and Kluender (2010) proposed that potential information is an important contributor to sentence intelligibility. Potential sensory information was inferred from cochlea-scaled spectral entropy (CSE), a measure of spectral change across time quantified as Euclidean distances between adjacent, psychoacoustically scaled spectral slices. They measured CSE in sentences, and then replaced segments with noise on the basis

^{a)}Author to whom correspondence should be addressed.

of having low, medium, or high amounts of CSE. Segment durations were 80 ms (approximate mean duration of consonants in the TIMIT database; Garofolo *et al.*, 1990) or 112 ms (approximate mean vowel duration), following methods in previous studies of sentence intelligibility amidst noise replacement (e.g., Kewley-Port *et al.*, 2007). Sentence intelligibility closely followed measures of CSE: as greater amounts of potential information were replaced by noise, performance worsened.

Estimates of sensory change prove important for speech perception by normal-hearing (NH) listeners, but implications for listeners with impaired hearing are unknown. Cochlear implants (CI) users receive dramatically different auditory input, leaving them to rely on restricted information for perception with potentially different weighting compared to that of NH listeners. As a result, while some can understand speech in quiet as well as NH listeners who are presented CI simulations (e.g., Friesen *et al.*, 2001), there is a wide range in speech perception performance of CI users. A significant challenge is to characterize information that high-performing CI listeners rely on to achieve such proficiency, and whether increasing or emphasizing this information would benefit lower-performing CI users.

While initial encoding widely differs in normal versus electric hearing, subsequent subcortical and cortical processing is similarly predicated on emphasizing changes in the input. Thus, similar to NH listeners, CI users may also exploit regions of high signal change for understanding speech. To explore this possibility, we investigate intelligibility of noise-vocoded sentences when segments are replaced by noise on the basis of acoustic changes available for normal (CSE) and simulated electric hearing (CSE_{CI}). These parameterizations allow comparison of the importance of potential information for speech perception as measured in signals available for acoustic versus simulated electric hearing.

2. Methods

Twenty undergraduates were recruited from the Department of Psychology at the University of Wisconsin–Madison. All participants reported being native English speakers with normal hearing and received course credit for their participation.

Experimental materials were 59 sentences from the TIMIT database (5–8 words, mean duration = 2046 ms; Garofolo *et al.*, 1990). Sentences were root-mean-square (RMS)-intensity-normalized, then information was measured either before (CSE; Stilp and Kluender, 2010) or after noise vocoding to estimate potential information available to CI users (CSE_{CI}).

To calculate CSE, sentences were divided into 16-ms slices, and each slice was passed through 33 simulated auditory filters (Patterson *et al.*, 1982) spaced 1 ERB apart (Glasberg and Moore, 1990) from 26 to 7743 Hz. Euclidean distances were calculated between spectral profiles (i.e., energy across all bands) for all pairs of adjacent 16-ms spectra. Distances were summed in boxcars of five successive slices (80 ms) then sorted into ascending (low entropy condition) or descending order (high entropy). All sentences were then processed by an eight-channel vocoder with a Gaussian noise carrier (center frequencies log-spaced from 300 to 5000 Hz, 2nd-order low-pass Butterworth filter with 150-Hz cutoff to extract amplitude envelopes, 4th-order bandpass Butterworth filter for channel analysis and synthesis). The boxcar that ranked first in the full-spectrum sentence (lowest or highest entropy) was replaced in the vocoded sentence by speech-shaped noise matched to mean sentence level (5-ms linear onset/offset ramps). Eighty-millisecond intervals immediately before and after replaced segments, as well as at the beginning of the sentence, were always left intact. The procedure proceeded iteratively to the next-highest-ranked boxcar, which was replaced only if its content had not already been replaced or preserved. Replacements continued until no more segments could be replaced (n.b. this maximum was matched for a given sentence across all experimental conditions).

To calculate CSE_{CI}, sentences were vocoded first before being divided into the same 16-ms slices as in CSE conditions. CSE_{CI} was parameterized as the Euclidean

distance between RMS amplitude profiles across eight vocoder channels for successive slices. Boxcar summation and noise replacement followed as described above. On average, 34.7% of sentence duration was replaced by noise, with the same duration of sentence replaced by noise in each condition.

Sentences were presented diotically at 72 dB SPL via circumaural headphones (Beyer-Dynamic DT-150, Farmingdale, NY). Listeners participated individually in a double-wall soundproof booth. Following acquisition of informed consent, listeners were given instructions and told to expect that some sentences would be difficult to understand, so guessing was encouraged. Listeners completed 9 practice sentences followed by 50 experimental sentences (10 trials in each of 4 experimental conditions plus 10 control trials presenting vocoded sentences without any noise replacement), one per trial, without hearing any sentence more than once. While listeners heard experimental sentences in the same order, the order of conditions was pseudo-randomized; within each of ten five-trial blocks, listeners heard one sentence in each of the five conditions in random order.

Responses were scored offline by three raters blind to experimental conditions using guidelines listed in Stilp *et al.*, (2010b). Inter-rater reliability, measured by intra-class correlation, was 0.99. Intelligibility was defined as the mean percent of words correctly identified in each sentence.

3. Results

Control sentences [mean = 69.86% of words correctly identified, standard error (s.e.) = 2.97] were included primarily as reference for performance in experimental conditions; therefore, results were analyzed in a 2 (measure of information-bearing acoustic change: CSE/CSE_{CI}) by 2 (level of information replaced by noise: low/high) repeated-measures analysis of variance. Replicating Stilp and Kluender (2010), performance was better when low-information segments were replaced by noise (mean = 28.32%, s.e. = 2.00) than when high-information segments were replaced (mean = 12.01%, s.e. = 1.22; $F_{1,19} = 118.24$, $p < 0.001$, $\eta_p^2 = 0.86$) (Fig. 1). Importantly, performance did not differ when information-bearing acoustic change was measured and replaced in sentences appropriate for acoustic hearing (CSE mean = 19.00%, s.e. = 1.58) versus

Fig. 1. (Color online) (a)–(c) Sample stimuli from CSE_{CI} conditions. (a) The sentence “They were already swollen to bursting” was noise-vocoded, and then 80-ms segments with low (b) or high (c) information-bearing acoustic change were replaced by speech-shaped noise. (d) Noise-vocoded sentence intelligibility is plotted as a function of experimental condition, reflecting when information was measured [before (CSE) or after noise-vocoding (CSE_{CI})] and level replaced by noise (low, high). Error bars indicate standard error. *** = $p < 0.001$ in Bonferroni-corrected paired-sample t -tests.

simulated electric hearing (CSE_{CI} mean = 21.32%, s.e. = 1.78; $F_{1,19} = 1.92$, n.s., $\eta_p^2 = 0.09$). The interaction did not reach statistical significance ($F_{1,19} = 0.29$, n.s., $\eta_p^2 = 0.01$). Closer examination reveals that intelligibility was higher for low-information-replaced sentences compared to high-information-replaced sentences for CSE (paired-sample t -test with Bonferroni correction for multiple comparisons; $t_{19} = 9.28$, $p < 0.001$) and CSE_{CI} ($t_{19} = 7.04$, $p < 0.001$), but performance did not differ for a given level of information replaced across CSE and CSE_{CI} ($t < 1.15$, n.s.).

For each sentence, slices that were replaced by noise were coded as 1 with all other slices coded as 0. Holding the level of information replaced constant, mean numbers of shared slices (replaced in both CSE and CSE_{CI} conditions) and unshared slices (replaced in only one condition) were calculated. Mean shared slices outnumbered unshared slices in both low-information [28.34 (s.e. = 1.05) versus 16.06 (s.e. = 0.96); paired-sample t -test: $t_{49} = 7.64$, $p < 0.001$] and high-information sentences [28.42 (s.e. = 1.21) versus 15.98 (s.e. = 1.07); $t_{49} = 6.43$, $p < 0.001$]. Broadly stated, measures of information-bearing acoustic change in acoustic (CSE) and simulated electric hearing (CSE_{CI}) are more similar to each other than they are different, and this supports highly comparable performance across these conditions.

4. Discussion

Results demonstrate the importance of potential information for speech perception via normal and simulated electric hearing. Parallel patterns of performance when information is measured in full-spectrum (CSE) versus vocoded (CSE_{CI}) materials suggest similar reliance on information-bearing spectral change for sentence intelligibility. This result offers great promise in better understanding perception by CI users, as relative change in the signal likely influences sentence intelligibility for both NH and CI listeners.

Normal-hearing listeners exploit relative change in the signal for perception of acoustically rich (Stilp and Kluender, 2010; Stilp *et al.*, 2010b) and vocoded sentences. This emphasis on change marks a promising avenue of exploration for investigating speech perception by CI users. Previous efforts to transmit more information in CIs have introduced unwanted drawbacks, such as current spread when stimulating more channels (e.g., Fishman *et al.*, 1997; Friesen *et al.*, 2001) or poorer intensity resolution and/or modulation detection when increasing stimulation rate (e.g., Galvin and Fu, 2005; Kreft *et al.*, 2004). By identifying information-bearing acoustic change, the present approach may provide a more efficient way to transmit information through CIs without stimulating across more channels or at higher rates.

Performance in experimental conditions (mean = 28% for conditions where low information was replaced by noise, 12% when high information was replaced) is comparable to investigations of spectrally degraded sentences interrupted with fixed duty cycle noise (Başkent and Chatterjee, 2010; Chatterjee *et al.*, 2010; Başkent, 2012). However, two key differences distinguish the present results from these reports. First, fixed duty cycle noise interrupts speech at regular intervals, while noise replacements in the present experiment are not strictly periodic. Second, 50% duty cycle noise results in 50% of overall sentence duration being replaced with noise, often in segments 100-ms or longer. The present experiment replaced 35% of overall sentence duration with noise using 80-ms segments, yet achieved comparable if not poorer levels of performance. Replacing less speech with noise using briefer segments and observing poorer performance further supports the suggestion that information-bearing acoustic change more closely corresponds to performance than physical measures of time (Stilp and Kluender, 2010; Stilp *et al.*, 2010b).

Results also inform investigations of the relative importance of amplitude envelope (E) versus temporal fine structure (TFS) cues for speech intelligibility. Speech perception relies more heavily upon E information (e.g., Smith *et al.*, 2002; Swaminathan and Heinz, 2012), and high levels of performance can be achieved with only E

information from a small number of frequency channels (e.g., Shannon *et al.*, 1995; Dorman *et al.*, 1997). CSE more closely corresponds to an E cue for speech recognition but with an important distinction: CSE measures changes in the spectral envelopes across time and not the presence or absence of this information *per se*. The present results suggest consideration of not only which acoustic cues are available (or unavailable) for speech perception, but how they change across time as well.

The present results encourage information-theoretic approaches to speech perception (e.g., Kluender and Alexander, 2007; Kluender *et al.*, 2013; Stilp and Kluender, 2010; Stilp *et al.*, 2010b). More importantly, results emphasize fundamental principles of perception that are shared across normal and simulated electric hearing. Information-bearing acoustic change offers great promise for investigations of speech perception by listeners with normal and impaired hearing.

Acknowledgments

This work was supported by grants from NIDCD (C.E.S., Grant No. F31 DC009532; M.J.G., Grant No. K99/R00 DC010206, and K.R.K., Grant No. R01 DC004072).

References and links

- Başkent, D. (2012). "Effect of speech degradation on top-down repair: Phonemic restoration with simulations of cochlear implants and combined electric-acoustic stimulation," *J. Assoc. Res. Otolaryngol.* **13**(5), 683–692.
- Başkent, D., and Chatterjee, M. (2010). "Recognition of temporally interrupted and spectrally degraded sentences with additional unprocessed low-frequency speech," *Hear. Res.* **270**, 127–133.
- Chatterjee, M., Peredo, F., Nelson, D., and Başkent, D. (2010). "Recognition of interrupted sentences under conditions of spectral degradation," *J. Acoust. Soc. Am.* **127**(2), EL37–EL41.
- Dorman, M. F., Loizou, P. C., and Rainey, D. (1997). "Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs," *J. Acoust. Soc. Am.* **102**(4), 2403–2411.
- Fishman, K. E., Shannon, R. V., and Slaterry, W. H. (1997). "Speech recognition as a function of the number of electrodes used in the SPEAK cochlear implant processor," *J. Speech Lang. Hear. Res.* **40**, 1201–1215.
- Friesen, L. M., Shannon, R. V., Başkent, D., and Wang, X. (2001). "Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants," *J. Acoust. Soc. Am.* **110**(2), 1150–1163.
- Galvin, J. J., III, and Fu, Q.-J. (2005). "Effects of stimulation rate, mode and level on modulation detection by cochlear implant users," *J. Assoc. Res. Otolaryngol.* **6**(3), 269–279.
- Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., and Dahlgren, N. (1990). "DARPA TIMIT acoustic-phonetic continuous speech corpus CDROM," NIST Order No. PB91-505065, National Institute of Standards and Technology, Gaithersburg, MD.
- Glasberg, B. R., and Moore, B. C. J. (1990). "Deviation of auditory filter shapes from notched-noise data," *Hear. Res.* **47**, 103–138.
- Kewley-Port, D., Burkle, T. Z., and Lee, J. H. (2007). "Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners," *J. Acoust. Soc. Am.* **122**(4), 2365–2375.
- Kluender, K. R., and Alexander, J. M. (2007). "Perception of speech sounds," in *The Senses: A Comprehensive Reference, Vol. 3, Audition*, edited by P. Dallos and D. Oertel (Academic, San Diego), pp. 829–860.
- Kluender, K. R., Coady, J. A., and Kiefte, M. (2003). "Sensitivity to change in perception of speech," *Speech Commun.* **41**(1), 59–69.
- Kluender, K. R., Stilp, C. E., and Kiefte, M. (2013). "Perception of vowel sounds within a biologically realistic model of efficient coding," in *Vowel Inherent Spectral Change*, edited by G. Morrison and P. Assmann (Springer, Berlin), pp. 117–151.
- Kreft, H. A., Donaldson, G. S., and Nelson, D. A. (2004). "Effects of pulse rate and electrode array design on intensity discrimination in cochlear implant users," *J. Acoust. Soc. Am.* **116**, 2258–2268.
- Patterson, R. D., Nimmo-Smith, I., Weber, D. L., and Milroy, R. (1982). "The deterioration of hearing with age: Frequency selectivity, the critical ratio, the audiogram, and speech threshold," *J. Acoust. Soc. Am.* **72**(6), 1788–1803.

- Shannon, C. E. (1948). "A mathematical theory of communication," *Bell Syst. Tech. J.* **27**, 379–423 and 623–656.
- Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (1995). "Speech recognition with primarily temporal cues," *Science* **270**, 303–304.
- Smith, Z. M., Delgutte, B., and Oxenham, A. (2002). "Chimaeric sounds reveal dichotomies in auditory perception," *Nature* **416**, 87–90.
- Stilp, C. E., Alexander, J. M., Kieft, M., and Kluender, K. R. (2010a). "Auditory color constancy: Calibration to reliable spectral properties across speech and nonspeech contexts and targets," *Attent., Percept. Psychophys.* **72**(2), 470–480.
- Stilp, C. E., Kieft, M., Alexander, J. M., and Kluender, K. R. (2010b). "Cochlea-scaled spectral entropy predicts rate-invariant intelligibility of temporally distorted sentences," *J. Acoust. Soc. Am.* **128**(4), 2112–2126.
- Stilp, C. E., and Kluender, K. R. (2010). "Cochlea-scaled spectral entropy, not consonants, vowels, or time, best predicts speech intelligibility," *Proc. Natl. Acad. Sci. U.S.A.* **107**(27), 12387–12392.
- Swaminathan, J., and Heinz, M. G. (2012). "Psychophysiological analyses demonstrate the importance of neural envelope coding for speech perception in noise," *J. Neurosci.* **32**(5), 1747–1756.