Speech intelligibility is best predicted by intensity, not cochlea-scaled entropy

Andrew J. Oxenham,[a] Jeffrey E. Boucher, and Heather A. Kreft
Department of Psychology, University of Minnesota, Elliott Hall, 75 East River Parkway,
Minneapolis, Minnesota 55455, USA
oxenham@umn.edu, bouch123@umn.edu, plumx002@umn.edu

Abstract: Cochlea-scaled entropy (CSE) is a measure of spectro-temporal change that has been reported to predict the contribution of speech segments to overall intelligibility. This paper confirms that CSE is highly correlated with intensity, making it impossible to determine empirically whether it is CSE or simply intensity that determines speech importance. A more perceptually relevant version of CSE that uses dB-scaled differences, rather than differences in linear amplitude, failed to predict speech intelligibility. Overall, a parsimonious account of the available data is that the importance of speech segments to overall intelligibility is best predicted by their relative intensity, not by CSE.

[Q-JF]
Date Received: June 23, 2017 **Date Accepted:** August 29, 2017

1. Introduction

What components of acoustic speech are most important for intelligibility? This question has been answered in many different ways. Stilp and Kluender (2010) recently argued that most information is transmitted during dynamic portions of the stimulus, and so postulated that the speech segments with the most change are also the most important for intelligibility. They tested this hypothesis by developing a metric termed "cochlea-scaled entropy" (CSE), which involved dividing the magnitude spectrum of 16-ms time slices into frequency subbands, scaled by the estimated equivalent rectangular bandwidths ($\mathrm{ERB_N}$) of auditory filters (Glasberg and Moore, 1990), taking the absolute differences in magnitude within each frequency subband between two adjacent time slices and then summing these absolute differences to calculate the Euclidean distance. Stilp and Kluender (2010) found that intelligibility was more strongly degraded when segments with high CSE were replaced with noise than when segments with low CSE were replaced. They concluded that CSE, not vowels or consonants, best predicts speech intelligibility. This approach is attractive because it is consistent with more general principles of perceptual coding that emphasize changing over static stimuli on information-theoretic grounds (e.g., Barlow, 1961), and because the CSE can be computed automatically, without the need to segment speech manually into categories, such as vowels or consonants.

Despite its appeal, one aspect makes the application of CSE to speech questionable: the differences in magnitude between consecutive time slices are measured using linear amplitude, or sound pressure, and not a logarithmic transform, such as dB. This is unusual for an auditory application, as a dB scale is more common and more easily relatable to perception. For instance, to a first approximation, a 1-dB change in the level of a broadband sound is just detectable across a wide range of sound pressure levels (SPLs) (e.g., Moore and Raab, 1975). In terms of CSE, a 1-dB change in amplitude from 70 to 71 dB SPL results in a CSE value that is 100 times greater than the CSE value resulting from the same 1-dB change in amplitude from 30 to 31 dB SPL. In other words, the CSE is likely dominated by higher sound intensities, such that CSE measured close to a talker would result in higher values than CSE measured further away from the same talker. Chen and Loizou (2012) compared CSE with root-mean-square (rms) level (dB), but failed to make a direct comparison using the same time window lengths and criteria. More recently Shu *et al.* (2016) directly compared CSE with rms sound level (in dB). They found a relatively high correlation between the two ($r = 0.79$), and reported that both measures provided equally good predictions of listener performance for Mandarin Chinese sentences. The present study confirms these observations for American English, and extends them by comparing the

[a] Author to whom correspondence should be addressed.

original CSE with a new dB-scaled CSE measure that is arguably more perceptually relevant because it takes into account the more logarithmic nature of auditory perception and discrimination.

2. Correlations between linear CSE, dB-CSE, and intensity measures

The speech test material was the AZBio sentences (Spahr *et al.*, 2012), with each sentence normalized to the same rms. The first and last 80 ms of each sentence were discarded to avoid potential onset or offset effects. Each sentence was then divided into contiguous segments of 112 ms. Contiguous segments were used to maximize the independence between segments. The segments were then processed by three separate algorithms. The first was the original CSE algorithm, implemented as described by Stilp and Kluender (2010); see also Stilp *et al.* (2010). Each 112-ms segment was divided into seven 16-ms time slices. A fast Fourier transform (FFT) was performed on each of the 16-ms time slices. The magnitudes of frequency bins were weighted by rounded-exponential (Roex) functions (Patterson *et al.*, 1982) to form 33 filters, each spaced one $\mathrm{ERB_N}$ apart between 26 and 7743 Hz (Glasberg and Moore, 1990). The absolute differences in magnitude within each filter between one time slice and the next were summed to calculate the Euclidean distance between the two time slices. The CSE measure for a given 112-ms segment was therefore the sum of the Euclidean distances between all adjacent 16-ms time slices within the segment. The second algorithm was the average intensity within each of the 112-ms segments. This was calculated simply by squaring the waveform within each segment and summing the squared values. The third algorithm was the dB-scaled CSE algorithm. This was the same as the original CSE algorithm, with the exception that the calculations were made on the log-transformed (dB) values of the FFT magnitudes, rather than the linear values.

Once the 112-ms segments within each sentence had been processed according to each of the three algorithms described above, the segments within each sentence were ranked from lowest to highest (in terms of CSE, intensity, or dB-CSE). Finally, a Spearman rank correlation was performed within each sentence to compare the CSE algorithm with intensity and with dB-CSE. The results of these correlations, pooled across all the sentences, are shown in Fig. 1. The mean correlation between the CSE and intensity rankings was very high (Spearman's rho = 0.926, 95% confidence interval: 0.916, 0.936), whereas the correlation between the CSE and the dB-CSE rankings was near zero and slightly negative (Spearman's rho = −0.168, 95% confidence interval: −0.206, −0.130).

This analysis confirms the finding of Shu *et al.* (2016) that CSE is closely related to sound intensity. Our correlation coefficients were generally higher than theirs (0.926 compared with 0.79), presumably because we used rank correlations, which may be more appropriate than Pearson's product-moment correlations, given the fact that CSE rank (rather than absolute value) is used to determine the relative contributions of speech segments within sentences. In contrast, the low correlation between the original measure of CSE and the dB-CSE measure suggests that it is possible to compare the predictive value of CSE with that of dB-CSE. As dB-CSE is arguably more

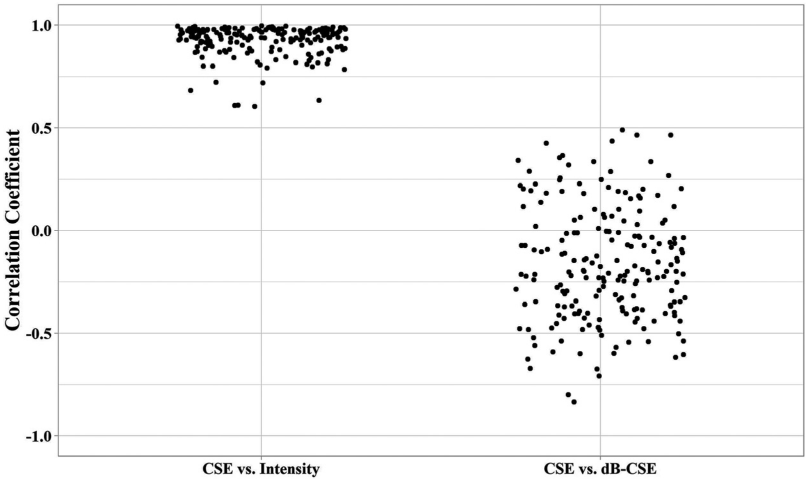

Fig. 1. Comparisons of CSE, Intensity, and dB-CSE measures of speech using the AZBio sentences. Each point represents the Spearman rank correlation between CSE and intensity (left) and CSE and CSE-dB (right) measures within a given sentence, based on contiguous 112-segments from each sentence.

perceptually relevant, evidence supporting the use of dB-CSE could be found if dB-CSE better predicted the effects on speech intelligibility of removing speech segments than either intensity or the original CSE. This prediction was tested in the following experiment.

3. Speech intelligibility after replacing segments of high, medium, or low CSE, intensity, or dB-CSE with noise

3.1 Methods

3.1.1 Listeners

Twenty-six (15 females, 11 males) native speakers of American English participated in the experiment. Participants ranged in age from 18 to 27 yrs ($M = 20.6$, standard deviation, s.d. = 2.7). All participants had normal hearing (thresholds less than 20 dB hearing level, HL) at octave frequencies between 250 and 8000 Hz and were compensated for their time. All experimental protocols were approved by the Institutional Review Board of the University of Minnesota, and all listeners provided informed written consent prior to participation.

3.1.2 Stimuli

Two sentence corpora were used for this experiment. The first corpus was comprised of 200 AZBio sentences (Spahr *et al.*, 2012), as tested in the correlations. It consists of sentences with relatively high semantic context, so that the later words in the sentences are predictable to some extent. These sentences, termed the context sentences, range in length from 3 to 11 words (mean 7.0, s.d. 1.6) and 1.5 to 4.3 s (mean 2.7, s.d. 0.5). An additional 20 AzBio sentences were used for practice. The sentences used had been recorded by four different speakers (two females and two males; 50 sentences each). The second corpus was comprised of 160 "nonsense" sentences, developed by Helfer (1997) and recorded by Freyman *et al.* (2012). These sentences were syntactically correct (subject/verb/object) but not semantically meaningful (e.g., "The goose kicked a street") and so provided no semantic context cues. They were employed as it was thought that sentences without context would provide a more critical test of the acoustic importance of individual speech segments. These sentences ranged in length from 5 to 7 words (1.34 to 2.12 s), with three keywords scored in each sentence. An additional 16 nonsense sentences were used for practice. The nonsense sentences were spoken by a single female talker. All speech materials were presented at an overall rms level of 57 dB SPL.

3.1.3 Signal processing

Ten conditions were tested in total. In the control condition, the sentences were presented to the listeners unprocessed. In the other nine conditions, each sentence was processed in MATLAB using one of three algorithms: (i) CSE, (ii) Intensity, or (iii) dB-CSE, as described in Sec. 2. Each 112-ms segment of each sentence was evaluated with a sliding window in steps of 16 ms. Stilp and Kluender (2010) used segment lengths of both 80 and 112 ms, corresponding roughly to consonant and vowel durations, respectively. Here, only segment lengths of 112 ms were used, as performance in a pilot study with 80-ms segments was close to 100% with the context sentences. Once the value (average intensity, CSE, or dB-CSE) within each of these segments was determined, one of three ranking procedures was undertaken. The segments were ranked from high to low (High conditions), from low to high (Low conditions), or from the smallest to the largest difference from the median (Median conditions). The rankings were used to replace some of the speech segments with noise, as follows: the top-ranked segment of each sentence was replaced with noise, and 80 ms before and after this segment were marked to be left intact (along with the first 80 ms of each sentence). The second ranked segment would then be replaced, but only if it did not overlap with either the first segment or the other marked sections, and again the neighboring 80 ms on either side of the segment were marked to be left intact. This procedure was then repeated until no further segments could be replaced. Using this method, the mean proportion of sentence replaced was 41.7% (s.d. 3.4%). The noise was created by passing a Gaussian white noise through a Butterworth lowpass filter with a cutoff frequency of 500 Hz and slope of 6 dB/oct. The noise segments were presented at the same rms level as the speech (57 dB SPL) and were gated on and off with 5-ms linear onset and offset ramps. The three processing algorithms (CSE, Intensity, and dB-CSE) and three replacement processes (High, Median, and Low) resulted in nine processing conditions.

3.1.4 Procedure

For each participant, the test sentences (200 AZBio and 160 nonsense sentences) were randomly and evenly distributed for use across the ten conditions (20 AZBio and 16 nonsense sentences per condition). Each participant completed one sentence corpus (context or nonsense sentences) before being tested on the other set. The order of testing was counterbalanced across the participants. Within each corpus, the presentation order of the ten conditions was randomized for each participant. Prior to each set of test sentences, participants completed a practice list for the respective sentence material.

The sentences were converted via a Lynx Studio Technology (Costa Mesa, CA) L22 soundcard at a sampling rate of 22 050 Hz and presented diotically via Sennheiser (Old Lyme, CT) HD650 headphones to listeners seated individually in a single-walled sound-attenuating booth. Listeners responded to sentences by typing what they heard via a computer keyboard. They were encouraged to guess individual words, even if they had not heard or understood the entire sentence. Sentences were scored for words correct as a proportion of the total number of words presented (context sentences) or total number of keywords (nonsense sentences). Scoring was first done automatically, and then errors were checked offline by a single rater for potential spelling errors or homophones. The proportion correct scores were converted to rationalized arcsine units (RAU) before statistical analysis (Studebaker, 1985). All reported analyses of variance (ANOVAs) include a Huynh-Feldt correction for lack of sphericity where applicable.

3.2 Results

Results from the sentence recognition experiment are shown in Fig. 2. Filled symbols represent results with the original CSE processing, gray symbols represent results with intensity processing, open symbols represent results with the dB-CSE processing, and red symbols at the far right represent results in the unprocessed conditions. Circles represent results with the context sentences and squares represent conditions with the nonsense sentences. As expected, performance was better for the context sentences than for the nonsense sentences. Also, in line with the results of Stilp and Kluender (2010), performance under CSE processing showed a progressive decline in performance from Low to Median to High CSE conditions for both sentence corpora. As expected based on the high rank correlations between CSE and intensity processing, the results under Intensity processing were very similar to those under CSE processing, again with a progressive decline in performance from Low to Median to High CSE conditions for both sentence corpora. In contrast, no systematic effects of processing from Low to High were observed under dB-CSE processing, suggesting that the dB-CSE measure does not successfully capture the relative importance of speech segments.

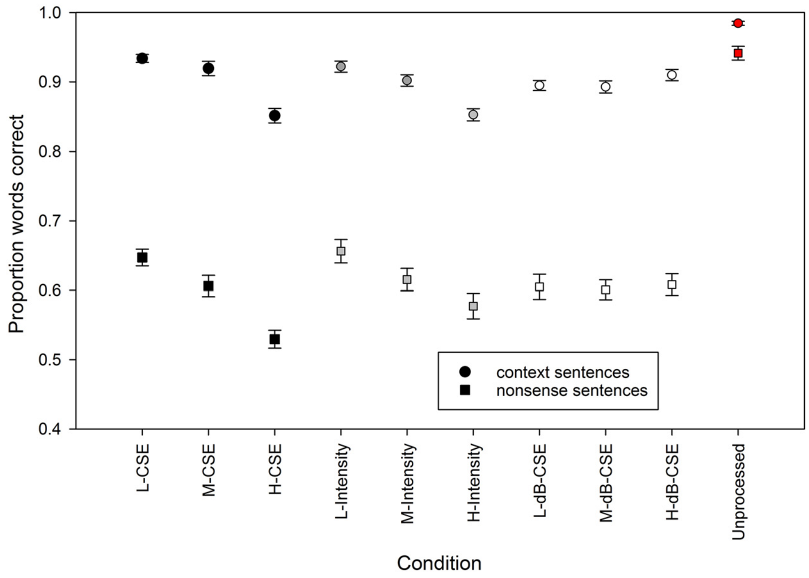

Fig. 2. (Color online) Speech intelligibility averaged across all 26 participants as a function of condition. Circles and squares represent the means of the context sentences (AzBio) and nonsense sentences, respectively. Each condition (CSE, Intensity, dB-CSE, and unprocessed) are filled with a different color (black, gray, white, red, respectively). Error bars represent ±1 standard error of the mean.

These observations were supported by a three-way repeated-measures ANOVA with RAU-transformed proportion of (key-) words correct as the dependent variable and factors of sentence corpus (context or nonsense), level of processing (High, Median, Low), and type of processing (CSE, Intensity, dB-CSE). As expected, there was a main effect of sentence corpus [$F(1,25) = 1744$, $p < 0.001$, partial-$\eta^2 = 0.986$]. There was also a main effect of level of processing [$F(2,50) = 49.9$, $p < 0.001$, partial-$\eta^2 = 0.666$], but no main effect of type of processing [$F(2,50) = 0.042$, $p = 0.959$, partial-$\eta^2 = 0.002$]. Significant interactions were found between corpus and type of processing [$F(1.66,41.5) = 4.23$, $p = 0.028$, partial-$\eta^2 = 0.145$], and between level and type of processing [$F(3.84,96.1) = 20.7$, $p < 0.001$, partial-$\eta^2 = 0.453$]. To further evaluate the interactions, a two-way (corpus × level) repeated-measures ANOVA was performed for each type of processing separately. There was a main effect of processing level for the CSE processing [$F(1.82,45.7) = 58.1$, $p < 0.001$, partial-$\eta^2 = 0.699$], with a significant linear trend from Low to Median to High [$F(2,153) = 5.337$, $p = 0.006$], and a main effect of level of processing for the Intensity processing [$F(2,50) = 29.7$, $\mathrm{p} < 0.001$, partial-$\eta^2 = 0.543$], also with a significant linear trend [$F(2,153) = 3.50$, $p = 0.033$]. However, there was no main effect of processing level for the dB-CSE processing [$F(2,50) = 1.49$, $p = 0.235$, partial-$\eta^2 = 0.056$].

4. Discussion

The correlational analyses confirm the findings of Shu *et al.* (2016) that the original measure of CSE is closely related to sound intensity or level. Therefore, based on the original CSE measure, it is not possible to determine whether it is truly change (entropy) or simply intensity that predicts the contributions of speech segments to word and sentence intelligibility. The newly proposed measure of dB-scaled CSE has the advantage of being more perceptually relevant, as it involves level differences in dB, rather than differences in linear units of amplitude or pressure, and is therefore insensitive to slow fluctuations in overall level, in line with perceptual results. In addition, the dB-scaled CSE is essentially uncorrelated with the original CSE measure, providing the opportunity to test the value of each measure empirically. Our speech-perception results confirm that a CSE-based approach is indistinguishable from an intensity-based approach (Shu *et al.*, 2016), but also show that dB-scaled CSE does not predict speech intelligibility. The results therefore suggest that it is relative intensity, and not CSE, that determines the contribution of speech segments to the intelligibility of words in a sentence. Clearly, a purely intensity-based measure also has some limitations: as with the original CSE, it can only be used effectively within a context where the overall rms level has been normalized. However, within its limits, the conclusion that intensity determines speech importance has some intuitive validity: from an ecological viewpoint, it is appropriate that the highest-intensity portions of speech, which are the least susceptible to interference and masking from other sources, are the most important for intelligibility. Similarly, the high-intensity portions are also those that are most likely to remain audible under reverberant conditions. In summary, intensity is a simple and intuitive measure that can explain the relative importance of individual speech segments within a sentence, without recourse to more complex change- or entropy-based measures.

Acknowledgments

This research was supported by NIH Grant No. R01 DC012262. We thank Ezleen Mustaffa and Cameron Eller for help with data collection.

References and links

Barlow, H. B. (**1961**). “Possible principles underlying the transformations of sensory messages,” in *Sensory Communication*, edited by W. A. Rosenblith (MIT Press, Cambridge, MA).

Chen, F., and Loizou, P. C. (**2012**). “Contributions of cochlea-scaled entropy and consonant-vowel boundaries to prediction of speech intelligibility in noise,” J. Acoust. Soc. Am. **131**, 4104–4113.

Freyman, R. L., Griffin, A. M., and Oxenham, A. J. (**2012**). “Intelligibility of whispered speech in stationary and modulated noise maskers,” J. Acoust. Soc. Am. **132**, 2514–2523.

Glasberg, B. R., and Moore, B. C. J. (**1990**). “Derivation of auditory filter shapes from notched-noise data,” Hear. Res. **47**, 103–138.

Helfer, K. S. (**1997**). “Auditory and auditory-visual perception of clear and conversational speech,” J. Speech. Lang. Hear. Res. **40**, 432–443.

Moore, B. C. J., and Raab, D. H. (**1975**). “Intensity discrimination for noise bursts in the presence of a continuous, bandstop background: Effects of level, width of the bandstop, and duration,” J. Acoust. Soc. Am. **57**, 400–405.

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**, 1788–1803.

Shu, Y., Feng, X. X., and Chen, F. (**2016**). "Comparing the perceptual contributions of cochlear-scaled entropy and speech level," J. Acoust. Soc. Am. **140**, EL517–EL521.

Spahr, A. J., Dorman, M. F., Litvak, L. M., Van Wie, S., Gifford, R. H., Loizou, P. C., Loiselle, L. M., Oakes, T., and Cook, S. (**2012**). "Development and validation of the AzBio sentence lists," Ear Hear. **33**, 112–117.

Stilp, C. E., Kiefte, M., Alexander, J. M., and Kluender, K. R. (**2010**). "Cochlea-scaled spectral entropy predicts rate-invariant intelligibility of temporally distorted sentences," J. Acoust. Soc. Am. **128**, 2112–2126.

Stilp, C. E., and Kluender, K. R. (**2010**). "Cochlea-scaled entropy, not consonants, vowels, or time, best predicts speech intelligibility," Proc. Natl. Acad. Sci. U.S.A. **107**, 12387–12392.

Studebaker, G. A. (**1985**). "A 'rationalized' arcsine transform," J. Speech Hear. Res. **28**, 455–462.